Proceedings of
23rd IEEE International Parallel and Distributed Processing Symposium

IPDPS 2009 Advance Program Abstracts

Abstracts for both contributed papers and all workshops have been compiled to allow authors to check accuracy and so that visitors to this website may preview the papers to be presented at the conference. Full proceedings of the conference will be published on a cdrom to be distributed to registrants at the conference.

Contents

Session 1: Algorithms - Scheduling I **2**

Session 2: Applications - Biological Applications **5**

Session 3: Architecture - Memory Hierarchy and Transactional Memory **8**

Session 4: Software - Fault Tolerance and Runtime Systems **11**

Session 5: Algorithms - Resource Management **14**

Session 6: Applications - System Software and Applications **17**

Session 7: Architecture - Power Efficiency and Process Variability **20**

Session 8: Software - Data Parallel Programming Frameworks **23**

Workshop 1: Heterogeneity in Computing Workshop **81**

Workshop 2: Reconfigurable Architectures Workshop **87**

Workshop 3: Workshop on High-Level Parallel Programming Models & Supportive Environments **103**

Workshop 7: Advances in Parallel and Distributed Computing Models **131**

Workshop 8: Communication Architecture for Clusters **141**

Workshop 9: High-Performance, Power-Aware Computing **147**

Workshop 10: High Performance Grid Computing **153**

IEEE International Parallel & Distributed Processing Symposium

IPDPS 2009

Session 1

Algorithms - Scheduling I

On Scheduling Dags to Maximize Area

Gennaro Cordasco
University of Salerno
cordasco@dia.unisa.it

Arnold L. Rosenberg
Colorado State University
rsnbrg@engr.colostate.edu

Abstract

A new quality metric, called *area*, is introduced for schedules that execute dags, i.e., computations having intertask dependencies. Motivated by the temporal unpredictability encountered when computing over the Internet, the goal under the new metric is to maximize the *average* number of tasks that are eligible for execution at each step of a computation. Area-maximization is a weakening of *IC-optimality,* which strives to maximize the number of eligible tasks at *every* step of the computation. In contrast to IC-optimal schedules, area-maximizing schedules exist for *every* dag. For dags that admit IC-optimal schedules, all area-maximizing schedules are IC-optimal, and vice versa. The basic properties of this metric are derived in this paper, and tools for efficiently crafting area-maximizing schedules for large classes of computationally significant dags are developed. Several of these results emerge from a close connection between area-maximizing scheduling and the *MAX Linear-Arrangement Problem for Dags*.

Efficient Scheduling of Task Graph Collections on Heterogeneous Resources

Matthieu Gallet[2,4,5], Loris Marchal[1,4,5] and Frédéric Vivien[3,4,5]
[1]CNRS [2]ENS Lyon [3]INRIA [4]Université de Lyon
[5]LIP laboratory, UMR 5668, ENS Lyon - CNRS - INRIA - UCBL, Lyon, France
{matthieu.gallet, loris.marchal, frederic.vivien}@ens-lyon.fr

Abstract

In this paper, we focus on scheduling jobs on computing Grids. In our model, a Grid job is made of a large collection of input data sets, which must all be processed by the same task graph or workflow, thus resulting in a collection of task graphs problem. We are looking for a competitive scheduling algorithm not requiring complex control. We thus only consider single-allocation strategies. In addition to a mixed linear programming approach to find an optimal allocation, we present different heuristic schemes. Then, using simulations, we compare the performance of our different heuristics to the performance of a classical scheduling policy in Grids, HEFT. The results show that some of our static-scheduling policies take advantage of their platform and application knowledge and outperform HEFT, especially under communication-intensive scenarios. In particular, one of our heuristics, DELEGATE, almost always achieves the best performance while having lower running times than HEFT.

Static Strategies for Worksharing with Unrecoverable Interruptions

A. Benoit[2,4,5], Y. Robert[2,4,5], A. L. Rosenberg[1] and F. Vivien[3,4,5]
[1]Colorado State University, USA [2]ENS Lyon, France [3]INRIA, France
[4]Université de Lyon, France 5 LIP, UMR 5668 ENS-CNRS-INRIA-UCBL, Lyon, France

Abstract

One has a large workload that is "divisible"—its constituent work's granularity can be adjusted arbitrarily—and one has access to p remote computers that can assist in computing the workload. The problem is that the remote computers are subject to interruptions of known likelihood that kill all work in progress. One wishes to orchestrate sharing the workload with the remote computers in a way that maximizes the expected amount of work completed. Strategies for achieving this goal, by balancing the desire to checkpoint often, in order to decrease the amount of vulnerable work at any point, vs. the desire to avoid the context-switching required to checkpoint, are studied. Strategies are devised that provably maximize the expected amount of work when there is only one remote computer (the case p=1). Results suggest the intractability of such maximization for higher values of p, which motivates the development of heuristic approaches. Heuristics are developed that replicate works on several remote computers, in the hope of thereby decreasing the impact of work-killing interruptions. The quality of these heuristics is assessed through exhaustive simulations.

On the Complexity of Mapping Pipelined Filtering Services on Heterogeneous Platforms

Anne Benoit, Fanny Dufossé and Yves Robert
LIP, École Normale Supérieure de Lyon, 46 allée d'Italie, 69364 Lyon Cedex 07, France
{Anne.Benoit, Fanny.Dufosse, Yves.Robert}@ens-lyon.fr

Abstract

In this paper, we explore the problem of mapping filtering services on large-scale heterogeneous platforms. Two important optimization criteria should be considered in such a framework. The period, which is the inverse of the throughput, measures the rate at which data sets can enter the system. The latency measures the response time of the system in order to process one single data set entirely. Both criteria are antagonistic. For homogeneous platforms, the complexity of period minimization is already known [12]; we derive an algorithm to solve the latency minimization problem in the general case with service precedence constraints; we also show that the bi-criteria problem (latency minimization without exceeding a prescribed value for the period) is of polynomial complexity. However, when adding heterogeneity to the platform, we prove that minimizing the period or the latency becomes NP-complete, and that these problems cannot be approximated by any constant factor (unless P=NP). The latter results hold true even for services without precedence constraints.

Session 2

Applications - Biological Applications

Sequence Alignment with GPU: Performance and Design Challenges

Gregory M. Striemer and Ali Akoglu
Department of Electrical and Computer Engineering
University of Arizona, 85721
Tucson, Arizona USA
{gmstrie, akoglu}@ece.arizona.edu

Abstract

In bioinformatics, alignments are commonly performed in genome and protein sequence analysis for gene identification and evolutionary similarities. There are several approaches for such analysis, each varying in accuracy and computational complexity. Smith-Waterman (SW) is by far the best algorithm for its accuracy in similarity scoring. However, execution time of this algorithm on general purpose processor based systems makes it impractical for use by life scientists. In this paper we take Smith-Waterman as a case study to explore the architectural features of Graphics Processing Units (GPUs) and evaluate the challenges the hardware architecture poses, as well as the software modifications needed to map the program architecture on to the GPU. We achieve a 23x speedup against the serial version of the SW algorithm. We further study the effect of memory organization and the instruction set architecture on GPU performance. For that purpose we analyze another implementation on an Intel Quad Core processor that makes use of Intel's SIMD based SSE2 architecture. We show that if reading blocks of 16 words at a time instead of 4 is allowed, and if 64KB of shared memory as opposed to 16KB is available to the programmer, GPU performance enhances significantly making it comparable to the SIMD based implementation. We quantify these observations to illustrate the need for studies on extending the instruction set and memory organization for the GPU.

Evaluating the use of GPUs in Liver Image Segmentation and HMMER Database Searches

John Paul Walters, Vidyananth Balu, Suryaprakash Kompalli[1] and Vipin Chaudhary
Department of Computer Science and Engineering
University at Buffalo, SUNY, Buffalo, NY
{waltersj, vbalu2, vipin}@buffalo.edu
[1]Hewlett-Packard Laboratories, Bangalore, India
kompalli@hp.com

Abstract

In this paper we present the results of parallelizing two life sciences applications, Markov random fieldsbased (MRF) liver segmentation and HMMER's Viterbi algorithm, using GPUs. We relate our experiences in porting both applications to the GPU as well as the techniques and optimizations that are most beneficial. The unique characteristics of both algorithms are demonstrated by implementations on an NVIDIA 8800 GTX Ultra using the CUDA programming environment. We test multiple enhancements in our GPU kernels in order to demonstrate the effectiveness of each strategy. Our optimized MRF kernel achieves over 130x speedup, and our hmmsearch implementation achieves up to 38x speedup. We show that the differences in speedup between MRF and hmmsearch is due primarily to the frequency at which the hmmsearch must read from the GPU's DRAM.

Improving MPI-HMMER's Scalability with Parallel I/O

John Paul Walters, Rohan Darole and Vipin Chaudhary
Department of Computer Science and Engineering
University at Buffalo, The State University of New York
{waltersj, rdarole, vipin}@buffalo.edu

Abstract

We present PIO-HMMER, an enhanced version of MPI-HMMER. PIO-HMMER improves on MPIHMMER's scalability through the use of parallel I/O and a parallel file system. In addition, we describe several enhancements, including a new load balancing scheme, enhanced post-processing, improved doublebuffering support, and asynchronous I/O for returning scores to the master node. Our enhancements to the core HMMER search tools, hmmsearch and hmmpfam, allow for scalability up to 256 nodes whereMPI-HMMER previously did not scale beyond 64 nodes. We show that our performance enhancements allow hmmsearch to achieve between 48x and 221x speedup using 256 nodes, depending on the size of the input HMM and the database. Further, we show that by integrating database caching with PIO-HMMER's hmmpfam tool we can achieve up to 328x performance using only 256 nodes.

Accelerating Leukocyte Tracking Using CUDA: A Case Study in Leveraging Manycore Coprocessors

Michael Boyer[1], David Tarjan[1], Scott T. Acton[2] and Kevin Skadron[1]
Departments of [1]Computer Science and [2]Electrical and Computer Engineering
University of Virginia, Charlottesville, VA 22904

Abstract

The availability of easily programmable manycore CPUs and GPUs has motivated investigations into how to best exploit their tremendous computational power for scientific computing. Here we demonstrate how a systems biology application—detection and tracking of white blood cells in video microscopy—can be accelerated by 200x using a CUDA-capable GPU. Because the algorithms and implementation challenges are common to a wide range of applications, we discuss general techniques that allow programmers to make efficient use of a manycore GPU.

Session 3

Architecture - Memory Hierarchy and Transactional Memory

Efficient Shared Cache Management through Sharing-Aware Replacement and Streaming-Aware Insertion Policy

Yu Chen[1], Wenlong Li[2], Changkyu Kim[2] and Zhizhong Tang[1]
[1]Department of Computer Science and Technology, Tsinghua University, Beijing, China
[2]Microprocessor Technology Lab, Intel Corp
chenyu00@gmail.com; {wenlong.li, changkyu.kim}@intel.com; tzz-dcs@tsinghua.edu.cn

Abstract

Multi-core processors with shared caches are now commonplace. However, prior works on shared cache management primarily focused on multi-programmed workloads. These schemes consider how to partition the cache space given that simultaneously-running applications may have different cache behaviors. In this paper, we examine policies for managing shared caches for running single multi-threaded applications. First, we show that the shared-cache miss rate can be significantly reduced by reserving a certain amount of space for shared data. Therefore, we modify the replacement policy to dynamically partition each set between shared and private data. Second, we modify the insertion policy to prevent streaming data (data not reused before eviction) from promoting to the MRU position. Finally, we use a low-overhead sampling mechanism to dynamically select the optimal policy. Compared to LRU policy, our scheme reduces the miss rate on average by 8.7% on 8MB caches and 20.1% on 16MB caches respectively.

Core-aware Memory Access Scheduling Schemes

Zhibin Fang, Xian-He Sun, Yong Chen and Surendra Byna Department of Computer Science
Illinois Institute of Technology, Chicago, IL 60616, USA
{zfang2, sun, chenyon1, sbyna}@iit.edu

Abstract

Multi-core processors have changed the conventional hardware structure and require a rethinking of system scheduling and resource management to utilize them efficiently. However, current multi-core systems are still using conventional single-core memory scheduling. In this study, we investigate and evaluate traditional memory access scheduling techniques, and propose a core-aware memory scheduling for multi-core environments. Since memory requests from the same source exhibit better locality, it is reasonable to schedule the requests by taking the source of the requests into consideration. Motivated from this principle of locality, we propose two core-aware policies based on traditional bank-first and row-first schemes. Simulation results show that the core-aware policies can effectively improve the performance. Compared with the bank-first and row-first policies, the proposed core-aware policies reduce the execution time of certain NAS Parallel Benchmarks by up to 20% in running the benchmarks separately, and by 11% in running them concurrently.

Using Hardware Transactional Memory for Data Race Detection

Shantanu Gupta
Department of EE and CS
University of Michigan
shangupt@umich.edu

Florin Sultan, Srihari Cadambi, Franjo Ivančić and Martin Rötteler
NEC Laboratories America
Princeton, NJ
{cadambi, ivancic, mroetteler}@nec-labs.com

Abstract

Widespread emergence of multicore processors will spur development of parallel applications, exposing programmers to degrees of hardware concurrency hitherto unavailable. Dependable multithreaded software will have to rely on the ability to dynamically detect nondeterministic and notoriously hard to reproduce synchronization bugs manifested through data races. Previous solutions to dynamic data race detection have required specialized hardware, at additional power, design and area costs. We propose RaceTM, a novel approach to data race detection that exploits hardware that will likely be present in future multiprocessors, albeit for a different purpose. In particular, we show how emerging hardware support for transactional memory can be leveraged to aid data race detection. We propose the concept of lightweight *debug transactions* that exploit the conflict detection mechanisms of transactional memory systems to perform data race detection. We present a proof-of-concept simulation prototype, and evaluate it on data races injected into applications from the SPLASH-2 suite. Our experiments show that this technique is effective at discovering data races and has low performance overhead.

Speculation-Based Conflict Resolution in Hardware Transactional Memory

Rubén Titos, Manuel E. Acacio, José M. García
Departamento de Ingeniería y Tecnología de Computadores
Universidad de Murcia
Murcia, Spain
{rtitos,meacacio,jmgarcia}@ditec.um.es

Abstract

Conflict management is a key design dimension of hardware transactional memory (HTM) systems, and the implementation of efficient mechanisms for detection and resolution becomes critical when conflicts are not a rare event. Current designs address this problem from two opposite perspectives, namely, lazy and eager schemes. While the former approach is based on an purely optimistic view that is not well-suited when conflicts become frequent, the latter results too pessimistic because resolves conflicts too conservatively, often limiting concurrency unnecessarily. In this paper, we present a hybrid, pseudo-optimistic scheme of conflict resolution for HTM systems that recaptures the concept of speculation to allow transactions to continue their execution past conflicting accesses. Simulation results show that our proposal is capable of combining the advantages of both classical approaches. For the STAMP transactional benchmarks, our hybrid scheme outperforms both eager and lazy systems with average reductions in execution time of 8 and 17%, respectively, and it decreases network traffic by another 17% compared to the eager policy.

Session 4

Software - Fault Tolerance and Runtime Systems

Compiler-Enhanced Incremental Checkpointing for OpenMP Applications

Greg Bronevetsky
Lawrence Livermore National Lab
greg@bronevetsky.com

Daniel Marques
Ballista Securities
marques77@gmail.com

Keshav Pingali
The University of Texas at Austin
pingali@cs.utexas.edu

Sally McKee
Chalmers University of Technology
mckee@chalmers.se

Radu Rugina
VMWare
radu.rugina@gmail.com

Abstract

As modern supercomputing systems reach the peta-flop performance range, they grow in both size and complexity. This makes them increasingly vulnerable to failures from a variety of causes. Checkpointing is a popular technique for tolerating such failures, enabling applications to periodically save their state and restart computation after a failure. Although a variety of automated system-level checkpointing solutions are currently available to HPC users, manual application-level checkpointing remains more popular due to its superior performance. This paper improves performance of automated checkpointing via a compiler analysis for incremental checkpointing. This analysis, which works with both sequential and OpenMP applications, significantly reduces checkpoint sizes and enables asynchronous checkpointing.

DMTCP: Transparent Checkpointing for Cluster Computations and the Desktop

Jason Ansel
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA
jansel@csail.mit.edu

Kapil Arya and Gene Cooperman
College of Computer and Information Science
Northeastern University
Boston, MA
{kapil,gene}@ccs.neu.edu

Abstract

DMTCP (Distributed MultiThreaded CheckPointing) is a transparent user-level checkpointing package for distributed applications. Checkpointing and restart is demonstrated for a wide range of over 20 well known applications, including MATLAB, Python, TightVNC, MPICH2, OpenMPI, and runCMS. RunCMS runs as a 680 MB image in memory that includes 540 dynamic libraries, and is used for the CMS experiment of the Large Hadron Collider at CERN. DMTCP transparently checkpoints general cluster computations consisting of many nodes, processes, and threads; as well as typical desktop applications. On 128 distributed cores (32 nodes), checkpoint and restart times are typically 2 seconds, with negligible run-time overhead. Typical checkpoint times are reduced to 0.2 seconds when using forked checkpointing. Experimental results show that checkpoint time remains nearly constant as the number of nodes increases on a medium-size cluster.

DMTCP automatically accounts for fork, exec, ssh, mutexes/ semaphores, TCP/IP sockets, UNIX domain sockets, pipes, ptys (pseudo-terminals), terminal modes, ownership of controlling terminals, signal handlers, open file descriptors, shared open file descriptors, I/O (including the readline library), shared memory (via mmap), parent-child process relationships, pid virtualization, and other operating system artifacts. By emphasizing an unprivileged, user-space approach, compatibility is maintained across Linux kernels from 2.6.9 through the current 2.6.28. Since DMTCP is unprivileged and does not require special kernel modules or kernel patches, DMTCP can be incorporated and distributed as a checkpoint-restart module within some larger package.

Elastic Scaling of Data Parallel Operators in Stream Processing

Scott Schneidery[1,2], Henrique Andrade[2], Buğra Gedik[2], Alain Biem[2] and Kun-Lung Wu[2]
[1]Virginia Tech
Department of Computer Science
Blacksburg, VA, USA
scschnei@cs.vt.edu
[2]Thomas J. Watson Research Center
IBM Research
Hawthorne, NY, USA
{hcma,bgedik,biem,klwu}@us.ibm.com

Abstract

We describe an approach to elastically scale the performance of a data analytics operator that is part of a streaming application. Our techniques focus on dynamically adjusting the amount of computation an operator can carry out in response to changes in incoming workload and the availability of processing cycles. We show that our elastic approach is beneficial in light of the dynamic aspects of streaming workloads and stream processing environments. Addressing another recent trend, we show the importance of our approach as a means to providing computational elasticity in multicore processor-based environments such that operators can automatically find their best operating point. Finally, we present experiments driven by synthetic workloads, showing the space where the optimizing efforts are most beneficial and a radioastronomy imaging application, where we observe substantial improvements in its performance-critical section.

Scalable RDMA performance in PGAS languages

Montse Farreras[†], George Almási[‡], Călin Caşcaval[‡], Toni Cortes[†]
[†] Department of Computer Architecture, Universitat Politecnica de Catalunya
Barcelona Supercomputing Center, Barcelona, Spain
{mfarrera, toni}@ac.upc.es
[‡] IBM T.J. Watson Research Center, Yorktown Heights, NY
{gheorghe, cascaval}@us.ibm.com

Abstract

Partitioned Global Address Space (PGAS) languages provide a unique programming model that can span shared-memory multiprocessor (SMP) architectures, distributed memory machines, or cluster of SMPs. Users can program large scale machines with easy-to-use, shared memory paradigms.

In order to exploit large scale machines efficiently, PGAS language implementations and their runtime system must be designed for scalability and performance. The IBM XLUPC compiler and runtime system provide a scalable design through the use of the Shared Variable Directory (SVD). The SVD stores meta-information needed to access shared data. It is dereferenced, in the worst case, for every shared memory access, thus exposing a potential performance problem.

In this paper we present a cache of remote addresses as an optimization that will reduce the SVD access overhead and allow the exploitation of native (remote) direct memory accesses. It results in a significant performance improvement while maintaining the run-time portability and scalability.

Session 5

Algorithms - Resource Management

Singular Value Decomposition on GPU using CUDA

Sheetal Lahabar and P. J. Narayanan
Center for Visual Information Technology
International Institute of Information Technology
Hyderabad, India
sheetal@students.iiit.ac.in, pjn@iiit.ac.in

Abstract

Linear algebra algorithms are fundamental to many computing applications. Modern GPUs are suited for many general purpose processing tasks and have emerged as inexpensive high performance co-processors due to their tremendous computing power. In this paper, we present the implementation of singular value decomposition (SVD) of a dense matrix on GPU using the CUDA programming model. SVD is implemented using the twin steps of bidiagonalization followed by diagonalization. It has not been implemented on the GPU before. Bidiagonalization is implemented using a series of Householder transformations which map well to BLAS operations. Diagonalization is performed by applying the implicitly shifted QR algorithm. Our complete SVD implementation outperforms the MATLAB and Intel ®Math Kernel Library (MKL) LAPACK implementation significantly on the CPU. We show a speedup of upto 60 over the MATLAB implementation and upto 8 over the Intel MKL implementation on a Intel Dual Core 2.66GHz PC on NVIDIA GTX 280 for large matrices. We also give results for very large matrices on NVIDIA Tesla S1070.

Coupled Placement in Modern Data Centers

Madhukar Korupolu
IBM Almaden Research Center
madhukar@us.ibm.com

Aameek Singh
IBM Almaden Research Center
aameek.singh@us.ibm.com

Bhuvan Bamba
Georgia Tech
bhuvan@cc.gatech.edu

Abstract

We introduce the coupled placement problem for modern data centers spanning placement of application computation and data among available server and storage resources. While the two have traditionally been addressed independently in data centers, two modern trends make it beneficial to consider them together in a coupled manner: (a) rise in virtualization technologies, which enable applications packaged as VMs to be run on any server in the data center with spare compute resources, and (b) rise in multi-purpose hardware devices in the data center which provide compute resources of varying capabilities at different proximities from the storage nodes.

We present a novel framework called CPA for addressing such coupled placement of application data and computation in modern data centers. Based on two well-studied problems C Stable Marriage and Knapsacks C the CPA framework is simple, fast, versatile and automatically enables high throughput applications to be placed on nearby server and storage node pairs. While a theoretical proof of CPA's worst-case approximation guarantee remains an open question, we use extensive experimental analysis to evaluate CPA on large synthetic data centers comparing it to Linear Programming based methods and other traditional methods. Experiments show that CPA is consistently and surprisingly within 0 to 4% of the Linear Programming based optimal values for various data center topologies and workload patterns. At the same time it is one to two orders of magnitude faster than the LP based methods and is able to scale to much larger problem sizes.

The fast running time of CPA makes it highly suitable for large data center environments where hundreds to thousands of server and storage nodes are common. LP based approaches are prohibitively slow in such environments. CPA is also suitable for fast interactive analysis during consolidation of such environments from physical to virtual resources.

An Upload Bandwidth Threshold for Peer-to-Peer Video-on-Demand Scalability

Yacine Boufkhad
Paris Diderot University, LIAFA, France.
boufkhad@liafa.jussieu.fr

Fabien Mathieu
Orange Labs, Issy-les-Moulineaux, France.
fabien.mathieu@orange-ftgroup.com

Fabien de Montgolfier
Paris Diderot University, LIAFA, France.
fm@liafa.jussieu.fr

Diego Perino
Orange Labs, Issy-les-Moulineaux, France.
diego.perino@orange-ftgroup.com

Laurent Viennot
INRIA Project-Team "GANG" between INRIA and LIAFA, France.
laurent.viennot@inria.fr

Abstract

We consider the fully distributed Video-on-Demand problem, where n nodes called *boxes* store a large set of videos and collaborate to serve simultaneously n videos or less between them. It is said to be *scalable* when $\Omega(n)$ videos can be distributively stored under the condition that any sequence of demands for these videos can always be satisfied. Our main result consists in establishing a threshold on the average upload bandwidth of a box, above which the system becomes scalable. We are thus interested in the normalized upload capacity $u = \frac{\textit{upload bandwidth}}{\textit{video bitrate}}$ of a box. The number m of distinct videos stored in the system is called its catalog size.

We show an upload capacity threshold of 1 for scalability in a homogeneous system, where all boxes have the same upload capacity. More precisely, a system with $u < 1$ has constant catalog size $m = O(1)$ (every box must store some data of every video). On the other hand, for $u > 1$, an homogeneous system where all boxes have same upload capacity at least u admits a static allocation of $m = \Omega(n)$ videos into the boxes such that any adversarial sequence of video demands can be satisfied. Moreover, such an allocation can be obtained randomly with high probability. This result is generalized to a system of boxes that have heterogeneous upload capacities under some balancing conditions.

Competitive Buffer Management with Packet Dependencies

Alex Kesselman
Google Inc.
Mountain View, CA, USA
alx@google.com

Boaz Patt-Shamir
School of Electrical Engineering
Tel Aviv University
Tel Aviv 69978, Israel
boaz@eng.tau.ac.il

Gabriel Scalosub
Department of Computer Science
University of Toronto
Toronto, ON, Canada
scalosub@cs.toronto.edu

Abstract

We introduce the problem of managing a FIFO buffer of bounded space, where arriving packets have dependencies among them. Our model is motivated by the scenario where large data frames must be split into multiple packets, because maximum packet size is limited by data-link restrictions. A frame is considered useful only if sufficiently many of its constituent packets are delivered. The buffer management algorithm decides, in case of overflow, which packets to discard and which to keep in the buffer. The goal of the buffer management algorithm is to maximize throughput of useful frames. This problem has a variety of applications, e.g., Internet video streaming, where video frames are segmented and encapsulated in IP packets sent over the Internet. We study the complexity of the above problem in both the offline and online settings. We give upper and lower bounds on the performance of algorithms using competitive analysis.

Session 6

Applications - System Software and Applications

Annotation-Based Empirical Performance Tuning Using Orio

Albert Hartono
Dept. of Computer Science and Engg.
Ohio State University
Columbus, Ohio 43210C1277
hartonoa@cse.ohio-state.edu

Boyana Norris
Mathematics and Computer Science Division
Argonne National Laboratory
Argonne, Illinois 60439C4844
norris@mcs.anl.gov

P. Sadayappan
Dept. of Computer Science and Engg.
Ohio State University
Columbus, Ohio 43210C1277
saday@cse.ohio-state.edu

Abstract

For many scientific applications, significant time is spent in tuning codes for a particular high-performance architecture. Tuning approaches range from the relatively nonintrusive (e.g., by using compiler options) to extensive code modifications that attempt to exploit specific architecture features. Intrusive techniques often result in code changes that are not easily reversible, and can negatively impact readability, maintainability, and performance on different architectures. We introduce an extensible annotation-based empirical tuning system called Orio that is aimed at improving both performance and productivity. It allows software developers to insert annotations in the form of structured comments into their source code to trigger a number of low-level performance optimizations on a specified code fragment. To maximize the performance tuning opportunities, the annotation processing infrastructure is designed to support both architecture-independent and architecture-specific code optimizations. Given the annotated code as input, Orio generates many tuned versions of the same operation and empirically evaluates the alternatives to select the best performing version for production use. We have also enabled the use of the Pluto automatic parallelization tool in conjunction with Orio to generate efficient OpenMP-based parallel code. We describe our experimental results involving a number of computational kernels, including dense array and sparse matrix operations.

Automatic Detection of Parallel Applications Computation Phases

Juan Gonzalez, Judit Gimenez and Jesus Labarta
BSC - UPC - Barcelona, Spain
{juan.gonzalez,judit,jesus}@bsc.es

Abstract

Analyzing parallel programs has become increasingly difficult due to the immense amount of information collected on large systems. The use of clustering techniques has been proposed to analyze applications. However, while the objective of previous works is focused on identifying groups of processes with similar characteristics, we target a much finer granularity in the application behavior.

In this paper, we present a tool that automatically characterizes the different computation regions between communication primitives in message-passing applications. This study shows how some of the clustering algorithms which may be applicable at a coarse grain are no longer adequate at this level. Density-based clustering algorithms applied to the performance counters offered by modern processors are more appropriate in this context. This tool automatically generates accurate displays of the structure of the application as well as detailed reports on a broad range of metrics for each individual region detected.

Handling OS Jitter on Multicore Multithreaded Systems

Pradipta and De Vijay Mann
IBM India Research Lab
New Delhi
{pradipta.de, vijamann}@in.ibm.com

Umang Mittaly
Indian Institute of Technology,
New Delhi
umang.k.mittal@gmail.com

Abstract

Various studies have shown that OS jitter can degrade parallel program performance considerably at large processor counts. Most sources of system jitter fall broadly into 5 categories - user space processes, kernel threads, interrupts, SMT interference and hypervisor activity. Solutions to OS jitter typically consist of a combination of techniques such as synchronization of jitter across nodes (co-scheduling or gang scheduling) and use of microkernels. Both techniques present several drawbacks. Multicore and Multithreaded systems present opportunities to handle OS jitter. They have multiple cores and threads, some of which can be used for handling OS jitter, while the application threads run on remaining cores and threads. However, they are also prone to risks such as inter-thread cache interference and process migration. In this paper, we present a holistic approach that aims to reduce jitter caused by various sources of jitter by utilizing the additional threads or cores in a system. Our approach handles jitter through reduction of kernel threads, intelligent interrupt handling, and switching of hardware SMT thread priorities. This helps in reducing jitter experienced by application threads in the user space, at the kernel level, and at the hardware level. We make use of existing features available in the Linux kernel and Power Architecture as well make enhancements to the Linux kernel. We demonstrate the efficacy of our techniques by reducing jitter on two different platforms and operating system versions. In the first case our approach helps in reducing periodic jitter that improves both average and worst case performance of a simulated parallel application. In the second case our approach helps in reducing infrequent very large jitter that helps the worst case performance of a real parallel application. Our experimental results show up to 30% reduction in slowdown in the average case at 16K OS images and up to 50% reduction in slowdown in the worst case at 8 OS images using this approach as compared to a baseline configuration.

Building a Parallel Pipelined External Memory Algorithm Library

Andreas Beckmann
Institut fur Informatik
Goethe-Universität Frankfurt am Main
beckmann@cs.uni-frankfurt.de

Roman Dementiev
Institut für Theoretische Informatik
Universität Karlsruhe (TH)
dementiev@ira.uka.de

Johannes Singler
Institut für Theoretische Informatik
Universität Karlsruhe (TH)
singler@ira.uka.de

Abstract

Large and fast hard disks for little money have enabled the processing of huge amounts of data on a single machine. For this purpose, the well-established STXXL library provides a framework for external memory algorithms with an easy-to-use interface. However, the clock speed of processors cannot keep up with the increasing bandwidth of parallel disks, making many algorithms actually compute-bound.

To overcome this steadily worsening limitation, we exploit today's multi-core processors with two new approaches. First, we parallelize the internal computation of the encapsulated external memory algorithms by utilizing the MCSTL library. Second, we augment the unique pipelining feature of the STXXL, to enable automatic task parallelization.

We show using synthetic and practical use cases that the combination of both techniques increases performance greatly.

Session 7

Architecture - Power Efficiency and Process Variability

On Reducing Misspeculations in a Pipelined Scheduler

R. Gran
University of Zaragoza-CPS
rgran@unizar.es

E. Morancho, A. Olivé and J.M. Llabería
Universitat Politècnica de Catalunya-DAC
{enricm,angel,llaberia}@ac.upc.edu

Abstract

Pipelining the scheduling logic, which exposes and exploits the instruction level parallelism, degrades processor performance. In a 4-issue processor, our evaluations show that pipelining the scheduling logic over two cycles degrades performance by 10% in SPEC-2000 integer benchmarks. Such a performance degradation is due to sacrificing the ability to execute dependent instructions in consecutive cycles.

Speculative selection is a previously proposed technique that boosts the performance of a processor with a pipelined scheduling logic. However, this new speculation source increases the overall number of misspeculated instructions, and this unuseful work wastes energy.

In this work we introduce a non-speculative mechanism named Dependence Level Scheduler (DLS) which not only tolerates the scheduling-logic latency but also reduces the number of misspeculated instructions with respect to a scheduler with speculative selection. In DLS, the selection of a group of one-cycle instructions (producer-level) is overlapped with the wake up in advance of its group of dependent instructions. DLS is not speculative because the group of woken in advance instructions will compete for selection only after issuing all producer-level instructions. On average, DLS reduces the number of misspeculated instructions with respect to a speculative scheduler by 17.9%. From the IPC point of view, the speculative scheduler outperforms DLS by 0.3%. Moreover, we propose two non-speculative improvements to DLS.

Efficient Microarchitecture Policies for Accurately Adapting to Power Constraints

Juan M. Cebrián[1], Juan L. Arag'on[1], José M. García[1], Pavlos Petoumenos[2] and Stefanos Kaxiras[2]
[1]Dept. of Computer Engineering, University of Murcia, Murcia, 30100, Spain
{jcebrian,jlaragon,jmgarcia}@ditec.um.es
[2]Dept. of Electrical and Computer Engineering, University of Patras, 26500, Greece
{kaxiras,ppetoumenos}@ee.upatras.gr

Abstract

In the past years Dynamic Voltage and Frequency Scaling (DVFS) has been an effective technique that allowed microprocessors to match a predefined power budget. However, as process technology shrinks, DVFS becomes less effective (because of the increasing leakage power) and it is getting closer to a point where DVFS won't be useful at all (when static power exceeds dynamic power). In this paper we propose the use of microarchitectural techniques to accurately match a power constraint while maximizing the energy efficiency of the processor. We will predict the processor power consumption at a basic block level, using the consumed power translated into tokens to select between different power-saving microarchitectural techniques. These techniques are orthogonal to DVFS so they can be simultaneously applied. We propose a two-level approach where DVFS acts as a coarse-grained technique to lower the average power while microarchitectural techniques remove all the power spikes efficiently. Experimental results show that the use of power-saving microarchitectural techniques in conjunction with DVFS is up to six times more precise, in terms of total energy consumed (area) over the power budget, than using DVFS alone for matching a predefined power budget. Furthermore, in a near future DVFS will become DFS because lowering the supply voltage will be too expensive in terms of leakage power. At that point, the use of power-saving microarchitectural techniques will become even more energy efficient.

An On/Off Link Activation Method for Low-Power Ethernet in PC Clusters

Michihiro Koibuchi[1], Tomohiro Otsuka[2], Hiroki Matsutani[2], and Hideharu Amano[2,1]

[1]National Institute of Informatics
2-1-2, Hitotsubashi, Chiyoda-ku, Tokyo,
JAPAN 101-8430
koibuchi@nii.ac.jp

[2]Keio University
3-14-1, Hiyoshi, Kohoku-ku, Yokohama,
JAPAN 223-8522
{terry, matutani, hunga}@am.ics.keio.ac.jp

Abstract

The power consumption of interconnects is increased as the link bandwidth is improved in PC clusters. In this paper, we propose an on/off link activation method that uses the static analysis of the traffic in order to reduce the power consumption of Ethernet switches while maintaining the performance of PC clusters. When a link whose utilization is low is deactivated, the proposed method renews the VLAN-based paths that avoid it without creating broadcast storms. Since each host does not need to process VLAN tags, the proposed method has advantages in both simple host configuration and high portability. Evaluation results using NAS Parallel Benchmarks show that the proposed method reduces the power consumption of switches by up to 37% without performance degradation.

A new mechanism to deal with process variability in NoC links

Carles Hernández, Federico Silla, Vicente Santonja, and José Duato
Parallel Architecture Group
Universidad Politécnica de Valencia
Camino de Vera s/n, 46022-Valencia, Spain
carherlu@gap.upv.es,{fsilla,visan,jduato}@disca.upv.es

Abstract

Associated with the ever growing integration scale of VLSI technologies is the increase in process variability, which makes silicon devices to become less predictable. In the context of network-on-chip (NoC), this variability affects the maximum frequency that could be sustained by each wire of the link that interconnects two cores in a CMP system.

Reducing the clock frequency so that all wires can properly work is a trivial solution but, as variability increases, this approach causes an unacceptable performance penalty. In this paper, we propose a new technique to deal with the effects of variability on the links of the NoC that interconnects cores in a CMP system. This technique, called Phit Reduction (PR), retrieves most of the bandwidth still available in links containing wires that are not able to operate at the designed operating frequency. More precisely, our mechanism discards these slow wires and uses all the wires that can work at the design frequency. Two implementations are presented: Local Phit Reduction (LPR), oriented to fabrication processes with very high variability, which requires more hardware but provides higher performance; and Global Phit Reduction (GPR), that requires less additional hardware but is not able to extract all the available bandwidth.

The performance evaluation presented in the paper confirms that LPR obtains good results both for low and high variability scenarios. Moreover, in most of our experiments LPR practically achieves the same performance than the ideal network. On the other hand, GPR is appropriate for systems where whithin-die variations are expected to be low.

Session 8

Software - Data Parallel Programming Frameworks

A framework for efficient and scalable execution of domain-specific templates on GPUs

Narayanan Sundaram[1,2], Anand Raghunathan[1,3] and Srimat T. Chakradhar[1]
[1]NEC Laboratories America, Princeton, NJ, USA
[2]Department of EECS, University of California at Berkeley, CA, USA
[3]School of ECE, Purdue University, IN, USA
narayans@eecs.berkeley.edu, raghunathan@purdue.edu, chak@nec-labs.com

Abstract

Graphics Processing Units (GPUs) have emerged as important players in the transition of the computing industry from sequential to multi- and many-core computing. We propose a software framework for execution of domainspecific parallel templates on GPUs, which simultaneously raises the abstraction level of GPU programming and ensures efficient execution with forward scalability to large data sizes and new GPU platforms. To achieve scalable and efficient GPU execution, our framework focuses on two critical problems that have been largely ignored in previous efforts - processing large data sets that do not fit within the GPU memory, and minimizing data transfers between the host and GPU. Our framework takes domain-specific parallel programming templates that are expressed as parallel operator graphs, and performs operator splitting, offload unit identification, and scheduling of off-loaded computations and data transfers between the host and the GPU, to generate a highly optimized execution plan. Finally, a code generator produces a hybrid CPU/GPU program in accordance with the derived execution plan, that uses lowerlevel frameworks such as CUDA. We have applied the proposed framework to templates from the recognition domain, specifically edge detection kernels and convolutional neural networks that are commonly used in image and video analysis. We present results on two different GPU platforms from NVIDIA (a Tesla C870 GPU computing card and a GeForce 8800 graphics card) that demonstrate 1.7 - 7.8X performance improvements over already accelerated baseline GPU implementations. We also demonstrate scalability to input data sets and application memory footprints of 6GB and 17GB, respectively, on GPU platforms with only 768MB and 1.5GB of memory.

A Cross-Input Adaptive Framework for GPU Program Optimizations

Yixun Liu, Eddy Z. Zhang and Xipeng Shen
Computer Science Department
College of William and Mary
{enjoywm,eddy,xshen}@cs.wm.edu

Abstract

Recent years have seen a trend in using graphic processing units (GPU) as accelerators for general-purpose computing. The inexpensive, single-chip, massively parallel architecture of GPU has evidentially brought factors of speedup to many numerical applications. However, the development of a high-quality GPU application is challenging, due to the large optimization space and complex unpredictable effects of optimizations on GPU program performance.

Recently, several studies have attempted to use empirical search to help the optimization. Although those studies have shown promising results, one important factorłprogram inputsłin the optimization has remained unexplored. In this work, we initiate the exploration in this new dimension. By conducting a series of measurement, we find that the ability to adapt to program inputs is important for some applications to achieve their best performance on GPU. In light of the findings, we develop an input-adaptive optimization framework, namely G-ADAPT, to address the influence by constructing cross-input predictive models for automatically predicting the (near-)optimal configurations for an arbitrary input to a GPU program. The results demonstrate the promise of the framework in serving as a tool to alleviate the productivity bottleneck in GPU programming.

CellMR: A Framework for Supporting MapReduce on Asymmetric Cell-Based Clusters

M. Mustafa Rafique[1], Benjamin Rose[1], Ali R. Butt[1]
[1]Dept. of Computer Science
Virginia Tech.
Blacksburg, Virginia, USA
{mustafa, bar234, butta, dsn}@cs.vt.edu

Dimitrios S. Nikolopoulos[1,2]
[2]Institute of Computer Science
Foundation for Research and Technology Hellas (FORTH)
GR 700 13, Heraklion Crete
dsn@ics.forth.gr

Abstract

The use of asymmetric multi-core processors with on-chip computational accelerators is becoming common in a variety of environments ranging from scientific computing to enterprise applications. The focus of current research has been on making efficient use of individual systems, and porting applications to asymmetric processors. In this paper, we take the next step by investigating the use of multi-core-based systems, especially the popular Cell processor, in a cluster setting. We present CellMR, an efficient and scalable implementation of the MapReduce framework for asymmetric Cell-based clusters. The novelty of CellMR lies in its adoption of a streaming approach to supporting MapReduce, and its adaptive resource scheduling schemes: Instead of allocating workloads to the components once, CellMR slices the input into small work units and streams them to the asymmetric nodes for efficient processing. Moreover, CellMR removes I/O bottlenecks by design, using a number of techniques, such as double-buffering and asynchronous I/O, to maximize cluster performance. Our evaluation of CellMR using typical MapReduce applications shows that it achieves 50.5% better performance compared to the standard non-streaming approach, introduces a very small overhead on the manager irrespective of application input size, scales almost linearly with increasing number of compute nodes (a speedup of 6.9 on average, when using eight nodes compared to a single node), and adapts effectively the parameters of its resource management policy between applications with varying computation density.

Message Passing on Data-Parallel Architectures

Jeff A. Stuart
Department of Computer Science
University of California, Davis
stuart@cs.ucdavis.edu

John D. Owens
Department of Electrical and Computer Engineering
University of California, Davis
jowens@ece.ucdavis.edu

Abstract

This paper explores the challenges in implementing a message passing interface usable on systems with data-parallel processors. As a case study, we design and implement the "DCGN" API on NVIDIA GPUs that is similar to MPI and allows full access to the underlying architecture. We introduce the notion of data-parallel thread-groups as a way to map resources to MPI ranks. We use a method that also allows the data-parallel processors to run autonomously from user-written CPU code. In order to facilitate communication, we use a sleep-based polling system to store and retrieve messages. Unlike previous systems, our method provides both performance and flexibility. By running a test suite of applications with different communication requirements, we find that a tolerable amount of overhead is incurred, somewhere between one and five percent depending on the application, and indicate the locations where this overhead accumulates. We conclude that with innovations in chipsets and drivers, this overhead will be mitigated and provide similar performance to typical CPU-based MPI implementations while providing fully-dynamic communication.

Session 9

Algorithms - Scheduling II

Online time constrained scheduling with penalties

Nicolas Thibault and Christian Laforest
IBISC, Université d'Évry/CNRS, 523 place des Terrasses, 91000 Evry, France.
nicolas.thibault@ibisc.fr,christian.laforest@isima.fr

Abstract

In this paper we prove the (constant) competitiveness of an online algorithm for scheduling jobs on multiple machines, supporting a mechanism of penalties for the scheduler/operator. Our context (online, multiple machines, supporting parameterizable penalties) is more general than in previous existing works. The main contribution of our paper is the (non trivial) analysis of our algorithm. Moreover, with our parameterizable penalties, the operator can find a trade-off between the attractiveness of its system and its own profit (gained with non canceled scheduled jobs).

Minimizing Total Busy Time in Parallel Scheduling with Application to Optical Networks

Michele Flammini[1], Gianpiero Monaco[1], Luca Moscardelli[2], Hadas Shachnai[3],
Mordechai Shalom[4], Tami Tamir [5] and Shmuel Zaks[3]
[1] Department of Computer Science, University of L'Aquila, L'Aquila, Italy.
{flammini,gianpiero.monaco}@di.univaq.it
[2] Department of Science, University of Chieti-Pescara, Pescara, Italy.
moscardelli@sci.unich.it
[3] Computer Science Department, The Technion, Haifa 32000, Israel.
{hadas,zaks}@cs.technion.ac.il
[4] Tel-Hai Academic College, 12210 Upper Gallilee, Israel.
cmshalom@telhai.ac.il
[5] School of Computer Science, The Interdisciplinary Center, Herzliya, Israel.
tami@idc.ac.il

Abstract

We consider a scheduling problem in which a bounded number of jobs can be processed simultaneously by a single machine. The input is a set of n jobs $\mathcal{J} = \{J_1, \ldots, J_n\}$. Each job, J_j, is associated with an interval $[s_j, c_j]$ along which it should be processed. Also given is the parallelism parameter $g \geq 1$, which is the maximal number of jobs that can be processed simultaneously by a single machine. Each machine operates along a contiguous time interval, called its *busy interval*, which contains all the intervals corresponding to the jobs it processes. The goal is to assign the jobs to machines such that the total busy time of the machines is minimized.

The problem is known to be NP-hard already for $g = 2$. We present a 4-approximation algorithm for general instances, and approximation algorithms with improved ratios for instances with bounded lengths, for instances where any two intervals intersect, and for instances where no interval is properly contained in another. Our study has important application in optimizing the switching costs of optical networks.

Energy Minimization for Periodic Real-Time Tasks on Heterogeneous Processing Units

Jian-Jia Chen, Andreas Schranzhofer and Lothar Thiele
Computer Engineering and Networks Laboratory (TIK)
Swiss Federal Institute of Technology (ETH) Zurich, Switzerland
{jchen, schranzhofer, thiele}@tik.ee.ethz.ch

Abstract

Adopting multiple processing units to enhance the computing capability or reduce the power consumption has been widely accepted for designing modern computing systems. Such configurations impose challenges on energy efficiency in hardware and software implementations. This work targets power-aware and energy-efficient task partitioning and processing unit allocation for periodic real-time tasks on a platform with a library of applicable processing unit types. Each processing unit type has its own power consumption characteristics for maintaining its activeness and executing jobs. This paper proposes polynomial-time algorithms for energy-aware task partitioning and processing unit allocation. The proposed algorithms first decide how to assign tasks onto processing unit types to minimize the energy consumption, and then allocate processing units to fit the demands. The proposed algorithms for systems without limitation on the allocated processing units are shown with an $(m+1)$-approximation factor, where m is the number of the available processing unit types. For systems with limitation on the number of the allocated processing units, the proposed algorithm is shown with bounded resource augmentation on the limited number of allocated units. Experimental results show that the proposed algorithms are effective for the minimization of the overall energy consumption.

Multi-Users Scheduling in Parallel Systems

Erik Saule and Denis Trystram
LIG, Grenoble University
51, avenue J. Kuntzmann
38330 Montbonnot St. Martin, France
{erik.saule,denis.trystram}@imag.fr

Abstract

We are interested in this paper to study scheduling problems in systems where many users compete to perform their respective jobs on shared parallel resources. Each user has specific needs or wishes for computing his/her jobs expressed as a function to optimize (among maximum completion time, sum of completion times and sum of weighted completion times). Such problems have been mainly studied through Game Theory. In this work, we focus on solving the problem by optimizing simultaneously each user's objective function independently using classical combinatorial optimization techniques. Some results have already been proposed for two users on a single computing resource. However, no generic combinatorial method is known for many objectives.

The analysis proposed in this paper concerns an arbitrarily fixed number of users and is not restricted to a single resource. We first derive inapproximability bounds; then we analyze several greedy heuristics whose approximation ratios are close to these bounds. However, they remain high since they are linear in the number of users. We provide a deeper analysis which shows that a slightly modified version of the algorithm is a constant approximation of a Pareto-optimal solution.

Session 10

Applications - Graph and String Applications

Input-independent, Scalable and Fast String Matching on the Cray XMT

Oreste Villa[1], Daniel Chavarría-Miranda[1], and Kristyn Maschhoff[2]
[1]High-Performance Computing
Pacific Northwest National Laboratory
{oreste.villa, daniel.chavarria}@pnl.gov
[2]Cray, Inc. kristyn@cray.com

Abstract

String searching is at the core of many security and network applications like search engines, intrusion detection systems, virus scanners and spam filters. The growing size of on-line content and the increasing wire speeds push the need for fast, and often real-time, string searching solutions. For these conditions, many software implementations (if not all) targeting conventional cache-based microprocessors do not perform well. They either exhibit overall low performance or exhibit highly variable performance depending on the types of inputs. For this reason, real-time state of the art solutions rely on the use of either custom hardware or Field- Programmable Gate Arrays (FPGAs) at the expense of overall system flexibility and programmability.

This paper presents a software based implementation of the Aho-Corasick string searching algorithm on the Cray XMT multithreaded shared memory machine. Our solution relies on the particular features of the XMT architecture and on several algorithmic strategies: it is fast, scalable and its performance is virtually contentindependent. On a 128-processor Cray XMT, it reaches a scanning speed of $\approx$ 28 Gbps with a performance variability below 10%. In the 10 Gbps performance range, variability is below 2.5%. By comparison, an Intel dual-socket, 8-core system running at 2.66 GHz achieves a peak performance which varies from 500 Mbps to 10 Gbps depending on the type of input and dictionary size.

Compact Graph Representations and Parallel Connectivity Algorithms for Massive Dynamic Network Analysis

Kamesh Madduri
Computational Research Division
Lawrence Berkeley National Laboratory
Berkeley, USA 94703

David A. Bader
College of Computing
Georgia Institute of Technology
Atlanta, USA 30332

Abstract

Graph-theoretic abstractions are extensively used to analyze massive data sets. Temporal data streams from socio-economic interactions, social networking web sites, communication traffic, and scientific computing can be intuitively modeled as graphs. We present the first study of novel high-performance combinatorial techniques for analyzing large-scale information networks, encapsulating dynamic interaction data in the order of billions of entities. We present new data structures to represent dynamic interaction networks, and discuss algorithms for processing parallel insertions and deletions of edges in small-world networks. With these new approaches, we achieve an average performance rate of 25 million structural updates per second and a parallel speedup of nearly 28 on a 64-way Sun UltraSPARC T2 multicore processor, for insertions and deletions to a small-world network of 33.5 million vertices and 268 million edges. We also design parallel implementations of fundamental dynamic graph kernels related to connectivity and centrality queries. Our implementations are freely distributed as part of the open-source SNAP (Small-world Network Analysis and Partitioning) complex network analysis framework.

Transitive Closure on the Cell Broadband Engine: A study on Self-Scheduling in a Multicore Processor

Sudhir Vinjamuri
Department of Electrical Engineering
University of Southern California
3740 McClintock Avenue EEB-244
Los Angeles USA 90007
sudhir.vinjamuri@usc.edu

Viktor K. Prasanna
Department of Electrical Engineering
University of Southern California
3740 McClintock Avenue EEB-200C
Los Angeles USA 90007
prasanna@ganges.usc.edu

Abstract

In this paper, we present a mappingmethodology and optimizations for solving transitive closure on the Cell multicore processor. Using our approach, it is possible to achieve near peak performance for transitive closure on the Cell processor. We first parallelize the Standard Floyd Warshall algorithm and show through analysis and experimental results that data communication is a bottleneck for performance and scalability. We parallelize a cache optimized version of Floyd Warshall algorithm to remove the memory bottleneck. As is the case with several scientific computing and industrial applications on a multicore processor, synchronization and scheduling of the cores plays a crucial role in determining the performance of this algorithm. We define a self-scheduling mechanism for the cores of a multicore processor and design a self-scheduler for Blocked Floyd Warshall algorithm on the Cell multicore processor to remove the scheduling bottleneck. We also present optimizations in scheduling order to remove synchronization points. Our implementations achieved up to 78GFLOPS.

Parallel Short Sequence Mapping for High Throughput Genome Sequencing

Doruk Bozdağ
The Ohio State University
Dept. of Biomedical Informatics
Columbus, OH 43210, USA
bozdag.1@osu.edu

Catalin C. Barbacioru
Applied Biosystems
850 Lincoln Center Drive
Foster City, CA 94404, USA
catalin@appliedbiosystems.com

Umit V. Catalyurek
The Ohio State University
Dept. of Biomedical Informatics
Dept. of Electrical & Computer Eng.
Columbus, OH 43210, USA
umit@bmi.osu.edu

Abstract

With the advent of next-generation high throughput sequencing instruments, large volumes of short sequence data are generated at an unprecedented rate. Processing and analyzing these massive data requires overcoming several challenges including mapping of generated short sequences to a reference genome. This computationally intensive process takes time on the order of days using existing sequential techniques on large scale datasets. In this work, we propose six parallelization methods to speedup short sequence mapping and to reduce the execution time under just a few hours for such large datasets. We comparatively present these methods and give theoretical cost models for each method. Experimental results on real datasets demonstrate the effectiveness of the parallel methods and indicate that the cost models help accurate estimation of parallel execution time. Based on these cost models we implemented a selection function to predict the best method for a given scenario. To the best of our knowledge this is the first study on parallelization of short sequence mapping problem.

Session 11

Architecture - Networks and Interconnects

TupleQ: Fully-Asynchronous and Zero-Copy MPI over InfiniBand

Matthew J. Koop, Jaidev K. Sridhar and Dhabaleswar K. Panda
Department of Computer Science and Engineering, The Ohio State University
{koop, sridharj, panda}@cse.ohio-state.edu

Abstract

The Message Passing Interface (MPI) is the defacto standard for parallel programming. As system scales increase, application writers often try to increase the overlap of communication and computation. Unfortunately, even on offloaded hardware such as InfiniBand, performance is not improved since the underlying protocols within MPI implementation require control messages that prevent overlap without expensive threads.

In this work we propose a fully-asynchronous and zerocopy design to allow full overlap of communication and computation. We design TupleQ with novel use of InfiniBand eXtended Reliable Connection (XRC) receive queues to allow zero-copy and asynchronous transfers for all message sizes. Our evaluation on 64 tasks reveals significant performance gains. By leveraging the network hardware we are able to provide fully-asynchronous progress. We show overlap of nearly 100% for all message sizes, compared to 0% for the traditional RPUT and RGET protocols. We also show a 27% improvement for NAS SP using our design over the existing designs.

Disjoint-Path Routing: Efficient Communication for Streaming Applications

DaeHo Seo
Intel Corporation
Austin, TX, USA
daeho.seo@intel.com

Mithuna Thottethodi
School of Electrical and Computer Engineering
Purdue University
West Lafayette, IN, USA
mithuna@purdue.edu

Abstract

Streaming is emerging as an important programming model for multicores. Streaming provides an elegant way to express task decomposition and inter-task communication, while hiding laborious orchestration details such as load balancing, assignment (of stream computation to nodes) and computation/communication scheduling from the programmer. This paper develops a novel communication optimization for streaming applications based on the observation that streaming computations typically involve large, systematic data transfers between known communicating pairs of nodes over extended periods of time. From the above observation, we advocate a family of routing algorithms that expend some over overheads to compute disjoint paths for stream communication. Disjoint-path routing is an attractive design point because (a) the overheads of discovering disjoint paths are amortized over large periods of time and (b) the benefits of disjoint path routing are significant for bandwidth-sensitive streaming applications. We develop one instance of disjoint-path routing called tentacle routing C a backtracking, besteffort technique. On a 4x4 (6x6) system, tentacle routing results in 55% (84%) and 28% (41%) mean throughput improvement for high-network-contention streaming applications, and for all streaming applications, respectively.

Performance Analysis of Optical Packet Switches Enhanced with Electronic Buffering

Zhenghao Zhang
Computer Science Department
Florida State University
Tallahassee, FL 32306, USA
zzhang@cs.fsu.edu

Yuanyuan Yang
Dept. Electrical & Computer Engineering
Stony Brook University
Stony Brook, NY 11794, USA
yang@ece.sunysb.edu

Abstract

Optical networks with Wavelength Division Multiplexing (WDM), especially Optical Packet Switching (OPS) networks, have attracted much attention in recent years. However, OPS is still not yet ready for deployment, which is mainly because of its high packet loss ratio at the switching nodes. Since it is very difficult to reduce the loss ratio to an acceptable level by only using all-optical methods, in this paper, we propose a new type of optical switching scheme for OPS which combines optical switching with electronic buffering. In the proposed scheme, the arrived packets that do not cause contentions are switched to the output fibers directly; other packets are switched to *shared* receivers and converted to electronic signals and will be stored in the buffer until being sent out by *shared* transmitters. We focus on performance analysis of the switch, and with both analytical models and simulations, we show that to dramatically improve the performance of the switch, for example, reducing the packet loss ratio from 10^{-2} to close to 10^{-6}, very few receivers and transmitters are needed to be added to the switch. Therefore, we believe that the proposed switching scheme can greatly improve the practicability of OPS networks.

An Approach for Matching Communication Patterns in Parallel Applications

Chao Ma[1,2], Yong Meng Teo[1,4], Verdi March[1,4], Naixue Xiong[2],
Ioana Romelia Pop[1,3], Yan Xiang He[2] and Simon See[4]
[1]Department of Computer Science, National University of Singapore
[2]College of Computer Science & Technology, Wuhan University
[3]Faculty of Automatic Control and Computer, Politechnica University of Bucharest
[4]Asia-Pacific Science and Technology Center, Sun Microsystems, Inc.
teoym@comp.nus.edu.sg

Abstract

Interprocessor communication is an important factor in determining the performance scalability of parallel systems. The communication requirements of a parallel application can be quantified to understand its communication pattern and communication pattern similarities among applications can be determined. This is essential for the efficient mapping of applications on parallel systems and leads to better interprocessor communication implementation among others. This paper proposes a methodology to compare the communication pattern of distributed-memory programs. Communication correlation coefficient quantifies the degree of similarity between two applications based on the communication metrics selected to characterize the applications. To capture the network topology requirements, we extract the communication graph of each applications and quantities this similarity. We apply this methodology to four applications in the NAS parallel benchmark suite and evaluate the communication patterns by studying the effects of varying problem size and the number of logical processes (LPs).

Session 12

Software - I/O and File Systems

Adaptable, Metadata Rich IO Methods for Portable High Performance IO

Jay Lofstead, Fang Zheng and Karsten Schwan
College of Computing
Georgia Institute of Technology
Atlanta, Georgia
{lofstead,schwan}@cc.gatech.edu,fzheng8@mail.gatech.edu

Scott Klasky
Oak Ridge National Laboratory
Oak Ridge, Tennessee
klasky@ornl.gov

Abstract

Since IO performance on HPC machines strongly depends on machine characteristics and configuration, it is important to carefully tune IO libraries and make good use of appropriate library APIs. For instance, on current petascale machines, independent IO tends to outperform collective IO, in part due to bottlenecks at the metadata server. The problem is exacerbated by scaling issues, since each IO library scales differently on each machine, and typically, operates efficiently to different levels of scaling on different machines. With scientific codes being run on a variety of HPC resources, efficient code execution requires us to address three important issues: (1) end users should be able to select the most efficient IO methods for their codes, with minimal effort in terms of code updates or alterations; (2) such performance-driven choices should not prevent data from being stored in the desired file formats, since those are crucial for later data analysis; and (3) it is important to have efficient ways of identifying and selecting certain data for analysis, to help end users cope with the flood of data produced by high end codes. This paper employs ADIOS, the ADaptable IO System, as an IO API to address (1)-(3) above. Concerning (1), ADIOS makes it possible to independently select the IO methods being used by each grouping of data in an application, so that end users can use those IO methods that exhibit best performance based on both IO patterns and the underlying hardware. In this paper, we also use this facility of ADIOS to experimentally evaluate on petascale machines alternative methods for high performance IO. Specific examples studied include methods that use strong file consistency vs. delayed parallel data consistency, as that provided by MPI-IO or POSIX IO. Concerning (2), to avoid linking IO methods to specific file formats and attain high IO performance, ADIOS introduces an efficient intermediate file format, termed BP, which can be converted, at small cost, to the standard file formats used by analysis tools, such as NetCDF and HDF-5. Concerning (3), associated with BP are efficient methods for data characterization, which compute attributes that can be used to identify data sets without having to inspect or analyze the entire data contents of large files.

Small-File Access in Parallel File Systems

Philip Carns, Sam Lang and Robert Ross
Mathematics and Computer Science Division
Argonne National Laboratory
Argonne, IL 60439
{carns,slang,rross}@mcs.anl.gov

Murali Vilayannur
VMware Inc.
3401 Hillview Ave.
Palo Alto, CA 94304
muraliv@vmware.com

Julian Kunkel and Thomas Ludwig
Institute of Computer Science
University of Heidelberg
{Julian.Kunkel,Thomas.Ludwig} @Informatik.uni-heidelberg.de

Abstract

Today's computational science demands have resulted in ever larger parallel computers, and storage systems have grown to match these demands. Parallel file systems used in this environment are increasingly specialized to extract the highest possible performance for large I/O operations, at the expense of other potential workloads. While some applications have adapted to I/O best practices and can obtain good performance on these systems, the natural I/O patterns of many applications result in generation of many small files. These applications are not well served by current parallel file systems at very large scale.

This paper describes five techniques for optimizing small-file access in parallel file systems for very large scale systems. These five techniques are all implemented in a single parallel file system (PVFS) and then systematically assessed on two test platforms. A microbenchmark and the mdtest benchmark are used to evaluate the optimizations at an unprecedented scale. We observe as much as a 905% improvement in small-file create rates, 1,106% improvement in small-file stat rates, and 727% improvement in small-file removal rates, compared to a baseline PVFS configuration on a leadership computing platform using 16,384 cores.

Making Resonance a Common Case: A High-Performance Implementation of Collective I/O on Parallel File Systems

Xuechen Zhang[1], Song Jiang[1], and Kei Davis[2]

[1]ECE Department
Wayne State University
Detroit, MI 48202, USA
{xczhang,sjiang}@eng.wayne.edu

[2] Computer and Computational Sciences
Los Alamos National Laboratory
Los Alamos, NM 87545, USA
kei.davis@lanl.gov

Abstract

Collective I/O is a widely used technique to improve I/O performance in parallel computing. It can be implemented as a client-based or as a server-based scheme. The client-based implementation is more widely adopted in the MPI-IO software such as ROMIO because of its independence from the storage system configuration and its greater portability. However, existing implementations of client-side collective I/O do not consider the actual pattern of file striping over multiple I/O nodes in the storage system. This can cause a large number of requests for non-sequential data at I/O nodes, substantially degrading I/O performance.

Investigating a surprisingly high I/O throughput achieved when there is an accidental match between a particular request pattern and the data striping pattern on the I/O nodes, we reveal the *resonance* phenomenon as the cause. Exploiting readily available information on data striping from the metadata server in popular file systems such as PVFS2 and Lustre, we design a new collective I/O implementation technique, named as *resonant I/O*, that makes resonance a common case. Resonant I/O rearranges requests from multiple MPI processes according to the presumed data layout on the disks of I/O nodes so that non-sequential access of disk data can be turned into sequential access, significantly improving I/O performance without compromising the independence of a client-based implementation. We have implemented our design in ROMIO. Our experimental results on a small- and medium-scale cluster show that the scheme can increase I/O throughput for some commonly used parallel I/O benchmarks such as *mpi-io-test* and *ior-mpi-io* over the existing implementation of ROMIO by up to 157%, with no scenario demonstrating significantly decreased performance.

Design, Implementation, and Evaluation of Transparent pNFS on Lustre

Weikuan Yu[†] Oleg Drokin[‡] Jeffrey S. Vetter [†]

Computer Science & Mathematics[†]
Oak Ridge National Laboratory
{wyu,vetter}@ornl.gov

Lustre Group [‡]
Sun Microsystems, Inc.
{Oleg.Drokin}@Sun.com

Abstract

Parallel NFS (pNFS) is an emergent open standard for parallelizing data transfer over a variety of I/O protocols. Prototypes of pNFS are actively being developed by industry and academia to examine its viability and possible enhancements. In this paper, we present the design, implementation, and evaluation of lpNFS, a Lustre-based parallel NFS. We achieve our primary objective in designing lpNFS as an enabling technology for transparent pNFS accesses to an opaque Lustre file system. We optimize the data flow paths in lpNFS by using two techniques: (a) fast memory coping for small messages, and (b) page sharing for zero-copy bulk data transfer. Our initial performance evaluation shows that the performance of lpNFS is comparable to that of original Lustre. Given these results, we assert that lpNFS is a promising approach to combining the benefits of pNFS and Lustre, and it exposes the underlying capabilities of Lustre file systems while transparently supporting pNFS clients.

Plenary Session

Best Papers

Crash Fault Detection in Celerating Environments

Srikanth Sastry, Scott M. Pike and Jennifer L. Welch
Department of Computer Science and Engineering
Texas A&M University
College Station, TX 77843-3112, USA
{sastry, pike, welch}@cse.tamu.edu

Abstract

In distributed systems, failure detectors are often provided as a system service to detect process crashes. Failure detectors provide (possibly incorrect) information about process crashes in the system. The eventually perfect failure detector – $\Diamond\mathcal{P}$ – is one such failure detector which can make mistakes initially, but eventually provides perfect information about process crashes. It is widely believed that $\Diamond\mathcal{P}$ can be implemented in partially synchronous systems with unknown upper bounds on message delay and relative process speeds. While this belief happens to be true, previous papers have failed to supply an adequate justification without making additional assumptions that bound absolute process speeds. Such implementations of $\Diamond\mathcal{P}$ have overlooked an important subtlety with respect to measuring the passage of time in celerating environments, wherein absolute process speeds can continually increase or decrease while maintaining bounds on relative process speeds. Existing implementations of $\Diamond\mathcal{P}$ use adaptive timeout mechanisms based on either an action clock or a real-time clock. In non-celerating environments, either clock is fine. However, an infinite number of failure detector mistakes can occur while: (1) using action clocks in accelerating environments, or (2) using real-time clocks in decelerating environments. We provide a much needed justification that $\Diamond\mathcal{P}$ can be implemented in such celerating environments. Our approach is based on bichronal clocks, which are a basic composition of action clocks and real-time clocks. As such, we provide a simple solution to a subtle problem which can be readily adopted to make existing implementations of $\Diamond\mathcal{P}$ robust to process celeration, and maintain a perfect suffix even under system volatility due to hardware upgrades, server overloads, denial-of-service attacks, and such.

HPCC RandomAccess Benchmark for Next Generation Supercomputers

Vikas Aggarwal, Yogish Sabharwal and Rahul Garg
IBM India Research Lab
Plot 4, Block C, Vasant Kunj Inst. Area
New Delhi 110070, India.
vicky.aggarwal@gmail.com,
ysabharwal@in.ibm.com, grahul@us.ibm.com

Philip Heidelberger
IBM T. J. Watson Research Center
1101 Kitchawan Rd, Rt. 134
Yorktown Heights, NY 10598, USA.
philiph@us.ibm.com

Abstract

In this paper we examine the key elements determining the performance of the HPC Challenge RandomAccess benchmark on next generation supercomputers. We find that the performance of this benchmark is closely related to the bisection bandwidth of the underlying communication network, performance of integer divide operation and details of benchmark specifications such as error tolerance and permissible multi-core mapping strategies. We demonstrate that seemingly small and innocuous changes in the benchmark can lead to significantly different system performance. We also present an algorithm to optimize RandomAccess benchmark for multi-core systems. Our algorithm uses aggregation and software routing and balances the load on the cores by specializing each of the cores for one specific routing or update function. This algorithm gives approximately a factor of 3 speedup on the Blue Gene/P system which is based on quad-core nodes.

Exploring the Multiple-GPU Design Space

Dana Schaa and David Kaeli
Department of Electrical and Computer Engineering
Northeastern University
{dschaa, kaeli}@ece.neu.edu

Abstract

Graphics Processing Units (GPUs) have been growing in popularity due to their impressive processing capabilities, and with general purpose programming languages such as NVIDIA's CUDA interface, are becoming the platform of choice in the scientific computing community. Previous studies that used GPUs focused on obtaining significant performance gains from execution on a single GPU. These studies employed low-level, architecture-specific tuning in order to achieve sizeable benefits over multicore CPU execution.

In this paper, we consider the benefits of running on multiple (parallel) GPUs to provide further orders of performance speedup. Our methodology allows developers to accurately predict execution time for GPU applications while varying the number and configuration of the GPUs, and the size of the input data set. This is a natural next step in GPU computing because it allows researchers to determine the most appropriate GPU configuration for an application without having to purchase hardware, or write the code for a multiple-GPU implementation. When used to predict performance on six scientific applications, our framework produces accurate performance estimates (11% difference on average and 40% maximum difference in a single case) for a range of short and long running scientific programs.

Accommodating Bursts in Distributed Stream Processing Systems

Yannis Drougas[1], Vana Kalogeraki[1,2]
[1]Department of Computer Science and Engineering, University of California-Riverside
[2]Department of Informatics, Athens University of Economics and Business
{drougas,vana}@cs.ucr.edu

Abstract

Stream processing systems have become important, as applications like media broadcasting, sensor network monitoring and on-line data analysis increasingly rely on realtime stream processing. Such systems are often challenged by the bursty nature of the applications. In this paper, we present BARRE (Burst Accommodation through Rate REconfiguration), a system to address the problem of bursty data streams in distributed stream processing systems. Upon the emergence of a burst, BARRE dynamically reserves resources dispersed across the nodes of a distributed stream processing system, based on the requirements of each application as well as the resources available on the nodes. Our experimental results over our Synergy distributed stream processing system demonstrate the efficiency of our approach.

Session 13

Algorithms - General Theory

Combinatorial Properties for Efficient Communication in Distributed Networks with Local Interactions

S. Nikoletseas
R.& A. Computer Technology Institute
and University of Patras
Patras, Greece
nikole@cti.gr

C. Raptopoulos
Heinz Nixdorf Institute
University of Paderborn
Paderborn, Germany
raptopox@hni.uni-paderborn.de

P. G. Spirakis
R.& A. Computer Technology Institute
and University of Patras
Patras, Greece
spirakis@cti.gr

Abstract

We investigate random intersection graphs, a combinatorial model that quite accurately abstracts distributed networks with local interactions between nodes blindly sharing critical resources from a limited globally available domain. We study important combinatorial properties (independence and hamiltonicity) of such graphs. These properties relate crucially to algorithmic design for important problems (like secure communication and frequency assignment) in distributed networks characterized by dense, local interactions and resource limitations, such as sensor networks. In particular, we prove that, interestingly, a small constant number of random, resource selections suffices to make the graph hamiltonian and we provide tight evaluations of the independence number of these graphs.

Remote-Spanners: What to Know beyond Neighbors

Philippe Jacquet INRIA
Rocquencourt, France
philippe.jacquet@inria.fr

Laurent Viennot INRIA
Paris, France
laurent.viennot@inria.fr

Abstract

Motivated by the fact that neighbors are generally known in practical routing algorithms, we introduce the notion of remote-spanner. Given an unweighted graph G, a sub-graph H with vertex set $V(H) = V(G)$ is an (α,β)-*remote-spanner* if for each pair of points u and v the distance between u and v in H_u, the graph H augmented by all the edges between u and its neighbors in G, is at most α times the distance between u and v in G plus β. We extend this definition to k-connected graphs by considering the minimum length sum over k disjoint paths as a distance. We then say that an (α,β)-remote-spanner is *k-connecting*.

In this paper, we give distributed algorithms for computing $(1+\varepsilon, 1-2\varepsilon)$-remote-spanners for any $\varepsilon > 0$, k-connecting $(1,0)$-remote-spanners for any $k \geq 1$ (yielding $(1,0)$-remote-spanners for $k = 1$) and 2-connecting $(2,-1)$-remote-spanners. All these algorithms run in constant time for any unweighted input graph. The number of edges obtained for k-connecting $(1,0)$-remote-spanner is within a logarithmic factor from optimal (compared to the best k-connecting $(1,0)$-remote-spanner of the input graph). Interestingly, sparse $(1,0)$-remote-spanners (i.e. preserving exact distances) with $O(n^{4/3})$ edges exist in random unit disk graphs. The number of edges obtained for $(1+\varepsilon, 1-2\varepsilon)$-remote-spanners and 2-connecting $(2,-1)$-remote-spanners is linear if the input graph is the unit ball graph of a doubling metric (even if distances between nodes are unknown). Our methodology consists in characterizing remote-spanners as sub-graphs containing the union of small depth tree sub-graphs dominating nearby nodes. This leads to simple local distributed algorithms.

A Fusion-based Approach for Tolerating Faults in Finite State Machines

Vinit Ogale
Parallel and Distributed Systems Laboratory,
Dept. of Electrical and Computer Engineering,
The University of Texas at Austin.
ogale@ece.utexas.edu

Bharath Balasubramanian
Parallel and Distributed Systems Laboratory,
Dept. of Electrical and Computer Engineering,
The University of Texas at Austin.
balasubr@ece.utexas.edu

Vijay K. Garg
IBM India Research Lab (IRL),
Delhi, India.
vijgarg1@in.ibm.com

Abstract

Given a set of n different deterministic finite state machines (DFSMs) modeling a distributed system, we examine the problem of tolerating f crash or Byzantine faults in such a system. The traditional approach to this problem involves replication and requires $n \cdot f$ backup DFSMs for crash faults and $2 \cdot n \cdot f$ backup DFSMs for Byzantine faults. For example, to tolerate two crash faults in three DFSMs, a replication based technique needs two copies of each of the given DFSMs, resulting in a system with six backup DFSMs. In this paper, we question the optimality of such an approach and present an approach called (f,m)-fusion that permits fewer backups than the replication based approaches. Given n different DFSMs, we examine the problem of tolerating f faults using just m additional DFSMs. We introduce the theory of fusion machines and provide an algorithm to generate backup DFSMs for both crash and Byzantine faults. We have implemented our algorithms in Java and have used them to automaticaly generate backup DFSMs for several examples.

The Weak Mutual Exclusion Problem

Paolo Romano, Luis Rodrigues and Nuno Carvalho
INESC-ID/IST

Abstract

In this paper we define the Weak Mutual Exclusion (WME) problem. Analogously to classical Distributed Mutual Exclusion (DME), WME serializes the accesses to a shared resource. Differently from DME, however, the WME abstraction regulates the access to a *replicated* shared resource, whose copies are locally maintained by every participating process. Also, in WME, processes suspected to have crashed are possibly ejected from the critical section

We prove that, unlike DME, WME is solvable in a partially synchronous model, i.e. a system where the bounds on communication latency and on relative process speeds are not known in advance, or are known but only hold after an unknown time.

Finally we demonstrate that $\diamond P$ is the weakest failure detector for solving WME, and present an algorithm that solves WME using $\diamond P$ with a majority of correct processes.

Session 14

Applications - Data Intensive Applications

Best-Effort Parallel Execution Framework for Recognition and Mining Applications

Jiayuan Meng[†‡], Srimat Chakradhar[†], and Anand Raghunathan[†§]
[†] NEC Laboratories America, Princeton, NJ
[‡] Department of Computer Science, University of Virginia, Charlottesville, VA
[§] School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN

Abstract

Recognition and mining (RM) applications are an emerging class of computing workloads that will be commonly executed on future multi-core and many-core computing platforms. The explosive growth of input data and the use of more sophisticated algorithms in RM applications will ensure, for the foreseeable future, a significant gap between the computational needs of RM applications and the capabilities of rapidly evolving multi- or many-core platforms. To address this gap, we propose a new parallel programming model that inherently embodies the notion of best-effort computing, wherein the underlying parallel computing environment is not expected to be perfect. The proposed best-effort programming model leverages three key characteristics of RM applications: (1) the input data is noisy and it often contains significant redundancy, (2) computations performed on the input data are statistical in nature, and (3) some degree of imprecision in the output is acceptable. As a specific instance of the best-effort parallel programming model, we describe an "iterative-convergence" parallel template, which is used by a significant class of RM applications. We show how best-effort computing can be used to not only reduce computational workload, but to also eliminate dependencies between computations and further increase parallelism. Our experiments on an 8-core machine demonstrate a speed-up of 3.5X and 4.3X for the K-means and GLVQ algorithms, respectively, over a conventional parallel implementation. We also show that there is almost no material impact on the accuracy of results obtained from best-effort implementations in the application context of image segmentation using K-means and eye detection in images using GLVQ.

Multi-Dimensional Characterization of Temporal Data Mining on Graphics Processors

Jeremy Archuleta, Yong Cao, Tom Scogland, Wu-chun Feng
Department of Computer Science, Virginia Tech
Blacksburg, VA, USA
{jsarch, yongcao, njustn, feng}@cs.vt.edu

Abstract

Through the algorithmic design patterns of data parallelism and task parallelism, the graphics processing unit (GPU) offers the potential to vastly accelerate discovery and innovation across a multitude of disciplines. For example, the exponential growth in data volume now presents an obstacle for high-throughput data mining in fields such as neuroscience and bioinformatics. As such, we present a characterization of a MapReduce-based data-mining application on a general-purpose GPU (GPGPU). Using neuroscience as the application vehicle, the results of our multi-dimensional performance evaluation show that a "one-size-fits-all" approach maps poorly across different GPGPU cards. Rather, a high-performance implementation on the GPGPU should factor in the 1) problem size, 2) type of GPU, 3) type of algorithm, and 4) data-access method when determining the type and level of parallelism. To guide the GPGPU programmer towards optimal performance within such a broad design space, we provide eight general performance characterizations of our data-mining application.

A Partition-based Approach to Support Streaming Updates over Persistent Data in an Active Data Warehouse

Abhirup Chakraborty, Ajit Singh
Department of Electrical and Computer Engineering
University of Waterloo, ON, Canada N2L 3G1
abhirupc@ieee.org, a.singh@ece.uwaterloo.ca

Abstract

Active warehousing has emerged in order to meet the high user demands for fresh and up-to-date information. Online refreshment of the source updates introduces processing and disk overheads in the implementation of the warehouse transformations. This paper considers a frequently occurring operator in active warehousing which computes the join between a fast, time varying or bursty update stream S and a persistent disk relation R, using a limited memory. Such a join operation is the crux of a number of common transformations (e.g., surrogate key assignment, duplicate detection etc) in an active data warehouse. We propose a partition-based join algorithm that minimizes the processing overhead, disk overhead and the delay in output tuples. The proposed algorithm exploits the spatio-temporal locality within the update stream, and improves the delays in output tuples by exploiting hot-spots in the range or domain of the joining attributes, and at the same time shares the I/O cost of accessing disk data of relation R over a volume of tuples from update stream S. We present experimental results showing the effectiveness of the proposed algorithm.

Architectural Implications for Spatial Object Association Algorithms

Vijay S. Kumar
Department of Computer Science and Engineering, The Ohio State University
Tahsin Kurc, Joel Saltz
Center for Comprehensive Informatics, Emory University
Ghaleb Abdulla, Scott R. Kohn, Celeste Matarazzo
Center for Applied Scientific Computing, Lawrence Livermore National Laboratory

Abstract

Spatial object association, also referred to as crossmatch of spatial datasets, is the problem of identifying and comparing objects in two or more datasets based on their positions in a common spatial coordinate system. In this work, we evaluate two crossmatch algorithms that are used for astronomical sky surveys, on the following database system architecture configurations: (1) Netezza Performance Server®, a parallel database system with active disk style processing capabilities, (2) MySQL Cluster, a high-throughput network database system, and (3) a hybrid configuration consisting of a collection of independent database system instances with data replication support. Our evaluation provides insights about how architectural characteristics of these systems affect the performance of the spatial crossmatch algorithms. We conducted our study using real use-case scenarios borrowed from a large-scale astronomy application known as the Large Synoptic Survey Telescope (LSST).

Session 15

Architecture - Emerging Architectures and Performance Modeling

vCUDA: GPU Accelerated High Performance Computing in Virtual Machines

Lin Shi, Hao Chen and Jianhua Sun
Advanced Internet and Media Lab
School of Computer and Communications
Hunan University, Chang Sha, 410082, China
{linshi,haochen,jhsun}@aimlab.org

Abstract

This paper describes vCUDA, a GPGPU (General Purpose Graphics Processing Unit) computing solution for virtual machines. vCUDA allows applications executing within virtual machines (VMs) to leverage hardware acceleration, which can be beneficial to the performance of a class of high performance computing (HPC) applications. The key idea in our design is: API call interception and redirection. With API interception and redirection, applications in VMs can access graphics hardware device and achieve high performance computing in a transparent way. We carry out detailed analysis on the performance and overhead of our framework. Our evaluation shows that GPU acceleration for HPC applications in VMs is feasible and competitive with those running in a native, non-virtualized environment. Furthermore, our evaluation also identifies the main cause of overhead in our current framework, and we give some suggestions for future improvement.

Understanding the Design Trade-offs among Current Multicore Systems for Numerical Computations

Seunghwa Kang, David A. Bader and Richard Vuduc
Georgia Institute of Technology
Atlanta, GA, 30332, USA

Abstract

In this paper, we empirically evaluate fundamental design trade-offs among the most recent multicore processors and accelerator technologies. Our primary aim is to aid application designers in better mapping their software to the most suitable architecture, with an additional goal of influencing future computing system design. We specifically examine five architectures, based on: the Intel quad-core Harpertown processor, the AMD quad-core Barcelona processor, the Sony-Toshiba-IBM Cell Broadband Engine processors (both the first-generation chip and the second-generation PowerXCell 8i), and the NVIDIA Tesla C1060 GPU. We illustrate the software implementation process on each platform for a set of widely-used kernels from computational statistics that are simple to reason about; measure and analyze the performance of each implementation; and discuss the impact of different architectural design choices on each implementation.

Parallel Data-Locality Aware Stencil Computations on Modern Micro-Architectures

Matthias Christen
High Performance and
Web Computing Group,
Computer Science Dept.,
University of Basel,
Switzerland
m.christen@unibas.ch

Olaf Schenk
High Performance and
Web Computing Group,
Computer Science Dept.,
University of Basel,
Switzerland
olaf.schenk@unibas.ch

Esra Neufeld
ITIS Foundation,
ETH Zurich,
Switzerland
neufeld@itis.ethz.ch

Peter Messmer
Tech-X Corporation,
Boulder CO,
USA
messmer@txcorp.com

Helmar Burkhart
High Performance and Web Computing Group,
Computer Science Dept.,
University of Basel, Switzerland
helmar.burkhart@unibas.ch

Abstract

Novel micro-architectures including the Cell Broadband Engine Architecture and graphics processing units are attractive platforms for compute-intensive simulations. This paper focuses on stencil computations arising in the context of a biomedical simulation and presents performance benchmarks on both the Cell BE and GPUs and contrasts them with a benchmark on a traditional CPU system.

Due to the low arithmetic intensity of stencil computations, typically only a fraction of the peak performance of the compute hardware is reached. An algorithm is presented, which reduces the bandwidth requirements and thereby improves performance by exploiting temporal locality of the data. We report on performance improvements over CPU implementations.

Performance Projection of HPC Applications Using SPEC CFP2006 Benchmarks

Sameh Sharkawi[1,2], Don DeSota[1], Raj Panda[1], Rajeev Indukuru[1], Stephen Stevens[1]
Valerie Taylor[2] and Xingfu Wu[2]
[1]Systems and Technology Group, IBM, Austin
[2]Department of Computer Science, Texas A&M University
{sssharka, desotad, panda, indukuru, sstevens}@us.ibm.com,
{sss1858, taylor, wuxf}@cs.tamu.edu

Abstract

Performance projections of High Performance Computing (HPC) applications onto various hardware platforms are important for hardware vendors and HPC users. The projections aid hardware vendors in the design of future systems, enable them to compare the application performance across different existing and future systems, and help HPC users with system procurement and application refinements. In this paper, we present a method for projecting the node level performance of HPC applications using published data of industry standard benchmarks, the SPEC CFP2006, and hardware performance counter data from one base machine. In particular, we project performance of eight HPC applications onto four systems, utilizing processors from different vendors, using data from one base machine, the IBM p575. The projected performance of the eight applications was within 7.2% average difference with respect to measured runtimes for IBM POWER6 systems and standard deviation of 5.3%. For two Intel based systems with different micro-architecture and Instruction Set Architecture (ISA) than the base machine, the average projection difference to measured runtimes was 10.5% with standard deviation of 8.2%.

Session 16

Software - Distributed Systems, Scheduling and Memory Management

Work-First and Help-First Scheduling Policies for Async-Finish Task Parallelism

Yi Guo, Rajkishore Barik, Raghavan Raman and Vivek Sarkar
Department of Computer Science
Rice University
{yguo, rajbarik, raghav, vsarkar}@cs.rice.edu

Abstract

Multiple programming models are emerging to address an increased need for dynamic task parallelism in applications for multicore processors and shared-address-space parallel computing. Examples include OpenMP 3.0, Java Concurrency Utilities, Microsoft Task Parallel Library, Intel Thread Building Blocks, Cilk, X10, Chapel, and Fortress. Scheduling algorithms based on work stealing, as embodied in Cilk's implementation of dynamic spawn-sync parallelism, are gaining in popularity but also have inherent limitations. In this paper, we address the problem of efficient and scalable implementation of X10's async-finish task parallelism, which is more general than Cilk's spawn-sync parallelism. We introduce a new work-stealing scheduler with compiler support for async-finish task parallelism that can accommodate both *work-first* and *help-first* scheduling policies. Performance results on two different multicore SMP platforms show significant improvements due to our new work-stealing algorithm compared to the existing *work-sharing* scheduler for X10, and also provide insights on scenarios in which the help-first policy yields better results than the work-first policy and vice versa.

Autonomic management of non-functional concerns in distributed & parallel application programming

Marco Aldinucci
Dept. Computer Science
University of Torino
Torino C Italy
aldinuc@di.unito.it

Marco Danelutto
Dept. Computer Science
University of Pisa
Pisa C Italy
marcod@di.unipi.it

Peter Kilpatrick
Dept. Computer Science
Queen's University of Belfast
Belfast C UK
p.kilpatrick@qub.ac.uk

Abstract

An approach to the management of non-functional concerns in massively parallel and/or distributed architectures that marries parallel programming patterns with autonomic computing is presented. The necessity and suitability of the adoption of autonomic techniques are evidenced. Issues arising in the implementation of autonomic managers taking care of multiple concerns and of coordination among hierarchies of such autonomic managers are discussed. Experimental results are presented that demonstrate the feasibility of the approach.

Scheduling Resizable Parallel Applications

Rajesh Sudarsan and Calvin J. Ribbens
Department of Computer Science
Virginia Tech, Blacksburg, VA 24061-0106
{sudarsar, ribbens}@vt.edu

Abstract

Most conventional parallel job schedulers only support static scheduling thereby restricting schedulers from being able to modify the number of processors allocated to parallel applications at runtime. The drawbacks of static scheduling can be overcome by using scheduling policies that can exploit dynamic resizability in distributed-memory parallel applications and a scheduler that supports these policies. The scheduler must be capable of adding and removing processors from a parallel application at runtime. This ability of a scheduler to resize parallel applications increases the possibilities for parallel schedulers to manage a large cluster. Our ReSHAPE framework includes an application scheduler that supports dynamic resizing of parallel applications. In this paper, we illustrate the impact of dynamic resizability on parallel scheduling. We propose and evaluate new scheduling policies made possible by our ReSHAPE framework. Experimental results show that these scheduling policies significantly improve individual application turn around time as well as overall cluster utilization.

Helgrind$^+$: An Efficient Dynamic Race Detector

Ali Jannesari, Kaibin Bao, Victor Pankratius and Walter F. Tichy
University of Karlsruhe
76131 Karlsruhe, Germany
{jannesari, bao, pankratius, tichy}@ipd.uni-karlsruhe.de

Abstract

Finding synchronization defects is difficult due to nondeterministic orderings of parallel threads. Current tools for detecting synchronization defects tend to miss many data races or produce an overwhelming number of false alarms. In this paper, we describe Helgrind$^+$, a dynamic race detection tool that incorporates correct handling of condition variables and a combination of the lockset algorithm and happens-before relation. We compare our techniques with Intel Thread Checker and the original Helgrind tool on two substantial benchmark suites. Helgrind$^+$ reduces the number of both false negatives (missed races) and false positives. The additional accuracy incurs almost no performance overhead.

Session 17

Algorithms - Wireless Networks

Sensor Network Connectivity with Multiple Directional Antennae of a Given Angular Sum

Binay Bhattacharya, Yuzhuang Hu and Qiaosheng Shi
School of Computing Science, Simon Fraser University
8888 University Drive, Burnaby, BC V5A1S6, Canada
{binay, yhu1, qshi1}@cs.sfu.ca
Evangelos Kranakis
School of Computer Science, Carleton University
1125 Colonel By Drive, Ottawa, Ontario K1S 5B6, Canada
kranakis@scs.carleton.ca
Danny Krizanc
Department of Mathematics and Computer Science, Wesleyan University
Middletown CT 06459, USA.
dkrizanc@wesleyan.edu

Abstract

We investigate the problem of converting sets of sensors into strongly connected networks of sensors using multiple directional antennae. Consider a set S of n points in the plane modeling sensors of an ad hoc network. Each sensor uses a fixed number, say $1 \leq k \leq 5$, of directional antennae modeled as a circular sector with a given spread (or angle) and range (or radius). We give algorithms for orienting the antennae at each sensor so that the resulting directed graph induced by the directed antennae on the nodes is strongly connected. We also study trade-offs between the total angle spread and range for maintaining connectivity.

Unit Disk Graph and Physical Interference Model: Putting Pieces Together

Emmanuelle Lebhar and Zvi Lotker

Abstract

Modeling communications in wireless networks is a challenging task, since it requires a simple mathematical object on which efficient algorithms can be designed but which must also reflect the complex physical constraints inherent in wireless networks, such as interferences, the lack of global knowledge, and purely local computations. As a tractable mathematical object, the unit disk graph (UDG) is a popular model that has enabled the development of efficient algorithms for crucial networking problems. In a ρ-UDG, two nodes are connected if and only if their distance is at most ρ, for some $\rho > 0$. However, such a connectivity requirement is basically not compatible with the reality of wireless networks due to the environment of the nodes as well as the constraints of radio transmission. For this purpose, the signal interference plus noise ratio model (SINR) is the more commonly used model. The SINR model focuses on radio interferences created over the network depending on the distance to transmitters. Nevertheless, due to its complexity, this latter model has been the subject of very few theoretical investigations and lacks of good algorithmic features.

In this paper, we demonstrate how careful scheduling of the nodes enables the two models to be combined to give the benefits of both the algorithmic features of the UDG and the physical validity of the SINR. Precisely, we show that it is possible to emulate a $1/\sqrt{n \ln n}$-UDG that satisfies the constraints of the SINR over any set of n wireless nodes distributed uniformly in a unit square, with only a $O(\ln^3 n)$ time and power stretch factor. The main strength of our contribution lies in the fact that the scheduling is set in a fully distributed way and considers non-uniform power ranges, and it can therefore fit the sensor network setting. Moreover, our scheduling is optimal up to a polylogarithmic factor in terms of throughput capacity according to the lower bound of Gupta and Kumar.

Path-Robust Multi-Channel Wireless Networks

Arnold L. Rosenberg
Dept. of Electrical & Computer Engineering
Colorado State University
Fort Collins, CO 80523, USA

Abstract

A mathematical-plus-conceptual framework is presented for studying problems such as the following. One wants to deploy an n-node multi-channel wireless network $\mathcal{N}$ in an environment that is inaccessible for repair and/or that contains malicious adversaries. (Example: One wants to "harden" a facility — say, a control base or organizational headquarters or hospital or computation center — against "destructive incidents" such as attacks by malicious adversaries or accidents of nature.) Given a (finite) set Ω of topologies that one wants to be able to "overlay" on (the surviving portion of) network $\mathcal{N}$, one wants to design $\mathcal{N}$ to be Ω-*robust,* in the following very strong sense. Even if *any* set of $m < n$ nodes is disabled, one still wants *all* of the surviving $n - m$ nodes to be able to organize themselves (logically, in the manner of an overlay network) into *any* topology $\mathcal{T} \in \Omega$ that has $\leq n - m$ nodes. A mathematical model for multi-channel wireless networks is presented and is used to develop a *scalable, deterministic* strategy for designing networks that are Ω-robust, for a very broad class of sets Ω. The strategy is illustrated for the simple case when Ω is the set of all paths of lengths $\leq n$. The resulting *path-robust* networks: (*a*) *are within a factor of* 2 *of optimal in complexity,* as measured by the number of node-channel access points; (*b*) *enable power-efficient communication,* in that a node's logical neighbors in the overlay path are its physically nearest nodes in the surviving portion of $\mathcal{N}$. It is suggested how the model and approach can extend to a much richer variety of topologies.

Information Spreading in Stationary Markovian Evolving Graphs

Andrea E.F. Clementi
Dipartimento di Matematica
Università di Roma "Tor Vergata"
clementi@mat.uniroma2.it

Angelo Monti
Dipartimento di Informatica
Università di Roma "La Sapienza"
monti@di.uniroma1.it

Francesco Pasquale
Dipartimento di Matematica
Università di Roma "Tor Vergata"
pasquale@mat.uniroma2.it

Riccardo Silvestri
Dipartimento di Informatica
Università di Roma "La Sapienza"
silver@di.uniroma1.it

Markovian evolving graphs are dynamic-graph models where the links among a fixed set of nodes change during time according to an arbitrary Markovian rule. They are extremely general and they can well describe important dynamic-network scenarios.

We study the speed of information spreading in the *stationary phase* by analyzing the completion time of the *flooding mechanism.* We prove a general theorem that establishes an upper bound on flooding time in *any* stationary Markovian evolving graph in terms of its node-expansion properties.

We apply our theorem in two natural and relevant cases of such dynamic graphs: *edge-Markovian evolving graphs* where the probability of existence of any edge at time t depends on the existence (or not) of the same edge at time $t - 1$; *geometric Markovian evolving graphs* where the Markovian behaviour is yielded by n mobile radio stations, with fixed transmission radius, that perform n independent random walks over a square region of the plane. In both cases, the obtained upper bounds are shown to be nearly tight and, in fact, they turn out to be tight for a large range of the values of the input parameters.

Session 18

Applications I - Cluster/Grid/P2P Computing

Multiple Priority Customer Service Guarantees in Cluster Computing

Kaiqi Xiong
Department of Computer Science
Texas A&M University
Commerce, TX 75429
kaiqi_xiong@tamu-commerce.edu

Abstract

Cluster computing is an efficient computing paradigm for solving large-scale computational problems. Resource management is an essential part in such a computing system. A service provider uses computational resources to process a customer's service request. In an effort to maximize a service provider's profit, it becomes commonplace and important to prioritize services in favor of customers who pay higher fees. In this paper, we present an approach for optimal resource management in cluster computing that minimizes the total cost of computer resources owned by a service provider while satisfying multiple priority customer service requirements. Simulation examples show that the proposed approach is efficient and accurate for resource management in a cluster computing system with multiple customer services.

Treat-Before-Trick : Free-riding Prevention for BitTorrent-like Peer-to-Peer Networks

Kyuyong Shin, Douglas S. Reeves and Injong Rhee
Department of Computer Science
North Carolina State University
Raleigh, NC 27695, USA
{kshin2, reeves, rhee}@ncsu.edu

Abstract

In P2P file sharing systems, free-riders who use others resources without sharing their own cause system-wide performance degradation. Existing techniques to counter freeriders are either complex (and thus not widely deployed), or easy to bypass (and therefore not effective). This paper proposes a simple yet highly effective free-rider prevention scheme using (t,n) threshold secret sharing. A peer must upload encrypted file pieces to obtain the subkeys necessary to decrypt a file which has been downloaded, i.e., subkeys are swapped for file pieces. No centralized monitoring or control is required. This scheme is called "treat-beforetrick" (TBeT). TBeT penalizes free-riding with increased file completion times (time to download file and necessary subkeys). TBeT counters known free-riding strategies, incentivizes peers to donate more upload bandwidth, and increases the overall system capacity for compliant peers. TBeT has been implemented as an extension to BitTorrent, and results of experimental evaluation are presented.

A Resource Allocation Approach for Supporting Time-Critical Applications in Grid Environments

Qian Zhu and Gagan Agrawal
Department of Computer Science and Engineering
Ohio State University, Columbus OH 43210
{zhuq,agrawal}@cse.ohio-state.edu

Abstract

There are many grid-based applications where a timely response to an important event is needed. Often such response can require a significant computation and possibly communication, and it can be very challenging to complete it within the time-frame the response is needed. At the same time, there could be application-specific flexibility in the computation that may be desired. We have been developing an *autonomic* middleware targeting this class of applications.

In this paper, we consider the resource allocation problem for such *adaptive* applications, which comprise services with adaptable service parameters in heterogeneous grid environments. Our goal is to optimize a benefit function while meeting a time deadline. We define an *efficiency value* to reflect how effectively a particular service can be executed on a particular node. We have developed a greedy scheduling algorithm, which is based on prioritizing the services considering their impact on the benefit function, and choosing resources using the computed *efficiency values*.

We have carefully evaluated our resource allocation approach using two applications from our target class, a volume rendering application and a Great Lake forecasting application. When compared to the resource allocation performed by an `Optimal` algorithm, which enumerates all mappings, the benefit achieved by our approach was an average of 87% and the average success-rate was 90%. Furthermore, the benefit we obtained was 32% higher than that of the `GrADS` algorithm, an existing approach we compared our approach with.

Session 19

Applications II - Multicore

High-Order Stencil Computations on Multicore Clusters

Liu Peng, Richard Seymour, Ken-ichi Nomura, Rajiv K. Kalia, Aiichiro Nakano and Priya Vashishta
Collaboratory for Advanced Computing and Simulations, Department of Computer Science,
Department of Physics & Astronomy, Department of Chemical Engineering & Material Science,
University of Southern California, Los Angeles, CA 90089-0242, USA
(liupeng, rseymour, knomura, rkalia, anakano, priyav)@usc.edu
Alexander Loddoch, Michael Netzband, William R. Volz and Chap C. Wong
Technical Computing, Chevron ETC, Houston, TX 77002, USA
(loddoch, mknetzband , Bill.Volz, ChapWong)@chevron.com

Abstract

Stencil computation (SC) is of critical importance for broad scientific and engineering applications. However, it is a challenge to optimize complex, highorder SC on emerging clusters of multicore processors. We have developed a hierarchical SC parallelization framework that combines: (1) spatial decomposition based on message passing; (2) multithreading using critical section-free, dual representation; and (3) single-instruction multiple-data (SIMD) parallelism based on various code transformations. Our SIMD transformations include translocated statement fusion, vector composition via shuffle, and vectorized data layout reordering (e.g. matrix transpose), which are combined with traditional optimization techniques such as loop unrolling. We have thereby implemented two SCs of different characteristics—diagonally dominant, lattice Boltzmann method (LBM) for fluid flow simulation and highly off-diagonal (6-th order) finite difference time-domain (FDTD) code for seismic wave propagation—on a Cell Broadband Engine (Cell BE) based system (a cluster of PlayStation3 consoles), a dual Intel quadcore platform, and IBM BlueGene/L and P. We have achieved high inter-node and intra-node (multithreading and SIMD) scalability for the diagonally dominant LBM: Weak-scaling parallel efficiency 0.978 on 131,072 BlueGene/P processors; strong-scaling multithreading efficiency 0.882 on 6 cores of Cell BE; and strong-scaling SIMD efficiency 0.780 using 4-element vector registers of Cell BE. Implementation of the high-order SC, on the contrary, is less efficient due to long-stride memory access and the limited size of the vector register file, which points out the need for further optimizations.

Dynamic Iterations for the Solution of Ordinary Differential Equations on Multicore Processors

Yanan Yu
Computer Science Department
Florida State University
Tallahassee FL 32306, USA
yu@cs.fsu.edu

Ashok Srinivasan
Computer Science Department
Florida State University
Tallahassee FL 32306, USA
asriniva@cs.fsu.edu

Abstract

In the past few years, there has been a trend of providing increased computing power through greater number of cores on a chip, rather than through higher clock speeds. In order to exploit the available computing power, applications need to be parallelized efficiently. We consider the solution of Ordinary Differential Equations (ODE) on multicore processors. Conventional parallelization strategies distribute the state space amongst the processors, and are efficient only when the state space of the ODE system is large. However, users of a desktop system with multicore processors may wish to solve small ODE systems. Dynamic iterations, parallelized along the time domain, appear promising for such applications. However, they have been of limited usefulness because of their slow convergence. They also have a high memory requirement when the number of time steps is large. We propose a hybrid method that combines conventional sequential ODE solvers with dynamic iterations. We show that it has better convergence and also requires less memory. Empirical results show a factor of two to four improvement in performance over an equivalent conventional solver on a single node. The significance of this paper lies in proposing a new method that can enable small ODE systems, possibly with long time spans, to be solved faster on multicore processors.

Efficient Large-Scale Model Checking

Kees Verstoep, Henri E. Bal
Dept. of Computer Science, Fac. of Sciences
VU University, Amsterdam, The Netherlands
{versto,bal}@cs.vu.nl

Jiří Barnat, Luboš Brim
Dept. of Computer Science, Fac. of Informatics
Masaryk University, Brno, Czech Republic
{barnat,brim}@fi.muni.cz

Abstract

Model checking is a popular technique to systematically and automatically verify system properties. Unfortunately, the well-known state explosion problem often limits the extent to which it can be applied to realistic specifications, due to the huge resulting memory requirements. Distributed-memory model checkers exist, but have thus far only been evaluated on small-scale clusters, with mixed results. We examine one well-known distributed model checker, DiVinE, in detail, and show how a number of additional optimizations in its runtime system enable it to efficiently check very demanding problem instances on a large-scale, multi-core compute cluster. We analyze the impact of the distributed algorithms employed, the problem instance characteristics and network overhead. Finally, we show that the model checker can even obtain good performance in a high-bandwidth computational grid environment.

Session 20

Software - Parallel Compilers and Languages

A Scalable Auto-tuning Framework for Compiler Optimization

Ananta Tiwari[1], Chun Chen[2], Jacqueline Chame[3],
Mary Hall[2] and Jeffrey K. Hollingsworth[1]

[1]University of Maryland
Department of Computer Science
College Park, MD 20740
{tiwari, hollings}@cs.umd.edu

[2]University of Utah
School of Computing
Salt Lake City, UT 84112
{chunchen, mhall}@cs.utah.edu

[3]University of Southern California
Information Sciences Institute
Marina del Ray, CA 90292
jchame@isi.edu

Abstract

We describe a scalable and general-purpose framework for auto-tuning compiler-generated code. We combine Active Harmony's parallel search backend with the CHiLL compiler transformation framework to generate in parallel a set of alternative implementations of computation kernels and automatically select the one with the best-performing implementation. The resulting system achieves performance of compiler-generated code comparable to the fully automated version of the ATLAS library for the tested kernels. Performance for various kernels is 1.4 to 3.6 times faster than the native Intel compiler without search. Our search algorithm simultaneously evaluates different combinations of compiler optimizations and converges to solutions in only a few tens of search-steps.

Taking the Heat off Transactions: Dynamic Selection of Pessimistic Concurrency Control

[1,2]Nehir Sönmez [3]Tim Harris [1]Adrián Cristal [1]Osman S. Ünsal [1,2]Mateo Valero
[1]BSC-Microsoft Research Centre
[2]Departament d'Arquitectura de Computadors - Universitat Politècnica de Catalunya, Spain
[3]Microsoft Research, Cambridge, UK
{nehir.sonmez, adrian.cristal, osman.unsal, mateo.valero}@bsc.es, tharris@microsoft.com

Abstract

In this paper we investigate feedback-directed dynamic selection between different implementations of atomic blocks. We initially execute atomic blocks using STM with optimistic concurrency control. At runtime, we identify "hot" variables that cause large numbers of transactions to abort. For these variables we selectively switch to using pessimistic concurrency control, in the hope of deferring transactions until they will be able to run to completion. This trades off a reduction in single-threaded speed (since pessimistic concurrency control is not as streamlined as our optimistic implementation), against a reduced amount of wasted work in aborted transactions. We describe our implementation in the Haskell programming language, and examine its performance with a range of micro-benchmarks and larger programs. We show that our technique is effective at reducing the amount of wasted work, but that for current workloads there is often not enough wasted work for an overall improvement to be possible. As we demonstrate, our technique is not appropriate for some workloads: the extra work introduced by lock-induced deadlock is greater than the wasted work saved from aborted transactions. For other workloads, we show that using mutual exclusion locks for "hot" variables could be preferable to multi-reader locks because mutual exclusion avoids deadlocks caused by concurrent attempts to upgrade to write access.

Packer: an Innovative Space-Time-Efficient Parallel Garbage Collection Algorithm Based on Virtual Spaces

Shaoshan Liu[1], Ligang Wang[2], Xiao-Feng Li[2] and Jean-Luc Gaudiot[1]
[1]EECS, University of California, Irvine
[2]Intel China Research Center
shaoshal@uci.edu, ligang.wang@intel.com, xiao.feng.li@intel.com, gaudiot@uci.edu

Abstract

The fundamental challenge of garbage collector (GC) design is to maximize the recycled space with minimal time overhead. For efficient memory management, in many GC designs the heap is divided into large object space (LOS) and non-large object space (non-LOS). When one of the spaces is full, garbage collection is triggered even though the other space may still have a lot of free room, thus leading to inefficient space utilization. Also, space partitioning in existing GC designs implies different GC algorithms for different spaces. This not only prolongs the pause time of garbage collection, but also makes collection not efficient on multiple spaces. To address these problems, we propose Packer, a space-and-time-efficient parallel garbage collection algorithm based on the novel concept of virtual spaces. Instead of physically dividing the heap into multiple spaces, Packer manages multiple virtual spaces in one physically shared space. With multiple virtual spaces, Packer offers the advantage of efficient memory management. At the same time, with one physically shared space, Packer avoids the problem of inefficient space utilization. To reduce the garbage collection pause time of Packer, we also propose a novel parallelization method that is applicable to multiple virtual spaces. We reduce the compacting GC parallelization problem into a tree traversal parallelization problem, and apply it to both normal and large object compaction.

Concurrent SSA for General Barrier-Synchronized Parallel Programs

Harshit Shah
School of Technology and Computer Science
Tata Institute of Fundamental Research
Mumbai 400005, India
harshit@tcs.tifr.res.in

R. K. Shyamasundar
School of Technology and Computer Science
Tata Institute of Fundamental Research
Mumbai 400005, India
shyam@tcs.tifr.res.in

Pradeep Varma
IBM India Research Laboratory
Plot No. 2, Phase-II, Block C, ISID Institutional Area
Vasant Kunj, New Delhi 110070, India
pvarma@in.ibm.com

Abstract

Static single assignment (SSA) form has been widely studied and used for sequential programs. This form enables many compiler optimizations to be done efficiently. Work on concurrent static single assignment form (CSSA) for concurrent programs is focused on languages that have limited, implicit barriers (e.g., `cobegin`/`coend` and `parallel do`). Recent programming languages for high-performance computing have general features for barrier/phase synchronization – this is essentially a dual of mutual exclusion and arises mainly in constructing synchronous systems from asynchronous systems. X10 is one such language that has features for general purpose barriers. In X10, barriers are provided through features such as *clocks* and *finish*. Since barriers provide explicit synchronization, they offer an opportunity for reducing π interferences needed for CSSA. This paper provides a means for computing improved CSSA form of a program taking advantage of the general barriers present in it. Our algorithm is based on constructing a control-flow graph of the program and flow equations. The efficiency of analysis and optimizations for parallel programs depends on the number and complexity of π assignments in their CSSA representations. We demonstrate that our approach of computing CSSA form for languages supporting general barrier synchronization can improve the precision of intermediate representation for computing global value numbering and loop invariant detection.

Session 21

Algorithms - Self-Stabilization

Optimal Deterministic Self-stabilizing Vertex Coloring in Unidirectional Anonymous Networks

Samuel Bernard[1], Stéphane Devismes[2], Maria Gradinariu Potop-Butucaru[1] and Sébastien Tixeuil[1]
[1]LIP6 - Université Pierre et Marie Curie - Paris, France
{samuel.bernard,maria.gradinariu,sebastien.tixeuil}@lip6.fr
[2]VERIMAG - Université Joseph Fourier - Grenoble, France
stephane.devismes@imag.fr

Abstract

A distributed algorithm is self-stabilizing if after faults and attacks hit the system and place it in some arbitrary global state, the systems recovers from this catastrophic situation without external intervention in finite time. Unidirectional networks preclude many common techniques in self-stabilization from being used, such as preserving local predicates. In this paper, we investigate the intrinsic complexity of achieving self-stabilization in unidirectional anonymous general networks, and focus on the classical vertex coloring problem. Specifically, we prove a lower bound of n states per process (where n is the network size) and a recovery time of at least $n(n-1)/2$ actions in total. We also provide a deterministic algorithm with matching upper bounds that performs in arbitrary unidirectional anonymous graphs.

Self-stabilizing minimum-degree spanning tree within one from the optimal degree

Lélia Blin
Université d'Evry, IBISC, CNRS, France
LIP6-CNRS UMR 7606, France
lelia.blin@ibisc.univ-evry.fr

Maria Gradinariu Potop-Butucaru
INRIA REGAL, France
Univ. Pierre & Marie Curie - Paris 6,
LIP6-CNRS UMR 7606, France
maria.gradinariu@lip6.fr

Stephane Rovedakis
Université d'Evry, IBISC, CNRS, France
stephane.rovedakis@ibisc.univ-evry.fr

Abstract

We propose a self-stabilizing algorithm for constructing a Minimum-Degree Spanning Tree (**MDST**) in undirected networks. Starting from an arbitrary state, our algorithm is guaranteed to converge to a legitimate state describing a spanning tree whose maximum node degree is at most $\Delta^* + 1$, where Δ^* is the minimum possible maximum degree of a spanning tree of the network.

To the best of our knowledge our algorithm is the first self-stabilizing solution for the construction of a minimum-degree spanning tree in undirected graphs. The algorithm uses only local communications (nodes interact only with the neighbors at one hop distance). Moreover, the algorithm is designed to work in any asynchronous message passing network with reliable FIFO channels. Additionally, we use a fine grained atomicity model (i.e. the send/receive atomicity). The time complexity of our solution is $O(mn^2 \log n)$ where m is the number of edges and n is the number of nodes. The memory complexity is $O(\delta \log n)$ in the send-receive atomicity model (δ is the maximal degree of the network).

A snap-stabilizing point-to-point communication protocol in message-switched networks

Alain Cournier
MIS Laboratory,
Université de Picardie Jules Verne
33 rue Saint Leu,
80039 Amiens Cedex 1 (France)
alain.cournier@u-picardie.fr

Swan Dubois
LIP6 - UMR 7606/INRIA Rocquencourt,
Project-team REGAL
Université Pierre et Marie Curie - Paris 6
104 Avenue du Président Kennedy,
75016 Paris (France)
swan.dubois@lip6.fr

Vincent Villain
MIS Laboratory,
Université de Picardie Jules Verne
33 rue Saint Leu,
80039 Amiens Cedex 1 (France)
vincent.villain@u-picardie.fr

Abstract

A *snap-stabilizing* protocol, starting from any configuration, always behaves according to its specification. In this paper, we present a snap-stabilizing protocol to solve the message forwarding problem in a message-switched network. In this problem, we must manage resources of the system to deliver messages to any processor of the network. In this purpose, we use informations given by a routing algorithm. By the context of stabilization (in particular, the system starts in any configuration), these informations can be corrupted. So, the existence of a snap-stabilizing protocol for the message forwarding problem implies that we can ask the system to begin forwarding messages even if routing informations are initially corrupted.

In this paper, we propose a snap-stabilizing algorithm (in the *state model*) for the following specification of the problem:

- Any message can be generated in a finite time.
- Any emitted message will be delivered to its destination once and only once in a finite time.

This implies that our protocol can deliver any emitted message regardless of the state of routing tables in the initial configuration.

An Asynchronous Leader Election Algorithm for Dynamic Networks

Rebecca Ingram
Trinity University

Patrick Shields
Vassar College

Jennifer E. Walter
Vassar College

Jennifer L. Welch
Texas A&M University

Abstract

An algorithm for electing a leader in an asynchronous network with dynamically changing communication topology is presented. The algorithm ensures that, no matter what pattern of topology changes occur, if topology changes cease, then eventually every connected component contains a unique leader. The algorithm combines ideas from the Temporally Ordered Routing Algorithm (TORA) for mobile ad hoc networks with a wave algorithm, all within the framework of a height-based mechanism for reversing the logical direction of communication links. It is proved that in certain well-behaved situations, a new leader is not elected unnecessarily.

Session 22

Applications - Scientific Applications

A Metascalable Computing Framework for Large Spatiotemporal-Scale Atomistic Simulations

Ken-ichi Nomura, Richard Seymour, Weiqiang Wang, Hikmet Dursun, Rajiv K. Kalia, Aiichiro Nakano,
Priya Vashishta
Collaboratory for Advanced Computing and Simulations, Department of Computer Science,
Department of Physics & Astronomy, Department of Chemical Engineering & Material Science,
University of Southern California, Los Angeles, CA 90089-0242, USA
(knomura, rseymour, wangweiq, hdursun, rkalia, anakano, priyav)@usc.edu
Fuyuki Shimojo
Department of Physics, Kumamoto University, Kumamoto 860-8555, Japan
shimojo@kumamoto-u.ac.jp
Lin H. Yang
Physics/H Division, Lawrence Livermore National Laboratory, Livermore, CA 94551, USA
lyang@llnl.gov

Abstract

A metascalable (or ”design once, scale on new architectures”) parallel computing framework has been developed for large spatiotemporal-scale atomistic simulations of materials based on spatiotemporal data locality principles, which is expected to scale on emerging multipetaflops architectures. The framework consists of: (1) an embedded divide-and-conquer (EDC) algorithmic framework based on spatial locality to design linear-scaling algorithms for high complexity problems; (2) a space-time-ensemble parallel (STEP) approach based on temporal locality to predict longtime dynamics, while introducing multiple parallelization axes; and (3) a tunable hierarchical cellular decomposition (HCD) parallelization framework to map these O(N) algorithms onto a multicore cluster based on hybrid implementation combining message passing and critical section-free multithreading. The EDC-STEP-HCD framework exposes maximal concurrency and data locality, thereby achieving: (1) inter-node parallel efficiency well over 0.95 for 218 billion-atom molecular-dynamics and 1.68 trillion electronic-degrees-of-freedom quantum-mechanical simulations on 212,992 IBM BlueGene/L processors (superscalability); (2) high intra-node, multithreading parallel efficiency (nanoscalability); and (3) nearly perfect time/ensemble parallel efficiency (eon-scalability). The spatiotemporal scale covered by MD simulation on a sustained petaflops computer per day (i.e. petaflops·day of computing) is estimated as NT = 2.14 (e.g. N = 2.14 million atoms for T = 1 microseconds).

Scalability Challenges for Massively Parallel AMR Applications

Brian Van Straalen[1], John Shalf[2], Terry Ligocki[1], Noel Keen[1,2] and Woo-Sun Yang[2]
[1]ANAG, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[2]NERSC, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

Abstract

PDE solvers using Adaptive Mesh Refinement on block structured grids are some of the most challeng- ing applications to adapt to massively parallel com- puting environments. We describe optimizations to the Chombo AMR framework that enable it to scale effi- ciently to thousands of processors on the Cray XT4. The optimization process also uncovered OS-related perfor- mance variations that were not explained by conven- tional OS interference benchmarks. Ultimately the variability was traced back to complex interactions between the application, system software, and the memory hier- archy. Once identified, software modifications to con- trol the variability improved performance by 20% and decreased the variation in computation time across pro- cessors by a factor of 3. These newly identified sources of variation will impact many applications and suggest new benchmarks for OS-services be developed.

Parallel Accelerated Cartesian Expansions for Particle Dynamics Simulations

M. Vikram[1], A. Baczewzki[1], B. Shanker[1] and S. Aluru[2]
[1]Dept. of ECE, Michigan State University
East Lansing, MI 48824, USA
vikramr@egr.msu.edu, baczewsk@egr.msu.edu, bshanker@egr.msu.edu
[2]Dept. of ECE, Iowa State University
Ames, IA 50011, USA
aluru@iastate.edu

Abstract

Rapid evaluation of potentials in large physical systems plays a crucial role in several fields and has been an intensely studied topic on parallel computers. Computational methods and associated parallel algorithms tend to vary depending on the potential being computed. Real applications often involve multiple potentials, leading to increased complexity and the need to strike a balance between competing data distribution strategies, ultimately resulting in low parallel efficiencies. In this paper, we present a parallel accelerated Cartesian expansion (PACE) method that enables rapid evaluation of multiple forms of potentials using a common Fast Multipole Method (FMM) type framework. In addition, our framework localizes potential dependent computations to one particular operator, allowing reuse of much of the computation across different potentials. We present an implicitly load balanced and communication efficient parallel algorithm and show that it can integrate multiple potentials, multiple time steps and address dynamically evolving physical systems. We demonstrate the applicability of the method by solving particle dynamics simulations using both long-range and Lennard-Jones potentials with parallel efficiencies of 97% on 512 to 1024 processors.

Parallel Implementation of Irregular Terrain Model on IBM Cell Broadband Engine

Yang Song and Ali Akoglu
Department of Electrical and Computer Engineering
The University of Arizona, Tucson, Arizona USA 85721
{yangsong,akoglu}@ece.arizona.edu

Jeffrey A Rudin
Mercury Computer Systems, Inc., USA
rudin@mc.com

Abstract

Prediction of radio coverage, also known as radio “hearability” requires the prediction of radio propagation loss. The Irregular Terrain Model (ITM) predicts the median attenuation of a radio signal as a function of distance and the variability of the signal in time and in space. Algorithm can be applied to a large amount of engineering problems to make area predictions for applications such as preliminary estimates for system design, surveillance, and land mobile systems. When the radio transmitters are mobile, the radio coverage changes dynamically, taking on a real-time aspect that requires thousands of calculations per second, which can be achieved through the use of recent advances in multicore processor technology. In this study, we evaluate the performance of ITM on IBM Cell Broadband Engine (BE). We first give a brief introduction to the algorithm of ITM and present both the serial and parallel execution manner of its implementation. Then we exploit how to map out the program on the target processor in detail. We choose message queues on Cell BE which offer the simplest possible expression of the algorithm while being able to fully utilize the hardware resources. Full code segment and a complete set of terrain profiles fit into each processing element without the need for further partitioning. Communications and memory management overhead is minimal and we achieve 90.2% processor utilization with 7.9x speed up compared to serial version. Through our experimental studies, we show that the program is scalable and suits very well for implementing on the CELL BE architecture based on the granularity of computation kernels and memory footprint of the algorithm.

Session 23

Software - Communications Systems

Phaser Accumulators: a New Reduction Construct for Dynamic Parallelism

J. Shirako, D. M. Peixotto, V. Sarkar and W. N. Scherer III
Department of Computer Science, Rice University
{shirako,dmp,vsarkar,scherer}@rice.edu

Abstract

A reduction is a computation in which a common operation, such as a sum, is to be performed across multiple pieces of data, each supplied by a separate task. We introduce *phaser accumulators*, a new reduction construct that meshes seamlessly with phasers to support dynamic parallelism in a phased (iterative) setting. By separating reduction computations into the parts of sending data, performing the computation itself, and retrieving the result, we enable overlap of communication and computation in a manner analogous to that of split-phase barriers. Additionally, this separation enables exploration of implementation strategies that differ as to when the reduction itself is performed: eagerly when the data is supplied, or lazily when a synchronization point is reached.

We implement accumulators as extensions to phasers in the Habanero dialect of the X10 programming language. Performance evaluations of the EPCC Syncbench, Spectral-norm, and CG benchmarks on AMD Opteron, Intel Xeon, and Sun UltraSPARC T2 multicore SMPs show superior performance and scalability over OpenMP reductions (on two platforms) and X10 code (on three platforms) written with atomic blocks, with improvements of up to 2.5× on the Opteron and 14.9× on the UltraSPARC T2 relative to OpenMP and 16.5× on the Opteron, 26.3× on the Xeon and 94.8× on the UltraSPARC T2 relative to X10 atomic blocks. To the best of our knowledge, no prior reduction construct supports the dynamic parallelism and asynchronous capabilities of phaser accumulators.

NewMadeleine: An Efficient Support for High-Performance Networks in MPICH2

Guillaume Mercier, Françis Trahay and Elisabeth Brunet
Bordeaux University
LaBRI - INRIA
F-33405 Talence, France
{mercier,trahay,brunet}@labri.fr
Darius Buntinas
Argonne National Laboratory
Mathematics and Computer Science Division
Argonne, IL 60439, USA
buntinas@mcs.anl.gov

Abstract

This paper describes how the NewMadeleine communication library has been integrated within the MPICH2 MPI implementation and the benefits brought. NewMadeleine is integrated as a Nemesis network module but the upper layers and in particular the CH3 layer has been modified. By doing so, we allow NewMadeleine to fully deliver its performance to an MPI application. NewMadeleine features sophisticated strategies for sending messages and natively supports multirail network configurations, even heterogeneous ones. It also uses a software element called PIOMan that uses multithreading in order to enhance reactivity and create more efficient progress engines. We show various results that prove that NewMadeleine is indeed well suited as a low-level communication library for building MPI implementations.

Scaling Communication-Intensive Applications on BlueGene/P Using One-Sided Communication and Overlap

Rajesh Nishtala[1], Paul H. Hargrove[2], Dan O. Bonachea[1] and Katherine A. Yelick[1,2]
[1]Computer Science Division, College of Engineering
University of California at Berkeley, Berkeley, CA, USA
[2]High Performance Computing Research Department
Lawrence Berkeley National Laboratory, Berkeley, CA, USA
{rajeshn,bonachea,yelick}@cs.berkeley.edu, phhargrove@lbl.gov

Abstract

In earlier work, we showed that the one-sided communication model found in PGAS languages (such as UPC) offers significant advantages in communication efficiency by decoupling data transfer from processor synchronization.

We explore the use of the PGAS model on IBM BlueGene/P, an architecture that combines low-power, quad-core processors with extreme scalability. We demonstrate that the PGAS model, using a new port of the Berkeley UPC compiler and GASNet one-sided communication layer, outperforms two-sided (MPI) communication in both microbenchmarks and a case study of the communication-limited benchmark, NAS FT. We scale the benchmark up to 16,384 cores of the BlueGene/P and demonstrate that UPC consistently outperforms MPI by as much as 66% for some processor configurations and an average of 32%. In addition, the results demonstrate the scalability of the PGAS model and the Berkeley implementation of UPC, the viability of using it on machines with multicore nodes, and the effectiveness of the BG/P communication layer for supporting one-sided communication and PGAS languages.

Dynamic High-Level Scripting in Parallel Applications

Filippo Gioachin and Laxmikant V. Kalé
Department of Computer Science
University of Illinois at Urbana-Champaign
gioachin@uiuc.edu, kale@cs.uiuc.edu

Parallel applications typically run in batch mode, sometimes after long waits in a scheduler queue. In some situations, it would be desirable to interactively add new functionality to the running application, without having to recompile and rerun it. For example, a debugger could upload code to perform consistency checks, or a data analyst could upload code to perform new statistical tests.

This paper presents a scalable technique to dynamically insert code into running parallel applications. We describe and evaluate an implementation of this idea that allows a user to upload Python code into running parallel applications. This uploaded code will run in concert with the main code. We prove the effectiveness of this technique in two case studies: parallel debugging to support introspection and data analysis of large cosmological datasets.

Session 24

Algorithms - Network Algorithms

Map Construction and Exploration by Mobile Agents Scattered in a Dangerous Network

Paola Flocchini[1], Matthew Kellett[2], Peter Mason[2] and Nicola Santoro[3]
[1]School of Information Technology and Engineering, University of Ottawa, Canada
[2]Defence R&D Canada C Ottawa, Government of Canada, Canada
[3]School of Computer Science, Carleton University, Canada
flocchin@site.uottawa.ca, {matthew.kellett, peter.mason}@drdc-rddc.gc.ca, santoro@scs.carleton.ca

Abstract

We consider the map construction problem in a simple, connected graph by a set of mobile computation entities or agents that start from scattered locations throughout the graph. The problem is further complicated by dangerous elements, nodes and links, in the graph that eliminate agents traversing or arriving at them. The agents working in the graph communicate using a limited amount of storage at each node and work asynchronously. We present a deterministic algorithm that solves the exploration and map construction problems. The end result is also a rooted spanning tree and the election of a leader. The total cost of the algorithm is $O(n_s\ m)$ total number of moves, where m is the number of links in the network and ns is the number of safe nodes, improving the existing $O(m^2)$ bound.

A General Approach to Toroidal Mesh Decontamination with Local Immunity

Fabrizio Luccio and Linda Pagli
Dipartimento di Informatica
Università di Pisa, Italy
luccio,pagli@di.unipi.it

Abstract

A General Approach to Toroidal Mesh Decontamination with Local Immunity Fabrizio Luccio and Linda Pagli Dipartimento di Informatica, University of Pisa, Italy

Network decontamination is studied on a k-dimensional torus $(n_1 \times \ldots \times n_k)$, with $k \geq 1$ and $2 \leq n_1 \leq \ldots \leq n_k$. The decontamination is done by a set of agents moving on the net according to a new cleaning model. After an agent leaves from a vertex, this vertex remains uncontaminated as long as m neighbors are uncontaminated. We propose algorithms valid for any $m \leq 2k$ (i.e., up to the vertex degree), proving that $A(k, m)$ synchronous agents suffice, with:

$A(k, 0) = 1$;

$A(k, m) = 2^{m-1}$, for $1 \leq m \leq k + 1$;

$A(k, m) = 2^{2k-m+1}\, n_1\, n_2 \ldots n_{m-k-1}$, for $k + 2 \leq m \leq 2k$.

We also study the total number $M(k, m)$ of agent moves, and prove matching lower bounds on $A(k, m)$ and $M(k, m)$ valid for $m = 3$ and any k, and for all $m \geq k + 1$. Our study can be simply extended to asynchronous functioning.

On the Tradeoff Between Playback Delay and Buffer Space in Streaming

Alix L.H. Chow[1], Leana Golubchik[1,3], Samir Khuller[2] and Yuan Yao[3]
[1]Computer Science Department
University of Southern California, Los Angeles, California 90089,
{lhchow,leana}@usc.edu
[2]Department of Computer Science
University of Maryland, College Park, Maryland 20742
samir@cs.umd.edu
[3]Department of Electrical Engineering-Systems
University of Southern California, Los Angeles, California 90089
yuanyao@usc.edu

Abstract

We consider the following basic question: a source node wishes to stream an ordered sequence of packets to a collection of receivers, which are distributed among a number of clusters. A node may send a packet to another node in its own cluster in one time step, whereas sending a packet to a node in a different cluster takes longer than one time step. Each cluster has two special nodes. We assume that the source and the special nodes in each cluster have a higher capacity and thus can send multiple packets at each step, while all other nodes can both send and receive a packet at each step. We construct two (intra-cluster) data communication schemes, one based on multi-trees (using a collection of interior-disjoint trees) and the other based on hypercubes. We use these approaches to explore the resulting playback delay, buffer space, and communication requirements.

Session 25

Applications - Sorting and FFTs

A Performance Model for Fast Fourier Transform

Yan Li[1], Li Zhao[2], Haibo Lin[1], Alex Chunghen Chow[3] and Jeffrey R Diamond[4]
[1]IBM China Research Lab, {liyancrl, linhb}@cn.ibm.com
[2]Chinese Academy of Science, zhaoli@lsec.cc.ac.cn
[3]IBM Systems Technology Group, achow@us.ibm.com
[4]University of Texas at Austin, jdiamond@cs.utexas.edu

Abstract

The Fast Fourier Transform (FFT) has been considered one of the most important computing algorithms for decades. Its vast application domain makes it an important performance benchmark for new computer architectures. The most common Cooley-Tukey FFT algorithm factorizes a large FFT into a combination of smaller ones. The choice of factors and the order in which they are applied are critical to the ultimate performance of the large FFT.

Traditional hand coded FFT libraries can immediately execute a given sized FFT applying constant heuristics to different kernel sizes, but are not always optimal. FFTW is a popular auto tuning FFT library which searches over the possible factorizations and empirically determines one with the best performance. This search method produces FFT kernels for a given size that are competitive with hand tuned libraries. Unfortunately, the search process for a large size takes hours on real hardware, and is completely infeasible to use when evaluating the FFT performance of new hardware which is still in the simulation phase. It is also less than ideal in environments where it is desirable to have a rapid response to a new sized FFT.

This paper introduces a novel performance model that allows the FFT performance of a given data size to be estimated to within 2% error without ever running the actual FFT. In addition, by recognizing more sophisticated patterns within the computation, this model reduces the search tree size from a permutation of the number of factors to a combination. Because typical FFT sizes contain a large number of similar factors, this effectively reduces the search by an order of magnitude. When given a set of computational kernels, this model can completely characterize the performance of a chosen target architecture by just running some short performance tests on each sized kernel, a process which takes a few minutes or less. Once characterized, an optimal FFT plan for a given input size can be determined in milliseconds instead of hours.

In this paper, we first derive our mathematical model. We then validate its accuracy by using it to improve the performance of a state of the art, hand tuned FFT library by 30%. Finally, we demonstrate its effectiveness by replacing FFTWs own planning stage with our model, resulting in the same FFT performance using FFTW's own kernels in as little as one millionth the computation time.

Designing Efficient Sorting Algorithms for Manycore GPUs

Nadathur Satish
Dept. of Electrical Engineering and Computer Sciences
University of California, Berkeley
Berkeley, CA
nrsatish@eecs.berkeley.edu
Mark Harris and Michael Garland
NVIDIA Corporation
Santa Clara, CA
mharris@nvidia.com, mgarland@nvidia.com

Abstract

We describe the design of high-performance parallel radix sort and merge sort routines for manycore GPUs, taking advantage of the full programmability offered by CUDA. Our radix sort is the fastest GPU sort and our merge sort is the fastest comparison-based sort reported in the literature. Our radix sort is up to 4 times faster than the graphics-based GPUSort and greater than 2 times faster than other CUDA-based radix sorts. It is also 23% faster, on average, than even a very carefully optimized multicore CPU sorting routine.

To achieve this performance, we carefully design our algorithms to expose substantial fine-grained parallelism and decompose the computation into independent tasks that perform minimal global communication. We exploit the high-speed on-chip shared memory provided by NVIDIA's GPU architecture and efficient data-parallel primitives, particularly parallel scan. While targeted at GPUs, these algorithms should also be well-suited for other manycore processors.

Minimizing Startup Costs for Performance-Critical Threading

Anthony M. Castaldo
Department of Computer Science
University of Texas at San Antonio
San Antonio, TX 78249
castaldo@cs.utsa.edu

R. Clint Whaley Department of Computer Science
University of Texas at San Antonio
San Antonio, TX 78249
whaley@cs.utsa.edu

Abstract

Using the well-known ATLAS and LAPACK dense linear algebra libraries, we demonstrate that the parallel management overhead (PMO) can grow with problem size on even statically scheduled parallel programs with minimal task interaction. Therefore, the widely held view that these thread management issues can be ignored in such computationally intensive libraries is wrong, and leads to substantial slowdown on today's machines. We survey several methods for reducing this overhead, the best of which we have not seen in the literature. Finally, we demonstrate that by applying these techniques at the kernel level, performance in applications such as LU and QR factorizations can be improved by almost 40% for small problems, and as much as 15% for large $O(N^3)$ computations. These techniques are completely general, and should yield significant speedup in almost any performance-critical operation. We then show that the lion's share of the remaining parallel inefficiency comes from bus contention, and, in the future work section, outline some promising avenues for further improvement.

Workshop 1
Heterogeneity in Computing Workshop
HCW 2009

Offer-based Scheduling of Deadline-Constrained Bag-of-Tasks Applications for Utility Computing Systems

Marco A. S. Netto and Rajkumar Buyya
Grid Computing and Distributed Systems Laboratory
Department of Computer Science and Software Engineering
The University of Melbourne, Australia
{netto, raj}@csse.unimelb.edu.au

Abstract

Metaschedulers can distribute parts of a Bag-of-Tasks (BoT) application among various resource providers in order to speed up its execution. When providers cannot disclose private information such as their load and computing power, which are usually heterogeneous, the metascheduler needs to make blind scheduling decisions. We propose three policies for composing resource offers to schedule deadline-constrained BoT applications. Offers act as a mechanism in which resource providers expose their interest in executing an entire BoT or only part of it without revealing their load and total computing power. We also evaluate the amount of information resource providers need to expose to the metascheduler and its impact on the scheduling. Our main findings are: (i) offer-based scheduling produces less delay for jobs that cannot meet deadlines in comparison to scheduling based on load availability (i.e. free time slots); thus it is possible to keep providers' load private when scheduling multi-site BoTs; and (ii) if providers publish their total computing power they can have more local jobs meeting deadlines.

Resource-aware allocation strategies for divisible loads on large-scale systems

Anne Benoit[2,4,5],Loris Marchal[1,4,5], Jean-François Pineau[6], Yves Robert[2,4,5] and Frédéric Vivien[3,4,5]
[1]CNRS [2]ENS Lyon[3]INRIA [4]Université de Lyon
[5]LIP laboratory, UMR 5668, ENS Lyon-CNRS-INRIA-UCBL, Lyon, France
[6]LIRMM laboratory, UMR 5506, CNRS-Université Montpellier 2, France
{anne.benoit,loris.marchal,jean-francois.pineau,yves.robert,frederic.vivien}@ens-lyon.fr

Abstract

In this paper, we deal with the large-scale divisible load problem studied in [12]. We show how to reduce this problem to a classical preemptive scheduling problem on a single machine, thereby establishing new complexity results, and providing new approximation algorithms and heuristics that subsume those presented in [12]. We also give some hints on how to extend the results to a more realistic framework where communication costs are taken into account.

Robust Sequential Resource Allocation in Heterogeneous Distributed Systems with Random Compute Node Failures

Vladimir Shestak[1], Edwin K. P. Chong[2,4], Anthony A. Maciejewski[2], and Howard Jay Siegel[2,3]
[1]InfoPrint Solutions Company
6300 Diagonal Highway Boulder, CO 80301
vshestak@us.ibm.com
[2]Department of Electrical and Computer Engineering
[3]Department of Computer Science
[4]Department of Mathematics
Colorado State University
Fort Collins, CO 80523.1373
{echong, aam, hj}@colostate.edu

Abstract

The problem of finding efficient workload distribution techniques is becoming increasingly important today for heterogeneous distributed systems where the availability of compute nodes may change spontaneously over time. Therefore, the resource-allocation policy must be designed to be robust with respect to absence and reemergence of compute nodes so that the performance of the system is maximized. Such a policy is developed in this work, and its performance is evaluated on a model of a dedicated system composed of a limited set of heterogeneous Web servers. Assuming that each HTML request results in a "reward" if completed before its hard deadline, the goal is to maximize a cumulative reward obtained in the system. A failure rate for each server is set relatively high to simulate its operation under harsh conditions. The results demonstrate that the proposed approach based on the concepts of the DermanCLiebermanCRoss theorem outperforms other policies compared in our experiments for inconsistent, processor-consistent, and task-processor-consistent types of heterogeneity.

Revisiting communication performance models for computational clusters

Alexey Lastovetsky, Vladimir Rychkov and Maureen O'Flynn
School of Computer Science and Informatics
University College Dublin
Dublin, Ireland
{alexey.lastovetsky, vladimir.rychkov, maureen.oflynn}@ucd.ie

Abstract

In this paper, we analyze restrictions of traditional models affecting the accuracy of analytical prediction of the execution time of collective communication operations. In particular, we show that the constant and variable contributions of processors and network are not fully separated in these models. Full separation of the contributions that have different nature and arise from different sources will lead to more intuitive and accurate models, but the parameters of such models cannot be estimated from only the point-to-point experiments, which are usually used for traditional models. We are making the point that all the traditional models are designed so that their parameters can be estimated from a set of point-to-point communication experiments. In this paper, we demonstrate that the more intuitive models allow for much more accurate analytical prediction of the execution time of collective communication operations on both homogeneous and heterogeneous clusters. We present in detail one such a point-to-point model and how it can be used for prediction of the execution time of scatter and gather. We describe a set of communication experiments sufficient for accurate estimation of its parameters, and we conclude with presentation of experimental results demonstrating that the model much more accurately predicts the execution time of collective operations than traditional models.

Cost-Benefit Analysis of Cloud Computing versus Desktop Grids

Derrick Kondo[1], Bahman Javadi[1], Paul Malecot[1], Franck Cappello[1], David P. Anderson[2]
[1]INRIA, France, [2]UC Berkeley, USA
Contact author: derrick.kondo@inria.fr

Abstract

Cloud Computing has taken commercial computing by storm. However, adoption of cloud computing platforms and services by the scientific community is in its infancy as the performance and monetary cost-benefits for scientific applications are not perfectly clear. This is especially true for desktop grids (aka volunteer computing) applications. We compare and contrast the performance and monetary cost-benefits of clouds for desktop grid applications, ranging in computational size and storage. We address the following questions: *(i)* What are the performance trade-offs in using one platform over the other? *(ii)* What are the specific resource requirements and monetary costs of creating and deploying applications on each platform? *(iii)* In light of those monetary and performance cost-benefits, how do these platforms compare? *(iv)* Can cloud computing platforms be used in combination with desktop grids to improve cost-effectiveness even further? We examine those questions using performance measurements and monetary expenses of real desktop grids and the Amazon elastic compute cloud.

Robust Data Placement in Urgent Computing Environments

Jason M. Cope [1], Nick Trebon [3], Henry M. Tufo [1], and Pete Beckman [2]
[1] Department of Computer Science, University of Colorado at Boulder
UCB 430, Boulder, CO 80309
{jason.cope, henry.tufo}@colorado.edu
[2] Mathematics and Computer Science Division, Argonne National Laboratory
9700 S. Cass Ave, Argonne, IL 60439
beckman@mcs.anl.gov
[3] Department of Computer Science, University of Chicago
1100 East 58th Street, Chicago, IL 60637
ntrebon@cs.uchicago.edu

Abstract

Distributed urgent computing workflows often require data to be staged between multiple computational resources. Since these workflows execute in shared computing environments where users compete for resource usage, it is necessary to allocate resources that can meet the deadlines associated with time-critical workflows and can tolerate interference from other users. In this paper, we evaluate the use of robust resource selection and scheduling heuristics to improve the execution of tasks and workflows in urgent computing environments that are dependent on the availability of data resources and impacted by interference from less urgent tasks.

A Robust Dynamic Optimization for MPI Alltoall Operation

Hyacinthe Nzigou Mamadou
Department of Informatics
Kyushu University
Momochihama, Fukuoka-shi 814-0001, Japan
hnzigoum@c.csce.kyushu-u.ac.jp

Takeshi Nanri
Research Institute for Information Technology
Kyushu University

Kazuaki Murakami
Faculty of Information Science
and Electrical Engineering
Kyushu University

Abstract

The performance of the Message Passing Interface collective communications is a critical issue to high performance computing widely discussed. In this paper we propose a mechanism that dynamically selects the most efficient MPI Alltoall algorithm for a given system/workload situation. This implementation method starts by grouping the fast algorithms based on respective performance prediction models that were obtained by using the point-to-point model P-LogP. The experiments performed on different parallel machines equipped with Infiniband and Gigabit Ethernet interconnects produced encouraging results, with negligible overhead to find the most appropriate algorithm to carry on the operation. In most cases, the dynamic Alltoall largely outperforms the traditional MPI implementations on different platforms.

Validating Wrekavoc: a Tool for Heterogeneity Emulation

Olivier Dubuisson
Felix Informatique
Laxou, France
olivier.dubuisson@free.fr

Jens Gustedt
INRIA, Nancy - Grand Est
Villers lès Nancy, France
Jens.Gustedt@loria.fr

Emmanuel Jeannot
INRIA, Nancy - Grand Est
Villers lès Nancy, France
Emmanuel.Jeannot@loria.fr

Abstract

Experimental validation and testing of solutions designed for heterogeneous environment is a challenging issue. Wrekavoc is a tool for performing such validation. It runs unmodified applications on emulated multisite heterogeneous platforms. Therefore it downgrades the performance of the nodes (CPU and memory) and the interconnection network in a prescribed way. We report on new strategies to improve the accuracy of the network and memory models. Then, we present an experimental validation of the tool that compares executions of a variety of application code. The comparison of a real heterogeneous platform is done against the emulation of that platform with Wrekavoc. The measurements show that our approach allows for a close reproduction of the real measurements in the emulator.

A Component-Based Framework for the Cell Broadband Engine

Timothy D. R. Hartley and Umit V. Catalyurek
Department of Biomedical Informatics,
Department of Electrical and Computer Engineering,
The Ohio State University, Columbus, OH, USA.
{hartleyt,umit}@bmi.osu.edu

Abstract

With the increasing trend of microprocessor manufacturers to rely on parallelism to increase their products' performance, there is an associated increasing need for simple techniques to leverage this hardware parallelism for good application performance. Unfortunately, many application developers do not have the benefit of long experience in programming parallel and distributed systems. While the filter-stream programming paradigm helps bridge the gap between developers of scientific applications and the performance they need, current and future high-performance multicore processor designs do not have a filter-stream programming library available. This work aims to fill that gap in the software world. This initial DataCutter-Lite implementation defines a powerful, but simple abstraction for carrying out complex computations in a filter-stream model. Additionally, the initial implementation shows that complex architectures such as the Cell Broadband Engine Architecture can make use of the filter-stream model, and give good application performance when doing so.

Portable Builds of HPC Applications on Diverse Target Platforms

Magdalena Slawinska, Jaroslaw Slawinski and Vaidy Sunderam
Dept. of Math and Computer Science
Emory University
400 Dowman Drive, Atlanta, GA 30322, USA
{magg,jaross,vss}@mathcs.emory.edu

Abstract

High-end machines at modern HPC centers are constantly undergoing hardware and system software upgrades – necessitating frequent rebuilds of application codes. The number of possible combinations of compilers, libraries, application build configurations, differing hardware architectures, etc, makes the process of building applications very onerous, requiring expert build knowledge from different domains. Our ongoing Harness Workbench Toolkit (HWT) project aims to foster and streamline the entire build process on heterogeneous computational platforms. This paper focuses on a key research issue of the HWT that regards facilitating and enhancement portability of build systems across multifarious machines, with particular respect to scientific software commonly used in the HPC community. The article presents a novel HWT approach based on the concept of *generic build systems* and *profiles* which encapsulate build knowledge provided independently by relevant experts. The paper describes profiles, the logistics of storing and retrieving build information, and interfacing to user-guided builds. We also report on experiences with applying the HWT approach to two scientific production codes (CPMD, GAMESS) on Cray XT4.

Workshop 2
Reconfigurable Architectures Workshop
RAW 2009

Evaluation of a Multicore Reconfigurable Architecture with Variable Core Sizes

Vu Manh Tuan, Naohiro Katsura, Hiroki Matsutani and Hideharu Amano
Graduate School of Science and Technology, Keio University
3-14-1 Hiyoshi, Kouhoku-ku, Yokohama, 223-8522 Japan

Abstract

A multicore architecture for processors has emerged as a dominant trend in the chip making industry. As reconfigurable devices gradually prove their capability in improving computation power while preserving flexibility, we are examining a multicore reconfigurable architecture consisting of multiple reconfigurable computational cores connected by an interconnection network. Using an NEC Electronics' DRP-1 as a core for the multicore architecture, a comparison with a tile-based architecture is performed by implementing several streaming applications with various versions. By using wider communication channels and assigning more resources for computations, it is possible to improve throughput over implementations for the tile-based architecture. Another evaluation with different core sizes is examined in order to see the effect of core size in a homogeneous multicore system on performance and internal fragmentation. Evaluation results show that the size of core is a trade-off between throughput and resource usage.

ARMLang: A Language and Compiler for Programming Reconfigurable Mesh Many-cores

Heiner Giefers and Marco Platzner
University of Paderborn
{hgiefers, platzner}@upb.de

Abstract

The reconfigurable mesh serves as a theoretical model for massively parallel computing, but has recently been investigated as a practical architecture for many-cores with light-weight, circuit-switched interconnects. There is a lack of programming environments, including languages, compilers, and debuggers for reconfigurable meshes. In this paper, we present the new language ARMLang for the specification of lockstep programs on regular processor arrays, in particular reconfigurable meshes. Lockstep synchronization is achieved by path equalization and barrier synchronization, both of which are supported by the new language. We further discuss the creation of an ARMLang compiler and a simulation environment that allows for debugging and visualization of the parallel programs.

Double Throughput Multiply-Accumulate Unit for FlexCore Processor Enhancements

Tung Thanh Hoang, Magnus Själander and Per Larsson-Edefors
Department of Computer Science and Engineering
Chalmers University of Technology
412 96 Gothenburg, Sweden
{hoangt,hms,perla}@chalmers.se

Abstract

As a simple five-stage General-Purpose Processor (GPP), the *baseline* FlexCore processor has a limited set of datapath units. By utilizing a flexible datapath interconnect and a wide control word, a FlexCore processor is explicitly designed to support integration of special units that, on demand, can accelerate certain data-intensive applications. In this paper, we propose the integration of a novel Double Throughput Multiply-Accumulate (DTMAC) unit, whose different operating modes allow for on-the-fly optimization of computational precision. For the two EEMBC benchmarks considered, the FlexCore processor performance is significantly enhanced when one DTMAC accelerator is included, translating into reduced execution time and energy dissipation. In comparison to the GPP reference, the accelerated FlexCore processor shows a 4.37x improvement in execution time and a 3.92x reduction in energy dissipation, for a benchmark with many consecutive MAC operations.

Energy Benefits of Reconfigurable Hardware for Use in Underwater Sensor Nets

Bridget Benson, Ali Irturk, Junguk Cho and Ryan Kastner
Computer Science and Engineering Department
University of California San Diego
La Jolla, CA, USA
{b1benson, airturk, jucho, kastner}@cs.ucsd.edu

Abstract

Small, dense underwater sensor networks have the potential to greatly improve undersea environmental and structural monitoring. However, few sensor nets exist because commercially available underwater acoustic modems are too costly and energy inefficient to be practical for this applications. Therefore, when designing an acoustic modem for sensor networks, the designer must optimize for low cost and low energy consumption at every level, from the analog electronics, to the signal processing scheme, to the hardware platform. In this paper we focus on the design choice of hardware platform: digital signal processors, microcontrollers, or reconfigurable hardware, to optimize for energy efficiency while keeping costs low. We implement one algorithm used in an acoustic modem design - Matching Pursuits for channel estimation - on all three platforms and perform a design space exploration to compare the timing, power and energy consumption of each implementation. We show that the reconfigurable hardware implementation can provide a maximum of 210X and 52X decrease in energy consumption over the microcontroller and DSP implementations respectively.

A Multiprocessor Self-reconfigurable JPEG2000 Encoder

Antonino Tumeo[1] Simone Borgio[1] Davide Bosisio[1] Matteo Monchiero[2]
Gianluca Palermo[1] Fabrizio Ferrandi[1] Donatella Sciuto[1]

[1]Politecnico di Milano - DEI
Via Ponzio 34/5
20133 Milano, Italy
{tumeo,gpalermo,ferandi,sciuto}@elet.polimi.it

[2]HP Labs
1501 Page Mill Rd.
Palo Alto 94304 CA, USA
matteo.monchiero@hp.com

Abstract

This paper presents a multiprocessor architecture prototype on a Field Programmable Gate Arrays (FPGA) with support for hardware and software multithreading. Thanks to partial dynamic reconfiguration, this system can, at run time, spawn both software and hardware threads, sharing not only the general purpose soft-cores present in the architecture but also area on the FPGA. While on a standard single processor architecture the partial dynamic reconfiguration requires the processor to stop working to instantiate the hardware threads, the proposed solution hides most of the reconfiguration latency through the parallel execution of software threads. We validate our framework on a JPEG 2000 encoder, showing how threads are spawned, executed and joined independently of their hardware or software nature. We also show results confirming that, by using the proposed approach, we are able to hide the reconfiguration time.

Reconfigurable Accelerator for WFS-Based 3D-Audio

Dimitris Theodoropoulos
D.Theodoropoulos@tudelft.nl

Georgi Kuzmanov
G.K.Kuzmanov@tudelft.nl

Georgi Gaydadjiev
g.n.gaydadjiev@tudelft.nl

Computer Engineering Laboratory
EEMCS, TU Delft
P.O. Box 5031, 2600 GA Delft, The Netherlands

Abstract

In this paper, we propose a reconfigurable and scalable hardware accelerator for 3D-audio systems based on the Wave Field Synthesis technology. Previous related work reveals that WFS sound systems are based on using standard PCs. However, two major obstacles are the relative low number of real-time sound sources that can be processed and the high power consumption. The proposed accelerator alleviates these limitations by its performance and energy efficient de- sign. We propose a scalable organization comprising multiple rendering units (RUs), each of them independently processing audio samples. The processing is done in an environment of continuously varying number of sources and speakers. We provide a comprehensive study on the design trade-offs with respect to this multiplicity of sources and speakers. A hardware prototype of our proposal was implemented on a Virtex4FX60 FPGA operating at 200 MHz. A single RU can achieve up to 7x WFS processing speedup compared to a software implementation running on a Pentium D at 3.4 GHz, while consuming, according to Xilinx XPower, approximately 3 W of power only.

A MicroBlaze specific Co-Processor for Real-Time Hyperelliptic Curve Cryptography on Xilinx FPGAs

Alexander Klimm, Oliver Sander and Jürgen Becker
Universität Karlsruhe (TH)
Institut für Technik der Informationsverarbeitung
Vincenz-Prienitz-Str. 1, 76131 Karlsruhe, Germany
{klimm,sander,becker}@itiv.uni-karlsruhe.de

Abstract

A Hardware/Software Codesign approach based on a MicroBlaze softcore processor and a GF2^n-coprocessor module to form a minimal hardware architecture for HECC on low-cost Xilinx FPGAs is described in this paper. Exploiting the features of the MicroBlaze's integrated interfaces instructions are streamed on-demand to the coprocessor to keep the controlflow highly flexible. At the same time the dataflow between hardware and software is minimized. Comparison with previous architectures shows high acceleration of HECC with minor increase in hardware resources. It is demonstrated that this speed-up can be used for countermeasures on algorithmic level against basic side-channel attacks while still keeping real-time constraints.

Implementing Protein Seed-Based Comparison Algorithm on the SGI RASC-100 Platform

Van-Hoa Nguyen
IRISA/INRIA
Rennes, France
vhnguyen@irisa.fr

Alexandre Cornu
IRISA/INRIA
Rennes, France
acornu@irisa.fr

Dominique Lavenier
ENS Cachan Bretagne/IRISA
Rennes, France
lavenier@irisa.fr

Abstract

This paper describes a parallel FPGA implementation of a genomic sequence comparison algorithm for finding similarities between a large set of protein sequences and full genomes. Results comparable to the `tblastn` program from the BLAST family are provided while the computation is improved by a factor 19. The performances are mainly due to the parallelization of a critical code section on the SGI RASC-100 accelerator.

Hardware Accelerated Montecarlo Financial Simulation over Low Cost FPGA Cluster

J. Castillo[1],José L. Bosque[2], E. Castillo[2], P. Huerta[1] and J.I. Martínez[1]
[1]Escuela Técnica Superior de Informática, Universidad Rey Juan Carlos, Madrid, Spain
[2]Departamento de Electrónica y Computadores. Universidad de Cantabria, Santander, Spain
[1]{javier.castillo,pablo.huerta,joseignacio.martinez}@urjc.es
[2]{emilio.castillo, joseluis.bosque}@unican.es

Abstract

The use of computational systems to help making the right investment decisions in financial markets is an open research field where multiple efforts have being carried out during the last few years. The ability of improving the assessment process and being faster than the rest of the players is one of the keys for the success on this competitive scenario. This paper explores different options to accelerate the computation of the option pricing problem (supercomputer, FPGA cluster or GPU) using the Montecarlo method to solve the Black-Scholes formula, and presents a quantitative study of their performance and scalability.

High Performance True Random Number Generator Based on FPGA Block RAMs

Tamas Györfi, Octavian Creţ and Alin Suciu
Technical University of Cluj-Napoca
Computer Science Department
tamasgy89@yahoo.co.uk, Octavian.Cret@cs.utcluj.ro, Alin.Suciu@cs.utcluj.ro

Abstract

This paper presents a new method for creating TRNGs in Xilinx FPGAs. Due to its simplicity and ease of implementation, the design constitutes a valuable alternative to existing methods for creating single-chip TRNGs. Its main advantages are the high throughput, the portability and the low amount of resources it occupies inside the chip. Therefore, it could further extend the use of FPGA chips in cryptography. Our primary source of entropy is a True Dual-Port Block- RAM operating at high frequency, which is used in a special architecture that creates a concurrent write conflict. The paper also describes the practical issues which make it possible to convert that conflict into a strong entropy source. Depending on the users' requirements, it is possible to connect many units of this generator in parallel on a single FPGA device, thus increasing the bit generation throughput up to the Gbps level. The generator has successfully passed the major statistical test batteries.

Design and implementation of the Quarc Network on-Chip

M. Moadeli[1], P. P. Maji[2] and W. Vanderbauwhede[1]
[1]Department of Computing Science
University of Glasgow
Glasgow, UK
{mahmoudm, wim}@dcs.gla.ac.uk
[2]Institute for System Level Integration
Livingston, UK
partha.maji@sli-institute.ac.uk

Abstract

Networks-on-Chip (NoC) have emerged as alternative to buses to provide a packet-switched communication medium for modular development of large Systems-on-Chip. However, to successfully replace its predecessor, the NoC has to be able to efficiently exchange all types of traffic including collective communications. The latter is especially important for e.g. cache updates in multicore systems. The Quarc NoC architecture [9] has been introduced as a Networks-on- Chip which is highly efficient in exchanging all types of traffic including broadcast and multicast. In this paper we present the hardware implementation of the switch architecture and the network adapter (transceiver) of the Quarc NoC. Moreover, the paper presents an analysis and comparison of the cost and performance between the Quarc and the Spidergon NoCs implemented in Verilog targeting the Xilinx Virtex FPGA family. We demonstrate a dramatic improvement in performance over the Spidergon especially for broadcast traffic, at no additional hardware cost.

Modeling Reconfiguration in a FPGA with a Hardwired Network on Chip

Muhammad Aqeel Wahlah[1] and Kees Goossens[12]
[1]Computer Engineering, Delft University of Technology, aqeel@ce.et.tudelft.nl
[2]NXP Semiconductors, The Netherlands, kees.goossens@nxp.com

Abstract

We propose that FPGAs use a *hardwired* network on chip (HWNOC) as a unified interconnect for *functional* communications (data and control) as well as *configuration* (bitstreams for soft IP). In this paper we model such a platform. Using the HWNOC applications mapped on hard or soft IPs are set up and removed using memory-mapped communications. Peer-to-peer streaming data is used to communicate data between IPs, and also to transport configuration bitstreams. The *composable* nature of the HWNOC ensures that applications can be dynamically configured, programmed, and can operate, without affecting other running (real-time) applications. We describe this platform and the steps required for dynamic reconfiguration of IPs. We then model the hardware, i.e. HWNOC and hard and soft IPs, in cycle-accurate transaction-level SystemC. Next, we model its dynamic behavior, including bitstream loading, HWNOC programming, dynamic (re)configuration, clocking, reset, and computation.

A Low Cost and Adaptable Routing Network for Reconfigurable Systems

Ricardo Ferreira and Marcone Laure
Departamento de Informatica
Universidade Federal de Vicosa
Vicosa, Brazil
ricardo@ufv.br
Antonio C. Beck, Thiago Lo, Mateus Rutzig and Luigi Carro
Instituto de Informatica
Universidade Federal do Rio Grande do Sul
Porto Alegre, Brazil
carro@inf.ufrgs.br

Abstract

Nowadays, scalability, parallelism and fault-tolerance are key features to take advantage of last silicon technology advances, and that is why reconfigurable architectures are in the spotlight. However, one of the major problems in designing reconfigurable and parallel processing elements concerns the design of a cost-effective interconnection network. This way, considering that Multistage Interconnection Network (MIN) has been successfully used in several computer system levels and applications in the past, in this work we propose the use of a MIN, at the word level, on a coarse-grained reconfigurable architecture. More precisely, this work presents a novel parallel self-placement and routing mechanism for MIN on the circuit-switching mode. We take into account one-to-one as well as multicast (one-to-many) permutations. Our approach is scalable and it is targeted to be used in run-time environments where dynamic routing among functional units is required. In addition, our algorithm is embedded in the switch structure, and it is independent of the interstage interconnection pattern. Our approach can handle blocking and non-blocking networks, symmetrical or asymmetrical topologies. As case study, we use the proposed technique in a dynamic reconfigurable system, showing a major area reduction of 30% without performance overhead.

Runtime decision of hardware or software execution on a heterogeneous reconfigurable platform

Vlad-Mihai Sima and Koen Bertels
Computer Engineering
Delft University of Technology
Mekelweg 4, 2628 CD Delft, The Netherlands

Abstract

In this paper, we present a runtime optimization targeting the speedup of applications running on a reconfigurable platform supporting the MOLEN programming paradigm. More specifically, for functions that have an execution time dependent on parameters, we propose an online adaptive decision algorithm to determine if the gain of running that function in hardware outweighs the overhead of transferring the parameters, managing the start and stop of the execution and obtaining the result. Our approach is dynamic in the sense it does not rely on compile time information.The algorithm is applied on a real video codec for which a function is implemented in hardware and we show improvements as big as 24% percent can be obtained for the specific kernel. We also determine the overhead and execution time ranges in which this optimisation is usefull and what other factors can influence it.

Impact of Run-Time Reconfiguration on Design and Speed - A Case Study Based on a Grid of Run-Time Reconfigurable Modules inside a FPGA

Jochen Strunk[1], Toni Volkmer[1], Klaus Stephan[1], Wolfgang Rehm[1] and Heiko Schick[2]
[1]Chemnitz University of Technology
Computer Architecture Group
{sjoc,tovo,stekl,rehm}@cs.tu-chemnitz.de
[2]IBM Deutschland Research & Development GmbH
schickhj@de.ibm.com

Abstract

This paper examines the feasibility of utilizing a grid of runtime reconfigurable (RTR) modules on a dynamically and partially reconfigurable (DPR) FPGA. The aim is to create a homogeneous array of RTR regions on a FPGA, which can be reconfigured on demand during run-time. We study its setup, implementation and performance in comparison with its static counterpart. Such a grid of partially reconfigurable regions (PRR) on a FPGA could be used as an accelerator for computers to offload compute kernels or as an enhancement of functionality in the embedded market which uses FPGAs. An in-depth look at the methodology of creating run-time reconfigurable modules and its tools is shown. Due to the lack of the tools in handling hundreds of dynamically reconfigurable regions a framework is presented which supports the user in the creation process of the design. A case study which uses state of the art Xilinx Virtex-5 FPGAs compares the run-time reconfigurable implementation and achievable clock speeds of a grid with up to 47 reconfigurable module regions with its static counterpart. For this examination a high performance module is used, which finds patterns in a bit stream (pattern matcher). This module is replicated for each partially reconfigurable region. Particularly, design considerations for the controller, which manages the modules, are introduced. Beyond this, the paper also addresses further challenges of the implementation of such a RTR grid and limitations of the reconfigurability of Xilinx FPGAs.

System-Level Runtime Mapping Exploration of Reconfigurable Architectures

Kamana Sigdel†, Mark Thompson‡, Andy D. Pimentel‡, Carlo Galuzzi† and Koen Bertels†

†Computer Engineering Laboratory
EEMCS, Delft University of Technology
The Netherlands
{K.Sigdel, K.L.M.Bertels, C.Galuzzi}@tudelft.nl

‡Computer Systems Architecture Group
University of Amsterdam
The Netherlands
{M.Thompson, A.D.Pimentel}@uva.nl

Abstract

Dynamic reconfigurable systems can evolve under various conditions due to changes imposed either by the architecture, or by the applications, or by the environment. In such systems, the design process becomes more sophisticated as all the design decisions have to be optimized in terms of runtime behaviors and values. Runtime mapping exploration allows to explore reconfigurable systems at runtime to optimize task mappings in order to adapt to the changing behavior of the application(s), the architecture, or the environment. Performing such explorations at runtime enables a system to be more efficient in terms of various design constraints such as performance, chip area, power consumption, etc. Towards this goal, in this paper, we present a model that facilitates runtime mapping exploration of reconfigurable architectures. A case study of an MJPEG application shows that the presented model can be used to perform runtime exploration of various functional and non-functional design parameters.

3D FPGA Resource Management and Fragmentation Metric for Hardware Multitasking

J. A. Valero, J. Septién, D. Mozos and H. Mecha
Dpto. Arquitectura de Computadores. Universidad Complutense de Madrid
jseptien@dacya.ucm.es

Abstract

This research work presents a novel proposal to get hardware multitasking in 3D FPGAs. Such architectures are still academic, but recent advances in 3D IC technologies allow foreseeing true 3D FPGAs in the near future. Starting from models for the 3D FPGA and for the tasks, an efficient technique for managing the 3D reconfigurable resources is proposed. This technique is based on a vertex-list structure in order to maintain information about the free space available on the FPGA at a given time moment. Moreover, a novel 3D fragmentation metric, based on cubeness of the free FPGA volume, is explained. And finally, several vertex-selection heuristics, a simpler one based on space adjacency and a more complex one based on space and time adjacency, are explained and their performance compared by some experiments.

RDMS: A Hardware Task Scheduling Algorithm for Reconfigurable Computing

Miaoqing Huang, Harald Simmler, Olivier Serres and Tarek El-Ghazawi
NSF Center for High-Performance Reconfigurable Computing (CHREC)
Department of Electrical and Computer Engineering, The George Washington University
{mqhuang,serres}@gwmail.gwu.edu, {simmler,tarek}@gwu.edu

Abstract

Reconfigurable Computers (RC) can provide significant performance improvement for domain applications. However, wide acceptance of today's RCs among domain scientist is hindered by the complexity of design tools and the required hardware design experience. Recent developments in HW/SW co-design methodologies for these systems provide the ease of use, but they are not comparable in performance to manual co-design. This paper aims at improving the overall performance of hardware tasks assigned to FPGA devices by minimizing both the communication overhead and configuration overhead, which are introduced by using FPGA devices. The proposed Reduced Data Movement Scheduling (RDMS) algorithm takes data dependency among tasks, hardware task resource utilization, and inter-task communication into account during the scheduling process and adopts a dynamic programming approach to reduce the communication between μP and FPGA co-processor and the number of FPGA configurations to a minimum. Compared to two other approaches that consider data dependency and hardware resource utilization only, RDMS algorithm can reduce inter-configuration communication time by 11% and 44% respectively based on simulation using randomly generated data flow graphs. The implementation of RDMS on a real-life application, N-body simulation, verifies the efficiency of RDMS algorithm against other approaches.

Flexible Pipelining Design for Recursive Variable Expansion

Zubair Nawaz, Thomas Marconi, Koen Bertels
Computer Engineering Lab
Delft University of Technology
The Netherlands
{z.nawaz, t.m.thomas, k.l.m.bertels}@tudelft.nl

Todor Stefanov
Leiden Embedded Research Center
Leiden University
The Netherlands
stefanov@liacs.nl

Abstract

Many image and signal processing kernels can be optimized for performance consuming a reasonable area by doing loops parallelization with extensive use of pipelining. This paper presents an automated flexible pipeline design algorithm for our unique acceleration technique called Recursive Variable Expansion. The preliminary experimental results on a kernel of real life application shows comparable performance to hand optimized implementation in reduced design time. This make it a good choice for generating high performance code for kernels which satisfy the given constraints, for which hand optimized codes are not available.

Generation Of Synthetic Floating-point Benchmark Circuits

Thomas C. P. Chau[1], Sam M. H. Ho[2] and Philip H.W. Leong[1]
[1]Department of Computer Science and Engineering,
[2]Department of Electronic Engineering,
The Chinese University of Hong Kong
{cpchau,phwl}@cse.cuhk.edu.hk mhho@ee.cuhk.edu.hk
Peter Zipf and Manfred Glesner
Institute of Microelectronic Systems,
Technische Universitt Darmstadt (TUD)
{zipf,glesner}@mes.tu-darmstadt.de

Abstract

Synthetic Floating-Point (SFP), a synthetic benchmark generator program for floating-point circuits is presented. SFP consists of two independent modules for characterisation and generation. The characterisation module extracts key dataflow statistics of an arbitrary software program. Generation involves producing randomised circuits with desired statistics which are either the output of the characterisation module or directly generated by the user. Using the basic linear algebra subprograms (BLAS) library, Whetstone benchmark and LINPACK benchmark, it is demonstrated that SFP can be used to generate floating-point benchmarks with different user-specified properties as well as benchmarks that mimic real computational programs.

The Radio Virtual Machine: A Solution for SDR Portability and Platform Reconfigurability

Riadh Ben Abdallah, Tanguy Risset and Antoine Fraboulet
Citi, Insa-Lyon,
6 av. des Arts,
69621 Villeurbanne Cedex, France
{riadh.ben-abdallah, tanguy.risset, antoine.fraboulet}@insa-lyon.fr

Yves Durand
CEA-LETI, MINATEC,
17 rue des Martyrs,
F-38054 Grenoble
yves.durand@cea.fr

Abstract

Instead of a single circuit dedicated to a particular physical (PHY) layer standard, a Software Defined Radio (SDR) platform embeds several hardware accelerators which enable it to support different modulation schemes. In this study we propose an architecture for a SDR PHY layer based on the Virtual Machine (VM) concept. Once a program is compiled in a portable byte-code, the VM can then execute it to manage the desired PHY layer. We demonstrate the feasibility of the proposed architecture through a case study and a proof-of-concept implementation.

Scheduling Tasks on Reconfigurable Hardware with a List Scheduler

Justin Teller and Füsun Özgüner
The Ohio State University, ECE Department
Columbus, Ohio 43210, USA
justin.s.teller@intel.com, ozguner@ece.osu.edu

Abstract

In this paper, we propose a static (compile-time) scheduling extension that considers reconfiguration and task execution together when scheduling tasks on reconfigurable hardware, designated as Mutually Exclusive Groups (-MEG), that can be used to extend any static list scheduler. In simulation, using -MEG generates higher quality schedules than those generated by the hardware-software co-scheduler proposed by Mei, et al. [6] and using a single configuration with the base scheduler. Additionally, we propose a dynamic (run-time), fault tolerant scheduler targeted to reconfigurable hardware. We present promising preliminary results using the proposed fault-tolerant dynamic scheduler, showing that application performance gracefully degrades when shrinking the available processing resources.

Software-Like Debugging Methodology for Reconfigurable Platforms

Loic Lagadec and Damien Picard
Architectures et Systèmes, Lab-STICC
Université de Bretagne Occidentale
{loic.lagadec, damien.picard}@univ-brest.fr

Abstract

This paper presents a new debugging methodology for applications targeting reconfigurable platforms. The key issue behind is that bringing software engineering techniques advantages to hardware design would reduce design cycles hence time-to-market. Our high-level synthesis framework supports probes insertion both in the behavioural description of the application and in its hierarchical netlist. Probe status can control the execution, and traced signals can be read back from software. Probes' conditions can be reassigned at runtime tackling the main disadvantage of modifications through re-synthesis and favours short debugging cycles similarly to software development.

Efficient Implementation of QRD-RLS Algorithm using Hardware-Software Co-design

Nupur Lodha[1], Nivesh Rai[1], Aarthy Krishnamurthy[2] and Hrishikesh Venkataraman[1,2]
[1]Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar, India
[2]Performance Engineering Laboratory, School of Electronic Engg., Dublin City University, Ireland
{nupur_lodha, nivesh_r, hrishikesh}@daiict.ac.in

Abstract

This paper presents the implementation of QR Decomposition based Recursive Least Square (QRD-RLS) algorithm on Field Programmable Gate Arrays (FPGA) using hardware-software co-design. The system has been implemented on Xilinx Spartan 3E FPGA with Microblaze soft core processor. The hardware part consists of a custom peripheral that solves the part of the algorithm with higher computational costs and the software part consists of an embedded soft core processor that manages the control functions and rest of the algorithm. The speed and flexibility of FPGAs render them viable for such computationally intensive application. This paper also presents the implementation results and their analysis.

Achieving Network on Chip Fault Tolerance by Adaptive Remapping

Cristinel Ababei
Electrical and Computer Engineering Dept.
North Dakota State University, Fargo ND, USA
cristinel.ababei@ndsu.edu

Rajendra Katti
Electrical and Computer Engineering Dept.
North Dakota State University, Fargo ND, USA
rajendra.katti@ndsu.edu

Abstract

This paper investigates achieving fault tolerance by adaptive remapping in the context of Networks on Chip. The problem of dynamic application remapping is formulated and an efficient algorithm is proposed to address single and multiple PE failures. The new algorithm can be used to dynamically react and recover from PE failures in order to maintain system functionality. The quality of results is similar to that achieved using simulated annealing but in significantly shorter runtimes.

On The Acceptance Tests of Aperiodic Real-Time Tasks for FPGAs

Ahmed A. El Farag, Hatem M. El-Boghdadi and Samir I. Shaheen
Computer Engineering Department, Faculty of Engineering,
Cairo University, Giza, Egypt.

Abstract

Partially Runtime-Reconfigurable devices allow tasks to be placed and removed dynamically at runtime. For real-time systems, tasks have to complete their work and also to meet their deadlines. It is important to decide at arrival time whether the real-time task could meet its deadline or not. Acceptance tests are concerned with determining whether the incoming task can meet its deadline or not. The utilization bound-based acceptance tests (UBTs) -that accept new tasks till certain utilization limit- were proposed to handle single processor with aperiodic tasks which do not apply to the reconfigurable environment. The *rejection ratio* was used as a measure for performance of acceptance tests when all the rejected tasks have the same failure cost. However, when acceptance tests are used, the tasks that are not accepted to run are diverted to other system resources and not actually rejected. In this paper, we modify the utilization bound acceptance test to cope with the reconfigurable platform. Although this test requires simple calculations, it may reject tasks that could have been accepted if they wait in the system. Then, we present an Exact Acceptance Test (EAT) for real-time non-preemptive tasks. This test decides exactly whether the incoming task can meet its deadline or not, at its arrival time. The test depends on a look-ahead placement (LAP) strategy. Finally, we propose a new factor, Acceptance Ratio Of Workload (AROW), to deal with systems that deploy acceptance tests. The AROW is suitable to measure the performance of acceptance tests as it takes into account sizes and computation times of accepted and diverted tasks. The increase in this ratio means an increase in the work done by the accepted tasks and vice versa. We compare the LAP strategy to the UBT and show its performance regarding the diversion ratio and the AROW measure. Our results show that the LAP strategy outperforms the UBT technique by over 80% using the AROW measure and also enhances the *diversion ratio* by around 40%.

High-Level Estimation and Trade-Off Analysis for Adaptive Real-Time Systems

Ingo Sander, Jun Zhu and Axel Jantsch
Royal Institute of Technology
Stockholm, Sweden
{ingo, junz, axel}@kth.se
Andreas Herrholz[†], Philipp A. Hartmann[†] and Wolfgang Nebel[‡]
[†]OFFIS Institute, [‡]Carl v. Ossietzky University
Oldenburg, Germany
{herrholz,hartmann,nebel}@offis.de

Abstract

We propose a novel design estimation method for adaptive streaming applications to be implemented on a partially reconfigurable FPGA. Based on experimental results we enable accurate design cost estimates at an early design stage. Given the size and computation time of a set of configurations, which can be derived through logic synthesis, our method gives estimates for configuration parameters, such as bitstream sizes, computation and reconfiguration times. To fulfil the system's throughput requirements, the required FIFO buffer sizes are then calculated using a hybrid analysis approach based on integer linear programming and simulation. Finally, we are able to calculate the total design cost as the sum of the costs for the FPGA area, the required configuration memory and the FIFO buffers. We demonstrate our method by analysing non-obvious trade-offs for a static and dynamic implementation of adaptivity.

Smith-Waterman Implementation on a FSB-FPGA module using the Intel Accelerator Abstraction Layer

Jeff Allred, Jack Coyne
William Lynch and Vincent Natoli
Stone Ridge Technology
2107 Laurel Bush Road
Bel Air, MD 21015
jallred@stoneridgetechnology.com

Joseph Grecco
Intel Corporation
77 Reed Road
Hudson, MA 01749
joe.grecco@intel.com

Joel Morrissette
Intel Corporation
5300 NE Elam Young Parkway
Hillsboro, OR 97124
joel.morrissette@intel.com

Abstract

The Smith-Waterman algorithm is employed in the field of Bioinformatics to find optimal local alignments of two DNA or protein sequences. It is a classic example of a dynamic programming algorithm. Because it is highly parallel both spatially and temporally and because the fundamental data structure is compact, Smith-Waterman lends itself very well to operation on an FPGA. Here we demonstrate an implementation of this important algorithm in a novel FSB module using the Intel Accelerator Abstraction Layer (AAL), a newly released software middleware layer. We have modified SSEARCH35, an industry standard open-source implementation of the Smith-Waterman algorithm, to transparently introduce a hardware accelerated option to users. We demonstrate performance of nine billion cell updates per second and discuss further opportunities for performance improvement.

High-Level Synthesis with Coarse Grain Reconfigurable Components

George Economakos and Sotiris Xydis
National Technical University of Athens
School of Electrical and Computer Engineering
Microprocessors and Digital Systems Laboratory
Heroon Polytechniou 9, GR-15780 Athens, Greece
geconom@microlab.ntua.gr

Abstract

High-level synthesis is the process of balancing the distribution of RTL components throughout the execution of applications. However, a lot of balancing and optimization opportunities exist below RTL. In this paper, a coarse grain reconfigurable RTL component that combines a multiplier and a number of additions is presented and involved in high-level synthesis. The gate-level synthesis methodology for this component imposes practically no extra hardware than a normal multiplier while involvement in high-level synthesis is performed with a scheduling postprocessor. Following this approach, components that would remain idle in certain control steps are working full-time in two different modes, without any reconfiguration overhead applied to the critical path of the application. The results obtained with different DSP benchmarks show a maximum performance gain of almost 70% with a 45% datapath area gain.

On-Line Task Management for a Reconfigurable Cryptographic Architecture

Ivan Beretta, Vincenzo Rana, Marco D. Santambrogio, Donatella Sciuto
Politecnico di Milano - Dipartimento di Elettronica e Informazione,
Via Ponzio 34/5 - 20133 Milano, Italy
ivan.beretta@dresd.org, {rana, santambr, sciuto}@elet.polimi.it

Abstract

The increasing amount of programmable logic provided by modern FPGAs makes it possible to execute multiple hardware applications on the same device. This approach is reinforced by dynamic reconfiguration, which allows a single part of the device to be configured with a single hardware module. The proposed solution is a Linux-based operating system to manage on-demand module configuration on an FPGA while providing a set of high-level abstractions to user applications. The proposed approach has been validated in a cryptographic context using the DES and the AES algorithms.

Workshop 3

Workshop on High-Level Parallel Programming Models & Supportive Environments

HIPS 2009

An Integrated Approach To Improving The Parallel Application Development Process

Gregory R. Watson
IBM T.J. Watson Research Center
grw@us.ibm.com

Craig E Rasmussen
Los Alamos National Laboratory
crasmussen@lanl.gov

Beth R. Tibbitts
IBM T.J. Watson Research Center
tibbitts@us.ibm.com

Abstract

The development of parallel applications is becoming increasingly important to a broad range of industries. Traditionally, parallel programming was a niche area that was primarily exploited by scientists trying to model extremely complicated physical phenomenon. It is becoming increasingly clear, however, that continued hardware performance improvements through clock scaling and feature-size reduction are simply not going to be achievable for much longer. The hardware vendors approach to addressing this issue is to employ parallelism through multi-processor and multi-core technologies. While there is little doubt that this approach produces scaling improvements, there are still many significant hurdles to be overcome before parallelism can be employed as a general replacement to more traditional programming techniques. The Parallel Tools Platform (PTP) Project was created in 2005 in an attempt to provide developers with new tools aimed at addressing some of the parallel development issues. Since then, the introduction of a new generation of peta-scale and multi-core systems has highlighted the need for such a platform. In this paper, we describe some of the challenges facing parallel application developers, present the current state of PTP, and provide a simple case study that demonstrates how PTP can be used to locate a potential deadlock situation in an MPI code.

MPIXternal: A Library for a Portable Adjustment of Parallel MPI Applications to Heterogeneous Environments

Carsten Clauss, Stefan Lankes, Thomas Bemmerl
Chair for Operating Systems, RWTH Aachen University
Kopernikusstr. 16, 52056 Aachen, Germany
{clauss, lankes, bemmerl}@lfbs.rwth-aachen.de

Abstract

Nowadays, common systems in the area of high performance computing exhibit highly hierarchical architectures. As a result, achieving satisfactory application performance demands an adaptation of the respective parallel algorithm to such systems. This, in turn, requires knowledge about the actual hardware structure even at the application level. However, the prevalent Message Passing Interface (MPI) standard (at least in its current version 2.1) intentionally hides heterogeneity from the application programmer in order to assure portability. In this paper, we introduce the MPIXternal library which tries to circumvent this obvious semantic gap within the current MPI standard. For this purpose, the library offers the programmer additional features that should help to adapt applications to today's hierarchical systems in a convenient and portable way.

A Lightweight Stream-processing Library using MPI

Alan Wagner and Camilo Rostoker
Department of Computer Science
University of British Columbia
Vancouver, British Columbia
{wagner, rostoker}@cs.ubc.ca

Abstract

We describe the design of a lightweight library using MPI to support stream-processing on acyclic process structures. The design can be used to connect together arbitrary modules where each module can be its own parallel MPI program. We make extensive use of MPI groups and communicators to increase the flexibility of the library, and to make the library easier and safer to use. The notion of a communication context in MPI ensures that libraries do not conflict where a message from one library is mistakenly received by another. The library is not required to be part of any larger workflow environment and is compatible with existing MPI execution environments. The library is part of MarketMiner, a system for executing financial workflows.

Sparse Collective Operations for MPI

Torsten Hoefler
Open Systems Lab
Indiana University
Bloomington, IN, 47405
htor@cs.indiana.edu

Jesper Larsson Träff
NEC Laboratories Europe, NEC Europe Ltd.
Rathausallee 10
D-53225 Sankt Augustin, Germany
traff@it.neclab.eu

Abstract

We discuss issues in designing *sparse (nearest neighbor) collective operations* for communication and reduction operations in small neighborhoods for the *Message Passing Interface* (MPI). We propose three such operations, namely a sparse gather operation, a sparse all-to-all, and a sparse reduction operation in both regular and irregular (vector) variants. By two simple experiments we show a) that a collective handle for message scheduling and communication optimization is necessary for any such interface, b) that the possibly different amount of communication between neighbors need to be taken into account by the optimization, and c) illustrate the improvements that are possible by schedules that posses global information compared to implementations that can rely on only local information. We discuss different forms the interface and optimization handles could take. The paper is inspired by current discussion in the MPI Forum.

Smart Read/Write for MPI-IO

Saba Sehrish and Jun Wang
School of Electrical Engineering and Computer Science
University of Central Florida
ssehrish, jwang@eecs.ucf.edu

Abstract

We present a case for automating the selection of MPI-IO performance optimizations, with an ultimate goal to relieve the application programmer from these details, thereby improving their productivity. Programmers productivity has always been overlooked as compared to the performance optimizations in high performance computing community. In this paper we present RFSA, a Reduced Function Set Abstraction based on an existing parallel programming interface (MPI-IO) for I/O. MPI-IO provides high performance I/O function calls to the scientists/engineers writing parallel programs; who are required to use the most appropriate optimization of a specific function, hence limits the programmer productivity. Therefore, we propose a set of reduced functions with an automatic selection algorithm to decide what specific MPI-IO function to use. We implement a selection algorithm for I/O functions like read, write, etc. RFSA replaces 6 different flavors of read and write functions by one read and write function. By running different parallel I/O benchmarks on both medium-scale clusters and NERSC supercomputers, we show that RFSA functions impose minimal performance penalties.

Triple-C: Resource-Usage Prediction for Semi-Automatic Parallelization of Groups of Dynamic Image-Processing Tasks

Rob Albers[1,2], Eric Suijs[2] and Peter H.N. de With[1,3]
[1]Eindhoven University of Technology, PO Box 513, 5600 MB Eindhoven, The Netherlands,
[2]Philips Healthcare, X-Ray, PO Box 10.000, 5680 DA Best, The Netherlands,
[3]CycloMedia Technology, PO Box 68, 4180 BB Waardenburg, The Netherlands.
r.albers@philips.com

Abstract

With the emergence of dynamic video processing, such as in image analysis, runtime estimation of resource usage would be highly attractive for automatic parallelization and QoS control with shared resources. A possible solution is to characterize the application execution using model descriptions of the resource usage. In this paper, we introduce Triple-C, a prediction model for Computation, Cache-memory and Communication-bandwidth usage with scenario-based Markov chains. As a typical application, we explore a medical imaging function to enhance objects of interest in X-ray angiography sequences. Experimental results show that our method can be successfully applied to describe the resource usage for dynamic image-processing tasks, even if the flow graph dynamically switches between groups of tasks. An average prediction accuracy of 97% is reached with sporadic excursions of the prediction error up to 20-30%. As a case study, we exploit the prediction results for semi-automatic parallelization. Results show that with Triple-C prediction, dynamic processing tasks can be executed in real-time with a constant low latency.

GPAW optimized for Blue Gene/P using hybrid programming

Mads Ruben Burgdorff Kristensen
eScience Centre
University of Copenhagen
Denmark

Hans Henrik Happe
eScience Centre
University of Copenhagen
Denmark

Brian Vinter
eScience Centre
University of Copenhagen
Denmark

Abstract

In this work we present optimizations of a Grid-based projector-augmented wave method software, GPAW [?] for the Blue Gene/P architecture. The improvements are achieved by exploring the advantage of shared and distributed memory programming also known as hybrid programming. The work focuses on optimizing a very time consuming operation in GPAW, the finite-different stencil operation, and different hybrid programming approaches are evaluated. The work succeeds in demonstrating a hybrid programming model which is clearly beneficial compared to the original flat programming model. In total an improvement of 1.94 compared to the original implementation is obtained. The results we demonstrate here are reasonably general and may be applied to other finite difference codes.

CellFS: Taking The “DMA” Out Of Cell Programming

Latchesar Ionkov
Los Alamos National Laboratory
lionkov@lanl.gov

Aki Nyrhinen
University of Helsinki
aki@helsinki.fi

Andrey Mirtchovski
Los Alamos National Laboratory
andrey@lanl.gov

Abstract

In this paper we present a new programming model for the Cell BE architecture called CellFS. CellFS aims to simplify the task of managing I/O between the local store of the synergistic processing units and main memory of the Cell. The CellFS support library provides the means for transferring data via simple file I/O operations therefore eliminating the need for programmers to handle DMA transfers explicitly. The CellFS programming also provides overlap between code execution and data transfer by means of a deterministic, lock-less concurrency model.

A Generalized, Distributed Analysis System for Optimization of Parallel Applications

Hung-Hsun Su, Max Billingsley III and Alan D. George
High-performance Computing and Simulation (HCS) Research Lab
ECE Department, University of Florida
Gainesville, Florida, USA
su@hcs.ufl.edu, billingsley@hcs.ufl.edu, george@hcs.ufl.edu

Abstract

Developing a high performance parallel application is difficult. An application must often be analyzed and optimized by the programmer before reaching an acceptable level of performance. Performance tools that collect and visualize performance data can reduce the effort needed by the user in the nontrivial optimization process. However, as the size of the performance dataset grows, it becomes nearly impossible for the user to manually examine the data and find performance issues. To address this problem, we have developed a new analysis system to automatically detect, diagnose, and possibly resolve bottlenecks. In this paper, we present the architecture and the distributed, peer-to-peer processing mechanism of a programming model-independent analysis system, which includes a range of useful analyses such as scalability analysis and common-bottleneck detection. We then describe the details of an initial sequential implementation of the system that has been integrated into our Parallel Performance Wizard (PPW) tool. Finally, we provide correctness and performance results for this initial version and demonstrate the effectiveness of the system through two case studies.

CuPP – A framework for easy CUDA integration

Jens Breitbart
Research Group Programming Languages / Methodologies
Universität Kassel
Kassel, Germany
jbreitbart@uni-kassel.de

Abstract

This paper reports on CuPP, our newly developed C++ framework designed to ease integration of NVIDIAs GPGPU system CUDA into existing C++ applications. CuPP provides interfaces to reoccurring tasks that are easier to use than the standard CUDA interfaces. In this paper we concentrate on memory management and related data structures. CuPP offers both a low level interface – mostly consisting of smartpointers and memory allocation functions for GPU memory – and a high level interface offering a C++ STL vector wrapper and the so-called type transformations. The wrapper can be used by both device and host to automatically keep data in sync. The type transformations allow developers to write their own data structures offering the same functionality as the CuPP vector, in case a vector does not conform to the need of the application. Furthermore the type transformations offer a way to have two different representations for the same data at host and device, respectively. We demonstrate the benefits of using CuPP by integrating it into an example application, the open-source steering library OpenSteer. In particular, for this application we develop a uniform grid data structure to solve the k-nearest neighbor problem that deploys the type transformations. The paper finishes with a brief outline of another CUDA application, the Einstein@Home client, which also requires data structure redesign and thus may benefit from the type transformations and future work on CuPP.

Fast Development of Dense Linear Algebra Codes on Graphics Processors

M. Jesús Zafont, Alberto Martín, Francisco Igual and Enrique S. Quintana-Ortí
Depto. de Ingeniería y Ciencia de los Computadores
Universidad Jaume I, Castellón (Spain)
al051631@uji.es, {martina, figual, quintana}@icc.uji.es

Abstract

We present an application programming interface (API) for the C programming language that facilitates the development of dense linear algebra algorithms on graphics processors applying the FLAME methodology. The interface, built on top of the NVIDIA CUBLAS library, implements all the computational functionality of the FLAMEC interface. In addition, the API includes data transference routines to explicitly handle communication between the CPU and GPU memory spaces. The flexibility and simplicity-of-use of this tool are illustrated using a complex operation of dense linear algebra: the Cholesky factorization. For this operation, we implement and evaluate all existing variants on an NVIDIA G80 processor, attaining speed-ups 7× compared with the CPU implementations.

Workshop 4

Workshop on Java and Components for Parallelism, Distribution and Concurrency

JAVAPDC 2009

Providing Security for MOCCA Component Environment

Michal Dyrda, Maciej Malawski and Marian Bubak
Institute of Computer Science, AGH,
Mickiewicza 30, 30-059 Krakow, Poland
ACC CYFRONET-AGH
Nawojki 11, 30-950 Krakow, Poland
{malawski,bubak}@agh.edu.pl

Syed Naqvi
CETIC
Rue des Freres Wright 29/3
6041 Charleroi, Belgium

Abstract

The subject of this paper is a detailed analysis and development of security in MOCCA, a CCA-compliant Grid component framework build over H2O, a Java-based distributed computing platform. The approach is to extend H2O with an authentication mechanism that will be both secure and compliant with solutions commonly used in modern Grid systems. The proposed authenticator is based on asymmetric cryptography with additional features provided by the Grid Security Infrastructure - proxy certificates that are used for Single Sign-On and delegation. The developed GSI Authenticator was subjected to threat analysis and performance tests, which proved its safety and usability.

Towards Efficient Shared Memory Communications in MPJ Express

Aamir Shafi
aamir.shafi@seecs.edu.pk
School of Electrical Engineering and Computer Science
National University of Sciences and Technology
Pakistan

Jawad Manzoor
jawad.manzoor@seecs.edu.pk
School of Electrical Engineering and Computer Science
National University of Sciences and Technology
Pakistan

Abstract

The need to increase performance while conserving energy lead to the emergence of multi-core processors. These processors provide a feasible option to improve performance of software applications by increasing the number of cores, instead of relying on increased clock speed of a single core. The uptake of multi-core processors by hardware vendors present variety of challenges to the software community. In this context, it is important that messaging libraries based on the Message Passing Interface (MPI) standard support efficient inter-core communication. Typically processing cores of today's commercial multi-core processors share the main memory. As a result, it is vital to develop devices to exploit this. MPJ Express is our implementation of the MPI-like Java bindings. The software has mainly supported communication with two devices; the first is based on Java New I/O (NIO) and the second is based on Myrinet. In this paper, we present two shared memory implementations meant for providing efficient communication of multi-core and SMP clusters. The first implementation is pure Java and uses Java threads to exploit multiple cores. Each Java thread represents an MPI level OS process and communication between these threads is achieved using shared data structures. The second implementation is based on the System V (SysV) IPC API. Our goal is to achieve better communication performance than already existing devices based on Transmission Control Protocol (TCP) and Myrinet on SMP and multi-core platforms. Another design goal is that existing parallel applications must not be modified for this purpose, thus relieving application developers from extra efforts of porting their applications to such modern clusters. We have benchmarked our implementations and report that threads-based device performs the best on an Intel quad-core Xeon cluster.

TM-STREAM: an STM Framework for Distributed Event Stream Processing

Heiko Sturzrehm
Institut d'informatique
Université de Neuchâtel
Neuchâtel, Switzerland
heiko.sturzrehm@unine.ch

Pascal Felber
Institut d'informatique
Université de Neuchâtel
Neuchâtel, Switzerland
pascal.felber@unine.ch

Christof Fetzer
Systems Engineering Group
Technische Universität Dresden
Dresden, Germany
christof.fetzer@inf.tu-dresden.de

Abstract

We extend DSTM2 with a combination of two techniques:

First, we applied speculative dependencies between transactions, as first introduced in [?]. Specifically, transactions may read data of earlier transactions that have completed their execution, but are not yet committed. This is the case, for instance, when transactions have to commit in a certain order and must wait for the completion of earlier transactions to detect possible conflicts (e.g., in stream processing systems).

Second, we expand speculation to distributed settings, by allowing not yet committed transactions to trigger execution of other speculative transactions on a remote machine. We use a simple notification mechanism to commit or abort remote speculative transactions once the outcome of all the transactions they depend on is known.

In this paper we describe our extensions to the DSTM2 framework to enable distributed speculation and evaluate their performance on a simple distributed application.

Is Shared Memory Programming Attainable on Clusters of Embedded Processors?

Konstantinos I. Karantasis and Eleftherios D. Polychronopoulos
High Performance Information Systems Laboratory
Computer Engineering & Informatics Department,
University of Patras
26500 Rio, Greece
kik@hpclab.ceid.upatras.gr, edp@hpclab.ceid.upatras.gr
http://pdsgroup.hpclab.ceid.upatras.gr

Abstract

The wide increase of total processing cores in commodity processors tends to lighten the need for computer performance by the classical scientific problems as well as by the modern multimedia and every day embedded applications. Nevertheless, the introduction of this powerful and promising technology seems to inherit and in some cases magnify all the classical problems that already exist on the programming side of these environments. In this work we present a portable environment based on the Java platform targeting to mitigate the performance loss associated with the programming models used in modern multicore systems. We propose a shared memory model for programming modern multicore and distributed environments, presuming only minor interventions by the application programmer. We finally prove that when we use shared memory programming with a widely accepted programming language like Java, we can achieve at least comparable performance with classical though more sophisticated technologies.

High Performance Computing Using ProActive Environment and The Asynchronous Iteration Model

Raphaël Couturier, David Laiymani and Sébastien Miquée
Laboratoire d'Informatique de Franche-Comté (LIFC)
University of Franche-Comté
IUT de Belfort-Montbéliard, Rue Engel Gros, BP 27, 90016 Belfort, France
Tel.: +33-3-84587781 Fax: +33-3-84587781
{raphael.couturier,david.laiymani,sebastien.miquee}@univ-fcomte.fr

Abstract

This paper presents a new library for the ProActive environment, called AIL-PA (Asynchronous Iterative Library for ProActive). This new library allows to execute programs for solving large scale problems on various architectures. Two models of algorithm can be used: the synchronous iteration model which is efficient on single clusters; the asynchronous iteration model which is more efficient on distributed clusters. Both approaches are tested on both architectures, using Kernel CG of the NAS Parallel Benchmarks on the Grid'5000 platform. These tests also allow us to compare ProActive with AIL-PA and with the Jace programming environment. The results show that the asynchronous iteration model with AIL-PA is more efficient on distributed clusters than the synchronous iteration model. Moreover, these experiments also show that AIL-PA does not involve additional overhead to ProActive.

Workshop 5

Workshop on Nature Inspired Distributed Computing

NIDISC 2009

Exact Pairwise Alignment of Megabase Genome Biological Sequences Using A Novel Z-align Parallel Strategy

Azzedine Boukerche[1], Rodolfo Bezerra Batista[2] and Alba Cristina Magalhaes Alves de Melo[1,2]
[1]School of Information Technology and Engineering (SITE), University of Ottawa, Canada
[2]Department of Computer Science, University of Brasilia (UnB), Brazil
boukerch@site.uottawa.ca, rodolfo@unb.br, albamm@cic.unb.br

Abstract

Pairwise Sequence Alignment is a basic operation in Bioinformatics that is performed thousands of times, in a daily basis. The exact methods proposed in the literature have quadratic time complexity. For this reason, heuristic methods such as BLAST are widely used. Nevertheless, it is known that exact methods present better sensitivity, leading to better results. To obtain exact results faster, many parallel strategies have been proposed but most of them fail to align huge biological sequences. This happens because not only the quadratic time must be considered but also the space should be reduced. In this paper, we evaluate the performance and sensibility of z-align, a parallel exact strategy that runs in user-restricted memory space. The results obtained in a 64-processor cluster show that two sequences of size 23MBP (Mega Base Pairs) and 24MBP, respectively, were successfully aligned with z-align. Also, in order to align two 3MBP sequences, a speedup of 34.35 was achieved. Finally, when comparing z-align with BLAST, we can see that the z-align alignments are longer and have a higher score.

Solving multiprocessor scheduling problem with GEO metaheuristic

Piotr Switalski[1], Franciszek Seredynski[23]
[1]Institute of Computer Science
University of Podlasie
3 Maja 54, 08-110 Siedlce, peter@ii.ap.siedlce.pl
[2]Institute of Computer Science
Polish Academy of Sciences
Ordona 21, 01-237 Warsaw, Poland, sered@ipipan.waw.pl
[3]Polish-Japanese Institute of Information Technology
Koszykowa 86, 02-008 Warsaw, Poland

Abstract

We propose a solution of the multiprocessor scheduling problem based on applying a relatively new metaheuristic called Generalized Extremal Optimization (GEO). GEO is inspired by a simple coevolutionary model known as Bak-Sneppen model. The model assumes existing of an ecosystem consisting of *N* species. Evolution in this model is driven by a process in which the weakest species in the ecosystem, together with its nearest neighbors is always forced to mutate. This process shows characteristic of a phenomenon called a punctuated equilibrium which is observed in evolutionary biology. We interpret the multiprocessor scheduling problem in terms of the Bak-Sneppen model and apply the GEO algorithm to solve the problem. We show that the proposed optimization technique is simple and yet outperforms both genetic algorithm (GA)-based and particle swarm optimization (PSO) algorithm-based approaches to the multiprocessor scheduling problem.

Using XMPP for ad-hoc grid computing - an application example using parallel ant colony optimisation

Gerhard Weis and Andrew Lewis

Abstract

XMPP (XML Messaging and Presence Protocol), also known as Jabber, is a popular instant messaging protocol that uses XML streams for communication. Due to it's high extensibility, XMPP is very easy to adapt to other uses than instant messaging. Furthermore, announcing of presence state makes it ideal for highly volatile environments. This paper outlines the use of XMPP for a grid-like computation environment. The biggest advantage of this setup was that available computing resources, such as laboratory computers, could be connected easily and used similarly to a grid. The application example described in this paper uses Ant Colony System (ACS) optimisation and the NEC-tool to optimise RFID antennas, involving computing the efficiency and resonant frequency of a large number of different antenna structures.

Hybridization of Genetic and Quantum Algorithm for Gene Selection and Classification of Microarray Data

Allani Abderrahim
Institut Supérieur de Gestion
41, Rue de la liberté,
Cité Bouchoucha 2000, Bardo, Tunisie
Allani.abderrahim@gmail.com

El-Ghazali Talbi
LIFL-INRIA Futurs
Bât M3, Cité Scientifique
59655 Villeneuve d'Ascq, France
El-ghazali.Talbi@lifl.fr

Mellouli Khaled
Institut des Hautes Etudes
Commerciales de Carthage
Carthage Présidence Carthage, Tunisie
Khaled.Mellouli@ihec.rnu.tn

Abstract

In this work, we hybridize the Genetic Quantum Algorithm with the Support Vector Machines classifier for gene selection and classification of high dimensional Microarray Data. We named our algorithm GQA_{SVM}. Its purpose is to identify a small subset of genes that could be used to separate two classes of samples with high accuracy.

A comparison of the approach with different methods of literature, in particular GA_{SVM} and PSO_{SVM} [2], was realized on six different datasets issued of microarray experiments dealing with cancer (leukemia, breast, colon, ovarian, prostate, and lung) and available on Web. The experiments clearified the very good performances of the method.

A first contribution shows that the algorithm GQA_{SVM} is able to find genes of interest and improve the classification on a meaningful way.

A second important contribution consists of the actual discovery of new and challenging results on datasets used.

Fine Grained Population Diversity Analysis for Parallel Genetic Programming

Stephan M. Winkler
Department for Medical and Bioinformatics
Upper Austria University of Applied Sciences
Hagenberg, Austria
stephan.winkler@fh-hagenberg.at
Michael Affenzeller and Stefan Wagner
Department for Software Engineering
Upper Austria University of Applied Sciences
Hagenberg, Austria
{michael.affenzeller,stefan.wagner}@fh-hagenberg.at

Abstract

In this paper we describe a formalism for estimating the structural similarity of formulas that are evolved by parallel genetic programming (GP) based identification processes. This similarity measurement can be used for measuring the genetic diversity among GP populations and, in the case of multi-population GP, the genetic diversity among sets of GP populations: The higher the average similarity among solutions becomes, the lower is the genetic diversity. Using this definition of genetic diversity for GP we test several different GP based system identification algorithms for analyzing real world measurements of a BMW Diesel engine as well as medical benchmark data taken from the UCI machine learning repository.

New sequential and parallel algorithm for Dynamic Resource Constrained Project Scheduling Problem

André Renato Villela da Silva and Luiz Satoru Ochi
Computing Institute
Federal Fluminense University
Niterói, Brazil
{avillela,satoru}@ic.uff.br

Abstract

This paper proposes a new Evolutionary Algorithm for the Dynamic Resource Constrained Project Scheduling Problem. This algorithm has new features that get around some problems like premature convergence and other ones. The indirect representation approach was used because it allows the construction of a feasible solution from any input priorities.

A parallel version is also proposed, making good use of multicore processors available nowadays. The results of sequential and parallel versions were very significant, improving in almost all ways the best results present in literature.

Interweaving Heterogeneous Metaheuristics Using Harmony Search

Young Choon Lee and Albert Y. Zomaya
Advanced Networks Research Group, School of Information Technologies
The University of Sydney
NSW 2006, Australia
{yclee,zomaya}@it.usyd.edu.au

Abstract

In this paper, we present a novel parallel-metaheuristic framework, which enables a set of heterogeneous metaheuristics to be effectively interwoven and coordinated. The key player of this framework is a harmony-search-based coordinator devised using a recent breed of soft computing paradigm called harmony search that mimics the improvisation process of musicians. For the applicability validation and the performance evaluation, we have implemented a parallel hybrid metaheuristic using the framework for the task scheduling problem on multiprocessor computing systems. Experimental results verify that the proposed framework is a compelling approach to parallelize heterogeneous metaheuristics.

Adaptative Clustering Particle Swarm Optimization

Salomão S. Madeiro, Carmelo J. A. Bastos-Filho, Member, IEEE and Fernando B. Lima Neto, Senior Member, IEEE, Elliackin M. N. Figueiredo

Abstract

The performance of Particle Swarm Optimization (PSO) algorithms depends strongly upon the interaction among the particles. The existing communication topologies for PSO (e.g. star, ring, wheel, pyramid, von Neumann, clan, four clusters) can be viewed as distinct means to coordinate the information flow within the swarm. Overall, each particle exerts some influence among others placed in its immediate neighborhood or even in different neighborhoods, depending on the communication schema (rules) used. The neighborhood of particles within PSO topologies is determined by the particles indexes that usually reflect a spatial arrangement. In this paper, in addition to position information of particles, we investigate the use of adaptive density-based clustering algorithm - ADACLUS - to create neighborhoods (i.e. clusters) that are formed considering velocity information of particles. Additionally, we suggest that the new clustering rationale be used in conjunction with Clan-PSO main ideas. The proposed approach was tested in a wide range of well known benchmark functions. The experimental results obtained indicate that this new approach can improve the global search ability of the PSO technique.

Metaheuristic Traceability Attack against SLMAP, an RFID Lightweight Authentication Protocol

Julio C. Hernandez-Castro[1], Juan E. Tapiador[3], Pedro Peris-Lopez[2], John A. Clark[3] and El-Ghazali Talbi[4]
[1]School of Computing, Portsmouth University
Julio.Hernandez-Castro@port.ac.uk
[2]Information and Communication Theory Group, Delft University of Technology
P.PerisLopez@tudelft.nl
[3]Department of Computer Science, University of York
jet@cs.york.ac.uk, jac@cs.york.ac.uk
[4]INRIA Futurs, Villeneuve d'Ascq, Lille
El-ghazali.Talbi@lifl.fr

Abstract

We present a metaheuristic-based attack against the traceability of an ultra-lightweight authentication protocol for RFID environments called SLMAP, and analyse its implications. The main interest of our approach is that it is a complete black-box technique that doesn't make any assumptions on the components of the underlying protocol and can thus be easily generalised to analyse many other proposals.

Parallel Nested Monte-Carlo Search

Tristan Cazenave
LAMSADE, Université Paris-Dauphine
Place Maréchal de Lattre de Tassigny, 75775 Paris Cedex 16, France
cazenave@lamsade.dauphine.fr
Nicolas Jouandeau
Université Paris 8, LIASD
2 rue de la liberté, 93526, Saint-Denis, France
n@ai.univ-paris8.fr

Abstract

We address the parallelization of a Monte-Carlo search algorithm. On a cluster of 64 cores we obtain a speedup of 56 for the parallelization of Morpion Solitaire. An algorithm that behaves better than a naive one on heterogeneous clusters is also detailed.

Combining Genetic Algorithm with Time-Shuffling in Order to Evolve Agent Systems More Efficiently

Patrick Ediger and Rolf Hoffmann
Technische Universität Darmstadt
FB Informatik, FG Rechnerarchitektur
Hochschulstraße 10, 64289 Darmstadt, Germany

Abstract

We have optimized a multi-agent system for all-to-all communication modeled in cellular automata. The agents' task is to solve the problem by communicating their initially mutually exclusive distributed information to all the other agents. We used a set of 20 environments (initial configurations), 10 with border, 10 with cyclic wrap-around to evolve the best behavior for agents with a uniform rule defined by a finite state machine. The state machine was evolved (1) directly by a genetic algorithm (GA) for all 20 environments and (2) indirectly by two separate GAs for the 10 environments with border and the 10 environments with wrap-around with a subsequent time-shuffling technique in order to integrate the good abilities from both of the separately evolved state machines. The time-shuffling technique alternates two state machines periodically. The results show that time-shuffling two separately evolved state machines is effective and much more efficient than the direct application of the GA.

Multi-thread integrative cooperative optimization for rich combinatorial problems

Teodor Gabriel Crainic[1], Gloria Cerasela Crisan[1,2], Michel Gendreau[3], Nadia Lahrichi[1] and Walter Rei[1]
[1]École des sciences de la gestion, U.Q.A.M.
Département de management et technologie
and
Centre Interuniversitaire de Recherche sur les Réseaux d'Entreprise, la Logistique et le Transport
C.P. 8888, Succursale Centre-ville, Montréal (QC), Canada H3C 3P8
(TeodorGabriel.Crainic,Nadia.Lahrichi,Walter.Rei)@cirrelt.ca
[2] University of Bacau, Romania
Cerasela.Crisan@cirrelt.ca
[3]Université de Montréal
Département d'informatique et de recherche opérationnelle
and
Centre Interuniversitaire de Recherche sur les Réseaux d'Entreprise, la Logistique et le Transport
C.P. 6128, Succursale Centre-ville, Montréal (QC), Canada H3C 3J7
Michel.Gendreau@cirrelt.ca

Abstract

Addressing multi-attribute, "rich" combinatorial optimization problems in a comprehensive manner presents significant methodological and computational challenges. In this paper, we present an integrative multi-thread cooperative optimization framework that can simultaneously deal with multiple dimensions of a rich problem. We present the basic concepts and detail the design and operating principles of the methodology. We illustrate the framework on a rich combinatorial problem, an extended version of the vehicle routing problem with the duration and capacity constraints as well as time windows, multiple periods and multiple depots.

The Effect of Population Density on the Performance of a Spatial Social Network Algorithm for Multi-Objective Optimisation

Andrew Lewis
the Institute for Integrated and Intelligent Systems,
Griffith University, Queensland, Australia
a.lewis@griffith.edu.au

Abstract

Particle Swarm Optimisation (PSO) is increasingly being applied to optimisation of multi-objective problems in engineering design and scientific investigation. This paper investigates the behaviour of a novel algorithm based on an extension of the concepts of spatial social networks using a model of the behaviour of locusts and crickets. In particular, observation of locust swarms suggests a specific dependence on population density for ordered behaviour. Computational experiments demonstrate that both the new, spatial, social network algorithm and a conventional MOPSO algorithm exhibit improved performance with increased swarm size and crowding. This observation may have particular significance for design of some forms of distributed PSO algorithms.

A Parallel Hybrid Genetic Algorithm-Simulated Annealing for Solving Q3AP on Computational Grid

Lakhdar Loukil[1], Malika Mehdi[2], Nouredine Melab[3],
El-Ghazali Talbi[3], and Pascal Bouvry[2]

[1]Université d'Oran
Faculté des Sciences
Département d'informatique
BP 1524 El M'Naouer Oran, Algérie
Loukil.Lakhdar@univ-oran.dz

[2]University of Luxembourg
Faculty of Sciences
Technology and Communication
6 rue de Coudenhove Kalergi
L-1359, Luxembour, Luxembourg
{Malika.Mehdi , Pascal.Bouvry}@uni.lu

[3]INRIA Futurs, Parc Scientifique de la Haute Borne
40, avenue Halley, Bt. A, Park Plaza
59650 Villeneuve d'Asq, France
{Nouredine.Melab , El-Ghazali.Talbi}@lifl.fr

Abstract

In this paper we propose a parallel hybrid genetic method for solving Quadratic 3-dimensional Assignment Problem (Q3AP). This problem is proved to be computationally NP-hard. The parallelism in our algorithm is of two hierarchical levels. The first level is an insular model where a number of GAs (genetic algorithms) evolve in parallel. The second level is a parallel transformation of individuals in each GA. Implementation has been done using ParadisEO framework, and the experiments have been performed on GRID5000, the French nation-wide computational grid. To evaluate our method, we used three benchmarks derived from QAP instances of QAPLIB and the results are compared with those reported in the literature. The preliminary results show that the method is promising. The obtained solutions are close to the optimal values and the execution is efficient.

Solving the industrial car sequencing problem in a Pareto sense

Arnaud Zinflou, Caroline Gagné and Marc Gravel
Université du Québec à Chicoutimi, Québec, Canada
arnaud_zinflou@uqac.ca, caroline_gagne@uqac.ca, marc_gravel@uqac.ca

Abstract

Until now, the industrial car sequencing problem, as defined during the ROADEF 2005 Challenge, has been tackled by organizing objectives in a hierarchy. In this paper, we suggest tackling this problem in a Pareto sense for the first time. We thus suggest the adaptation of the PMSMO, an elitist evolutionary algorithm which distinguishes itself through a fitness calculation that takes into account the history of solutions found so as to diversify the compromise solutions along the Pareto frontier. A comparison of the performance is carried out using a well-known published algorithm, the NSGAII, and proves an advantage for the PMSMO. As well, we aim to demonstrate the relevance of handling applied problems such as the car sequencing problem using a multi-objective approach.

A Multi-objective Strategy for Concurrent Mapping and Routing in Networks on Chip

Rafael Tornero[*] Valentino Sterrantino[†] Maurizio Palesi[†] Juan M. Orduña[*]
[*] Departamento de Informática
Universidad de Valencia, Spain
{Rafael.Tornero, Juan.Orduna}@uv.es
[†] Dipartimento di Ingegneria Informatica e delle Telecomunicazioni
Unversità di Catania, Italy
{vster, mpalesi}@diit.unict.it

Abstract

The design flow of network-on-chip (NoCs) include several key issues. Among other parameters, the decision of where cores have to be topologically mapped and also the routing algorithm represent two highly correlated design problems that must be carefully solved for any given application in order to optimize several different performance metrics. The strong correlation between the different parameters often makes that the optimization of a given performance metric has a negative effect on a different performance metric. In this paper we propose a new strategy that simultaneously refines the mapping and the routing function to determine the Pareto optimal configurations which optimize average delay and routing robustness. The proposed strategy has been applied on both synthetic and real traffic scenarios. The obtained results show how the solutions found by the proposed approach outperforms those provided by other approaches proposed in literature, in terms of both performance and fault tolerance.

Evolutionary Game Theoretical Analysis of Reputation-based Packet Forwarding in Civilian Mobile Ad Hoc Networks

Marcin Seredynski and Pascal Bouvry
Faculty of Sciences, Technology and Communication, University of Luxembourg
6, rue Coudenhove Kalergi, L-1359, Luxembourg, Luxembourg
{marcin.seredynski, pascal.bouvry}@uni.lu

Abstract

A mobile wireless ad hoc network (MANET) consists of a number of devices that form a temporary network operating without support of a fixed infrastructure. The correct operation of such a network requires its users to cooperate on the level of packet forwarding. However, a distributed nature of MANET, lack of a single authority, and limited battery resources of participating devices may lead to a noncooperative behavior of network users, resulting in a degradation of the network throughput. Thus, a cooperation enforcement system specifying certain packet forwarding strategies is a necessity is such networks. In this work we investigate general properties of such a system. We introduce a Prisoner's Dilemma-based model of packet forwarding and next using an evolutionary game-theoretical approach we demonstrate that cooperation very likely to be developed on the basis of conditionally cooperative strategies similar to the TIT-FOR-TAT strategy.

Workshop 6

Workshop on High Performance Computational Biology

HiCOMB 2009

Parallel Reconstruction of Neighbor-Joining Trees for Large Multiple Sequence Alignments using CUDA

Yongchao Liu, Bertil Schmidt and Douglas L. Maskell
School of Computer Engineering, Nanyang Technological University, Singapore 639798
{liuy0039, asbschmidt, asdouglas}@ntu.edu.sg

Abstract

Computing large multiple protein sequence alignments using progressive alignment tools such as ClustalW requires several hours on state-of-the-art workstations. ClustalW uses a three-stage processing pipeline: (i) pairwise distance computation; (ii) phylogenetic tree reconstruction; and (iii) progressive multiple alignment computation. Previous work on accelerating ClustalW was mainly focused on parallelizing the first stage and achieved good speedups for a few hundred input sequences. However, if the input size grows to several thousand sequences, the second stage can dominate the overall runtime. In this paper, we present a new approach to accelerating this second stage using graphics processing units (GPUs). In order to derive an efficient mapping onto the GPU architecture, we present a parallelization of the neighbor-joining tree reconstruction algorithm using CUDA. Our experimental results show speedups of over 26× for large datasets compared to the sequential implementation.

Accelerating Error Correction in High-Throughput Short-Read DNA Sequencing Data with CUDA

Haixiang Shi, Bertil Schmidt, Weiguo Liu and Wolfgang Müller-Wittig
School of Computer Engineering, Nanyang Technological University, Singapore 639798,
{hxshi,asbschmidt,liuweiguo,askwmwittig}@ntu.edu.sg

Abstract

Emerging DNA sequencing technologies open up exciting new opportunities for genome sequencing by generating read data with a massive throughput. However, produced reads are significantly shorter and more error-prone compared to the traditional Sanger shotgun sequencing method. This poses challenges for de-novo DNA fragment assembly algorithms in terms of both accuracy (to deal with short, error-prone reads) and scalability (to deal with very large input data sets). In this paper we present a scalable parallel algorithm for correcting sequencing errors in high-throughput short-read data. It is based on spectral alignment and uses the CUDA programming model. Our computational experiments on a GTX 280 GPU show runtime savings between 10 and 19 times (for different error-rates using simulated datasets as well as real Solexa/Illumina datasets).

Parallel Monte Carlo Study on Caffeine-DNA Interaction in Aqueous Solution

M.D. Kalugin[1] and A.V. Teplukhin[2]

[1]Institute of System Programming, Russian
Academy of Sciences, Moscow, Russia
shaman@isp.ras.ru

[2]Institute of Mathematical Problems in Biology,
Russian Academy of Sciences, Pushchino, Russia
tepl@impb.psn.ru

Abstract

Monte Carlo simulation of the caffeine-DNA interaction in aqueous solution at room temperature was carried out using parallel calculations on supercomputer. Very large simulation boxes were used containing superhelical B-DNA fragment surrounded by caffeine and water molecules. The most probable binding sites of caffeine molecules on the DNA surface as well as structural features of the respective caffeine- DNA complexes were revealed for several solutions' concentrations.

Dynamic Parallelization for RNA Structure Comparison

Eric Snow, Eric Aubanel, and Patricia Evans
Faculty of Computer Science
University of New Brunswick
Fredericton, New Brunswick, Canada E3B 5A3
{eric.snow, aubanel, pevans} @unb.ca

Abstract

In this paper we describe the parallelization of a dynamic programming algorithm used to find common RNA secondary structures including pseudoknots and similar structures. The sequential algorithm is recursive and uses memoization and data-driven selective allocation of the tables, in order to cope with the high space and time demands. These features, in addition to the irregular nature of the data access pattern, present particular challenges to parallelization. We present a new *manager-worker* approach, where workers are responsible for task creation and the manager's sole responsibility is overseeing load balancing. Special considerations are given to the management of distributed, dynamic task creation and data structures, along with general inter-process communication and load balancing on a heterogeneous computational platform. Experimental results show a modest level of speedup with a highly-scalable level of memory usage, allowing the comparison of much longer RNA molecules than is possible in the sequential implementation.

Accelerating HMMer on FPGAs Using Systolic Array Based Architecture

Yanteng Sun[1], Peng Li[2], Guochang Gu[1], Yuan Wen[1], Yuan Liu[2] and Dong Liu[2]
[1]College of Computer Science and Technology, Harbin Engineering University
[2]Intel China Research Center
{sunyanteng, guguochang, wenyuan}@hrbeu.edu.cn, {peng.p.li, yuan.y.liu, dong.liu}@intel.com

Abstract

HMMer is a widely-used bioinformatics software package that uses profile HMMs (Hidden Markov Models) to model the primary structure consensus of a family of protein or nucleic acid sequences. However, with the rapid growth of both sequence and model databases, it is more and more time-consuming to run HMMer on traditional computer architecture. In this paper, the computation kernel of HMMer, P7Viterbi, is selected to be accelerated by FPGA. There is an infrequent feedback loop in P7Viterbi to update the value of beginning state (B state), which limits further parallelization. Previous work either ignored the feedback loop or serialized the process, leading to loss of either precision or efficiency. Our proposed syslolic array based architecture with a parallel data providing unit can exploit maximum parallelism of the full version of P7Viterbi. The proposed architecture speculatively runs with fully parallelism assuming that the feedback loop does not take place. If the rare feedback case actually occurs, a rollback mechanism is used to ensure correctness. Results show that by using Xilinx Virtex-5 110T FPGA, the proposed architecture with 20 PEs can achieve about a 56.8 times speedup compared with that of Intel Core2 Duo 2.33GHz CPU.

A Resource-Efficient Computing Paradigm for Computational Protein Modeling Applications

Yaohang Li and Douglas Wardell
Department of Computer Science
North Carolina A&T State University
{yaohang,dwardell}@ncat.edu

Vincent Freeh
Department of Computer Science
North Carolina State University
vin@csc.ncsu.edu

Abstract

Many computational protein modeling applications using numerical methods such as Molecular Dynamics (MD), Monte Carlo (MC), or Genetic Algorithms (GA) require a large number of energy estimations of the protein molecular system. A typical energy function describing the protein energy is a combination of a number of terms characterizing various interactions within the protein molecule as well as the protein-solvent interactions. Evaluating the energy function of a relatively large protein molecule is rather computationally costly and usually occupies the major computation time in the protein simulation process. In this paper, we present a resource-efficient computing paradigm based on "consolidation" to reduce the computational time of evaluating the energy function of large protein molecule. The fundamental idea of consolidation is to increase computational density to a computer in order to increase the CPU utilizations. Consolidation will be particularly efficient when the consolidated computations have heterogeneous resource demands. In computational protein modeling applications with costly energy function evaluation, we advocate the use of "thread consolidation," which is to spawn concurrent threads to carry out parallel energy function terms computations. Our computational results show that 7% 11% speedup in a protein loop structure prediction program on various hardware architectures where memory-intensive and computation-intensive terms coexist in the energy function. For an MD protein simulation program where computation-intensive energy function evaluations are divided and carried out by concurrent threads, we also find slight performance improvement when the thread consolidation technique is applied.

Exploring FPGAs for Accelerating the Phylogenetic Likelihood Function

N. Alachiotis[1,2], E. Sotiriades[1], A. Dollas[1], A. Stamatakis[2]
[1]Department of Electronic and Computer Engineering,
Technical University of Crete, Chania, Crete, Greece
[2]The Exelixis Lab, Department of Computer Science,
Technische Universität München, Germany

Abstract

Driven by novel biological wet lab techniques such as pyrosequencing there has been an unprecedented molecular data explosion over the last 2-3 years. The growth of biological sequence data has significantly outpaced Moore's law. This development also poses new computational and architectural challenges for the field of phylogenetic inference, i.e., the reconstruction of evolutionary histories (trees) for a set of organsims which are represented by respective molecular sequences. Phylogenetic trees are currently increasingly reconstructed from multiple genes or even whole genomes. The recently introduced term "phylogenomics" reflects this development. Hence, there is an urgent need to deploy and develop new techniques and computational solutions to calculate the computationally intensive scoring functions for phylogenetic trees.

In this paper, we propose a dedicated computer architecture to compute the phylogenetic Maximum Likelihood (ML) function. The ML criterion represents one of the most accurate statistical models for phylogenetic inference and accounts for 85% to 95% of total execution time in all state-of-the-art ML-based phylogenetic inference programs. We present the implementation of our architecture on an FPGA (Field Programmable Gate Array) and compare the performance to an efficient C implementation of the ML function on a high-end multi-core architecture with 16 cores.

Our results are two-fold: (i) the initial exploratory implementation of the ML function for trees comprising 4 up to 512 sequences on an FPGA yields speedups of a factor 8.3 on average compared to execution on a single-core and is faster than the OpenMP-based parallel implementation on up to 16 cores in all but one case; and (ii) we are able to show that current FPGAs are capable to efficiently execute floating point intensive computational kernels.

Long time-scale simulations of *in vivo* diffusion using GPU hardware

Elijah Roberts[1], John E. Stone[2], Leonardo Sepúlveda[1], Wen-Mei W. Hwu[3] and Zaida Luthey-Schulten[4]

[1]Center for Biophysics and
Computational Biology
University of Illinois
Urbana, IL, USA
{erobert3,lsepulv2}@illinois.edu

[2]Beckman Institute
University of Illinois
Urbana, IL, USA
johns@ks.uiuc.edu

[3]Department of Electrical and
Computer Engineering
University of Illinois
Urbana, IL, USA
w-hwu@illinois.edu

[4]Department of Chemistry
University of Illinois, Urbana, IL, USA
zan@illinois.edu

Abstract

To address the problem of performing long time simulations of biochemical pathways under in vivo cellular conditions, we have developed a lattice-based, reaction-diffusion model that uses the graphics processing unit (GPU) as a computational co-processor. The method has been specifically designed from the beginning to take advantage of the GPU's capacity to perform massively parallel calculations by not only executing a core set of mathematical calculations, but also running much of the underlying algorithmic logic on the GPU. In this study we present our three-dimensional model for in vivo diffusion that exploits the calculation capabilities of the GPU. The implementation of the diffusion operator on the GPU is subject to architectural constraints, and we discuss its structure and the trade-offs made to accommodate the GPU hardware.

An Efficient Implementation of Smith Waterman Algorithm on GPU Using CUDA, for Massively Parallel Scanning of Sequence Databases

Łukasz Ligowski and Witold Rudnicki
Interdisciplinary Centre for Mathematical and Computational Modelling
University of Warsaw
Warsaw, Poland
W.Rudnicki@icm.edu.pl

Abstract

The Smith Waterman algorithm for sequence alignment is one of the main tools of bioinformatics. It is used for sequence similarity searches and alignment of similar sequences. The high end Graphical Processing Unit (GPU), used for processing graphics on desktop computers, deliver computational capabilities exceeding those of CPUs by an order of magnitude. Recently these capabilities became accessible for general purpose computations thanks to CUDA programming environment on Nvidia GPUs and ATI Stream Computing environment on ATI GPUs. Here we present an efficient implementation of the Smith Waterman algorithm on the Nvidia GPU. The algorithm achieves more than 3.5 times higher per core performance than previously published implementation of the Smith Waterman algorithm on GPU, reaching more than 70% of theoretical hardware performance. The differences between current and earlier approaches are described showing the example for writing efficient code on GPU.

Stochastic Multi-particle Brownian Dynamics Simulation of Biological Ion Channels: A Finite Element Approach

May Siksik
Electrical and Computer Engineering
University of British Columbia
Vancouver, Canada
mays@ece.ubc.ca

Vikram Krishnamurthy
Electrical and Computer Engineering
University of British Columbia
Vancouver, Canada
vikramk@ece.ubc.ca

Abstract

Biological ion channels are protein tubes that span the cell membrane. They provide a conduction pathway and regulate the flow of ions though the low dielectric membrane. Modeling the dynamics of these channels is crucial in understanding their functionality. This paper proposes a novel simulation framework for modeling ion channels that is based on Finite Element Method (FEM). By using FEM, this is the first framework to allow the use of multiple dielectric constants inside the channel thus providing a more realistic model of the channel. Due to the run-time complexity of the problem, lookup tables must be constructed in memory to store precalculated electric potential information. Because of the large number of elements involved in FEM and channel resolution requirements there is the potential for very large lookup tables leading to a performance "bottleneck". This paper discusses strategies for minimizing table size and shows that currently available personal computers are sufficient for attaining reasonable levels of accuracy. For the framework proposed, results show diminishing returns in accuracy with tables sized greater than 2.2 GB.

High-throughput protein structure determination using grid computing

Jason W. Schmidberger[1], Blair Bethwaite[4], Colin Enticott[4], Mark A. Bate[1], Steve G. Androulakis[1], Noel Faux[5], Cyril F. Reboul[1,2], Jennifer M. N. Phan[1], James C. Whisstock[1,2], Wojtek J. Goscinski[3], Slavisa Garic[4], David Abramson[3,4], and Ashley M. Buckle[1,2]
[1]Department of Biochemistry and Molecular Biology,
[2]ARC Centre of Excellence in Structural and Functional Microbial Genomics,
[3]Monash eResearch Centre,
[4]Clayton School of Information Technology,
Monash University, Victoria 3800, Australia.
[5]NICTA Victoria Research Laboratory at The University of Melbourne, Australia.

Abstract

Determining the X-ray crystallographic structures of proteins using the technique of molecular replacement (MR) can be a time and labor-intensive trial-and-error process, involving evaluating tens to hundreds of possible solutions to this complex 3D jigsaw puzzle. For challenging cases indicators of success often do not appear until the later stages of structure refinement, meaning that weeks or even months could be wasted evaluating MR solutions that resist refinement and do not lead to a final structure. In order to improve the chances of success as well as decrease this timeframe, we have developed a novel grid computing approach that performs many MR calculations in parallel, speeding up the process of structure determination from weeks to hours. This high-throughput approach also allows parameter sweeps to be performed in parallel, improving the chances of MR success.

Folding@home: Lessons From Eight Years of Volunteer Distributed Computing

Adam L. Beberg[1], Daniel L. Ensign[2], Guha Jayachandran[1], Siraj Khaliq[1], Vijay S. Pande[2]
[1]Computer Science Dept, Stanford University
[2]Chemistry Department, Stanford University
{beberg@cs., densign@, guha@cs., siraj@cs., pande@}stanford.edu

Abstract

Accurate simulation of biophysical processes requires vast computing resources. Folding@home is a distributed computing system first released in 2000 to provide such resources needed to simulate protein folding and other biomolecular phenomena. Now operating in the range of 5 PetaFLOPS sustained, it provides more computing power than can typically be gathered and operated locally due to cost, physical space, and electrical/cooling load. This paper describes the architecture and operation of Folding@home, along with some lessons learned over the lifetime of the project.

Workshop 7

Advances in Parallel and Distributed Computing Models

APDCM 2009

Graph Orientation to Maximize the Minimum Weighted Outdegree

Yuichi Asahiro
Department of Social Information Systems,
Kyushu Sangyo University,
Higashi-ku, Fukuoka 813-8503, Japan.
asahiro@is.kyusan-u.ac.jp

Jesper Jansson
Ochanomizu University,
Bunkyo-ku, Tokyo 112-8610, Japan.
Jesper.Jansson@ocha.ac.jp

Eiji Miyano
Department of Systems Design and Informatics,
Kyushu Institute of Technology,
Iizuka, Fukuoka 820-8502, Japan.
miyano@ces.kyutech.ac.jp

Hirotaka Ono
Department of computer Science
and Communication Engineering,
Kyushu University,
Nishi-ku, Fukuoka 819-0395, Japan.
ono@csce.kyushu-u.ac.jp

Abstract

We study a new variant of the graph orientation problem called MaxMinO, where the input is an undirected, edge-weighted graph and the objective is to assign a direction to each edge so that the minimum weighted outdegree (taken over all vertices in the resulting directed graph) is maximized. All edge weights are assumed to be positive integers. This problem is closely related to the job scheduling on parallel machines, called the machine covering problem, where its goal is to assign jobs to parallel machines such that each machine is covered as much as possible. First, we prove that MaxMinO is strongly NP-hard and cannot be approximated within a ratio of $2 - \epsilon$ for any constant $\epsilon > 0$ in polynomial time unless P=NP, even if all edge weights belong to $\{1, 2\}$, every vertex has degree at most three, and the input graph is bipartite or planar. Next, we show how to solve MaxMinO exactly in polynomial time for the special case in which all edge weights are equal to 1. This technique gives us a simple polynomial-time $\frac{w_{max}}{w_{min}}$-approximation algorithm for MaxMinO, where w_{max} and w_{min} denote the maximum and minimum weights among all the input edges. Furthermore, we also observe that this approach yields an exact algorithm for the general case of MaxMinO whose running time is polynomial whenever the number of edges having weight larger than w_{min} is at most logarithmic in the number of vertices. Finally, we show that MaxMinO is solvable in polynomial time if the input is a cactus graph.

Uniform Scattering of Autonomous Mobile Robots in a Grid

Lali Barrière[1], Paola Flocchini[2], Eduardo Mesa-Barrameda[3] and Nicola Santoro[4]
[1]Universitat Politcnica de Catalunya [2]University of Ottawa [3]Universidad de la Habana
[4]Carleton University

Abstract

We consider the uniform scattering problem for a set of autonomous mobile robots deployed in a grid network: starting from an arbitrary placement in the grid, using purely localized computations, the robots must move so to reach in finite time a state of static equilibrium in which they cover uniformly the grid. The theoretical quest is on determining the minimal capabilities needed by the robots to solve the problem.

We prove that uniform scattering is indeed possible even for very weak robots. The proof is constructive. We present a provably correct protocol for uniform self-deployment in a grid. The protocol is fully localized, collision-free, and it makes minimal assumptions; in particular: (1) it does not require any direct or explicit communication between robots; (2) it makes no assumption on robots synchronization or timing, hence the robots can be fully asynchronous in all their actions; (3) it requires only a limited visibility range; (4) it uses at each robot only a constant size memory, hence computationally the robots can be simple Finite-State Machines; (5) it does not need a global localization system but only orientation in the grid (e.g., a compass); (6) it does not require identifiers, hence the robots can be anonymous and totally identical.

Resource Allocation Strategies for Constructive In-Network Stream Processing

Anne Benoit[1], Henri Casanova[2], Veronika Rehn-Sonigo[1] and Yves Robert[1]
[1]École Normale Supérieure de Lyon, France
[2]University of Hawaii at Manoa, Honolulu, USA
{Anne.Be|Veronika.Rehn|Yves.Robert}@ens-lyon.fr,
henric@hawaii.edu

Abstract

We consider the operator mapping problem for innetwork stream processing, i.e., the application of a tree of operators in steady-state to multiple data objects that are continuously updated at various locations in a network. Examples of in-network stream processing include the processing of data in a sensor network, or of continuous queries on distributed relational databases. Our aim is to provide the user a set of processors that should be bought or rented in order to ensure that the application achieves a minimum steady-state throughput, and with the objective of minimizing platform cost. We prove that even the simplest variant of the problem is NP-hard, and we design several polynomial time heuristics, which are evaluated via extensive simulations and compared to theoretical bounds.

Filter placement on a pipelined architecture

Anne Benoit, Fanny Dufossé and Yves Robert
École Normale Supérieure de Lyon, France
{Anne.Benoit|Fanny.Dufosse|Yves.Robert}@ens-lyon.fr

Abstract

In this paper, we explore the problem of mapping filtering query services on chains of heterogeneous processors. Two important optimization criteria should be considered in such a framework. The period, which is the inverse of the throughput, measures the rate at which data sets can enter the system. The latency measures the response time of the system in order to process one single data set entirely. We provide a comprehensive set of complexity results for period and latency optimization problems, with proportional or arbitrary computation costs, and without or with communication costs. We present polynomial algorithms for problems whose dependence graph is a linear chain (hence a fixed ordering of the filtering services). For independent services, the problems are all NP-complete except latency minimization with proportional computation costs, which was shown polynomial in [6].

Crosstalk-Free Mapping of Two-dimensional Weak Tori on Optical Slab Waveguides

Hatem M. El-Boghdadi
Computer Engineering Department
Faculty of Engineering, Cairo University
Giza, Egypt
helboghdadi@eng.cu.edu.eg

Abstract

While optical slab waveguides can deliver a huge bandwidth for the communication need through offering a huge number of communication channels, they require a large number of high speed lasers and photodetectors. This makes a limited use of the offered huge bandwidth. Some trials were proposed to implement communication networks on the slab waveguides [7]. However, the proposed mappings suffer from the possibility of crosstalk among different channels if they are to be used simultaneously.

In this paper, we consider solving the problem of crosstalk when mapping weak two-dimensional tori on optical slab waveguides. We introduce the notion of diagonal pair and use it in the proposed mapping. The approach assigns edges to channels such that the mapping guarantees a crosstalk free communication between nodes (no two adjacent channels are used at the same communication step.) We also consider the cost of the mapping in terms of the number of lasers and the number of photodetectors. Our results show that the cost is within constants to the cost lower bound.

Combining Multiple Heuristics on Discrete Resources

Marin Bougeret , Pierre-François Dutot, Alfredo Goldman, Yanik Ngoko and Denis Trystram
LIG, Grenoble University, France

Abstract

In this work we study the portfolio problem which is to find a good combination of multiple heuristics to solve given instances on parallel resources in minimum time. The resources are assumed to be discrete, it is not possible to allocate a resource to more than one heuristic. Our goal is to minimize the average completion time of the set of instances, given a set of heuristics on homogeneous discrete resources. This problem has been studied in the continuous case (Sayag et al, ”Combining multiple heuristics”). We first show that the problem is hard and that there is no constant ratio polynomial approximation unless $P = NP$ in the general case. Then, we design several approximation schemes for a restricted version of the problem where each heuristic must be used at least once. These results are obtained by using oracle with several guesses, leading to various tradeoff between the size of required information and the approximation ratio. Some additional results based on simulations are finally reported using a benchmark of instances on SAT solvers.

A Distributed Approach for the Problem of Routing and Wavelength Assignment in WDM Networks

Simone Cintra Chagas[1], Eber Huanca Cayo[2], Koji Nakano [3] and Jacir Luiz Bordim [4]
[1]School of Electrical Engineering , Faculty of Technology,
University of Brasília, Brasília-DF, Brazil
simone_cintra@unb.br
[2]School of Mechanic Engineering, Faculty of Technology, University of Brasília , Brasília-DF, Brazil
Email:eber@unb.br
[3]Department of Information Engineering,
School of Engineering, Hiroshima University , Higashi-Hiroshima, JAPAN
nakano@hiroshima-u.ac.jp
[4]Department of Computer Science,
Campus Universitario - Asa Norte, University of Brasília, Brasília-DF, Brazil
bordim@unb.br

Abstract

The main contribution of this work is to propose a distributed on-demand routing and wavelength assignment algorithm for WDM networks. The proposed algorithm, termed WDM-DSR, is capable to select routes and establish light-paths via message exchanges without imposing a major overhead on the network. Also, we show that the proposed scheme can be used to balance the load in a WDM network. The simulation results show that the proposed solution is comparable with the other algorithms that demands for a much higher computational and message costs.

Self-Stabilizing *k*-out-of-*l* Exclusion on Tree Networks

Ajoy K. Datta[1], Stéphane Devismes[2], Florian Horn[3], and Lawrence L. Larmore[1]
[1]School of Computer Science - University of Nevada - Las Vegas, USA
lastname@cs.unlv.edu
[2]VERIMAG - Université Joseph Fourier - Grenoble, France
stephane.devismes@imag.fr
[3]LIAFA - Université Paris Denis Diderot- Paris, France
florian.horn@liafa.jussieu.fr

Abstract

In this paper, we address the problem of k-out-of-l, a generalization of the mutual exclusion problem, in which there are l units of a shared resource, and any process can request up to k units ($1 \leq k \leq l$). We propose the first deterministic self-stabilizing distributed k-out-of-l protocol in message-passing systems for asynchronous oriented tree networks which assumes bounded local memory for each process.

Improving Accuracy of Host Load Predictions on Computational Grids by Artificial Neural Networks

Truong Vinh Truong Duy[1],Yukinori Sato and Yasushi Inoguchi
[1]Graduate School of Information Science,
[2]Center for Information Science,
Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa, 923-1292 Japan
{duytvt, yukinori, inoguchi}@jaist.ac.jp

Abstract

The capability to predict the host load of a system is significant for computational grids to make efficient use of shared resources. This paper attempts to improve the accuracy of host load predictions by applying a neural network predictor to reach the goal of best performance and load balance. We describe feasibility of the proposed predictor in a dynamic environment, and perform experimental evaluation using collected load traces. The results show that the neural network achieves a consistent performance improvement with surprisingly low overhead. Compared with the best previously proposed method, the typical 20:10:1 network reduces the mean and standard deviation of the prediction errors by approximately 60% and 70%, respectively. The training and testing time is extremely low, as this network needs only a couple of seconds to be trained with more than 100,000 samples in order to make tens of thousands of accurate predictions within just a second.

Computation with a constant number of steps in membrane computing

Akihiro Fujiwara and Takeshi Tateishi
Department of Computer Science and Electronics,
Kyushu Institute of Technology
680-4 Kawazu, Iizuka, Fukuoka 820-8502, JAPAN
fujiwara@cse.kyutech.ac.jp

Abstract

In the present paper, we propose P systems that work in a constant number of steps. We first propose two P systems for computing multiple input logic functions. An input of the logic functions is a set of n binary numbers of m bits, and an output is a binary number defined by the logic functions. The first and second P systems compute AND and EX-OR functions for the input, and both of the P systems work in a constant number of steps using $O(mn)$ types of objects, a constant number of membranes, and evolution rules of size $O(mn)$.

Next, we propose the P system for the addition of two binary numbers of m bits. The P system works in a constant number of steps using $O(m)$ types of objects, a constant number of membranes and evolution rules of size $O(m^2)$. We also introduce a P system that computes the addition of two vectors of size n using the above P system as a sub-system. The P system for vector addition works in a constant number of steps using $O(mn)$ types of objects, a constant number of membranes, and evolution rules of size $O(m^2n)$.

New Implementation of a BSP Composition Primitive with Application to the Implementation of Algorithmic Skeletons

Frédéric Gava
Laboratory of Algorithms, Complexity and Logic (LACL)
University of Paris-East
Créteil-Paris, France
gava@univ-paris12.fr

Ilias Garnier
LIST laboratory
CEA Saclay, Essonne, France
ilias.garnier@cea.fr

Abstract

BSML is a ML based language designed to code Bulk Synchronous Parallel (BSP) algorithms. It allows an estimation of execution time, avoids deadlocks and non-determinism. BSML proposes an extension of ML programming with a small set of primitives. One of these primitives, called parallel superposition, allows the parallel composition of two BSP programs. Nevertheless, its past implementation used system threads and have unjustified limitations. This paper presents a new implementation of this primitive based on a continuation-passing-style (CPS) transformation guided by a flow analysis. To test it and show its usefulness, we also have implemented the OCamlP3l algorithmic skeletons and compared their efficiencies with the original ones.

Distributed Selfish Bin Packing

Flávio K. Miyazawa, André L. Vignatti
Institute of Computing
University of Campinas
Campinas, Brazil 13084-971C6176
{fkm,vignatti}@ic.unicamp.br

Abstract

We consider a game-theoretic *bin packing problem* with identical items, and we study the convergence time to a *Nash equilibrium*. In the model proposed, users choose their strategy simultaneously. We deal with two bins and multiple bins cases. We consider the case when users know the load of all bins and a case with less information. We consider two approaches, depending if the system can undo movements that lead to infeasible states. In the two bins case, we show an $O(\log\log n)$ bound when undo movements are allowed. In multiple bins case, we show an $O(\log n)$ and an $O(nm)$ bounds when undo movements are allowed and when they are not allowed, resp. In the case with less information, we show an $O(m\log n)$ and an $O(n^3m)$ bounds when undo movements are allowed and when they are not allowed, resp.

Predictive Analysis and Optimisation of Pipelined Wavefront Computations

G.R. Mudalige, S.D. Hammond, J.A. Smith, S.A. Jarvis
High Performance Systems Group, University of Warwick
Coventry, CV4 7AL, UK
{g.r.mudalige, sdh, jas, saj}@dcs.warwick.ac.uk

Abstract

Pipelined wavefront computations are a ubiquitous class of parallel algorithm used for the solution of a number of scientific and engineering applications. This paper investigates three optimisations to the generic pipelined wavefront algorithm, which are investigated through the use of predictive analytic models. The modelling of potential optimisations is supported by a recently developed reusable LogGP-based analytic performance model, which allows the speculative evaluation of each optimisation within the context of an industry-strength pipelined wavefront benchmark developed and maintained by the United Kingdom Atomic Weapons Establishment (AWE). The paper details the quantitative and qualitative benefits of: (1) parallelising computation blocks of the wavefront algorithm using OpenMP; (2) a novel restructuring/shifting of computation within the wavefront code and, (3) performing simultaneous multiple sweeps through the data grid.

RSA Encryption and Decryption using the Redundant Number System on the FPGA

Koji Nakano, Kensuke Kawakami and Koji Shigemoto
Department of Information Engineering, Hiroshima University
Kagamiyama 1-4-1, Higashi-Hiroshima, JAPAN

Abstract

The main contribution of this paper is to present efficient hardware algorithms for the modulo exponentiation P^E mod M used in RSA encryption and decryption, and implement them on the FPGA. The key ideas to accelerate the modulo exponentiation are to use the Montgomery modulo multiplication on the redundant radix-64K number system in the FPGA, and to use embedded 18 × 18-bit multipliers and embedded 18k-bit block RAMs in effective way. Our hardware algorithms for the modulo exponentiation for R-bit numbers P, E, and M can run in less than $(2R + 4)(R/16 + 1)$ clock cycles and in expected $(1.5R+4)(R/16+1)$ clock cycles. We have implemented our modulo exponentiation hardware algorithms on Xilinx VirtexII Pro family FPGA XC2VP30-6. The implementation results shows that our hardware algorithm for 1024-bit modulo exponentiation can be implemented to run in less than 2.521ms and in expected 1.892ms.

Table-based Method for Reconfigurable Function Evaluation

María Teresa Signes Pont, Higinio Mora
Mora and Juan Manuel García Chamizo
Departamento de Tecnología Informática y
Computación
University of Alicante, 03690,
San Vicente del Raspeig, Alicante, Spain
{teresa; hmora; juanma}@dtic.ua.es

Gregorio de Miguel Casado,
Departamento de Informática e Ingeniería
de Sistemas.
Centro Politécnico Superior.
Edf. Ada Byron, C/ María de Luna, 1,
University of Zaragoza, Spain
gmiguel@unizar.es

Abstract

This paper presents a new approach to function evaluation using tables. The proposal argues for the use of a more complete primitive, namely a weighted sum, which converts the calculation of the function values into a recursive operation defined by a two input table. This weighted sum can be tuned for different values of the weighting parameters holding the features of the specific function to be evaluated. A parametric architecture for reconfigurable FPGA-based hardware implements the design. Our method has been tested in the calculation of the function sine. The comparison with other well-known proposals reveals the advantages of our approach, because it provides memory and hardware resource saving as well as a good trade-off between speed and error.

Analytical Model of Inter-Node Communication under Multi-Versioned Coherence Mechanisms

Shigero Sasaki
NEC Corporation

Atsuhiro Tanaka
NEC Corporation

Abstract

Our goal is to predict the performance of multi-node systems consisting of identical processing nodes based on single node profiles. The performance of multi-node systems significantly depends on the amount of inter-node communication. Therefore, we built an analytical model of the communication amount, i.e., the number of transfers of cached copies, on multi-node systems with coherence mechanisms that support multi-versioning. Multi-versioned mechanisms are assumed because databases are most likely to be the bottleneck and because a typical clustered database has one of these mechanisms. In our model, the number of transfers of copies of a block per write access is expressed as a function of a write ratio, the number of nodes, and the lock down factor which denotes how many versions of copies can exist. To empirically verify our analytical and theoretical model, we compared the number of transfers that predicted by our model and that counted by a toy simulator of multi-versioned mechanisms.

Deciding Model of Population Size in Time-Constrained Task Scheduling

Wei Sun
System Platform Research Laboratories, NEC Corporation
1753, Shimonumabe, Nakahara-ku, Kawasaki, Kanagawa, 211-8666, Japan
w-sun@ap.jp.nec.com

Abstract

Genetic algorithms (GAs) have been well applied in solving scheduling problems and their performance advantages have also been recognized. However, practitioners are often troubled by parameters setting when they are tuning GAs. Population Size (PS) has been shown to greatly affect the efficiency of GAs. Although some population sizing models exist in the literature, reasonable population sizing for task scheduling is rarely observed. In this paper, based on the PS deciding model in [8], we present a model to predict the optimal PS for the GA applied in time-constrained task scheduling, where the efficiency of GAs is more necessitated than in solving other kinds of problems. In the experimental evaluation, our deciding model can well predict the success ratio of the GA, given different population sizes.

Performance Study of Interference on Sharing GPU and CPU Resources with Multiple Applications

Shinichi Yamagiwa
INESC-ID
Rua Alves Redol, 9
1000-029 Lisboa, Portugal
yama@inesc-id.pt

Koichi Wada
Department of Computer Science
University of Tsukuba
Tenno-dai 1-1-1, Tsukuba, Ibaraki, Japan
wada@cs.tsukuba.ac.jp

Abstract

In the last years, the performance and capabilities of Graphics Processing Units (GPUs) improved drastically, mostly due to the demands of the entertainment market, with consumers and companies alike pushing for improvements in the level of visual fidelity, which is only achieved with high performing GPU solutions. Beside the entertainment market, there is an ongoing global research effort for using such immense computing power for applications beyond graphics, such as the domain of general purpose computing. Efficiently combining these GPUs resources with existing CPU resources is also an important and open research task. This paper is a contribution to that effort, focusing on analysis of performance factors of combining both resource types, while introducing also a novel job scheduler that manages these two resources. Through experimental performance evaluation, this paper reports what are the most important factors and design considerations that must be taken into account while designing such job scheduler.

Workshop 8
Communication Architecture for Clusters
CAC 2009

A Power-Aware, Application-Based Performance Study Of Modern Commodity Cluster Interconnection Networks

Torsten Hoefler, Timo Schneider and Andrew Lumsdaine
Open Systems Laboratory
Indiana University
Bloomington IN 47405, USA
{htor,timoschn,lums}@cs.indiana.edu

Abstract

Microbenchmarks have long been used to assess the performance characteristics of high-performance networks. It is generally assumed that microbenchmark results indicate the parallel performance of real applications. This paper reports the results of performance studies using real applications in a strictly controlled environment with different networks. In particular, we compare the performance of Myrinet and InfiniBand, and analyze them with respect to microbenchmark performance, real application performance and power consumption.

An analysis of the impact of multi-threading on communication performance

François Trahay, Élisabeth Brunet and Alexandre Denis
INRIA, LABRI, Université Bordeaux 1
351 cours de la Libération
F-33405 TALENCE, FRANCE
{trahay,brunet,denis}@labri.fr

Abstract

Although processors become massively multicore and therefore new programming models mix message passing and multi-threading, the effects of threads on communication libraries remain neglected. Designing an efficient modern communication library requires precautions in order to limit the impact of thread-safety mechanisms on performance. In this paper, we present various approaches to building a thread-safe communication library and we study their benefit and impact on performance. We also describe and evaluate techniques used to exploit idle cores to balance the communication library load across multicore machines.

RI2N/DRV: Multi-link Ethernet for High-Bandwidth and Fault-Tolerant Network on PC Clusters

Shinichi Miura[1], Toshihiro Hanawa[1,2], Taiga Yonemoto[2]
Taisuke Boku[1,2] and Mitsuhisa Sato[1,2]
[1]Center for Computational Sciences, University of Tsukuba
[2]Graduate School of Systems and Information Engineering, University of Tsukuba
1-1-1 Tennodai, Tsukuba, Ibaraki 305-8577, Japan
{miura, yonemoto}@hpcs.cs.tsukuba.ac.jp
hanawa@ccs.tsukuba.ac.jp
{taisuke, msato}@cs.tsukuba.ac.jp

Abstract

Although recent high-end interconnection network devices and switches provide a high performance to cost ratio, most of the small to medium sized PC clusters are still built on the commodity network, Ethernet. To enhance performance on commonly used Gigabit Ethernet networks, link aggregation or binding technology is used. Currently, Linux kernels are equipped with software named Linux Channel Bonding (LCB), which is based IEEE802.3ad Link Aggregation technology. However, standard LCB has the disadvantage of mismatch with the TCP protocol; consequently, both large latency and bandwidth instability can occur. Fault-tolerance feature is supported by LCB, but the usability is not sufficient. We developed a new implementation similar to LCB named Redundant Interconnection with Inexpensive Network with Driver (RI2N/DRV) for use on Gigabit Ethernet. RI2N/DRV has a complete software stack that is very suitable for TCP, an upper layer protocol. Our algorithm suppresses unnecessary ACK packets and retransmission of packets, even in imbalanced network traffic and link failures on multiple links. It provides both high-bandwidth and fault-tolerant communication on multi-link Gigabit Ethernet. We confirmed that this system improves the performance and reliability of the network, and our system can be applied to ordinary UNIX services such as network file system (NFS), without any modification of other modules.

Efficient and Deadlock-Free Reconfiguration for Source Routed Networks

Åshild Grønstad Solheim[12], Olav Lysne[12], Aurelio Bermúdez[3], Rafael Casado[3], Thomas Sødring[1],
Tor Skeie[12] and Antonio Robles-Gómez[3]
[1]Networks and Distributed Systems Group, Simula Research Laboratory, Lysaker, Norway
aashig@simula.no
[2]Department of Informatics, University of Oslo, Oslo, Norway
[3]Computing Systems Department, University of Castilla-La Mancha, Albacete, Spain

Abstract

Overlapping Reconfiguration is currently the most efficient method to reconfigure an interconnection network, but is only valid for systems that apply distributed routing. This paper proposes a solution which enables utilization of Overlapping Reconfiguration in a source routed environment. We demonstrate how a synchronized injection of tokens has a significant impact on the performance of the method. Furthermore, we propose and evaluate an optimization of the original algorithm that reduces (and in some cases even eliminates) performance issues caused by the token forwarding regime, such as increased latency and decreased throughput.

Deadlock Prevention by Turn Prohibition in Interconnection Networks

Lev Levitin
Dept. of Computer Engineering
Boston University
Boston, MA 02215, USA
levitin@bu.edu

Mark Karpovsky
Dept. Computer Engineering
Boston University
Boston, MA 02215, USA
markkar@bu.edu

Mehmet Mustafa
Dept. of Computer Engineering
Boston University
Boston, MA 02215, USA
mmustafa@bu.edu

Abstract

In this paper we consider the problem of constructing minimal cycle-breaking sets of turns for graphs that model communication networks, as a method to prevent deadlocks in the networks. We present a new cycle-breaking algorithm called Simple Cycle-Breaking or SCB algorithm that is considerably simpler than earlier algorithms. The SCB algorithm guarantees that the fraction of prohibited turns does not exceed 1/3. Experimental simulation results for the SCB algorithm are shown.

Implementation and Analysis of Nonblocking Collective Operations on SCI Networks

Christian Kaiser
Dolphin Interconnect Solutions
Siebengebirgsblick 26,
53343 Wachtberg, Germany
kaiser@dolphinics.com

Torsten Hoefler
Open Systems Laboratory,
Indiana University
150 S Woodlawn Ave,
Bloomigton, IN 47405, USA
htor@cs.indiana.edu

Boris Bierbaum, Thomas Bemmerl
Chair for Operating Systems,
RWTH Aachen University
Kopernikusstr. 16,
52056 Aachen, Germany
{bierbaum,bemmerl}@lfbs.rwth-aachen.de

Abstract

Nonblocking collective communication operations are currently being considered for inclusion into the MPI standard and are an area of active research. The benefits of such operations are documented by several recent publications, but so far, research concentrates on InfiniBand clusters. This paper describes an implementation of nonblocking collectives for clusters with the Scalable Coherent Interface (SCI) interconnect. We use synthetic and application kernel benchmarks to show that with nonblocking functions for collective communication performance enhancements can be achieved on SCI systems. Our results indicate that for the implementation of these nonblocking collectives data transfer methods other than those usually used for the blocking version should be considered to realize such improvements.

Designing Multi-Leader-Based Allgather Algorithms for Multi-Core Clusters

Krishna Kandalla, Hari Subramoni, Gopal Santhanaraman, Matthew Koop and Dhabaleswar K. Panda
Department of Computer Science and Engineering, The Ohio State University
{kandalla, subramon, santhana, koop, panda}@cse.ohio-state.edu

Abstract

The increasing demand for computational cycles is being met by the use of multi-core processors. Having large number of cores per node necessitates multi-core aware designs to extract the best performance. The Message Passing Interface (MPI) is the dominant parallel programming model on modern high performance computing clusters. The MPI collective operations take a significant portion of the communication time for an application. The existing optimizations for collectives exploit shared memory for intranode communication to improve performance. However, it still would not scale well as the number of cores per node increase. In this work, we propose a novel and scalable multi-leader-based hierarchical Allgather design. This design allows better cache sharing for Non-Uniform Memory Access (NUMA) machines and makes better use of the network speed available with high performance interconnects such as InfiniBand. The new multi-leader-based scheme achieves a performance improvement of up to 58% for small messages and 70% for medium sized messages.

Using Application Communication Characteristics to Drive Dynamic MPI Reconfiguration

Manjunath Gorentla Venkata, Patrick G. Bridges and Patrick M. Widener
Department of Computer Science
University of New Mexico
Albuquerque NM 87131
{manjugv,bridges,pmw} @cs.unm.edu

Abstract

Modern HPC applications, for example adaptive mesh refinement and multi-physics codes, have dynamic communication characteristics which result in poor performance on current MPI implementations. Current MPI implementations do not change transport protocols or allocate resources based on the application characteristics, resulting in degraded application performance. In this paper, we describe PRO-MPI, a Protocol Reconfiguration and Optimization system for MPI that we are developing to meet the needs of dynamic modern HPC applications. PRO-MPI uses profiles of past application communication characteristics to dynamically reconfigure MPI protocol choices. We show that such dynamic reconfiguration can improve the performance of important MPI applications significantly when exact communication profiles are known. We also present preliminary data showing that profiles from past application runs with different (but related) inputs can be used to optimize the performance of later application runs.

Decoupling Memory Pinning from the Application with Overlapped on-Demand Pinning and MMU Notifiers

Brice Goglin
INRIA – LaBRI – Université Bordeaux 1
351 cours de la Libération, F-33405 Talence – France
Brice.Goglin@inria.fr

Abstract

High-performance cluster networks achieve very high throughput thanks to zero-copy techniques that require pinning of application buffers in physical memory. The Open-MX stack implements message passing over generic Ethernet hardware with similar needs.

We present the design of an innovative pinning model in Open-MX based on the decoupling of memory pinning from the application. This idea eases the implementation of a reliable pinning cache in the kernel and enables full overlap of pinning with communication. The pinning cache enables performance improvement when the application reuses the same buffers multiple times, while overlapped pinning is also applicable to other applications.

Performance evaluation shows that both these optimizations bring from 5 up to 20 % throughput improvements depending on the host and network performance.

Improving RDMA-based MPI Eager Protocol for Frequently-used Buffers

Mohammad J. Rashti and Ahmad Afsahi
Department of Electrical and Computer Engineering
Queen's University, Kingston, ON, CANADA K7L 3N6
mohammad.rashti@ece.queensu.ca ahmad.afsahi@queensu.ca

Abstract

MPI is the main standard for communication in high-performance clusters. MPI implementations use the Eager protocol to transfer small messages. To avoid the cost of memory registration and prenegotiation, the Eager protocol involves a data copy to intermediate buffers at both sender and receiver sides. In this paper, however, we propose that when a user buffer is used frequently in an application, it is more efficient to register the sender buffer and avoid the sender-side data copy. The performance results of our proposed Eager protocol on MVAPICH2 over InfiniBand indicate that up to 14% improvement can be achieved in a single medium-size message latency, comparable to a maximum 15% theoretical improvement on our platform. We also show that collective communications such as broadcast can benefit from the new protocol by up to 19%. In addition, the communication time in MPI applications with high buffer reuse is improved using this technique.

Workshop 9

High-Performance, Power-Aware Computing

HPPAC 2009

On the Energy Efficiency of Graphics Processing Units for Scientific Computing

S. Huang, S. Xiao and W. Feng
Department of Computer Science
Virginia Tech
{huangs, shucai, wfeng}@vt.edu

Abstract

The graphics processing unit (GPU) has emerged as a computational accelerator that dramatically reduces the time to discovery in high-end computing (HEC). However, while today's state-of-the-art GPU can easily reduce the execution time of a parallel code by many orders of magnitude, it arguably comes at the expense of significant power and energy consumption. For example, the NVIDIA GTX 280 video card is rated at 236 watts, which is as much as the rest of a compute node, thus requiring a 500-W power supply. As a consequence, the GPU has been viewed as a "non-green" computing solution.

This paper seeks to characterize, and perhaps debunk, the notion of a "power-hungry GPU" via an empirical study of the performance, power, and energy characteristics of GPUs for scientific computing. Specifically, we take an important biological code that runs in a traditional CPU environment and transform and map it to a hybrid CPU+GPU environment. The end result is that our hybrid CPU+GPU environment, hereafter referred to simply as GPU environment, delivers an energy-delay product that is multiple orders of magnitude better than a traditional CPU environment, whether unicore or multicore.

Power-Aware Dynamic Task Scheduling for Heterogeneous Accelerated Clusters

Tomoaki Hamano, Toshio Endo and Satoshi Matsuoka
Tokyo Institute of Technology/JST, Japan

Abstract

Recent accelerators such as GPUs achieve better cost-performance and watt-performance ratio, while the range of their application is more limited than general CPUs. Thus heterogeneous clusters and supercomputers equipped both with accelerators and general CPUs are becoming popular, such as LANL's Roadrunner and our own TSUBAME supercomputer. Under the assumption that many applications will run both on CPUs and accelerators but with varying speed and power consumption characteristics, we propose a task scheduling scheme that optimize overall energy consumption of the system. We model task scheduling in terms of the scheduling makespan and energy to be consumed for each scheduling decision. We define *acceleration factor* to normalize the effect of acceleration per each task. The proposed scheme attempts to improve energy efficiency by effectively adjusting the schedule based on the acceleration factor. Although in the paper we adopted the popular EDP (Energy-Delay Product) as the optimization metric, our scheme is agnostic on the optimization function. Simulation studies on various sets of tasks with mixed acceleration factors, the overall makespan closely matched the theoretical optimal, while the energy consumption was reduced up to 13.8%.

Clock Gate on Abort: Towards Energy-Efficient Hardware Transactional Memory

Sutirtha Sanyal[1], Sourav Roy[2], Adrian Cristal[1], Osman S. Unsal[1], Mateo Valero[1]
[1]Barcelona Supercomputing Center, Barcelona, Spain; [2]Freescale Semiconductors, India
[1]{sutirtha.sanyal,adrian.cristal,osman.unsal,mateo.valero}@bsc.es;[2]sourav.roy@freescale.com

Abstract

Transactional Memory (TM) is an emerging technology which promises to make parallel programming easier compared to earlier lock based approaches. However, as with any form of speculation, Transactional Memory too wastes a considerable amount of energy when the speculation goes wrong and transaction aborts. For Transactional Memory this wastage will typically be quite high because programmer will often mark a large portion of the code to be executed transactionally[**?**].

We are proposing to turn-off a processor dynamically by gating all its clocks, whenever any transaction running in it is aborted. We have described a novel protocol which can be used in the Scalable-TCC like Hardware Transactional Memory systems. Also in the protocol we are proposing a gating-aware contention management policy to set the duration of the clock gating period precisely so that both performance and energy can be improved.

With our proposal we got an average 19% savings in the total consumed energy and even an average speed-up of 4%.

Power-Aware Load Balancing Of Large Scale MPI Applications

Maja Etinski[†]
maja.etinski@bsc.es

Julita Corbalan[†]
julita.corbalan@bsc.es

Jesus Labarta[†]
jesus.labarta@bsc.es

Mateo Valero[†]
mateo.valero@bsc.es

Alex Veidenbaum[‡]
alexv@ics.uci.edu

[†]Barcelona Supercomputing Center
Jordi Girona 31, 08034 Barcelona, Spain

[‡]Department of Computer Science
University of California, Irvine CA

Abstract

Power consumption is a very important issue for HPC community, both at the level of one application or at the level of whole workload. Load imbalance of a MPI application can be exploited to save CPU energy without penalizing the execution time. An application is load imbalanced when some nodes are assigned more computation than others. The nodes with less computation can be run at lower frequency since otherwise they have to wait for the nodes with more computation blocked in MPI calls. A technique that can be used to reduce the speed is Dynamic Voltage Frequency Scaling (DVFS). Dynamic power dissipation is proportional to the product of the frequency and the square of the supply voltage, while static power is proportional to the supply voltage. Thus decreasing voltage and/or frequency results in power reduction. Furthermore, over-clocking can be applied in some CPUs to reduce overall execution time. This paper investigates the impact of using different gear sets , over-clocking, and application and platform propreties to reduce CPU power. A new algorithm applying DVFS and CPU over-clocking is proposed that reduces execution time while achieving power savings comparable to prior work. The results show that it is possible to save up to 60% of CPU energy in applications with high load imbalance. Our results show that six gear sets achieve, on average, results close to the continuous frequency set that has been used as a baseline.

The GREEN-NET Framework: Energy Efficiency in Large Scale Distributed Systems

Georges Da Costa[3], Jean-Patrick Gelas[1], Yiannis Georgiou[2], Laurent Lefèvre[1]
Anne-Cécile Orgerie[1], Jean-Marc Pierson[3], Olivier Richard[2], Kamal Sharma[4]

Abstract

The question of energy savings has been a matter of concern since a long time in the mobile distributed systems and battery-constrained systems. However, for large-scale non-mobile distributed systems, which nowadays reach impressive sizes, the energy dimension (electrical consumption) just starts to be taken into account.

In this paper, we present the GREEN-NET framework which is based on 3 main components: an ON/OFF model based on an Energy Aware Resource Infrastructure (EARI), an adapted Resource Management System (OAR) for energy efficiency and a trust delegation component to assume network presence of sleeping nodes.

Analysis of Trade-Off Between Power Saving and Response Time in Disk Storage Systems

E. Otoo, D. Rotem, S. C. Tsao
Lawrence Berkeley National Laboratory
University of California
Berkeley, CA 94720
{ejotoo, d_rotem, weafon}@lbl.gov

Abstract

It is anticipated that in the near future disk storage systems will surpass application servers and will become the primary consumer of power in the data centers. Shutting down of inactive disks is one of the more widespread solutions to save power consumption of disk systems. This solution involves spinning down or completely shutting off disks that exhibit long periods of inactivity and placing them in standby mode. A file request from a disk in standby mode will incur an I/O cost penalty as it takes time to spin up the disk before it can serve the file. In this paper, we address the problem of designing and implementing file allocation strategies on disk storage that save energy while meeting performance requirements of file retrievals. We present an algorithm for solving this problem with guaranteed bounds from the optimal solution. Our algorithm runs in $O(nlogn)$ time where n is the number of files allocated. Detailed simulation results and experiments with real life workloads are also presented.

Enabling Autonomic Power-Aware Management of Instrumented Data Centers

Nanyan Jiang and Manish Parashar
Center for Autonomic Computing
Department of Electrical and Computer Engineering
Rutgers University, Piscataway NJ 08855, USA
{nanyanj, parashar}@rutgers.edu

Abstract

Sensor networks support flexible, non-intrusive and fine-grained data collection and processing and can enable online monitoring of data center operating conditions as well as autonomic data center management. This paper describes the architecture and implementation of an autonomic power aware data center management framework, which is based on the integration of embedded sensors with computational models and workload schedulers to improve data center performance in terms of energy consumption and throughput. Specifically, workload schedulers use online information about data center operating conditions obtained from the sensors to generate appropriate management policies. Furthermore, local processing within the sensor network is used to enable timely responses to changes in operating conditions and determine job migration strategies. Experimental results demonstrate that the framework achieves near optimal management, and in-network analysis enables timely response while reducing overheads.

Modeling and Evaluating Energy-Performance Efficiency of Parallel Processing on Multicore Based Power Aware Systems

Rong Ge
Marquette University
rong.ge@marquett.edu

Xizhou Feng
Virginia Tech
fengx@vt.edu

Kirk W. Cameron
Virginia Tech
kcameron@cs.vt.edu

Abstract

In energy efficient high end computing, a typical problem is to find an energy-performance efficient resource allocation for computing a given workload. An analytical solution to this problem includes two steps: first estimating the performances and energy costs for the workload running with various resource allocations, and second searching the allocation space to identify the optimal allocation according to an energy-performance efficiency measure. In this paper, we develop analytical models to approximate performance and energy cost for scientific workloads on multicore based power aware systems. The performance models extend Amdahl's law and poweraware speedup model to the context of multicore-based power aware computing. The power and energy models describe the power effects of resource allocation and workload characteristics. As a proof of concept, we show model parameter derivation and model validation using performance, power, and energy profiles collected on a prototype multicore based power aware cluster.

Time-Efficient Power-Aware Scheduling for Periodic Real-Time Tasks

Da-Ren Chen[1], Chiun-Chieh Hsu[2] and Ming-Fong Lai[3]
[1]Dept. of Information Management Hwa Hsia Institute of Technology, Taiwan, R.O.C.
[2]Dep. of Information Management National Taiwan University of Science and Technology, Taiwan, R.O.C.
[3]Science & Technology Research and Information Center National Applied Research Laboratories

Abstract

In this paper, we pay attention to the inter-task dynamic voltage scaling (DVS) algorithms for periodic real-time task systems. We propose a fast dynamic reclaiming scheme for power-aware hard real-time systems and discuss their performances and time complexities against other inter-task DVS algorithms. The time complexity of our off-line and on-line algorithms are $O(nlogn)$ and $O(n)$, respectively, where n denotes the number of task.

The Green500 List: Year One

W. Feng and T. Scogland
{feng,njustn}@cs.vt.edu
Department of Computer Science
Virginia Tech

Abstract

The latest release of the Green500 List in November 2008 marked its one-year anniversary. As such, this paper aims to provide an analysis and retrospective examination of the Green500 List in order to understand how the list has evolved and what trends have emerged. In addition, we present community feedback on the Green500 List, particularly from two Green500 birds-of-a-feather (BoF) sessions at the International Supercomputing Conference in June 2008 and SC|08 in November 2008, respectively.

Workshop 10
High Performance Grid Computing
HPGC 2009

INFN-CNAF activity in the TIER-1 and GRID for LHC experiments

M. Bencivenni, M. Canaparo, F. Capannini, L. Carota, M. Carpenè, A. Cavalli, A. Ceccanti,
M. Cecchi, D. Cesini, A. Chierici, V. Ciaschini, A. Cristofori, S. Dal Pra, L. dell'Agnello,
D. De Girolamo, M. Donatelli, D. N. Dongiovanni, E. Fattibene, T. Ferrari, A. Ferraro, A. Forti,
A. Ghiselli, D. Gregori, G. Guizzunti, A. Italiano, L. Magnoni, B. Martelli, M. Mazzucato, G. Misurelli,
M. Onofri, A. Paolini, A. Prosperini, P. P. Ricci, E. Ronchieri, F. Rosso, D. Salomoni, V. Sapunenko,
V. Venturi, R. Veraldi, P. Veronesi, C. Vistoli, D. Vitlacil, S. Zani, R. Zappi
INFN-CNAF V.le Berti Pichat 6/2 40127 Bologna, Italy

Abstract

The four High Energy Physics (HEP) detectors at the Large Hadron Collider (LHC) at the European Organization for Nuclear Research (CERN) are among the most imp ortant experiments where the National Institute of Nuclear Physics (INFN) is bei ng actively involved. A Grid infrastructure of the World LHC Computing Grid (WLCG) has been developed by the HEP community leveragi ng on broader initiatives (e.g. EGEE in Europe, OSG in northen America) as a fra mework to exchange and maintain data storage and provide computing infrastructur e for the entire LHC community. INFN-CNAF in Bologna hosts the Italian Tier-1 site, which represents the biggest italian center in the WLCG distributed computing.

In the first part of this paper we will describe on the building of the Italian Tier-1 to cope with the WLCG computing requirements focusing on some peculiarit ies; in the second part we will analyze the INFN-CNAF contribution for the devel opement of the grid middleware, stressing in particular the characteristics of t he Virtual Organization Membership Service (VOMS), the de facto standard for aut horization on a grid, and StoRM, an implementation of the Storage Resource Manag er (SRM) specifications for POSIX file systems. In particular StoRM is used at INFN-CNAF in conjunction with General Parallel File System (GPFS) and we are als o testing an integration with Tivoli Storage Manager (TSM) to realize a complete Hierarchical Storage Management (HSM).

Ibis: Real-World Problem Solving using Real-World Grids

H.E. Bal, N. Drost, R. Kemp, J. Maassen,
R.V. van Nieuwpoort, C. van Reeuwijk, and F.J. Seinstra
Department of Computer Science, Vrije Universiteit,
De Boelelaan 1081A, 1081 HV Amsterdam, The Netherlands
{bal, ndrost, rkemp, jason, rob, reeuwijk, fjseins}@cs.vu.nl

Abstract

Ibis is an open source software framework that drastically simplifies the process of programming and deploying large-scale parallel and distributed grid applications. Ibis supports a range of programming models that yield efficient implementations, even on distributed sets of heterogeneous resources. Also, Ibis is specifically designed to run in hostile grid environments that are inherently dynamic and faulty, and that suffer from connectivity problems.

Recently, Ibis has been put to the test in two competitions organized by the IEEE Technical Committee on Scalable Computing, as part of the CCGrid 2008 and Cluster/Grid 2008 international conferences. Each of the competitions' categories focused either on the aspect of scalability, efficiency, or fault-tolerance. Our Ibis-based applications have won the first prize in all of these categories. In this paper we give an overview of Ibis, and — to exemplify its power and flexibility — we discuss our contributions to the competitions, and present an overview of our lessons learned.

A Semantic-aware Information System for Multi-Domain Applications over Service Grids

Carmela Comito[1], Carlo Mastroianni[2], Domenico Talia[1,2]
[1]DEIS, University of Calabria
[2]ICAR-CNR, Italy
{ccomito, talia}@deis.unical.it, mastroianni@icar.cnr.it

Abstract

Service-oriented Grid frameworks offer resources and facilities to support the design and execution of distributed applications in different domains, ranging from scientific applications and public computing projects to commercial and industrial applications. A critical issue in such a context is the management of the heterogeneity of resources and services offered by a Grid, including computers, data, and software tools provided by different organizations. This paper presents a general architecture of a service-oriented information system, which exploits the characteristics of a multi-domain and semantically enriched metadata model. The main objective of the information system is to uniformly manage service-oriented applications and basic resources by assuring metadata persistence through an XML distributed database, without merely relying on the functionalities of persistent Grid services. The information system has been implemented on the basic services of the WSRF-based Globus Toolkit 4 and its performance has been evaluated in a testbed.

Managing the Construction and Use of Functional Performance Models in a Grid Environment

Robert Higgins and Alexey Lastovetsky
School of Computer Science
University College Dublin
Ireland
{robert.higgins, alexey.lastovetsky}@ucd.ie

Abstract

This paper presents a tool, the Performance Model Manager, which addresses the complexity of the construction and management of a set of Functional Performance Models on a computing server in a Grid environment. The operation of the tool and the features it implements to achieve this goal are described. Integration of Functional Performance Models with a GridRPC middleware, using the tool's interfaces is illustrated. Finally, an example application is used to demonstrate the construction of the models and experiments that show the benefit of using the detailed models are presented.

Modelling Memory Requirements for Grid Applications

Tanvire Elahi, Cameron Kiddle and Rob Simmonds
Grid Research Centre, University of Calgary, Canada
{telahi,kiddlec,simmonds}@cpsc.ucalgary.ca

Abstract

Automating the execution of applications in grid computing environments is a complicated task due to the heterogeneity of computing resources, resource usage policies, and application requirements. Applications differ in memory usage, performance, scalability and storage usage. Having knowledge of this information can aid in matching jobs to resources and in selecting appropriate configuration parameters such as the number of processors to run on and memory requirements for those resources.

This paper presents an application memory usage model that can be used to aid in selecting appropriate job configurations for different resources. The model can be used to represent how memory scales with the number of processors, the memory usage of different types of processes, and changes in memory usage during execution. It builds on a previously developed information model used for describing resources, resource usage policies and limited information on applications. An analysis of the memory usage model illustrating its use towards automating job execution in grid computing environments is also presented.

Improving GridWay with Network Information: Tuning the Monitoring Tool

Luis Tomás, Agustín Caminero, Blanca Caminero and Carmen Carrión
Department of Computing Systems
University of Castilla La Mancha. Campus Universitario s/n, 02071
Albacete, Spain
{luistb, agustin, blanca, carmen}@dsi.uclm.es

Abstract

The aggregation of heterogeneous and geographically distributed resources for new science and engineering applications has been made possible thanks to the deployment of Grid technologies. These systems have communication network requirements which should be taken into account when performing usual tasks such as scheduling, migrating or monitoring of jobs. Recall the network should be used in an efficient and fault-tolerant way since the users, services, and data use it to communicate with each other. However, most of the existing efforts to provide QoS in Grids do not take network issues into account, and focus instead on processor workload and disk access. Authors have previously proposed a proof-of-concept implementation of a network-aware metascheduler, developed as an extension to GridWay. This extension harnesses network information to perform scheduling tasks. In this work, a tuning of the network monitoring tool is presented and evaluated. Results show that the overhead caused by this monitoring tool varies depending on the monitoring parameters, and these ones, in turn, depend on the actual Grid environment.

Using a Market Economy to Provision Compute Resources Across Planet-wide Clusters

Murray Stokely, Jim Winget,
Ed Keyes and Carrie Grimes
Google, Inc.
Mountain View, CA
{mstokely,winget,edkeyes,cgrimes}@google.com
Benjamin Yolken
Department of Management Science and Engineering
Stanford University
Stanford, CA
yolken@stanford.edu

Abstract

We present a practical, market-based solution to the resource provisioning problem in a set of heterogeneous resource clusters. We focus on provisioning rather than immediate scheduling decisions to allow users to change long-term job specifications based on market feedback. Users enter bids to purchase quotas, or bundles of resources for long-term use. These requests are mapped into a simulated clock auction which determines uniform, fair resource prices that balance supply and demand. The reserve prices for resources sold by the operator in this auction are set based on current utilization, thus guiding the users as they set their bids towards under-utilized resources. By running these auctions at regular time intervals, prices fluctuate like those in a real-world economy and provide motivation for users to engineer systems that can best take advantage of available resources.

These ideas were implemented in an experimental resource market at Google. Our preliminary results demonstrate an efficient transition of users from more congested resource pools to less congested resources. The disparate engineering costs for users to reconfigure their jobs to run on less expensive resource pools was evidenced by the large price premiums some users were willing to pay for more expensive resources. The final resource allocations illustrated how this framework can lead to significant, beneficial changes in user behavior, reducing the excessive shortages and surpluses of more traditional allocation methods.

Evaluation of Replication and Fault Detection in P2P-MPI

Stéphane Genaud[1] and Choopan Rattanapoka[2]

[1]AlGorille Team - LORIA
Campus Scientifique - BP 239,
F-54506 Vandoeuvre-lès-Nancy, France
stephane.genaud@loria.fr

[2]College of Industrial Technology,
King Mongkut's University of Technology North Bangkok,
Bangkok, Thailand,
choopanr@kmutnb.ac.th

Abstract

We present in this paper an evaluation of fault management in the grid middleware P2P-MPI. One of P2P-MPI's objective is to support environments using commodity hardware. Hence, running programs is failure prone and a particular attention must be paid to fault management. The fault management covers two issues: fault-tolerance and fault detection. P2P-MPI provides a transparent fault tolerance facility based on replication of computations. Fault detection concerns the monitoring of the program execution by the system. The monitoring is done through a distributed set of modules called failure detectors. In this paper, we report results from several experiments which show the overhead of replication, and the cost of fault detection.

Grid-Enabled Hydropad: a Scientific Application for Benchmarking GridRPC-Based Programming Systems

Michele Guidolin and Alexey Lastovetsky
School of Computer Science and Informatics
University College Dublin
Belfield, Dublin 4, Ireland
{michele.guidolin, alexey.lastovetsky}@ucd.ie

Abstract

GridRPC is a standard API that allows an application to easily interface with a Grid environment. It implements a remote procedure call with a single task map and client-server communication model. In addition to nonperformance-related benefits, scientific applications having large computation and small communication tasks can also obtain important performance gains by being implemented in GridPRC. However, such convenient applications are not representative of the majority of scientific applications and therefore cannot serve as fair benchmarks for comparison of the performance of different GridRPC-based systems. In this paper, we present Hydropad, a real life astrophysical simulation, which is composed of tasks that have a balanced ratio between computation and communication. While Hydropad is not the ideal application for performance benefits from its implementation with GridRPC middleware, we show how even its performance can be improved by using GridSolve and SmartGridSolve. We believe that the Grid-enabled Hydropad is a good candidate application to benchmark GridRPC-based programming systems in order to justify their use for high performance scientific computing.

Assessing the Impact of Future Reconfigurable Optical Networks on Application Performance

Jason Maassen, Kees Verstoep and Henri E. Bal
VU University Amsterdam, The Netherlands, {jason,versto,bal}@cs.vu.nl
Paola Grosso and Cees de Laat
University of Amsterdam, The Netherlands, {p.grosso,delaat}@uva.nl

Abstract

The introduction of optical private networks (lightpaths) has significantly improved the capacity of long distance network links, making it feasible to run large parallel applications in a distributed fashion on multiple sites of a computational grid. Besides offering bandwidths of 10 Gbit/s or more, lightpaths also allow network connections to be dynamically reconfigured. This paper describes our experiences with running data-intensive applications on a grid that offers a (manually) reconfigurable optical wide-area network. We show that the flexibility offered by such a network is useful for applications and that it is often possible to estimate the necessary network configuration in advance.

Workshop 11

Workshop on System Management Techniques, Processes, and Services

SMTPS 2009

Performability Evaluation of EFT Systems for SLA Assurance

Erica Sousa, Paulo Maciel, Carlos Araújo and Fábio Chicout
Federal University of Pernambuco
Center of Informatics - Performance Evaluation Laboratory Itautec
Recife, PE, Brazil
etgs,prmm,cjma,fcfmc@cin.ufpe.br

Abstract

The performance evaluation of Electronic Funds Transfer (EFT) Systems has an enormous importance for Electronic Transactions providers, since the computing resources must be efficiently used in order to attain requirements defined in Service Level Agreements (SLA). Among such requirements, we may stress agreements on availability, reliability, scalability and security. In the EFT system, faults can cause severe degradation on system performance. The modeling of these systems, ignoring the dependability effects on performance can lead to incomplete or inaccurate results. This paper presents an expolinomial stochastic model for evaluating the performance of the EFT system processing and storage infrastructure considering load variation range. This work also presents a model for evaluating the dependability of the EFT system infrastructures and combines both models (dependability and performance) for evaluating the impact of dependability issues on the system performance. The performability analysis employs a hierarchical decomposition method aiming at reducing the computational effort and avoiding stiffness problems. Finally, a case study is presented and the respective results are depicted, stressing important aspects of dependability and performance for EFT system planning.

A Global Scheduling Framework for Virtualization Environments

Yoav Etsion[1,2], Tal Ben-Nun[1] and Dror G. Feitelson[1]

[1]School of Computer Science and Engineering
The Hebrew University of Jerusalem
91904 Jerusalem, Israel
{talbn,feit}@cs.huji.ac.il

[2]Barcelona Supercomputing Center (BSC)
08034 Barcelona, Spain
yoav.etsion@bsc.es

Abstract

A premier goal of resource allocators in virtualization environments is to control the relative resource consumption of the different virtual machines, and moreover, to be able to change the relative allocations at will. However, it is not clear what it means to provide a certain fraction of the machine when multiple resources are involved. We suggest that a promising interpretation is to identify the system bottleneck at each instant, and to enforce the desired allocation on that device. This in turn induces an efficient allocation of the other devices.

Symmetric Mapping: an Architectural Pattern for Resource Supply in Grids and Clouds

Xavier Gréhant
Institut Telecom, Telecom ParisTech, CNRS, LTCI, and CERN openlab
Isabelle Demeure
Institut Telecom, Telecom ParisTech, CNRS, LTCI

Abstract

This paper presents the *Symmetric Mapping* pattern, an architectural pattern for the design of resource supply systems. The focus of Symmetric Mapping is on separation of concerns for cost-effective resource allocation. It divides resource supply in three functions: (1) Users and providers match and engage in resource supply agreements, (2) users place tasks on subscribed resource containers, and (3) providers place supplied resource containers on physical resources. The pattern relies on stakeholders to act for their own interest. The efficiency of the whole system is determined by the degree of freedom left to the three functions and the efficiency of the associated decision systems. We propose a formalism of the Symmetric Mapping pattern, we observe to what extend existing grid and cloud systems follow it, and we propose elements of an original implementation.

Application Level I/O Caching on Blue Gene/P Systems

Seetharami Seelam, I-Hsin Chung, John Bauer, Hao Yu and Hui-Fang Wen
IBM Thomas J. Watson Research Center
Yorktown Heights, New York 10598 USA
{sseelam,ihchung,bauerj,yuh,hfwen}@us.ibm.com

Abstract

In this paper, we present an application level aggressive I/O caching and prefetching system to hide I/O access latency experienced by out-of-core applications. Without the application level prefetching and caching capability, users of I/O intensive applications need to rewrite them with asynchronous I/O calls or restructure their code with MPI-IO calls to efficiently use the large scale system resources. Our proposed solution of user controllable aggressive caching and prefetching system maintains a file-IO cache in the user space of the application, analyzes the I/O access patterns, prefetches requests, and performs write-back of dirty data to storage asynchronously. So each time the application needs the data it does not have to pay the full I/O latency penalty in going to the storage and getting the required data.

We have implemented this aggressive caching and asynchronous prefetching on the Blue Gene/P (BGP) system. The preliminary experiment evaluates the caching performance using the WRF benchmark. The results on BGP system demonstrate that our method improves application I/O throughput.

Low Power Mode in Cloud Storage Systems

Danny Harnik, Dalit Naor and Itai Segall
IBM Haifa Research Labs
Haifa, Israel
{dannyh, dalit, itais}@il.ibm.com

Abstract

We consider large scale, distributed storage systems with a redundancy mechanism; cloud storage being a prime example. We investigate how such systems can reduce their power consumption during low-utilization time intervals by operating in a low-power mode. In a low power mode, a subset of the disks or nodes are powered down, yet we ask that each data item remains accessible in the system; this is called *full coverage*. The objective is to incorporate this option into an existing system rather than redesign the system. When doing so, it is crucial that the low power option should not affect the performance or other important characteristics of the system during full-power (normal) operation. This work is a comprehensive study of what can or cannot be achieved with respect to full coverage low power modes.

The paper addresses this question for generic distributed storage systems (where the key component under investigation is the *placement function* of the system) as well as for specific popular system designs in the realm of storing data in the cloud. Our observations and techniques are instrumental for a wide spectrum of systems, ranging from distributed storage systems for the enterprise to cloud data services. In the cloud environment where low cost is imperative, the effects of such savings are magnified by the large scale.

Predicting Cache Needs and Cache Sensitivity for Applications in Cloud Computing on CMP Servers with Configurable Caches

Jacob Machina
School of Computer Science
University of Windsor
Windsor, Canada
machina@uwindsor.ca

Angela Sodan
School of Computer Science
University of Windsor
Windsor, Canada
acsodan@uwindsor.ca

Abstract

QoS criteria in cloud computing require guarantees about application runtimes, even if CMP servers are shared among multiple parallel or serial applications. Performance of computation-intensive application depends significantly on memory performance and especially cache performance. Recent trends are toward configurable caches that can dynamically partition the cache among cores. Then, proper cache partitioning should consider the applications' different cache needs and their sensitivity towards insufficient cache space. We present a simple, yet effective and therefore practically feasible black-box model that describes application performance in dependence on allocated cache size and only needs three descriptive parameters. Learning these parameters can therefore be done with very few sample points. We demonstrate with the SPEC benchmarks that the model adequately describes application behavior and that curve fitting can accomplish very high accuracy, with mean relative error of 2.8% and maximum relative error of 17%.

Resource Monitoring and Management with OVIS to Enable HPC in Cloud Computing Environments

Jim Brandt*, Ann Gentile°, Jackson Mayo*, Philippe Pébay*,
Diana Roe°, David Thompson* and Matthew Wong°
Sandia National Laboratories
MS *9159 / °9152
P.O. Box 969, Livermore, CA 94551 U.S.A.
{brandt,gentile,jmayo,pppebay,dcroe,dcthomp,mhwong}@sandia.gov

Abstract

Using the cloud computing paradigm, a host of companies promise to make huge compute resources available to users on a pay-as-you-go basis. These resources can be configured on the fly to provide the hardware and operating system of choice to the customer on a large scale. While the current target market for these resources in the commercial space is web development/hosting, this model has the lure of savings of ownership, operation, and maintenance costs, and thus sounds like an attractive solution for people who currently invest millions to hundreds of millions of dollars annually on High Performance Computing (HPC) platforms in order to support large-scale scientific simulation codes. Given the current interconnect bandwidth and topologies utilized in these commercial offerings, however, the only current viable market in HPC would be small-memory-footprint embarrassingly parallel or loosely coupled applications, which inherently require little to no inter-processor communication. While providing the appropriate resources (bandwidth, latency, memory, etc.) for the HPC community would increase the potential to enable HPC in cloud environments, this would not address the need for scalability and reliability, crucial to HPC applications. Providing for these needs is particularly difficult in commercial cloud offerings where the number of virtual resources can far outstrip the number of physical resources, the resources are shared among many users, and the resources may be heterogeneous. Advanced resource monitoring, analysis, and configuration tools can help address these issues, since they bring the ability to dynamically provide and respond to information about the platform and application state and would enable more appropriate, efficient, and flexible use of the resources key to enabling HPC. Additionally such tools could be of benefit to non-HPC cloud providers, users, and applications by providing more efficient resource utilization in general.

Distributed Management of Virtual Cluster Infrastructures

Michael A. Murphy, Michael Fenn, Linton Abraham,
Joshua A. Canter, Benjamin T. Sterrett and Sebastien Goasguen
School of Computing, Clemson University
Clemson, South Carolina 29634-0974 USA
{mamurph, mfenn, labraha, jcanter, bsterre, sebgoa}@clemson.edu

Abstract

Cloud services that provide virtualized computational clusters present a dichotomy of systems management challenges, as the virtual clusters may be owned and administered by one entity, while the underlying physical fabric may belong to a different entity. On the physical fabric, scalable tools that “push” configuration changes and software updates to the compute nodes are effective, since the physical system administrators have complete system access. However, virtual clusters executing atop federated Grid sites may not be directly reachable for management purposes, as network or policy limitations may prevent unsolicited connections to virtual compute nodes. For these systems, a distributed middleware solution could permit the compute nodes to “pull” updates from a centralized server, thereby permitting the management of virtual compute nodes that are inaccessible to the system administrator. This paper compares both models of system administration and describes emerging software utilities for managing both the physical fabric and the virtual clusters.

Blue Eyes: Scalable and Reliable System Management for Cloud Computing

Sukhyun Song[1], Kyung Dong Ryu[2] and Dilma Da Silva[2]
[1]Department of Computer Science
University of Maryland, College Park, MD
shsong@cs.umd.edu

[2]IBM T.J. Watson Research Center
Yorktown Heights, NY
{kryu, dilmasilva}@us.ibm.com

Abstract

With the advent of cloud computing, massive and automated system management has become more important for successful and economical operation of computing resources. However, traditional monolithic system management solutions are designed to scale to only hundreds or thousands of systems at most. In this paper, we present Blue Eyes, a new system management solution to handle hundreds of thousands of systems. Blue Eyes enables highly scalable and reliable system management with a multi-server scaleout architecture. In particular, we structure the management servers into a hierarchical tree to achieve scalability, and management information is replicated into secondary servers to provide reliability and high availability. In addition, Blue Eyes is designed to extend the existing single server implementation without significantly restructuring the code base. Several experimental results with the prototype have demonstrated that Blue Eyes can reliably handle typical management tasks for a large scale of endpoints with dynamic load-balancing across the servers, near linear performance gain with server additions, and an acceptable network overhead.

Desktop to Cloud Transformation Planning

Kirk Beaty, Andrzej Kochut, Hidayatullah Shaikh
IBM T.J. Watson Research Center
19 Skyline Drive
Hawthorne, NY, 10532, USA
{kirkbeaty,akochut,hshaikh}@us.ibm.com

Abstract

Traditional desktop delivery model is based on a large number of distributed PCs executing operating system and desktop applications. Managing traditional desktop environments is incredibly challenging and costly. Tasks like installations, configuration changes, security measures require time-consuming procedures and dedicated deskside support. Also these distributed desktops are typically underutilized, resulting in low ROI for these assets. Further, this distributed computing model for desktops also creates a security concern as sensitive information could be compromised with stolen laptops or PCs. Desktop virtualization, which moves computation to the data center, allows users to access their applications and data using stateless “thin-client” devices and therefore alleviates some of the problems of traditional desktop computing. Enterprises can now leverage the flexibility and cost-benefits of running users’ desktops on virtual machines hosted at the data center to enhance business agility and reduce business risks, while lowering TCO. Recent research and development of cloud computing paradigm opens new possibilities of mass hosting of desktops and providing them as a service.

However, transformation of legacy systems to desktop clouds as well as proper capacity provisioning is a challenging problem. Desktop cloud needs to be appropriately designed and provisioned to offer low response time and good working experience to desktop users while optimizing back-end resource usage and therefore minimizing provider’s costs. This paper presents tools and approaches we have developed to facilitate fast and accurate planning for desktop clouds. We present desktop workload profiling and benchmarking tools as well as desktop to cloud transformation process enabling fast and accurate transition of legacy systems to new cloud-based model.

Workshop 12

Workshop on Parallel and Distributed Scientific and Engineering Computing

PDSEC 2009

Optimization Techniques for Concurrent STM-Based Implementations: A Concurrent Binary Heap as a Case Study

Kristijan Dragičević and Daniel Bauer
IBM Zurich Research Laboratory
Zurich, Switzerland
{kdr, dnb}@zurich.ibm.com

Abstract

Much research has been done in the area of software transactional memory (STM) as a new programming paradigm to help ease the implementation of parallel applications. While most research has been invested for answering the question of how STM should be implemented, there is less work about how to use STM efficiently. This paper is focused on the challenge of how to use STM for efficient and scalable implementations of non-trivial applications. We present a fine-grained STM-based concurrent binary heap, an application of STM for a data structure that is notoriously difficult to parallelize. We describe extensions to the basic STM approach and also the benefits of our proposal. Our results show that the fine-grained STM-based binary heap provides very good scalability compared to the naive approach. Nevertheless, we reach a point where the complexity of some fine- grained techniques do not justify its use for the increase in performance that can be obtained.

Optimizing the execution of a parallel meteorology simulation code

Sonia Jerez and Juan Pedro Montávez
University of Murcia, 30071 Murcia, Spain.
Departamento de Física,
sonia.jerez@gmail.com, montavez@um.es,
http://chubasco.inf.um.es.

Domingo Giménez
University of Murcia, 30071 Murcia, Spain.
Departamento de Informática y Sistemas,
domingo@um.es,
http://dis.um.es/˜domingo.

Abstract

Climate simulations are very computational time consuming tasks which are usually solved in parallel systems. However, to reduce the time needed for the simulations, a set of parameters must be optimally selected. This paper presents a methodology to select such parameters for a particular simulation code (the MM5 mesoescalar model). When the code is installed in a computational system its behaviour when executing the code is characterized by a set of parameters. The values obtained are included in a model of the execution time of the code, and the simulation is carried out at running time with the running configuration with which the lowest theoretical time is obtained. An important reduction in the execution time is achieved. In the experiments the reduction is between 25% and 40%. The methodology proposed could be applied to other problems in which the code to be optimized is considered as a black box.

NUMA-ICTM: A Parallel Version of ICTM Exploiting Memory Placement Strategies for NUMA Machines

Márcio Castro, Luiz Gustavo Fernandes
GMAP, PPGCC
Pontifícia Universidade Católica do Rio Grande do Sul
Porto Alegre - Brazil
{mcastro, gustavo}@inf.pucrs.br

Christiane Pousa, Jean-François Méhaut
Laboratoire d'Informatique Grenoble
Grenoble Université
Grenoble - France
{christiane.pousa, mehaut}@imag.fr

Marilton Sanchotene de Aguiar
GMFC, PPGInf
Universidade Católica de Pelotas
Pelotas - Brazil
marilton@atlas.ucpel.tche.br

Abstract

In geophysics, the appropriate subdivision of a region into segments is extremely important. ICTM (Interval Categorizer Tesselation Model) is an application that categorizes geographic regions using information extracted from satellite images. The categorization of large regions is a computational intensive problem, what justifies the proposal and development of parallel solutions in order to improve its applicability. Recent advances in multiprocessor architectures lead to the emergence of NUMA (Non-Uniform Memory Access) machines. In this work, we present NUMA-ICTM: a parallel solution of ICTM for NUMA machines. First, we parallelize ICTM using OpenMP. After, we improve the OpenMP solution using the MAI (Memory Affinity Interface) library, which allows a control of memory allocation in NUMA machines. The results show that the optimization of memory allocation leads to significant performance gains over the pure OpenMP parallel solution.

Distributed Randomized Algorithms for Low-Support Data Mining

Alfredo Ferro, Rosalba Giugno, Misael Mongiovì and Alfredo Pulvirenti
Department of Mathematics and Computer Science
University of Catania
Catania, Italy
{ferro,giugno,mongiovi,apulvirenti}@dmi.unict.it

Abstract

Data mining in distributed systems has been facilitated by using high-support association rules. Less attention has been paid to distributed low-support/high-correlation data mining. This has proved useful in several fields such as computational biology, wireless networks, web mining, security and rare events analysis in industrial plants. In this paper we present distributed versions of efficient algorithms for low-support/high-correlation data mining such as Min-Hashing, K-Min-Hashing and Locality-Sensitive-Hashing. Experimental results on real data concerning scalability, speed-up and network traffic are reported.

Towards a framework for automated performance tuning

G. Cong, S. Seelam, I. Chung, H. Wen and D. Klepacki
IBM T.J. Watson Research Center
Yorktown Heights, NY 10598, USA
{gcong, sseelam, ihchung, hfwen,klepacki}@us.ibm.com

Abstract

As part of the DARPA sponsored High Productivity Computing Systems (HPCS) program, IBM is building petaflop supercomputers that will be fast, power-efficient, and easy to program. In addition to high performance, high productivity to the end user is another prominent goal. The challenge is to develop technologies that bridge the *productivity gap* – the gap between the hardware complexity and the software limitations. In addition to language, compiler, and runtime research, powerful and user-friendly performance tools are critical in debugging performance problems and tuning for maximum performance. Traditional tools have either focused on specific performance aspects (e.g., communication problems) or provided limited diagnostic capabilities, and using them alone usually do not pinpoint accurately performance problems. Even fewer tools attempt to provide solutions for problems detected. In our study, we develop an open framework that unifies tools, compiler analysis, and expert knowledge to automatically analyze and tune the performance of an application. Preliminary results demonstrated the efficiency of our approach

Parallel Numerical Asynchronous Iterative Algorithms: large scale experimentations

Jean-Claude Charr, Raphaël Couturier and David Laiymani
Laboratory of computer sciences, University of Franche-Comté (LIFC)
IUT de Belfort-Montbéliard, Rue Engel Gros, BP 527, 90016 Belfort, France
Tel.: +33-3-84587785 Fax: +33-3-84587781
{jean-claude.charr,raphael.couturier,david.laiymani}@univ-fcomte.fr

Abstract

This paper presents many typical problems that are encountered when executing large scale scientific applications over distributed architectures. The causes and effects of these problems are explained and a solution for some classes of scientific applications is also proposed. This solution is the combination of the asynchronous iteration model with JACEP2P-V2 which is a fully decentralized and fault tolerant platform dedicated to executing parallel asynchronous applications over volatile distributed architectures. We explain in detail how our approach deals with each of these problems. Then we present two large scale numerical experiments that prove the efficiency and the robustness of our approach.

Exploring the Effect of Block Shapes on the Performance of Sparse Kernels

Vasileios Karakasis, Georgios Goumas and Nectarios Koziris
Computing Systems Laboratory
National Technical University of Athens
{bkk,goumas,nkoziris}@cslab.ece.ntua.gr

Abstract

In this paper we explore the impact of the block shape on blocked and vectorized versions of the Sparse Matrix-Vector Multiplication (SpMV) kernel and build upon previous work by performing an extensive experimental evaluation of the most widespread blocking storage format, namely Block Compressed Sparse Row (BCSR) format, on a set of modern commodity microarchitectures. We evaluate the merit of vectorization on the memory-bound blocked SpMV kernel and report the results for single- and multithreaded (both SMP and NUMA) configurations. The performance of blocked SpMV can significantly vary with the block shape, despite similar memory bandwidth demands for different blocks. This is further accentuated when vectorizing the kernel. When moving to multiple cores, the memory wall problem becomes even more evident and may overwhelm any benefit from optimizations targeting the computational part of the kernel. In this paper we explore and discuss the architectural characteristics of modern commodity architectures that are responsible for these performance variations between block shapes.

Coupled Thermo-Hydro-Mechanical Modelling: A New Parallel Approach

Vardon, P.J.
Geoenvironmental Research Centre, Cardiff University, UK
VardonPJ@Cardiff.ac.uk

Banicescu, I.
Department of Computer Science and Centre for Computational Sciences, Mississippi State University, USA
ioana@CSE.MsState.Edu

Cleall, P.J.
Geoenvironmental Research Centre, Cardiff University, UK
Cleall@Cardiff.ac.uk

Thomas, H.R.
Geoenvironmental Research Centre, Cardiff University, UK
ThomasHR@Cardiff.ac.uk

Philp, R.N.
Computer Science, Cardiff University, UK
Roger.Philp@astro.cf.ac.uk

Abstract

A hybrid MPI/OpenMP method of parallelising a bi-conjugate gradient iterative solver for coupled thermo-hydro-mechanical finite-element simulations in unsaturated soil is implemented and found to be efficient on modern parallel computers. In particular, a new method of parallelisation using a hybrid multi-threaded and message-passing approach depending on calculation size was implemented yielding better performance over more processing units. This was tested on both an Opteron 2218 2.6GHz Dual-Core processor based system with a Gigabit Ethernet interconnect and an Intel Xeon (Harpertown / Seaburg) 3.0GHz Quad-Core processor based system with an InfiniBand Connect-X interconnect. The impact of the experimental results reflect on the scalability of field-scale simulations with a higher resolution both spatially and temporally.

Concurrent Scheduling of Parallel Task Graphs on Multi-Clusters Using Constrained Resource Allocations

Tchimou N'Takpé
Nancy University / LORIA
Nancy, France
Tchimou.Ntakpe@loria.fr

Frédéric Suter
IN2P3 Computing Center, CNRS/IN2P3
Lyon-Villeurbanne, France
Frederic.Suter@cc.in2p3.fr

Abstract

Scheduling multiple applications on heterogeneous multi-clusters is challenging as the different applications have to compete for resources. A scheduler thus has to ensure a fair distribution of resources among the applications and prevent harmful selfish behaviors while still trying to minimize their respective completion time. In this paper we consider mixed-parallel applications, represented by graphs whose nodes are data-parallel tasks, that are scheduled in two steps: allocation and mapping. We investigate several strategies to constrain the amount of resources the scheduler can allocate to each application and evaluate them over a wide range of scenarios.

Solving "Large" Dense Matrix Problems on Multi-Core Processors

Mercedes Marqués, Gregorio Quintana-Ortí, Enrique S. Quintana-Ortí
Depto. de Ingeniería y Ciencia de Computadores
Universidad Jaume I
12.071-Castellón, Spain
{mmarques,gquintan,quintana}@icc.uji.es
Robert A. van de Geijn
Department of Computer Sciences
The University of Texas at Austin
Austin, TX 78712
rvdg@cs.utexas.edu

Abstract

Few realize that for large matrices dense matrix computations achieve nearly the same performance when the matrices are stored on disk as when they are stored in a very large main memory. Similarly, few realize that, given the right programming abstractions, coding Out-of-Core (OOC) implementations of dense linear algebra operations (where data resides on disk and has to be explicitly moved in and out of main memory) is no more difficult than programming high-performance implementations for the case where the matrix is in memory. Finally, few realize that on a contemporary eight core architecture one can solve a $100,000 \times 100,000$ dense symmetric positive definite linear system in about an hour. Thus, for problems that used to be considered large, it is not necessary to utilize distributed-memory architectures with massive memories if one is willing to wait longer for the solution to be computed on a fast multithreaded architecture like an SMP or multi-core computer. This paper provides evidence in support of these claims.

Parallel Solvers for Dense Linear Systems for Heterogeneous Computational clusters

Ravi Reddy
School of Computer Science
and Informatics
University College Dublin
manumachu.reddy@ucd.ie

Alexey Lastovetsky
School of Computer Science
and Informatics
University College Dublin
alexey.lastovetsky@ucd.ie

Pedro Alonso
Department of Information
Systems and Computation
Polytechnic University of
Valencia
palonso@dsic.upv.es

Abstract

This paper describes the design and the implementation of parallel routines in the Heterogeneous ScaLAPACK library that solve a dense system of linear equations. This library is written on top of HeteroMPI and ScaLAPACK whose building blocks, the de facto standard kernels for matrix and vector operations (BLAS and its parallel counterpart PBLAS) and message passing communication (BLACS), are optimized for heterogeneous computational clusters.

We show that the efficiency of these parallel routines is due to the most important feature of the library, which is the automation of the difficult optimization tasks of parallel programming on heterogeneous computing clusters. They are the determination of the accurate values of the platform parameters such as the speeds of the processors and the latencies and bandwidths of the communication links connecting different pairs of processors, the optimal values of the algorithmic parameters such as the total number of processes, the 2D process grid arrangement and the efficient mapping of the processes executing the parallel algorithm to the executing nodes of the heterogeneous computing cluster.

We describe this process of automation followed by presentation of experimental results on a local heterogeneous computing cluster demonstrating the efficiency of these solvers.

Concurrent Adaptive Computing in Heterogeneous Environments (CACHE)

John U Duselis, Isaac D. Scherson
The Donald Bren School of Information and Computer Science
University of California, Irvine
Irvine, CA 92697
{jduselis, isaac}@uci.edu

Abstract

We introduce a computational framework for concurrent adaptive computing in heterogeneous environments for computationally intensive applications. This framework considers the presence of inter-connected computational resources which are discoverable and a workload which needs to be executed either by concurrent means or on a singular resource. The selection of resources, using a novel measurement of performance, leads to the adaptive inclusion/exclusion of resources to be used in the efficient execution of workload computations. The adaptive approach is that it makes a determination to include a proper subset of resources for system inclusion to execute a workload, which is contrary (except when the subset is all-inclusive) to a greedy approach where all the resources are seized for the workload application. The selection of a subset of resources may be more efficient due to the high level of heterogeneity of the resources, where, for some resources, certain resource selections may be detrimental or have no value to send work there. Furthermore, this framework aims to lessen the unpredictability and uncontrollability of heterogeneous systems by using this analysis for resource selection.

Toward Adjoinable MPI

Jean Utke[1 2], Laurent Hascoët[3], Patrick Heimbach[4], Chris Hill[4], Paul Hovland[2] and Uwe Naumann[5]
[1]University of Chicago, Chicago, IL, USA
[2]Argonne National Laboratory, Argonne, IL, USA, [utke|hovland]@mcs.anl.gov
[3]INRIA Sophia-Antipolis, Valbonne, France, laurent.hascoet@sophia.inria.fr
[4]EAPS, MIT, Cambridge, MA, USA, [heimbach|cnh]@mit.edu
[5]Department of Computer Science, RWTH Aachen University, Aachen, Germany,
naumann@stce.rwth-aachen.de

Abstract

Automatic differentiation is the primary means of obtaining analytic derivatives from a numerical model given as a computer program. Therefore, it is an essential productivity tool in numerous computational science and engineering domains. Computing gradients with the adjoint (also called reverse) mode via source transformation is a particularly beneficial but also challenging use of automatic differentiation. To date only ad hoc solutions for adjoint differentiation of MPI programs have been available, forcing automatic differentiation tool users to reason about parallel communication dataflow and dependencies and manually develop adjoint communication code. Using the communication graph as a model we characterize the principal problems of adjoining the most frequently used communication idioms. We propose solutions to cover these idioms and consider the consequences for the MPI implementation, the MPI user and MPI-aware program analysis. The MIT general circulation model serves as a use case to illustrate the viability of our approach.

Parallelization and Optimization of a CBVIR System on Multi-Core Architectures

Qiankun Miao† ‡, Yurong Chen‡, Jianguo Li‡, Qi Zhang†, Yimin Zhang‡ and Guoliang Chen†
† Department of Computer Science, University of Science and Technology of China,
Hefei, 230027, Anhui, P.R. China.
‡ Intel China Research Center, Intel Corporation, Beijing, 100190, P.R. China.
yurong.chen@intel.com

Abstract

Technique advances have made image capture and storage very convenient, which results in an explosion of the amount of visual information. It becomes difficult to find useful information from these tremendous data. Content-based Visual Information Retrieval (CBVIR) is emerging as one of the best solutions to this problem. Unfortunately, CBVIR is a very compute-intensive task. Nowadays, with the boom of multi-core processors, CBVIR can be accelerated by exploiting multi-core processing capability. In this paper, we propose a parallelization implementation of a CBVIR system facing to server application and use some serial and parallel optimization techniques to improve its performance on an 8-core and on a 16-core systems. Experimental results show that optimized implementation can achieve very fast retrieval on the two multi-core systems. We also compare the performance of the application on the two multi-core systems and give an explanation of the performance difference between the two systems. Furthermore, we conduct detailed scalability and memory performance analysis to identify possible bottlenecks in the application. Based on these experimental results and performance analysis, we gain many insights into developing efficient applications on future multi-core architectures.

EHGrid: an emulator of heterogeneous computational grids

Basile Clout and Eric Aubanel
Faculty of Computer Science
University of New Brunswick
Fredericton, NB
Canada, E3B5A3
{basile.clout, aubanel}@unb.ca

Abstract

Heterogeneous distributed computing is found in a variety of fields including scientific computing, Internet and mobile devices. Computational grids focusing primarily on computationally-intensive operations have emerged as a new infrastructure for high performance computing. Specific algorithms such as scheduling, load balancing and data redistribution have been devised to overcome the limitations of these systems and take full advantage of their processing power. However, experimental validation and fine-tuning of such algorithms require multiple heterogeneous platforms and configurations. We present EHGrid, a computational grid emulator based on the heterogeneous emulator Wrekavoc. EHGrid reshapes the virtual topology of a homogeneous cluster, degrades the performance of the processors and modifies the characteristics of the network links in an accurate, independent and reproducible way. We demonstrate its utility using two parallel matrix-vector programs and selected NAS parallel benchmarks on a series of four emulated grids.

Optimizing Assignment of Threads to SPEs on the Cell BE Processor

C.D. Sudheer, T. Nagaraju and P.K. Baruah
Dept. of Mathematics and Computer Science
Sri Sathya Sai University
Prashanthi Nilayam, India
sudheer@sssu.edu.in

Ashok Srinivasan
Dept. of Computer Science
Florida State University
Tallahassee, USA
asriniva@cs.fsu.edu

Abstract

The Cell is a heterogeneous multicore processor that has attracted much attention in the HPC community. The bulk of the computational workload on the Cell processor is carried by eight co-processors called SPEs. The SPEs are connected to each other and to main memory by a high speed bus called the Element Interconnect Bus (EIB), which is capable of 204.8 GB/s. However, access to the main memory is limited by the performance of the Memory Interface Controller (MIC) to 25.6 GB/s. It is, therefore, advantageous for the algorithms to be structured such that SPEs communicate directly between themselves over the EIB, and make less use of memory. We show that the actual bandwidth obtained for inter-SPE communication is strongly influenced by the assignment of threads to SPEs (thread-SPE affinity) in many realistic communication patterns. We identify the bottlenecks to optimal performance and use this information to determine good affinities for common communication patterns. Our solutions improve performance by up to a factor of two over the default assignment. We also discuss the optimization of affinity on a Cell blade consisting of two Cell processors, and provide a software tool to help with this. Our results will help Cell application developers choose good affinities for their applications.

Guiding Performance Tuning for Grid Schedules

Jörg Keller, Wolfram Schiffmann
FernUniversität in Hagen
Dept. of Mathematics and Computer Science
58084 Hagen, Germany
{Joerg.Keller,wolfram.schiffmann}@fernuni-hagen.de

Abstract

Grid jobs often consist of a large number of tasks. If the performance of a statically scheduled grid job is unsatisfactory, one must decide which code of which task should be improved. We propose a novel method to guide grid users as to which tasks of their grid job they should accelerate in order to reduce the makespan of the complete job. The input we need is the task schedule of the grid job, which can be derived from traces of a previous run of the job. We provide several algorithms depending on whether only one or several tasks can be improved, or whether task improvement is achieved by improvement of one processor.

Design and Analysis of An Active Predictive Algorithm in Wireless Multicast Networks

Naixue Xiong[1]
[1]Dept of Comp. Scie.
Georgia State Univ., US
{nxiong, pan}@cs.gsu.edu

Laurence T. Yang[2], Yi Pan[1]
[2]Dept of Comp. Scie.
St. Francis Xavier Univ., CA
lyang@stfx.ca

Athanasios V. Vasilakos[3], Jing He[1]
[3]Dept of Comp. and Telec. Engi.
Univ. of Western Macedonia, GR
vasilako@ath.forthnet.gr

Abstract

With the ever-increasing wireless multicast data applications recently, considerable efforts have focused on the large scale heterogeneous wireless multicast, especially those with large propagation delays, which means the feedbacks arriving at the source node are somewhat outdated and harmful to the control actions. To attack the above problem, this paper describes a novel, autonomous, and predictive wireless multicast flow control scheme, the so-called proportional, integrative plus neural network (PINN) predictive technique, which includes two components: the PI flow controller located at the wireless multicast source has explicit rate algorithm to regulate the transmission rate; and the neural network part located at the middle branch node predicts the available buffer occupancy for those longer delay receivers. The ultimate sending rate of the multicast source is the expected receiving rates computed by PI controller based on the consolidated feedback information, and it can be accommodated by its participating branches. This network-assisted property is different from the existing control schemes in that neural network controller can predict the buffer occupancy caused by those long delay receivers, which probably cause irresponsiveness of a wireless multicast flow. This active scheme makes the control more responsive to the network status, therefore, the rate adaptation can be in a timely manner for the sender to react to network congestion quick. We analyze the theoretical aspects of the proposed algorithm, show how the control mechanism can be used to design a controller to support wireless multi-rate multicast transmission based on feedback of explicit rates.

Workshop 13

Performance Modeling, Evaluation, and Optimisation of Ubiquitous Computing and Networked Systems

PMEO 2009

Performance Evaluation of Gang Scheduling in a Two-Cluster System with Migrations

Zafeirios C. Papazachos and Helen D. Karatza
Department of Informatics
Aristotle University of Thessaloniki
54124 Thessaloniki, Greece
{zpapazac, karatza}@csd.auth.gr

Abstract

Gang scheduling is considered to be a highly effective task scheduling policy for distributed systems. In this paper we present a migration scheme which reduces the fragmentation in the schedule caused by gang scheduled jobs which cannot start. Furthermore, the existence of high priority jobs in the workload is addressed by the proposed strategy. High priority jobs need to be started immediately, which can in turn lead to the interruption of a parallel job's execution. A distributed system consisting of two homogeneous clusters is simulated to evaluate the performance. Our simulation results indicate that the proposed strategy can result in a performance boost.

Performance Evaluation of a Resource Discovery Scheme in a Grid Environment Prone to Resource Failures

Konstantinos I. Karaoglanoglou and Helen D. Karatza
Aristotle University of Thessaloniki
kkaraogl@csd.auth.gr

Abstract

This paper studies the problem of discovering the most suitable resource for a specific request in a Grid system. A Grid can be seen as an environment comprised by routers and resources, where each router is in charge of its local resources. In our previous works we enhanced the routers of the system with matchmaking capabilities in order to determine an appropriate set of resources capable of satisfying a specific request. Moreover, we presented an efficient resource discovery mechanism called Re-routing Tables that directs the requests to the resources capable of satisfying them in a dynamical Grid system, where resources are not statically online. In this paper, we present an expansion of our resource discovery scheme in order to cover the cases of consecutive resource failures, and we emphasize in the performance evaluation of our resource discovery scheme by providing new sets of simulation tests in Grid environments that are prone to resource failures.

A Novel Information Model for Efficient Routing Protocols in Delay Tolerant Networks

Xiao Chen
Dept. of Comp. Sci.
Texas State Univ.
San Marcos, TX 78666
xc10@txstate.edu

Jian Shen
Dept. of Math
Texas State Univ.
San Marcos, TX 78666
js48@txstate.edu

Jie Wu
Dept. of Comp. Sci. and Eng.
Florida Atlantic Univ.
Boca Raton, FL 33431
jie@cse.fau.edu

Abstract

Delay tolerant networks (DTNs) are wireless mobile networks that do not guarantee the existence of a path between a source and a destination at any time. When two nodes move within each other's transmission range during a period of time, they can contact each other. The contact of nodes can be periodical, predictable and nonpredictable. In this paper, we assume the contact of nodes is nonpredictable so that it can reflect the most flexible way of nodes movement. Due to the uncertainty and time-varying nature of DTNs, routing poses special challenges. Some existing schemes use utility functions to steer the routing in the right direction. We find that these schemes do not capture enough information about the network and their information processing is inadequate. We develop an information model that can capture more contact information and use regression functions for data processing. Simulation results show that our routing algorithms based on our information model can increase the delivery ratio of the messages and reduce the delivery latency of routing compared with existing ones.

Accurate Analytical Performance Model of Communications in MPI Applications

D. R. Martínez, J. C. Cabaleiro, T. F. Pena, F. F. Rivera
Dept. Electronic and Computer Science
University of Santiago de Compostela
Santiago de Compostela, Spain
{diego.rodriguez, jc.cabaleiro, tf.pena, ff.rivera}@usc.es

V. Blanco
Dept. Statistics and Computer Science
La Laguna University
La Laguna, Spain
Vicente.Blanco@ull.es

Abstract

This paper presents a new LogP-based model, called LoOgGP, which allows an accurate characterization of MPI applications based on microbenchmark measurements. This new model is an extension of LogP for long messages in which both overhead and gap parameters perform a linear dependency with message size. The LoOgGP model has been fully integrated into a modelling framework to obtain statistical models of parallel applications, providing the analyst with an easy and automatic tool for LoOgGP parameter set assessment to characterize communications. The use of LoOgGP model to obtain a statistical performance model of an image deconvolution application is illustrated as a case of study.

Prolonging Lifetime via Mobility and Load-balanced Routing in Wireless Sensor Networks

Zuzhi Fan
College of Information Sci. & Tec., Jinan University
Guangzhou, China
zuzhifan@gmail.com

Abstract

One of the main challenges for a sensor network is conserving the available energy at each sensor node and then prolonging the network lifetime. Many energy efficient/conserving routing protocols have been proposed to the issue; however, the "funnelling effect" in multi-hop communications which describes the convergence of data traffic towards the static sinks (Base Stations) remains a major threat to the network lifetime. This is because the sensor nodes located near a base station have to relay data for those nodes that are farther away. In this paper, we introduce a few mobile elements, named aggregators into network and study their mobility strategies. In particular, we propose a Local Aggregator Deployment Protocol for Energy Conservation (LADPEC) and consider the integration of mobility and routing algorithms for lifetime elongation. Based on the simulation results, we show that joint mobility and routing would significantly increase the lifetime of network.

A Performance Model of Multicast Communication in Wormhole-Routed Networks on-Chip

Mahmoud Moadeli and Wim Vanderbauwhede
Department of Computing Science
University of Glasgow
Glasgow, UK
{mahmoudm, wim}@dcs.gla.ac.uk

Abstract

Collective communication operations form a part of overall traffic in most applications running on platforms employing direct interconnection networks. This paper presents a novel analytical model to compute communication latency of multicast as a widely used collective communication operation. The novelty of the model lies in its ability to predict the latency of the multicast communication in wormhole-routed architectures employing asynchronous multi-port routers scheme. The model is applied to the Quarc [17] NoC and its validity is verified by comparing the model predictions against the results obtained from a discrete-event simulator developed using OMNET++.

Reduction of Quality (RoQ) Attacks on Structured Peer-to-Peer Networks

Yanxiang He, Qiang Cao, Yi Han, Libing Wu and Tao Liu
School of Computer, Wuhan University, Wuhan 430079
yxhe@whu.edu.cn, caoqiang.cs@gmail.com

Abstract

In contrast to traditional brute-force attacks, RoQ (Reduction of Quality) attacks are periodic, stealthy, yet potent, which exploit the vulnerability of adaptation mechanisms to undermine certain services. As the application-level peer-to-peer (p2p) protocols depend on a recovery-adjustment process to maintain global consistency of routing information when peers join and leave the systems, we propose a novel breed of RoQ attacks in structured p2p systems: (1) We induce a general model for RoQ attacks, and then derive in structured p2p networks a new attack form that RoQ attackers periodically create concurrent failure through manipulation of massive nodes, degrading the system performance repeatedly. (2) We explore the impacts of RoQ attacks on Chord with detailed analysis and theoretical estimations, and confirm them by simulation results on p2psim, including successful lookup ratio and lookup latency. Moreover, we also discuss the detection and defense against such attacks and the improvements of protocols for attack tolerance.

New Adaptive Counter Based Broadcast Using Neighborhood Information in MANETS

M. Bani Yassein,
Department of Computer Science
Jordan University of Science and Technology
Irbid 22110, Jordan
masadeh@just.edu.jo

A. Al-Dubai
School of Computing
Napier University,
Merchiston Campus, 10 Colinton Road
a.aldubai@napier.ac.uk

M. Ould Khaoua
Department of Electrical &
Computer Engineering
Sultan Qaboos University
Al-Khodh, 123, Muscat, Oman
mok@squ.edu.om

Omar M. Al-jarrah
Computer Engineering Department
Jordan University of
Science and Technology
Irbid 22110, Jordan
aljarrah@just.edu.jo

Abstract

Broadcasting in MANETs is a fundamental data dissemination mechanism, with important applications, e.g., route query process in many routing protocols, address resolution and diffusing information to the whole network. Broadcasting in MANETs has traditionally been based on flooding, which overwhelm the network with large number of rebroadcast packets. Fixed counter-based flooding has been one of the earliest suggested approaches to overcome blind-flooding or the "broadcast storm problem". As the topological characteristics of mobile networks varies instantly, the need of an adapted counter-based broadcast emerge. This research argues that neighbouring information could be used to better estimate the counter-based threshold value at a given node. Additionally results of extensive simulation experiments performed in order to determine the minimum and maximum number of neighbours for a given node is shown. This is done based on locally available information and without requiring any assistance of distance measurements or exact location determination devices.

A Distributed Filesystem Framework for Transparent Accessing Heterogeneous Storage Services

Yutong Lu, Huajian Mao and Jie Shen
National University of Defense and Technology
Changsha, China
{ytlu, huajianmao, jies}@nudt.edu.cn

Abstract

This paper introduces an extensible distributed file system framework, YaFS, using heterogeneous online storage services as its back-ends. It provides a configurable solution for simplifying the usage of multiple storage resources and accessing data ubiquitously and safely. YaFS is POSIX compliant, so that it could support most of the existing applications seamlessly. An offline mode is used to cope with the challenged unreliable network environment. We implement a storage abstraction layer and a plug-in mechanism for uniformly accessing different storage services transparently, and it makes the system expanded easily. YaFS could effectively support for storing large object to size limited services and achieving high aggregate bandwidth by striping data on multiple servers with bandwidth-saving method. The evaluation on a prototype implementation with email services as its storage back-end shows that the performance and usability of the framework is viable.

Dynamic Adaptive Redundancy for Quality-of-Service Control in Wireless Sensor Networks

Ing-Ray Chen and Anh Ngoc Speer
Department of Computer Science
Virginia Tech
{irchen,nphan}@vt.edu

Mohamed Eltoweissy
Department of Electrical and Computer Engineering
Virginia Tech
toweissy@vt.edu

Abstract

In this paper, we develop and evaluate a new concept of adaptive optimal redundancy to efficiently provide wireless sensor network (WSN) users with QoS-aware information services. Our approach to satisfying application QoS requirements while maximizing the system lifetime is to determine dynamically the optimal level of redundancy at the “source” and “path” levels in response to network dynamics and query QoS requirements on a query by query basis. We develop a mathematical model to analyze the performance characteristics of our proposed solution. The obtained results show that dynamically and adaptively identifying and managing optimal redundancy for per-hop data delivery lead to satisfying query QoS requirements in terms of reliability and timeliness, while maximizing the useful lifetime of WSNs. We validate the analytical results with extensive simulation.

The Effect of Heavy-Tailed Distribution on the Performance of Non-Contiguous Allocation Strategies in 2D Mesh Connected Multicomputers

Saad Bani Mohammad
Department of Computer Science,
Prince Hussein Bin Abdullah College for Information Technology,
Al al-Bayt University, Mafraq 25113, Jordan.
bani@aabu.edu.jo

Abstract

The performance of non-contiguous allocation strategies has been evaluated under the assumption that the number of messages sent by jobs, which is one of the factors that the job execution times depend on, follow an exponential distribution. However, many measurement studies have convincingly demonstrated that the execution times of certain computational applications are best characterized by heavy-tailed job execution times. In this paper, the performance of existing non-contiguous allocation strategies is re-visited in the context of heavy-tailed distributions. The strategies are evaluated and compared using simulation experiments for both First-Come-First-Served (FCFS) and Shortest-Service-Demand (SSD) scheduling under a variety of system loads and system sizes. The results show that the performance of the non-contiguous allocation strategies degrades considerably when the number of messages sent follow a heavy-tailed distribution against that of the exponential distribution. Moreover, SSD copes much better than FCFS scheduling in the presence of heavy-tailed job execution times.

Energy Efficient and Seamless Data Collection with Mobile Sinks in Massive Sensor Networks

Taisoo Park, Daeyoung Kim, Seonghun Jang, Seong-eun Yoo and Yohhan Lee
Real Time Embedded Systems Laboratory,
School of Engineering, Information and Communications University
Daejeon, South Korea
taisoo_park@tmax.co.kr, {kimd, exigent, seyoo, yohhanlee}@icu.ac.kr

Abstract

Wireless Sensor Networks (WSNs) enable the surveillance and reconnaissance of a particular area with low cost and less manpower. However, the biggest problem against the commercialization of the WSN is the limited lifetime of the battery-operated sensor node. Taking this problem into account, a mobile sink is deployed as a robot, vehicle or portable device to only activate the sensor nodes that are interesting to the sink and leaves other nodes deactivated for. This can considerably extend the lifetime of the sensor nodes compared to existing power management algorithm of using all nodes. However, in this environment, the mobility of the sink raises new issues of energy efficiency and connectivity in communications. To solve these issues, we propose a DRMOS (Dynamic Routing protocol for Mobile Sink) method that includes a designated wake-up-zone to make sensor nodes prepare for an incoming sink. The shape of the wake-up-zone is dynamically changing to reflect the past moving patterns of the sinks. Moreover, we present the extensive simulation results and recommend parameters for practical use of DRMOS from the simulation analysis.

Priority-based QoS MAC Protocol for Wireless Sensor Networks

Hoon Kim
Dept. of Computer & Radio
Communications Engineering
Korea University
Seoul, Korea
gegol@korea.ac.kr

Sung-Gi Min
Dept. of Computer & Radio
Communications Engineering
Korea University
Seoul, Korea
sgmin@korea.ac.kr

Abstract

The media access control (MAC) protocol in wireless sensor networks provides a periodic listen/sleep state for protection from overhearing and idle listening. However, many scenarios and applications exist in which sensor nodes must send data quickly to destination nodes. This paper proposes the priority-based quality-of-service MAC (PQMAC) protocol for wireless sensor networks. We use data priority levels to differentiate among data transmissions, and propose a MAC protocol based on these levels. This protocol manages scheduling by adaptively controlling network traffic and the priority level. We focused on reducing the latency of the message transmission from the source to the destination. Simulation results showed that PQMAC reduces latency problems in wireless sensor networks while maintaining energy efficiency.

Experimental Evaluation of a WSN Platform Power Consumption

Ch. Antonopoulos[1,2], A. Prayati[2], T. Stoyanova[1,2], C. Koulamas[2], G. Papadopoulos[1,2]
[1]Applied Electronics Laboratory, University of Patras, 26500 Rion-Patras, Greece
[2]Industrial Systems Institute, Stadiu Str., Platani, Patras, Greece
{cantonop, tsstoyanova, papadopoulos}@ee.upatras.gr
{prayati, koulamas}@isi.gr

Abstract

Critical characteristics of wireless sensor networks, as being autonomous and comprising small or miniature devices are achieved at the expense of very strict available energy related limitations. Therefore, it is apparent that optimal resource management is among the most important challenges in WSNs development and success.

However, energy management requires in depth knowledge and detailed insight concerning the factors contributing to the overall power consumption of a WSN mote. To achieve such awareness, appropriate measuring test-beds and methodologies are needed, enabling reliable and accurate power consumption measurements of critical functionalities.

Moving towards that direction, the contribution of this paper is twofold. On one hand, the design and implementation of a system is presented, capable of accurately measuring and displaying a wide range of power consumption thresholds. On the other hand, the elementary functionalities of a WSN platform are identified, isolated and measured with respect to their contribution to the overall mote power consumption. Valuable conclusions are extracted and analyzed.

Throughput-Fairness Tradeoff in Best Effort Flow Control for On-Chip Architectures

Fahimeh Jafari[1,2], Mohammad S. Talebi[2], Mohammad H. Yaghmaee[1], Ahmad Khonsari[3,2]
and Mohamed Ould-Khaoua[4,5]
[1]Ferdowsi University of Mashhad, Mashhad, Iran
[2]School of Computer Science, IPM, Tehran, Iran
[3]ECE Department, University of Tehran, Tehran, Iran
[4]Department of Electrical and Computer Engineering, Sultan Qaboos University, Oman
[5]Department of Computing Science, University of Glasgow
jafari@ipm.ir, mstalebi@ipm.ir, hyaghmae@ferdowsi.um.ac.ir, ak@ipm.ir, mohamed@dcs.gla.ac.uk

Abstract

We consider two flow control schemes for Best Effort traffic in on-chip architectures, which can be deemed as the solutions to the boundary extremes of a class of utility maximization problem. At one extreme, we consider the so-called Rate-Sum flow control scheme, which aims at improving the performance of the underlying system by roughly maximizing throughput while satisfying capacity constraints. At the other extreme, we deem the Max-Min flow control, whose concern is to maintain Max-Min fairness in rate allocation by fairly sacrificing the throughput. We then elaborate our argument through a weighting mechanism in order to achieve a balance between the orthogonal goals of performance and fairness. Moreover, we investigate the implementation facets of the presented flow control schemes in on-chip architectures. Finally, we validate the proposed flow control schemes and the subsequent arguments through extensive simulation experiments.

Analysis of Data Scheduling Algorithms in Supporting Real-time Multi-item Requests in On-demand Broadcast Environments

Jun Chen[1,2], Kai Liu[2] and Victor C.S.Lee[2]
[1]School of Information Management, Wuhan University, Wuhan, Hubei, China
[2]Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong
junchen2@cityu.edu.hk, kevin.liu@student.cityu.edu.hk, csvlee@cityu.edu.hk

Abstract

On-demand broadcast is an effective wireless data dissemination technique to enhance system scalability and capability to handle dynamic data access patterns. Previous studies on time-critical on-demand data broadcast were under the assumption that each client requests only one data item at a time. With rapid growth of time-critical information dissemination services in emerging applications, there is an increasing need for systems to support efficient processing of real-time multi-item requests. Little work, however, has been done. In this work, we study the behavior of six representative single-item request based scheduling algorithms in time-critical multi-item request environments. The results show that the performance of all algorithms deteriorates when dealing with multi-item requests. We observe that data popularity, which is an effective factor to save bandwidth and improve performance in scheduling single-item requests, becomes a hindrance to performance in multi-item request environments. Most multi-item requests scheduled by these algorithms suffer from a starvation problem, which is the root of performance deterioration.

Network Processing Performability Evaluation on Heterogeneous Reliability Multicore Processors using SRN Model

Peter D. Ungsunan, Chuang Lin and Yang Wang
Department of Computer Science and Technology
Tsinghua University, Beijing, China
{hongsunan, clin, ywang}@csnet1.cs.tsinghua.edu.cn
Yi Gai
Ming Hsieh Department of Electrical Engineering
University of Southern California
Los Angeles, CA, USA
ygai@usc.edu

Abstract

Future network systems and embedded infrastructure devices in ubiquitous environments will need to consume low power and process large amounts of network packet traffic. In order to meet necessary high processing efficiency requirements, future processors will have many heterogeneous cores with reduced reliability due to low voltage, small transistor sizes, semiconductor wearout, and environmental factors such as noise and interference. It will be necessary for multi-core network infrastructure software to mitigate transient hardware faults to maintain acceptable system reliability. Applications such as packet processing can benefit from the reliability versus performance tradeoff. We propose a model based on Stochastic Reward Nets to evaluate the performance vs. reliability tradeoff of unreliable embedded multi-core network processors, and apply this model to a multi-core packet processing application.

A Statistical Study on the Impact of Wireless Signals' Behavior on Location Estimation Accuracy in 802.11 Fingerprinting Systems

Reza Farivar, David Wiczer, Alejandro Gutierrez and Roy H. Campbell
University of Illinois at Urbana Champaign
{farivar2, dwiczer, agutier3, rhc}@illinois.edu

Abstract

Much of the recent interest in location estimation systems has focused on 802.11 fingerprinting. Unlike GPS systems, 802.11 based systems can accurately estimate a user's location inside buildings. Moreover, users don't need any special equipment to carry around, as their WiFi enabled cell phone can already act as the receiver in WiFi fingerprinting systems. However, wireless access points in buildings are placed mostly according to another criteria, namely to increase the network coverage inside the building. But optimal coverage may not necessarily result in optimal location discovery. In this paper, we provide analyses on data gathered for a real WiFi location estimation system, and show what makes it perform inaccurately in some parts of a building while it is more accurate in other parts. We have defined two new metrics for quantifying the wireless signal behavior of multiple access points in small neighborhoods in a building. Finally, we identify the properties that differentiate well behaving and poorly behaving neighborhoods.

Performance Prediction for Running Workflows under Role-based Authorization Mechanisms

Ligang He, Mark Calleja[1] Mark Hayes[1] and Stephen A. Jarvis
Department of Computer Science, University of Warwick, Coventry CV4 7AL, UK
[1]Cambridge eScience Centre, Centre for Mathematical Sciences, Cambridge CB3 0WA, UK
liganghe@dcs.warwick.ac.uk

Abstract

When investigating the performance of running scientific/commercial workflows in parallel and distributed systems, we often take into account only the resources allocated to the tasks constituting the workflow, assuming that computational resources will accept the tasks and execute them to completion once the processors are available. In reality, and in particular in Grid or e-business environments, security policies may be implemented in the individual organisations in which the computational resources reside. It is therefore expedient to have methods to calculate the performance of executing workflows under security policies. Authorisation control, which specifies who is allowed to perform which tasks when, is one of the most fundamental security considerations in distributed systems such as Grids. Role-Based Access Control (RBAC), under which the users are assigned to certain roles while the roles are associated with prescribed permissions, remains one of the most popular authorisation control mechanisms. This paper presents a mechanism to theoretically compute the performance of running scientific workflows under RBAC authorisation control. Various performance metrics are calculated, including both system-oriented metrics, (such as system utilisation, throughput and mean response time) and user-oriented metrics (such as mean response time of the workflows submitted by a particular client). With this work, if a client informs an organisation of the workflows they are going to submit, the organisation is able to predict the performance of these workflows running in its local computational resources (e.g. a high-performance cluster) enforced with RBAC authorisation control, and can also report client-oriented performance to each individual user.

Routing, Data Gathering, and Neighbor Discovery in Delay-Tolerant Wireless Sensor Networks

Abbas Nayebi[1], Hamid Sarbazi-Azad[1] and Gunnar Karlsson[2]
[1]Sharif University of Technology, Tehran, Iran
[2]Royal Institute of Technology (KTH), Stockholm, Sweden
nayebi@ripi.ir, azad@sharif.edu, gk@kth.se

Abstract

This paper investigates a class of mobile wireless sensor networks that are not connected most of the times. The characteristics of these networks is inherited from both delay tolerate networks (DTN) and wireless sensor networks. First, delay-tolerant wireless sensor networks (DTWSN) are introduced. Then, three main problems in the design space of these networks are discussed: Routing, data gathering, and neighbor discovery. An approach is proposed for deployment of DTWSNs based on the traditional opportunistic broadcast in delay tolerant networks with on-off periods. The delay and the throughput of the routing scheme were investigated in the DTN literature. However, the energy consumption was not studied thoroughly, which is focused here. Neighbor discovery in a sparse network could be a major source of energy consumption. Therefore, energy per contact measure is evaluated analytically based on the distribution of physical link duration. The results for 2D constant velocity model and random waypoint model are reported and the average PLD is suggested as an appropriate choice of beacon interval.

A Service Discovery Protocol for Vehicular Ad Hoc Networks: A Proof of Correctness

Azzedine Boukerche and Kaouther Abrougui
SITE university of Ottawa
{boukerch, abrougui}@site.uottawa.ca

Abstract

Recently, vehicle networks are gaining great deal of attention from the research community. In order to provide efficient and pervasive road communication, Next Generation Vehicular Networks (NVN) are considered a promising solution. NVNs have unique characteristics and face challenging problems. Consequently, it is hard to use the traditional mechanisms and protocols in this type of network. Service discovery is a very challenging problem for NVN-based applications. Furthermore, to the best of our knowledge, very little work has been done to deal with the service discovery problem in NVNs. Due to the high mobility and density of vehicles, traditional discovery techniques do not perform well. To solve this problem, we propose a novel class of service discovery protocol that would allow vehicles to discover services through the vehicular wireless network. Our hybrid proposed technique combines both proactive and reactive discovery approaches. It is also adaptive because it adapts to the vehicular network conditions, thus enabling efficient discovery characterized by low overhead and a high success rate. In this paper, we present the proof of correctness and the message and time complexities computation of our protocol.

A QoS Aware Multicast Algorithm for Wireless Mesh Networks

Liang Zhao[1], Ahmed Yassin Al-Dubai[1] and Geyong Min[2]
[1]School of Computing, Napier University
Edinburgh, EH10 5DT, UK
{l.zhao, a.al-dubai}@napier.ac.uk
[2]Department of Computing, University of Bradford
Bradford, BD7 1DP, U.K.
g.min@brad.ac.uk

Abstract

Wireless mesh networks have been attracting significant attention due to its promising technology. It is becoming a major avenue for the fourth generation of wireless mobility. Communication in large-scale wireless networks can create bottlenecks for scalable implementations of computationally intensive applications. A class of crucially important communication patterns that have already received considerable attention in this regard are group communication operations, since these inevitably place a high demand on network bandwidth and have a consequent impact on algorithm execution times. Multicast communication has been among the most primitive group capabilities of any message passing networks. It is central to many important distributed applications in Science and Engineering and fundamental to the implementation of higher-level communication operations such as gossip, gather, and barrier synchronisation. Existing solutions offered for providing multicast communications in WMN have severe restriction in terms of almost all performance characteristics. Consequently, there is a need for the design and analysis of new efficient multicast communication schemes for this promising network technology. Hence, the aim of this study is to tackle the challenges posed by the continuously growing need for delivering efficient multicast communication over WMN. In particular, this study presents a new load balancing aware multicast algorithm with the aim of enhancing the QoS in the multicast communication over WMNs.

Design and implemention of a novel MAC layer handoff protocol for IEEE 802.11 wireless networks

Zhenxia Zhang and Azzedine Boukerche
PARADISE Research Laboratory
SITE, University of Ottawa, Canada
{zzhan036, boukerch}@site.uottawa.ca

Abstract

In recent years, IEEE 802.11 wireless networks become one of the most important components in wireless networks, since compared with other wireless technologies, IEEE 802.11 devices are inexpensive and easier to be configured. To provide seamless roaming in the IEEE 802.11 wireless networks, MAC layer handoff latency should be minimized to support real-time applications. This paper proposes a novel MAC layer handoff protocol over IEEE 802.11 wireless networks by using an advertisement message. The experiment results illustrate that our solution can reduce MAC layer handoff latency to less than 50 *ms* required by real-time applications.

Workshop 14

Dependable Parallel, Distributed and Network-Centric Systems

DPDNS 2009

Robust CDN Replica Placement Techniques

Samee Ullah Khan
Department of Electrical and Computer Engineering
North Dakota State University
Fargo, ND 58108
samee.khan@ndsu.edu
Anthony A. Maciejewski[1] and Howard Jay Siegel[1,2]
[1]Department of Electrical and Computer Engineering
[2]Department of Computer Science
Colorado State University
Fort Collins, CO 80523
{aam, hj}@colostate.edu

Abstract

Creating replicas of frequently accessed data objects across a read-intensive Content Delivery Network (CDN) can result in reduced user response time. Because CDNs often operate under volatile conditions, it is of the utmost importance to study replica placement techniques that can cope with uncertainties in the system parameters. We propose four CDN replica placement heuristics that guarantee a robust performance under the uncertainty of arbitrary CDN server failures. By robust performance we mean the solution quality that a heuristic guarantees given the uncertainties in system parameters. The simulation results reveal interesting characteristics of the studied heuristics. We report these characteristics with a detailed discussion on which heuristics to utilize for robust CDN data replication given a specific scenario.

A flexible and robust lookup algorithm for P2P systems

Mauro Andreolini and Riccardo Lancellotti
University of Modena and Reggio Emilia

Abstract

One of the most critical operations performed in a P2P system is the lookup of a resource. The main issues to be addressed by lookup algorithms are: (1) support for flexible search criteria (e.g., wildcard or multi-keyword searches), (2) effectiveness – i.e., ability to identify all the resources that match the search criteria, (3) efficiency – i.e. low overhead, (4) robustness with respect to node failures and churning. Flood-based P2P networks provide flexible lookup facilities and robust performance at the expense of high overhead, while other systems (e.g. DHT) provide a very efficient lookup mechanism, but lacks flexibility.

In this paper, we propose a novel resource lookup algorithm, namely *fuzzy-DHT*, that solves this trade-off by introducing a flexible and robust lookup criteria based on multiple keywords on top of a distributed hash table algorithm. We demonstrate that the fuzzy-DHT algorithm satisfies all the requirements of P2P lookup systems combining the flexibility of flood-based mechanisms while preserving high efficiency, effectiveness ad robustness.

Extending SRT for Parallel Applications in Tiled-CMP Architectures

Daniel Sánchez, Juan L. Aragón and José M. García
Departamento de Ingeniería y Tecnología de Computadores
Universidad de Murcia, Spain
{dsanchez, jlaragon, jmgarcia}@ditec.um.es

Abstract

Reliability has become a first-class consideration issue for architects along with performance and energy-efficiency. The increasing scaling technology and subsequent supply voltage reductions are increasing the susceptibility of architectures to soft errors. However, mechanisms to achieve full coverage to errors usually degrade performance in an unacceptable way for the majority of common users.

Simultaneous and Redundantly Threaded (SRT) is a fault tolerant architecture in which pairs of threads in a SMT core redundantly execute the same program instructions. In this paper, we study the under-explored architectural support of SRT to reliably execute shared-memory applications. We show how atomic operations induce a serialization point between master and slave threads. This bottleneck has an impact of 34% in execution speed for several parallel scientific benchmarks. We propose an alternative mechanism in which the L1 cache is updated by master's stores before verification reducing the overhead up to 21%. Our approach also outperforms other recent proposals such as DCC with a decrease of 8% in execution speed.

Byzantine Fault-Tolerant Implementation of a Multi-Writer Regular Register

Khushboo Kanjani
Oracle Corporation
khushboo.kanjani@gmail.com

Hyunyoung Lee and Jennifer L. Welch
Dept. of Computer Science & Engineering
Texas A&M University
{hlee, welch}@cse.tamu.edu

Abstract

Distributed storage systems have become popular for handling the enormous amounts of data in network-centric systems. A distributed storage system provides client processes with the abstraction of a shared variable that satisfies some consistency and reliability properties. Typically the properties are ensured through a replication-based implementation. This paper presents an algorithm for a replicated read-write register that can tolerate Byzantine failures of some of the replica servers. The targeted consistency condition is a version of regularity that supports multiple writers. Although regularity is weaker than the more frequently supported condition of atomicity, it is still strong enough to be useful in some important applications. By weakening the consistency condition, the algorithm can support multiple writers more efficiently than the known multi-writer algorithms for atomic consistency.

APART+: Boosting APART Performance via Optimistic Pipelining of Output Events

Paolo Romano
INESC-ID, Lisbon, Portugal

Francesco Quaglia and Bruno Ciciani
Sapienza Università di Roma, Italy

Abstract

APART (A Posteriori Active ReplicaTion) is a recently proposed active replication protocol specifically tailored for multi-tier data acquisition systems. It ensures consistency of middle-tier sink replicas by means of an a-posteriori synchronization phase based on reconciliation, which is activated only in case replicas react to an input message from the sensors by generating an output event destined to the back-end tier.

This paper enhances APART via a novel non-blocking synchronization scheme which prevents replicas from stalling while waiting for the outcome of an on-going synchronization phase. Contrarily, replicas are allowed to optimistically process data from the sensors, and to immediately propagate any output event towards the back-end tier. The removal of the blocking synchronization phase from the critical path gives rise to striking performance gains via an effective overlapping of event processing and synchronization. On the other hand, system consistency is ensured by enhancing the back-end tier synchronization logic in order to filter out optimistically produced output events that are incompatible with the reconciled state trajectory.

Message-Efficient Omission-Tolerant Consensus with Limited Synchrony

C. Delporte-Gallet, H. Fauconnier and A. Tielmann
LIAFA
Université Paris Diderot, France
{cd,hf,tielmann}@liafa.jussieu.fr
F. C. Freiling and M. Kilic
Laboratory for Dependable Distributed Systems
University of Mannheim, Germany
freiling@informatik.uni-mannheim.de, kilic_mahir@hotmail.com

Abstract

We study the problem of consensus in the *general omission failure* model, i.e., in systems where processes can crash and omit messages while sending or receiving. This failure model is motivated from a smart card-based security framework in which certain security problems can be reduced to consensus in that model. We propose an algorithm that solves consensus based on very weak timing assumptions. More precisely, we show that consensus is solvable using an *eventual bisource* and a majority of fault-free processes. An eventual bisource is a fault-free process that can eventually communicate with all other processes in a timely manner. In contrast to previous work, we use timing assumptions directly in the algorithm and do not employ the notion of a failure detector. We argue that this is helpful in reducing the message complexity, a critical aspect of algorithms which run on smart cards.

AVR-INJECT: a Tool for Injecting Faults in Wireless Sensor Nodes

Marcello Cinque[1], Domenico Cotroneo[1], Catello Di Martino[1], Stefano Russo[1,2], Alessandro Testa[1]
[1]Dipartimento di Informatica e Sistemistica, Universita' di Napoli Federico II,
Via Claudio 21, 80125 Napoli, Italy - ph: +39 081 7683812, fax: +39 0817683816
[2]Laboratorio ITEM Carlo Savy - Consorzio Interuniversitario Nazionale per l'Informatica
Via Cinthia 1, 80124 Naples, Italy - ph: +39 081 679942, fax: +39 081 676574
{macinque, cotroneo, catello.dimartino, sterusso}@unina.it

Abstract

As the incidence of faults in real Wireless Sensor Networks (WSNs) increases, fault injection is starting to be adopted to verify and validate their design choices. Following this recent trend, this paper presents a tool, named AVR-INJECT, designed to automate the fault injection, and analysis of results, on WSN nodes. The tool emulates the injection of hardware faults, such as bit flips, acting via software at the assembly level. This allows to attain simplicity, while preserving the low level of abstraction needed to inject such faults. The potential of the tool is shown by using it to perform a large number of fault injection experiments, which allow to study the reaction to faults of real WSN software.

Dependable QoS support in Mesh Networks

M. Fazio, M. Paone, D. Bruneo and A. Puliafito
Faculty of Engineering
University of Messina
Contrada di Dio, S. Agata, 98166 Messina, Italy
mfazio{mpaone, dbruneo, apuliafito}@unime.it

Abstract

Wireless networks are a very challenging communication technology since their ability to be set everywhere and whenever. Among the several types of wireless systems, a new class of networks is gradually emerging: Wireless Mesh Networks (WMNs). A WMN is a distributed communication infrastructure organized in a mesh topology, which handles multihops connections and is capable of provide dependable services by dynamically updating and optimizing communications. This paper presents a new cross-layer architecture for supporting QoS in WMNs. It integrates the DiffServ paradigm with the admission control signaling and a multipath routing to route QoS traffic along reserved paths, while making use of alternative paths for Best Effort load. Performance measurements, based on simulative techniques are carried out to test the reliability of the proposed system.

Storage Architecture with Integrity, Redundancy and Encryption

Henning Klein
Fujitsu Siemens Computers GmbH
Buergermeister-Ulrich-Strasse 100
86199 Augsburg, Germany
Henning.Klein@fujitsu-siemens.com

Jörg Keller
FernUniversität in Hagen
Dept. of Mathematics and Computer Science
58084 Hagen, Germany
Joerg.Keller@fernuni-hagen.de

Abstract

We propose a storage system that treats confidentiality, integrity and availability of data in a unified manner. Extending RAID6, it allows for failures of multiple disks, encrypts data on disk, and stores checksums to detect faulty data without disks failing, which occurs e.g. in solid state disks due to wear out of cells. By handling encryption and integrity check together, the probability of undetected faulty data is reduced further. We provide an implementation, i.e. a driver, which encapsulates all these features and uses parallel algorithms exploiting multicore processor performance to match the bandwidth available from multiple disks. We present performance figures of our experiments.

Pre-calculated Equation-based Decoding in Failure-tolerant Distributed Storage

Peter Sobe
University of Luebeck, Germany
Institute of Computer Engineering
sobe@iti.uni-luebeck.de

Abstract

Data distribution together with erasure-tolerant codes allow to store data reliably, even with failed or temporarily disconnected storage resources. The encoding algorithm, i.e. the calculation of the codewords is expressed by XOR equations. Even decoding is the execution of a failure-specific set of equations that are build code-specifically and with knowledge of the failure situation. A new concept for a storage system is to provide encoding equations and decoding equations in advance, as a full description of the code which eliminates the calculations to obtain the recovery strategy. This concept includes that also decoding equations have to be provided in advance, for many different failure situations. This results in a large number of equations and may require a considerable amount of memory, but still a moderate amount - which can be traded for the gained flexibility and simplicity. In this paper we analyze the storage consumption of such a preprocessed decoding equation set. Furthermore, a data structure to access the required equations is proposed. It is shown that codes can be translated into equation sets that are used as parameter set by a storage system.

Workshop 15

International Workshop on Security in Systems and Networks

SSN 2009

Intrusion detection and tolerance for transaction based applications in wireless environments

Yacine Djemaiel and Noureddine Boudriga
CN&S Research Lab., University of the 7th of November at Carthage, Tunisia.
yacine.djemaiel@laposte.net, nab@supcom.rnu.tn

Abstract

Nowadays, many intrusion detection and tolerance systems have been proposed in order to detect attacks in both wired and wireless networks. Even if these solutions have shown some efficiency by detecting a set of complex attacks in wireless environments, they are unable to detect attacks using transaction based traffic in wireless environments. In this context, we propose an intrusion detection and tolerance scheme that is able to monitor heterogeneous traffic and to detect and tolerate attacks targeting transaction based applications interoperating in wireless environments. A case study is given to illustrate the proposed system capabilities against a complex attack scenario targeting a multi-player wireless gaming service.

A Topological Approach to Detect Conflicts in Firewall Policies

Subana Thanasegaran, Yi Yin, Yuichiro Tateiwa, Yoshiaki Katayama and Naohisa Takahashi
Department of Computer Science Engineering, Nagoya Institute of Technology
Gokiso, Showa, Nagoya 466-8555, Japan
{subana, yinyi, tateiwa, katayama, naohisa}@moss.elcom.nitech.ac.jp

Abstract

Packet filtering provides initial layer of security based upon set of ordered filters called firewall policies. It examines the network packets and decides whether to accept or deny them. But when a packet matches two or more filters conflicts arise. Due to the conflicts, some filters are never executed and some filters are occasionally executed. It may results into unintended traffic and it is a tedious job for administrator to detect conflicts. Detection of conflicts through geometrical approach provides a systematic and powerful error classification, but as the filters and key fields of header increase, it demands high memory and computation time. To solve this problem, we propose a topological approach called BISCAL (Bit-vector based spatial calculus) to detect the conflicts in the firewall policies. As because of our approach preserves only the topology of the filters, it can reduce memory usage and computation time to a great extend.

Automated Detection of Confidentiality Goals

Anders Moen Hagalisletto
Norwegian Computing Center
Postbox 1080 Blindern, 0316 Oslo, Norway
anders.moen@nr.no

Abstract

The security goals of an authentication protocol specify the high level properties of a protocol. Despite the importance of goals, these are rarely specified explicitly. Yet, a qualified analysis of a security protocol requires that the goals are stated explicitly. We propose a novel approach to find confidentiality goals in an automated way, based only on the protocol specification. The benefits of the method are: (i) Manual specification of goals is replaced by fully automated methods, (ii) the algorithm constructs the entire protection domain of a protocol, that is, all private and shared secrets, and (iii) the goal of an attack can be found, explaining which compromised entities are shared between the attacker and the honest principals.

Performance Analysis of Distributed Intrusion Detection Protocols for Mobile Group Communication Systems

Jin-Hee Cho
Computational & Information Sciences Directorate
U.S. Army Research Laboratory
jinhee.cho@us.army.mil

Ing-Ray Chen
Department of Computer Science
Virginia Tech
irchen@vt.edu

Abstract

Under highly security vulnerable, resource restricted, and dynamically changing mobile ad hoc environments, it is critical to be able to maximize the system lifetime while bounding the communication response time for mission-oriented mobile groups. In this paper, we analyze the tradeoff of security versus performance for distributed intrusion detection protocols employed in mobile group communication systems (GCSs). We investigate a distributed voting-based intrusion detection protocol for GCSs in multi-hop mobile ad hoc networks and examine the effect of intrusion detection on system survivability measured by the mean time to security failure (MTTSF) metric and efficiency measured by the communication cost metric. We identify optimal design settings under which the MTTSF metric can be best traded off for the communication cost metric or vice versa.

A New RFID Authentication Protocol with Resistance to Server Impersonation

Mete Akgün
Tübitak UEKAE
41470, Kocaeli, Turkey
makgun@uekae.tubitak.gov.tr

M. Ufuk Çağlayan
Computer Engineering Department
Boğaziçi University, İstanbul, Turkey
caglayan@boun.edu.tr

Emin Anarim
Electrical Engineering Department
Boğaziçi University, İstanbul, Turkey
anarim@boun.edu.tr

Abstract

Security is one of the main issues to adopt RFID technology in daily use. Due to resource constraints of RFID systems, it is very restricted to design a private authentication protocol based on existing cryptographic functions. In this paper, we propose a new RFID authentication protocol. The proposed protocol provides better protection against privacy and security threats than those before. Our proposed protocol is resistant to server impersonation attack introduced in [17]. Former proposal assumes that the adversary should miss any reader-to-tag communication flows and claims that their protocol is secure against forward traceability only in such communication environment. We show that even under such an assumption, the former proposed protocol is not secure. Our proposed protocol is secure against forward traceability, if the adversary misses any reader-to-tag communication flows. Our protocol also has low computational load on both the tag and the server side.

TLS Client Handshake with a Payment Card

David J. Boyd
Information Security Group, Royal Holloway, University of London, United Kingdom
D.Boyd@rhul.ac.uk

Abstract

Transport Layer Security (TLS) is the *de facto* standard for preventing eavesdropping, tampering or message forgery of higher-risk Internet communications, for example when making a payment. At heart TLS is a stateful cryptographic protocol built around a Public Key Infrastructure (PKI). However TLS is configurable; at one extreme it provides little protection and at the other end of the scale it provides protection against most threats to an Internet communication. In practice the “I” part of PKI is often not available at the client end so only the server end is authenticated. In this paper an optional TLS extension is proposed that dispenses with the need for the client to be registered with a PKI registration authority and instead uses a payment card to authenticate the user. This facilitates wider use of the available TLS services and can provide additional security services: enhanced privacy and certain non-repudiation services, for example.

Combating Side-Channel Attacks Using Key Management

Donggang Liu and Qi Dong
iSec Laboratory, Department of Computer Science and Engineering
The University of Texas at Arlington

Abstract

Embedded devices are widely used in military and civilian operations. They are often unattended, publicly accessible, and thus vulnerable to physical capture. Tamper-resistant modules are popular for protecting sensitive data such as cryptographic keys in these devices. However, recent studies have shown that adversaries can effectively extract the sensitive data from tamper-resistant modules by launching *semi-invasive side-channel* attacks such as power analysis and laser scanning. This paper proposes an effective key management scheme to harden embedded devices against side-channel attacks. This technique leverages the bandwidth limitation of side channels and employs an effective updating mechanism to prevent the keying materials from being exposed. This technique forces attackers to launch much more expensive and invasive attacks to tamper embedded devices and also has the potential of defeating *unknown* semi-invasive side-channel attacks.

Design of a Parallel AES for Graphics Hardware using the CUDA framework

Andrea Di Biagio, Alessandro Barenghi and Giovanni Agosta
Politecnico di Milano
{dibiagio,barenghi,agosta}@elet.polimi.it
Gerardo Pelosi
Universit ‘a degli Studi di Bergamo
gerardo.pelosi@unibg.it

Abstract

Web servers often need to manage encrypted transfers of data. The encryption activity is computationally intensive, and exposes a significant degree of parallelism. At the same time, cheap multicore processors are readily available on graphics hardware, and toolchains for development of general purpose programs are being released by the vendors. In this paper, we propose an effective implementation of the AES-CTR symmetric cryptographic primitive using the CUDA framework. We provide quantitative data for different implementation choices and compare them with the common CPU-based OpenSSL implementation on a performance-cost basis. With respect to previous works, we focus on optimizing the implementation for practical application scenarios, and we provide a throughput improvement of over 14 times. We also provide insights on the programming knowledge required to efficiently exploit the hardware resources by exposing the different kinds of parallelism built in the AES-CTR cryptographic primitive.

Security Analysis of Micali's Fair Contract Signing Protocol by Using Coloured Petri Nets : Multi-session case

Panupong Sornkhom[1] and Yongyuth Permpoontanalarp[2]
[1]Department of Electrical and Computer Engineering, Faculty of Engineering,
Naresuan University, Phitsanulok, Thailand
[2]Logic and Security Laboratory, Department of Computer Engineering, Faculty of Engineering
King Mongkut's University of Technology Thonburi, Bangkok, Thailand
yongyuth@cpe.kmutt.ac.th

Abstract

Micali proposed a simple and practical optimistic fair exchange protocol, called ECS1, for contract signing. Bao et al. found some message replay attacks in both the original ECS1 and a modified ECS1 where the latter aims to solve an ambiguity in the former. Furthermore, Bao et al. proposed an improved ECS1 which aims to prevent all those attacks. In this paper, we present a systematic method to analyze the security of Micali's ECS1 by using Coloured Petri Nets (CPN). By using CPN, we found two new attacks in the original protocol, five new attacks in Bao's modified protocol and surprisingly one new attack in Bao's improved protocol. All these new attacks occur when multiple sessions of protocol execution are performed concurrently.

Modeling and Analysis of Self-stopping BT Worms Using Dynamic Hit List in P2P Networks

Jiaqing Luo, Bin Xiao, Guobin Liu and Qingjun Xiao
Department of Computing
The Hong Kong Polytechnic University
Hong Kong
{csjluo, csbxiao, csgliu, and csqjxiao}@comp.polyu.edu.hk
Shijie Zhou
School of Computer Science and Engineering
University of Electronic Science and Technology of China
Chengdu, P.R. China
sjzhou@uestc.edu.cn

Abstract

Worm propagation analysis, including exploring mechanisms of worm propagation and formulating effects of network/worm parameters, has great importance for worm containment and host protection in P2P networks. Previous work only focuses on topological worm propagation where worms search a hosts neighbor-list to find new victims. In BitTorrent (BT) networks, the information from servers or trackers, however, could be fully exploited to design effective worms. In this paper, we propose a new approach for worm propagation in BT-like P2P networks. The worm, called Dynamic Hit-List (DHL) worm, locates new victims and propagates itself by requesting a tracker to build a dynamic hit list, which is a self-stopping BT worm to be stealthy. We construct an analytical model to study the propagation of such a worm: breadth-first propagation and depth-first propagation. The analytical results provide insights of the worm design into choosing parameters that enable the worm to stop itself after compromising a large fraction of vulnerable peers in a P2P network. We finally evaluate the performance of DHL worm through simulations. The simulation results verify the correctness of our model and show the effectiveness of the worm by comparing it with the topological worm.

SFTrust: A Double Trust Metric Based Trust Model in Unstructured P2P System

Yunchang Zhang
Nanjing University of Posts
and Telecommunications,
Nanjing, China
tinyyun@gmail.com

Shanshan Chen
Nanjing University of Posts
and Telecommunications,
Nanjing, China
moist_css@163.com

Geng Yang
Nanjing University of Posts
and Telecommunications,
Nanjing, China
yangg@njupt.edu.cn

Abstract

The P2P system is an anonymous and dynamic system, which offers enormous opportunities, and also presents potential threats and risks. In order to restrain malicious behaviors in P2P system, previous studies try to establish efficient trust models on P2P system. However, most of the trust models use a single trust metric, which can not reflect the practical trust values of the peers effectively. In this paper, we propose a trust model called SFTrust based on a double trust metric. SFTrust separates trust between the service providing and feedbacking. The system is based on the topology adaptation protocol, which was proposed in the unstructured P2P system. Simulation results show that SFTrust can efficiently resist general attacks. Compared with the single trust metric model, our mechanism can take full advantages of all the peers' service abilities for high performance.

Workshop 16

International Workshop on Hot Topics in Peer-to-Peer Systems

HOTP2P 2009

Robust vote sampling in a P2P media distribution system

Rameez Rahman, David Hales, Michel Meulpolder, Vincent Heinink, Johan Pouwelse and Henk Sips
Dept. of Computer Science
Technical University of Delft, Netherlands
rrameez@gmail.com

Abstract

The explosion of freely available media content through BitTorrent file sharing networks over the Internet means that users need guides or recommendations to find the right, high quality, content. Current systems rely on centralized servers to aggregate, rate and moderate metadata for this purpose. We present the design and simulations, using real BitTorrent traces, for a method combining fully decentralized metadata dissemination, vote sampling and ranking for deployment in the Tribler.org BitTorrent media client. Our design provides robustness to spam attacks, where metadata does not reflect the content it is attached to, by controlling metadata spreading and by vote sampling based on a collusion proof *experience* function. Our design is light-weight, fully decentralized and offers good performance and robustness under realistic conditions.

Reliable P2P Networks: TrebleCast and TrebleCast⋆

Ivan Hernandez-Serrano, Shadanan Sharma and Alberto Leon-Garcia
{ivan.hernandez.serrano, shadanan.sharma, alberto.leongarcia}@utoronto.ca
Department of Electrical and Computer Engineering
University of Toronto

Abstract

Node churn can have a severe impact on the performance of P2P applications. In this paper, we consider the design of reliable P2P networks that can provide predictable performance. We exploit the experimental finding that the age of a node can be a reliable predictor of longer residual lifetime to develop mechanisms that organize the network around these more reliable nodes. We propose two protocols, TrebleCast and TrebleCast⋆, to implement reliable overlay networks. These protocols dynamically create reliable layers of peers by moving nodes with higher expected lifetime to the center of the overlay. These more reliable layers can then be called upon to deliver predictable performance in the presence of churn.

Ten weeks in the life of an eDonkey server

Frédéric Aidouni, Matthieu Latapy and Clémence Magnien
LIP6 - CNRS and University Pierre & Marie Curie
104 avenue du president Kennedy, 75016 Paris, France
Firstname.Lastname@lip6.fr

Abstract

This paper presents a capture of the queries managed by an *eDonkey* server during almost 10 weeks, leading to the observation of almost 9 billion messages involving almost 90 million users and more than 275 million distinct files. Acquisition and management of such data raises several challenges, which we discuss as well as the solutions we developed. We obtain a very rich dataset, orders of magnitude larger than previously avalaible ones, which we provide for public use. We finally present basic analysis of the obtained data, which already gives evidence of non-trivial features.

Study on Maintenance Operations in a Chord-based Peer-to-Peer Session Initiation Protocol Overlay Network

Jouni Mäenpää and Gonzalo Camarillo
Ericsson
{jouni.maenpaa, gonzalo.camarillo}@ericsson.com

Abstract

Peer-to-Peer Session Initiation Protocol (P2PSIP) is a new technology being standardized in the Internet Engineering Task Force. A P2PSIP network consists of a collection of nodes organized in a peer-to-peer fashion for the purpose of enabling real-time communication using the Session Initiation Protocol (SIP). In this paper, we present experimental results obtained by running a P2PSIP prototype in PlanetLab. Our prototype uses the Chord Distributed Hash Table (DHT) to organize the P2PSIP overlay and Peer-to- Peer Protocol (P2PP) as the protocol spoken between the peers. In the experiments, the performance of the system is studied under different churn rates and using different DHT maintenance intervals.

Resource Advertising in PROSA P2P Network

Vincenza Carchiolo, Antonio Lima and Giuseppe Mangioni
Dipartimento di Ingegneria Informatica e delle Telecomunicazioni
Facoltà di Ingegneria - Università degli Studi di Catania
{vcarchiolo, gmangioni}@diit.unict.it, alima@studenti.unict.it

Abstract

P2P communication paradigm is a successful solution to the problem of resources sharing as shown by the numerous real overlay networks present on Internet. One of the issue of P2P networks is how a resource shared by a peer can be made known to the other peers or, in other words, how to advertise a resource on the network. In this paper we propose an advertising method for PROSA, a P2P architecture inspired by social relationships. In the paper we show that the introduction of our resources advertising method improves PROSA performance with a low overhead.

Relaxed-2-Chord: Efficiency, Flexibility and provable Stretch

Gennaro Cordasco
University of Salerno
cordasco@dia.unisa.it

Francesca Della Corte
University of Salerno,
francescadellacorte@virgilio.it

Alberto Negro
University of Salerno
alberto@dia.unisa.it

Alessandra Sala
University of California at Santa Barbara
alessandra@cs.ucsb.edu

Vittorio Scarano
University of Salerno
vitsca@dia.unisa.it

Abstract

Several proposals have been presented to supplement the traditional measure of routing efficiency in P2P networks, i.e. the (average) number of hops for lookup operations, with measures of the latency incurred in the underlying network. So far, no solution has been presented to this "latency" problem without incurring in extra and heavy management costs. We propose Relaxed-2-Chord, a new design of the traditional Chord protocol, that is able to fit the routing tables with low latency nodes, doing a parasitic measurement of nodes' latency without adding any overhead. The solution that we present is a Distributed Hash Table system whose aim is to combine the routing efficiency and flexibility of the Chord protocol – i.e. a good degree/diameter tradeoff – and a provable optimal hop by hop latency. Our work is inspired by the recent Lookup-parasitic random sampling (LPRS) strategies which allow to improve the network stretch, that is, the ratio between the latency of two nodes on the overlay network and the unicast latency between those nodes. Relaxed-2-Chord reaches the same results as LPRS without introducing any overhead.

Measurement of *eDonkey* Activity with Distributed Honeypots

Oussama Allali, Matthieu Latapy and Clémence Magnien
LIP6 C CNRS and Université Pierre et Marie Curie (UPMC C Paris 6)
104, avenue du Président Kennedy
75016 Paris – France
firstname.lastname@lip6.fr

Abstract

Collecting information about user activity in peer-to-peer systems is a key but challenging task. We describe here a distributed platform for doing so on the *eDonkey* network, relying on a group of honeypot peers which claim to have certain files and log queries they receive for these files. We then conduct some measurements with typical scenarios and use the obtained data to analyze the impact of key parameters like measurement duration, number of honeypots involved, and number of advertised files. This illustrates both the possible uses of our measurement system, and the kind of data one may collect using it.

Network Awareness of P2P Live Streaming Applications

Delia Ciullo[1], Maria Antonieta Garcia[1], Ákos Horváth[2] Emilio Leonardi[1], Marco Mellia[1], Dario Rossi[3]
Miklós Telek[2]and Paolo Veglia[3]
[1]Politecnico di Torino, lastname@tlc.polito.it
[2]Budapest University of Technology and Economics, lastname@hit.bme.hu
[3]TELECOM-ParisTech, firstname.lastname@enst.fr

Abstract

Early P2P-TV systems have already attracted millions of users, and many new commercial solutions are entering this market. Little information is however available about how these systems work. In this paper we present large scale sets of experiments to compare three of the most successful P2P-TV systems, namely PPLive, SopCast and TVAnts. Our goal is to assess what level of "network awareness" has been embedded in the applications, i.e., what parameters mainly drive the peer selection and data exchange. By using a general framework that can be extended to other systems and metrics, we show that all applications largely base their choices on the peer bandwidth, i.e., they prefer high-bandwidth users, which is rather intuitive. Moreover, TVAnts and PPLive exhibits also a preference to exchange data among peers in the same Autonomous System the peer belongs to. However, no evidence about preference versus peers in the same subnet or that are closer to the considered peer emerges. We believe that next-generation P2P live streaming applications definitively need to improve the level of network-awareness, so to better localize the traffic in the network and thus increase their network-friendliness as well.

BarterCast: A practical approach to prevent lazy freeriding in P2P networks

M. Meulpolder, J.A. Pouwelse, D.H.J. Epema and H.J. Sips
Parallel and Distributed Systems Group
Delft University of Technology, The Netherlands

Abstract

A well-known problem in P2P systems is *freeriding*, where users do not share content if there is no incentive to do so. In this paper, we distinguish *lazy freeriders* that are merely reluctant to share but follow the protocol, versus *die-hard freeriders* that employ sophisticated methods to subvert the protocol. Existing incentive designs often provide theoretically attractive resistance against die-hard freeriding, yet are rarely deployed in real networks because of practical infeasibility. Meanwhile, real communities benefit greatly from prevention of lazy freeriding, but have only centralized technology available to do so. We present a lightweight, fully distributed mechanism called BARTERCAST that prevents lazy freeriding and is deployed in practice. BarterCast uses a *maxflow* reputation algorithm based on a peer's private history of its data exchanges as well as indirect information received from other peers. We assess different reputation policies under realistic, trace-based community conditions and show that our mechanism is consistent and effective, even when significant fractions of peers spread false information. Furthermore, we present results of the deployment of BarterCast in the BitTorrent-based Tribler network which currently has thousands of users worldwide.

Underlay Awareness in P2P Systems: Techniques and Challenges

Osama Abboud, Aleksandra Kovacevic, Kalman Graffi, Konstantin Pussep and Ralf Steinmetz
Multimedia Communications Lab1, Technische Universität Darmstadt
{abboud, sandra, graffi, pussep, steinmetz}@kom.tu-darmstadt.de

Abstract

Peer-to-peer (P2P) applications have recently attracted a large number of Internet users. Traditional P2P systems however, suffer from inefficiency due to lack of information from the underlay, i.e. the physical network. Although there is a plethora of research on underlay awareness, this aspect of P2P systems is still not clearly structured. In this paper, we provide a taxonomic survey that outlines the different steps for achieving underlay awareness. The main contribution of this paper is presenting a clear picture of what underlay awareness is and how it can be used to build next generation P2P systems. Impacts of underlay awareness and open research issues are also discussed.

Analysis of PPLive through active and passive measurements

Salvatore Spoto, Rossano Gaeta, Marco Grangetto and Matteo Sereno
Dipartimento di Informatica, Università di Torino
Corso Svizzera 185, 10149 Torino, Italia
{spoto, rossano, grangetto, matteo}@di.unito.it

Abstract

The P2P-IPTV is an emerging class of Internet applications that is becoming very popular. The growing popularity of these rather bandwidth demanding multimedia streaming applications has the potential to flood the Internet with a huge amount of traffic.

In this paper we present an investigation of the popular P2PIPTV application PPLive exploiting a measurement strategy that combines both active and passive measures.

To this end, we use a crawler that allows the study of the topological characteristics of the overlay of one of the PPLive channels; concurrently, we perform passive measures on a PPlive client we run to join the crawled channel. We successively cross correlate information we obtained from the two measurements to assess the accuracy of the data captured by the crawler.

Our results reveal the potentials and the limits of PPLive active measures strategies.

A DDS-Compliant P2P Infrastructure for Reliable and QoS-Enabled Data Dissemination

Antonio Corradi and Luca Foschini
Dipartimento di Elettronica Informatica e Sistemistica
Università di Bologna
Viale Risorgimento, 2 – 40136 Bologna – ITALY
Telephone: +39-051-2093001
Fax: +39-051-2093073
{acorradi,lfoschini}@deis.unibo.it

Abstract

Recent trends in data-centric systems have motivated significant standardization efforts, such as the Data Distribution Service (DDS) to support data dis-semination with guaranteed Quality of Service (QoS) in heterogeneous Internet environments. Notwithstand-ing the central relevance of DDS in that scenario, DDS-based pub/sub solutions still exhibit limited sup-port for reliability, by omitting advanced techniques to reduce/eliminate QoS-degradations and data losses due to possible network and DDS system faults. We propose an original solution for fault- tolerance and prompt recovery of DDS-based pub/sub systems based on a DDS-compliant P2P routing substrate that con-tinuously achieves a guaranteed data delivery with expected QoS-levels. In contrast with similar solutions in the field, our proposal neither requires support for data persistency nor implies heavy client-side opera-tions. We exploit a DDS-compliant data dispatching infrastructure to reliably disseminate events and to balance data distribution load. The reported experi-mental results point out that our solution can guaran-tee desired requirements together with a limited over-head: the paper reports also performance indicators for our proposal CPU and memory resource usage.

Peer-to-Peer Beyond File Sharing: Where are P2P Systems Going?

Renato Lo Cigno
DISI, Univ. of Trento
Trento, Italy
locigno@disi.unitn.it

Tommaso Pecorella
DET, Univ. of Firenze
Firenze, Italy
tommaso.pecorella@unifi.it

Matteo Sereno
DI, Univ. of Torino
Torino, Italy
matteo@di.unito.it

Luca Veltri
DII, Univ. of Parma
Parma, Italy
luca.veltri@unipr.it

Abstract

Are P2P systems and applications here to stay? Or are they a bright meteor whose destiny is to disappear soon? In this paper we try to give a positive answer to the first question, highlighting reasons why the P2P paradigm should become an integral part of computing and communication services and not only oddities for Cyber-geeks.

Workshop 17

Workshop on Large-Scale, Volatile Desktop Grids

PCGRID 2009

An Analysis of Resource Costs in a Public Computing Grid

John A. Chandy
Department of Electrical and Computer Engineering
University of Connecticut
Storrs, CT USA
john.chandy@uconn.edu

Abstract

Public resource computing depends on the availability of computing resources that have been contributed by individuals. The amount of resources can be increased by incentivizing resource providers through payment for resources. However, there are costs associated with providing resources for public grid computing. Without an understanding of these costs, it is impossible for a provider to judge if the payment is sufficient to overcome those costs. In this paper, we present a provider cost model that considers all resource provider costs including opportunity costs, future-value costs, penalties, utility costs, and fixed costs. This model helps set a cost structure that a resource provider can use to determine whether it is profitable to participate in a public resource computing market.

MGST: A Framework for Performance Evaluation of Desktop Grids

Majd Kokaly, Issam Al-Azzoni and Douglas G. Down
Department of Computing and Software
McMaster University
Hamilton, Ontario, Canada
kokalym@mcmaster.ca, alazzoi@mcmaster.ca, downd@mcmaster.ca

Abstract

Desktop Grids are rapidly gaining popularity as a cost-effective computing platform for the execution of applications with extensive computing needs. As opposed to grids and clusters, these systems are characterized by having a non-dedicated infrastructure. These unique characteristics need to be considered in developing resource management strategies for Desktop Grids. Several frameworks for the performance evaluation of resource management strategies have been suggested for grids. However, similar projects for Desktop Grids are still lacking. This paper presents MGST, the first performance testing framework for Desktop Grids. We discuss the design of the tool and show how it can be used to analyze and improve the performance of an existing Desktop Grid scheduling policy.

Evaluating the Performance and Intrusiveness of Virtual Machines for Desktop Grid Computing

Patricio Domingues
School of Technology and Management -
Polytechnic Institute of Leiria, Portugal
patricio@estg.ipleiria.pt

Filipe Araujo, Luis Silva
CISUC, Dept. of Informatics Engineering,
University of Coimbra, Portugal
{filipius, luis}@dei.uc.pt

Abstract

We experimentally evaluate the performance overhead of the virtual environments VMware Player, QEMU, VirtualPC and VirtualBox on a dual-core machine. Firstly, we assess the performance of a Linux guest OS running on a virtual machine by separately benchmarking the CPU, file I/O and the network bandwidth. These values are compared to the performance achieved when applications are run on a Linux OS directly over the physical machine. Secondly, we measure the impact that a virtual machine running a volunteer @home project worker causes on a host OS. Results show that performance attainable on virtual machines depends simultaneously on the virtual machine software and on the application type, with CPU-bound applications much less impacted than IO-bound ones. Additionally, the performance impact on the host OS caused by a virtual machine using all the virtual CPU, ranges from 10% to 35%, depending on the virtual environment.

EmBOINC: An Emulator for Performance Analysis of BOINC Projects

Trilce Estradaand Michela Taufer
University of Delaware
{estrada, taufer}@udel.edu

Kevin Reed
IBM
knreed@ibm.us.com

David P. Anderson
University of Berkeley
davea@ssl.berkeley.edu

Abstract

BOINC is a platform for volunteer computing. The server component of BOINC embodies a number of scheduling policies and parameters that have a large impact on the projects' throughput and other performance metrics. We have developed a system, EmBOINC, for studying these policies and parameters. EmBOINC uses a hybrid approach: it simulates a population of volunteered clients (including heterogeneity, churn, availability, reliability) and it emulates the server component; that is, it uses the actual server software and its associated database. This paper describes the design of EmBOINC and validates its results based on trace data from an existing BOINC project.

GenWrapper: A Generic Wrapper for Running Legacy Applications on Desktop Grids

Attila Csaba Marosi, Zoltán Balaton and Péter Kacsuk
MTA SZTAKI Computer and Automation Research Institute of
Hungarian Academy of Sciences H-1528 Budapest, P.O.Box 63, Hungary
{atisu,balaton,kacsuk}@sztaki.hu

Abstract

Desktop Grids represent an alternative trend in Grid computing using the same software infrastructure as Volunteer Computing projects, such as BOINC. Applications to be deployed on a BOINC infrastructure need special preparations. However, there are many legacy applications, that have either no source code available or would require too much effort to port. For these applications BOINC provides a wrapper. This wrapper can handle the simple cases and it is configurable, but it can only be used to execute a list of legacy executables (tasks) one after the other. GenWrapper aims to provide a generic solution for wrapping and executing an arbitrary set of legacy applications by utilizing a POSIX like shell scripting environment to describe how the application is to be run and how the work unit should be processed. This is realized by an extended version of BusyBox providing the most common UNIX commands and a POSIX shell interpreter in a single executable with a special applet (BusyBox extension) to make BOINC API functions accessible from the shell on Windows, Linux and Mac OS X platforms. In this paper we present how GenWrapper works and how it can be used to port legacy applications to Desktop Grid systems.

Towards a Formal Model of Volunteer Computing Systems

WANG Yu[1,2], Haiwu HE[1,2] and WANG ZhiJian[2]
[1] INRIA, LIP, ENS Lyon, 46 avenue d'Italie, 69364 Lyon Cedex 07, France
[2] College of Computer and Information Engineering, Hohai University, Nanjing 210098, China
won9805@gmail.com, haiwu.he@inria.fr, zhjwang@hhu.edu.cn

Abstract

Volunteer Computing is a form of distributed computing in which the general public offers processing power and storage to scientific research projects. A large variety of Volunteer Computing Systems (VCS) have been proposed in the literature which use different architectures from client/server to P2P. This paper aims to provide a formal abstraction of VCS. At first, we identify three key roles played by VCS computing resources. Then, a formal model and related methods concerning Volunteer Computing are introduced. Relationships among elements are also characterized, based on set theory and operational reduction rules. We apply this model to describe a part of the XtremWeb protocol. Our results can help to lay a substantial foundation for the research on formalisms of Volunteer Computing.

Monitoring the EDGeS Project Infrastructure

Filipe Araujo[1], David Santiago[1], Diogo Ferreira[2], Jorge Farinha[2],
Patricio Domingues[3], Luis Moura Silva[1], Etienne Urbah[4], Oleg Lodygensky[4],
Haiwu He[5], Attila Csaba Marosi[6], Gabor Gombas[6], Zoltan Balaton[6],
Zoltan Farkas[6] and Peter Kacsuk[6]

[1] CISUC, Dept. of Informatics Engineering
University of Coimbra, Portugal
{filipius, demanuel, luis}@dei.uc.pt
[2] {defer, jfar}@student.dei.uc.pt

[3] School of Technology and Management
Polytechnic Institute of Leiria, Portugal
patricio@estg.ipleiria.pt

[4] LAL Universite Paris Sud, CNRS, IN2P3, France
{urbah, lodygens}@lal.in2p3.fr

[5] INRIA, LIP, ENS Lyon, France
haiwu.he@inria.fr

[6] MTA SZTAKI, Computer and Automation Research Institute
of the Hungarian Academy of Sciences
H-1528 Budapest, P.O.Box 63, Hungary
{atisu, gombasg, balaton, zfarkas, kacsuk}@sztaki.hu

Abstract

EDGeS is an European funded Framework Program 7 project that aims to connect desktop and service grids together. While in a desktop grid, personal computers pull jobs when they are idle, in service grids there is a scheduler that pushes jobs to available resources. The work in EDGeS goes well beyond conceptual solutions to bridge these grids together: it reaches as far as actual implementation, standardization, deployment, application porting and training.

One of the work packages of this project concerns monitoring the overall EDGeS infrastructure. Currently, this infrastructure includes two types of desktop grids, BOINC and XtremWeb, the EGEE service grid, and a couple of bridges to connect them. In this paper, we describe the monitoring effort in EDGeS: our technical approaches, the goals we achieved, and the plans for future work.

Thalweg: A Framework For Programming 1,000 Machines With 1,000 Cores

Adam L. Beberg
Department of Computer Science
Stanford University
beberg@cs.stanford.edu

Vijay S. Pande
Department of Chemistry
Stanford University
pande@stanford.edu

Abstract

While modern large-scale computing tasks have grown to span many machines, each with many cores, traditional programming models have not kept up with these advancements, resulting in difficulty exploiting these computing resources with only modest programmer effort. Thalweg seeks to address this breakdown in several ways. It provides a model for designing algorithms that have the potential to scale to multiple cores and machines, with subsequent optimization by software engineers. Based on this concept, Thalweg presents an API for handling these algorithms, for transferring data to and from nodes and coprocessors, and for verifying the correct operation of the hardware. Finally, Thalweg presents a set of concepts and a laboratory framework for pedagogical use that will educate the next generation of software engineers to operate in a world in which multi-core and distributed computing are everywhere.

BonjourGrid: Orchestration of Multi-instances of Grid Middlewares on Institutional Desktop Grids

Heithem Abbes[1,2]
[1]LIPN/UMR 7030—
CNRS/Université Paris 13,
99, avenue Jean-Baptiste Clément,
93430 Villetaneuse, FRANCE
heithem.abbes@lipn.univ-paris13.fr

Christophe Cérin
LIPN/UMR 7030—
CNRS/Université Paris 13
99, avenue Jean-Baptiste Clément,
93430 Villetaneuse, FRANCE
christophe.cerin@lipn.univ-paris13.fr

Mohamed Jemni
[2]Research Unit UTIC
ESSTT/Université de Tunis,
5, Av. Taha Hussein, B.P. 56,
Bab Mnara,Tunis,TUNISIA
mohamed.jemni@fst.rnu.tn

Abstract

While the rapidly increasing number of users and applications running on Desktop Grid (DG) systems does demonstrate its inherent potential, current DG implementations follow the traditional master-worker paradigm and DG middlewares do not cooperate. To extend the DG architecture, we propose a novel system, called BonjourGrid, capable of 1) creating, for each user, a specific execution environment in a decentralized fashion and 2) contrarily to classical DG, of orchestrating multiple and various instances of Desktop Grid middlewares. This will enable us to construct, on demand, specific execution environments (combinations of XtremWeb, Condor, Boinc middlewares). BonjourGrid is a software which aims to link a discovery service based on publish/subscribe protocol with the upper layer of a Desktop Grid middleware bridging the gap to meta-grid. Our experimental evaluation proves that BonjourGrid is robust and able to orchestrate more than 400 instances of XtremWeb middleware in a concurrent fashion on a 1000 host cluster. This experiment demonstrates the concept of BonjourGrid as well as its potential and shows that, comparing to a classical Desktop Grid with one central master, BonjourGrid suffers from an acceptable overhead that can be explained.

PyMW - a Python Module for Desktop Grid and Volunteer Computing

Eric M. Heien, Yusuke Takata and Kenichi Hagihara
Graduate School of Information Science and Technology
Osaka University, Toyonaka, Osaka 560-8531, Japan
{e-heien, y-takata, hagihara}@ist.osaka-u.ac.jp
Adam Kornafeld
Laboratory of Parallel and Distributed Systems
Computer and Automation Research Institute
Hungarian Academy of Sciences
H-1132 Victor Hugo u. 18-22, Budapest, Hungary
kadam@sztaki.hu

Abstract

We describe a general purpose master-worker parallel computation Python module called PyMW. PyMW is intended to support rapid development, testing and deployment of large scale master-worker style computations on a desktop grid or volunteer computing environment. This module targets nonexpert computer users by hiding complicated task submission and result retrieval procedures behind a simple interface. PyMW also provides a unified interface to multiple computing environments with easy extension to support additional environments. In this paper, we describe the internal structure and external interface to the PyMW module and its support for the Condor computing environment and the Berkeley Open Infrastructure for Network Computing (BOINC) platform. We demonstrate the effectiveness and scalability of PyMW by performing master-worker style computations on a desktop grid using Condor and a BOINC volunteer computing project.

Workshop 18

Workshop on Multi-Threaded Architectures and Applications

MTAAP 2009

Implementing OpenMP on a high performance embedded multicore MPSoC

Barbara Chapman and Lei Huang
University of Houston
Houston, TX, USA
{chapman,leihuang}@cs.uh.edu
Eric Biscondi, Eric Stotzer, Ashish Shrivastava and Alan Gatherer
Texas Instruments
Houston, TX, USA
{eric-biscondi,estotzer,ashu,gatherer}@xyz.com

Abstract

In this paper we discuss our initial experiences adapting OpenMP to enable it to serve as a programming model for high performance embedded systems. A high-level programming model such as OpenMP has the potential to increase programmer productivity, reducing the design/development costs and time to market for such systems. However, OpenMP needs to be extended if it is to meet the needs of embedded application developers, who require the ability to express multiple levels of parallelism, real-time and resource constraints, and to provide additional information in support of optimization. It must also be capable of supporting the mapping of different software tasks, or components, to the devices configured in a given architecture.

Multi-Threaded Library for Many-Core Systems

Allan Porterfield, Nassib Nassar, Rob Fowler
Renaissance Computing Institute (RENCI)
Chapel Hill, NC
{akp,nassar,rjf}@renci.org

Abstract

MAESTRO is a prototype runtime designed to provide simple, very light threads and synchronization between those threads on modern commodity (x86) hardware. The MAESTRO threading library is designed to be a target for a high-level language compiler or source-to-source translator, not for user-level programming. It provides parallel programming environments with a straight forward hardware model which can be mapped to available hardware dynamically. MAESTRO separates the size of the hardware system being used from the amount of parallelism available in an application. By separating the problem of locating parallelism from the problem of effectively using parallelism, both problems can be made easier. To the extent possible, the programming environment should be responsible for finding parallelism and the runtime should manage resource allocation and assignment.

Parallel regions and parallel loops are implemented. Several simple benchmarks have been ported from OpenMP to use the MAESTRO threading interface. Two synchronization mechanisms have been implemented, one for general synchronization and one for producer-consumer relationships. We have started building a level of 'virtualization' between the programming environment and the actual hardware, which will allow better hardware utilization and support new parallel programming languages.

Implementing a Portable Multi-threaded Graph Library: the MTGL on Qthreads

Brian W. Barrett[1], Jonathan W. Berry[1],
Richard C. Murphy[1] and Kyle B. Wheeler[1,2]

[1]Sandia National Laboratories
Albuquerque, NM USA
{bwbarre, jberry, rcmurph}@sandia.gov

[2]University of Notre Dame
Computer Science and Engineering
Notre Dame, IN USA
kwheeler@cse.nd.edu

Abstract

Graph-based Informatics applications challenge traditional high-performance computing (HPC) environments due to their unstructured communications and poor load-balancing. As a result, such applications have typically been relegated to either poor efficiency or specialized platforms, such as the Cray MTA/XMT series. The multi-threaded nature of the Cray MTA architecture presents an ideal platform for graph-based informatics applications. As commodity processors adopt features to enable greater levels of multi-threaded programming and higher memory densities, the ability to run these multi-threaded algorithms on less expensive, more available hardware becomes attractive. We present results from the Multi-Threaded Graph Library with the Qthreads portable threading package on a variety of commodity processors.

A Super-Efficient Adaptable Bit-Reversal Algorithm for Multithreaded Architectures

Anne C. Elster and Jan C. Meyer
Department of Computer and Information Science
Norwegian University of Science and Technology
Trondheim, Norway
elster@idi.ntnu.no and janchris@idi.ntnu.no

Abstract

Fast bit-reversal algorithms have been of strong interest for many decades, especially after Cooley and Tukey introduced their FFT implementation in 1965. Many recent algorithms, including FFTW try to avoid the bit-reversal all together by doing in-place algorithms within their FFTs. We therefore motivate our work by showing that for FFTs of up to 65.536 points, a minimally tuned Cooley-Tukey FFT in C using our bit-reversal algorithm performs comparable or better than the default FFTW algorithm.

In this paper, we present an extremely fast linear bit-reversal adapted for modern multithreaded architectures. Our bit-reversal algorithm takes advantage of recursive calls combined with the fact that it only generates pairs of indices for which the corresponding elements need to be exchanges, thereby avoiding any explicit tests. In addition we have implemented an adaptive approach which explores the trade-off between compile time and run-time work load. By generating look-up tables at compile time, our algorithm becomes even faster at run-time. Our results also show that by using more than one thread on tightly coupled architectures, further speed-up can be achieved.

Implementing and Evaluating Multithreaded Triad Census Algorithms on the Cray XMT

George Chin Jr., Andres Marquez and Sutanay Choudhury
High-Performance Computing
Pacific Northwest National Laboratory
Richland, WA USA
{George.Chin, Andres.Marquez, Sutanay.Choudhury}@pnl.gov

Kristyn Maschhoff
Cray, Inc.
Seattle, WA USA
kristyn@cray.com

Abstract

Commonly represented as directed graphs, social networks depict relationships and behaviors among social entities such as people, groups, and organizations. Social network analysis denotes a class of mathematical and statistical methods designed to study and measure social networks. Beyond sociology, social network analysis methods are being applied to other types of data in other domains such as bioinformatics, computer networks, national security, and economics. For particular problems, the size of a social network can grow to millions of nodes and tens of millions of edges or more. In such cases, researchers could benefit from the application of social network analysis algorithms on high-performance architectures and systems.

The Cray XMT is a third generation multithreaded system based on the Cray XT-3/4 platform. Like most other multithreaded architectures, the Cray XMT is designed to tolerate memory access latencies by switching context between threads. The processors maintain multiple threads of execution and utilize hardware-based context switching to overlap the memory latency incurred by any thread with the computations from other threads. Due to its memory latency tolerance, the Cray XMT has the potential of significantly improving the execution speed of irregular data-intensive applications such as those found in social network analysis.

In this paper, we describe our experiences in developing and optimizing three implementations of a social network analysis method known as triadic analysis to execute on the Cray XMT. The three implementations possess different execution complexities, qualities, and characteristics. We evaluate how the various attributes of the codes affect their performance on the Cray XMT. We also explore the effects of different compiler options and execution strategies on the different triadic analysis implementations and identify general XMT programming issues and lessons learned.

A Faster Parallel Algorithm and Efficient Multithreaded Implementations for Evaluating Betweenness Centrality on Massive Datasets

Kamesh Madduri
Computational Research Division
Lawrence Berkeley
National Laboratory
Berkeley, CA, USA

David Ediger,Karl Jiang
and David A. Bader
College of Computing
Georgia Institute of Technology
Atlanta, GA, USA

Daniel Chavarría-Miranda
High Performance Computing
Pacific Northwest
National Laboratory
Richland, WA, USA

Abstract

We present a new lock-free parallel algorithm for computing betweenness centrality of massive complex networks that achieves better spatial locality compared with previous approaches. Betweenness centrality is a key kernel in analyzing the importance of vertices (or edges) in applications ranging from social networks, to power grids, to the influence of jazz musicians, and is also incorporated into the DARPA HPCS SSCA#2, a benchmark extensively used to evaluate the performance of emerging high-performance computing architectures for graph analytics. We design an optimized implementation of betweenness centrality for the massively multithreaded Cray XMT system with the Threadstorm processor. For a small-world network of 268 million vertices and 2.147 billion edges, the 16-processor XMT system achieves a TEPS rate (an algorithmic performance count for the number of edges traversed per second) of 160 million per second, which corresponds to more than a 2× performance improvement over the previous parallel implementation. We demonstrate the applicability of our implementation to analyze massive real-world datasets by computing approximate betweenness centrality for the large IMDb movie-actor network.

Accelerating Numerical Calculation on the Cray XMT

Chad Scherrer, Tim Shippert and Andres Marquez
Computational Science and Mathematics Division
Pacific Northwest National Laboratory
Richland, WA, USA
{chad.scherrer, tim.shippert, andres.marquez}@pnl.gov

Abstract

The Cray XMT provides hardware support for parallel algorithms that would be communication- or memory-bound on other machines. Unfortunately, even if an algorithm meets these criteria, performance suffers if the algorithm is too numerically intensive. We present a lookup-based approach that achieves a significant performance advantage over explicit calculation. We describe an approach to balancing memory bandwidth against on-chip floating point capabilities, leading to further speedup. Finally, we provide table lookup algorithms for a number of common functions.

Early Experiences on Accelerating Dijkstra's Algorithm Using Transactional Memory

Nikos Anastopoulos, Konstantinos Nikas, Georgios Goumas and Nectarios Koziris
National Technical University of Athens
School of Electrical and Computer Engineering
Computing Systems Laboratory
{anastop,knikas,goumas,nkoziris}@cslab.ece.ntua.gr

Abstract

In this paper we use Dijkstra's algorithm as a challenging, hard to parallelize paradigm to test the efficacy of several parallelization techniques in a multicore architecture. We consider the application of Transactional Memory (TM) as a means of concurrent accesses to shared data and compare its performance with straightforward parallel versions of the algorithm based on traditional synchronization primitives. To increase the granularity of parallelism and avoid excessive synchronization, we combine TM with Helper Threading (HT). Our simulation results demonstrate that the straightforward parallelization of Dijkstra's algorithm with traditional locks and barriers has, as expected, disappointing performance. On the other hand, TM by itself is able to provide some performance improvement in several cases, while the version based on TM and HT exhibits a significant performance improvement that can reach up to a speedup of 1.46.

Early Experiences with Large-Scale Cray XMT Systems

David Mizell and Kristyn Maschhoff
Cray Inc.
Seattle, WA, USA
dmizell@cray.com; kristyn@cray.com

Abstract

Several 64-processor XMT systems have now been shipped to customers and there have been 128-processor, 256- processor and 512-processor systems tested in Cray's development lab. We describe some techniques we have used for tuning performance in hopes that applications continued to scale on these larger systems. We discuss how the programmer must work with the XMT compiler to extract maximum parallelism and performance, especially from multiply nested loops, and how the performance tools provide vital information about whether or how the compiler has parallelized loops and where performance bottlenecks may be occurring. We also show data that indicate that the maximum performance of a given application on a given size XMT system is limited by memory or network bandwidth, in a way that is somewhat independent of the number of processors used.

Linear Optimization on Modern GPUs

Daniele G. Spampinato and Anne C. Elster
Department of Computer and Information Science
Norwegian University of Science and Technology
Trondheim, Norway
daniele.spampinato@gmail.com, elster@idi.ntnu.no

Abstract

Optimization algorithms are becoming increasingly more important in many areas, such as finance and engineering. Typically, real problems involve several hundreds of variables, and are subject to as many constraints. Several methods have been developed trying to reduce the theoretical time complexity. Nevertheless, when problems exceed reasonable sizes they end up being very computationally intensive. Heterogeneous systems composed by coupling commodity CPUs and GPUs are becoming relatively cheap, highly performing systems. Recent developments of GPGPU technologies give even more powerful control over them.

In this paper, we show how we use a revised simplex algorithm for solving linear programming problems originally described by Dantzig for both our CPU and GPU implementations. Previously, this approach has showed not to scale beyond around 200 variables. However, by taking advantage of modern libraries such as ATLAS for matrix-matrix multiplication, and the NVIDIA CUDA programming library on recent GPUs, we show that we can scale to problem sizes up to at least 2000 variables in our experiments for both architectures. On the GPU, we also achieve an appreciable precision on large problems with thousands of variables and constraints while achieving between 2X and 2.5X speed-ups over the serial ATLAS-based CPU version. With further tuning of both the algorithm and its implementations, even better results should be achievable for both the CPU and GPU versions.

Enabling High-Performance Memory Migration for Multithreaded Applications on Linux

Brice Goglin[1], Nathalie Furmento[2]
[1]INRIA, [2]CNRS
LaBRI – 351 cours de la Libération, F-33405 TALENCE – FRANCE
[1]Brice.Goglin@inria.fr — [2]Nathalie.Furmento@labri.fr

Abstract

As the number of cores per machine increases, memory architectures are being redesigned to avoid bus contention and sustain higher throughput needs. The emergence of Non-Uniform Memory Access (NUMA) constraints has caused affinities between threads and buffers to become an important decision criterion for schedulers.

Memory migration dynamically enables the joint distribution of work and data across the machine but requires high-performance data transfers as well as a convenient programming interface. We present improvements of the LINUX migration primitives and the implementation of a *Next-touch* policy in the kernel to provide multithreaded applications with an easy way to dynamically maintain thread-data affinity.

Microbenchmarks show that our work enables a high-performance, synchronous and lazy memory migration within multithreaded applications. A threaded LU factorization then reveals the large improvement that our *Next-touch* policy model may bring in applications with complex access patterns.

Exploiting DMA to enable non-blocking execution in Decoupled Threaded Architecture

Roberto Giorgi, Zdravko Popovic and Nikola Puzovic
Department of Information Engineering
University of Siena - Siena, Italy
http://www.dii.unisi.it/ {giorgi, popovic, puzovic}

Abstract

DTA (Decoupled Threaded Architecture) is designed to exploit fine/medium grained Thread Level Parallelism (TLP) by using a distributed hardware scheduling unit and relying on existing simple cores (in-order pipelines, no branch predictors, no ROBs).

In DTA, the local variables and synchronization data are communicated via a fast frame memory. If the compiler can not remove global data accesses, the threads are excessively fragmented. Therefore, in this paper, we present an implementation of a pre-fetching mechanism (for global data) that complements the original DTA pre-load mechanism (for consumer-producer data patterns) with the aim of improving non-blocking execution of the threads.

Our implementation is based on an enhanced DMA mechanism to prefetch global data. We estimated the benefit and identified the required support of this proposed approach, in an initial implementation. In case of longer latency to access memory, our idea can reduce execution time greatly (i.e., 11x for the zoom benchmark on 8 processors) compared to the case of no-prefetching.

Workshop 19

Workshop on Parallel and Distributed Computing in Finance

PDCoF 2009

Pricing American Options with the SABR Model

M.H. Vellekoop
Department of Applied Mathematics, University of Twente, the Netherlands,
The Derivatives Technology Foundation, Amsterdam.
Tel +31 53 489 2087
m.h.vellekoop@math.utwente.nl
G. Vlaming
Saen Options, Amsterdam, The Netherlands.

Abstract

We introduce a simple and flexible method to price derivative securities on assets with volatilities which are stochastic. As a special case we treat the SABR model in more detail. Our approach is based on the construction of recombining trees using interpolation methods on probability measures, and this makes it very suitable for the application of parallel computing techniques. We show how one can easily incorporate features which are characteristic for practical option pricing problems, such as a term structure of interest, early exercise possibilities and the payment of cash dividends.

High Dimensional Pricing of Exotic European Contracts on a GPU Cluster, and Comparison to a CPU Cluster

Lokman A. Abbas-Turki[1], Stephane Vialle[2,3], Bernard Lapeyre[1], Patrick Mercier[2]
[1]ENPC-CERMICS, Applied Probability Research Group, 77455 Champs-sur-Marne, France
[2]SUPELEC, IMS group, 2 rue Edouard Belin, 57070 Metz, France
[3]AlGorille INRIA Project Team, 615, rue du Jardin Botanique 54600 Villers-les-Nancy France, France

Abstract

The aim of this paper is the efficient use of CPU and GPU clusters for a general path-dependent exotic European pricing, and their comparison in terms of speed and energy consumption. To reach our goal, we propose a parallel random number generator which is well suited to the parallelization paradigm, then, we implement a multidimensional Asian contract as a benchmark using g++/OpenMP/OpenMPI on CPUs and CUDA-nvcc/OpenMPI on GPUs. Finally, we give the detailed results of the two architectures for different size problems using 1-16 GPUs and 1-256 dual-core CPUs.

Using Premia and Nsp for Constructing a Risk Management Benchmark for Testing Parallel Architecture

Jean-Philippe Chancelier and Bernard Lapeyre
Université Paris-Est, CERMICS, École des Ponts
Champs sur Marne, 77455 Marne la Vallée Cedex 2, France
jpc@cermics.enpc.fr, bl@cermics.enpc.fr
Jérôme Lelong
Ecole Nationale Supérieure de Techniques Avancées
ParisTech
Unité de Mathématiques Appliquées
42 bd Victor 75015 Paris
jerome.lelong@ensta.fr

Abstract

Financial institutions have massive computations to carry out overnight which are very demanding in terms of the consumed CPU. The challenge is to price many different products on a cluster-like architecture. We have used the Premia software to valuate the financial derivatives. In this work, we explain how Premia can be embedded into Nsp, a scientific software like Matlab, to provide a powerful tool to valuate a whole portfolio. Finally, we have integrated an MPI toolbox into Nsp to enable to use Premia to solve a bunch of pricing problems on a cluster. This unified framework can then be used to test different parallel architectures.

Towards the Balancing Real-Time Computational Model: Example of Pricing and Risk Management of Exotic Derivatives

Grzegorz Gawron
Foreign Exchange and Precious Metals Options Pricing IT
HSBC IB

Abstract

Instant pricing and risk calculation of exotic financial derivative instruments is essential in the process of risk management and trading performed by financial institutions. Due to the lack of analytical solutions for pricing of such instruments, systems require the use of computationally intensive Monte-Carlo methods. Despite using extensive computational power of clusters or grids, these calculations are usually difficult to complete in real-time, as the rate of the incoming market data is too high to handle.

The objective of this paper is to present a certain phenomenon existing in the pricing and risk management systems. The phenomenon is based on an interplay of intense computational requirements for single calculation, with frequent change in the environment state. A suggested abstraction leads to a definition of a Balancing Real-time Computational Model.

An implementation of the solution to the problem is presented as an optimalisation task. It is based on a distance function quantifying the degree of the imbalance of the system.

Advanced Risk Analytics on the Cell Broadband Engine

Ciprian Docan and Manish Parashar
Center for Autonomic Computing, ECE Department
Rutgers University, Piscataway NJ, USA
{docan,parashar}@caip.rutgers.edu

Christopher Marty
Bloomberg LP
New York, NY, USA
cmarty@bloomberg.net

Abstract

This paper explores the effectiveness of using the CBE platform for Value-at-Risk (VaR) calculations. Specifically, it focuses on the design, optimization and evaluation of pricing European and American stock options across Monte-Carlo VaR scenarios. This analysis is performed on two distinct platforms with CBE processors, i.e., IBM Q22 blade server and the Playstation3 gaming console.

A High Performance Pair Trading Application

Jieren Wang
Department of Mathematics
University of British Columbia
Vancouver, British Columbia
cwang@math.ubc.ca

Camilo Rostoker and Alan Wagner
Department of Computer Science
University of British Columbia
Vancouver, British Columbia
{rostokec,wagner}@cs.ubc.ca

Abstract

This paper describes a high-frequency pair trading strategy that exploits the power of MarketMiner, a high-performance analytics platform that enables a real-time, marketwide search for short-term correlation breakdowns across multiple markets and asset classes. The main theme of this paper is to discuss the computational requirements of model formulation and back-testing, and how a scalable solution built using a modular, MPI-based infrastructure can assist quantitative model and strategy developers by increasing the scale of their experiments or decreasing the time it takes to thoroughly test different parameters. We describe our work to date which is the design of a canonical pair trading algorithm, illustrating how fast and efficient backtesting can be performed using MarketMiner. Preliminary results are given based on a small set of stocks, parameter sets and correlation measures.

Option Pricing with COS method on Graphics Processing Units

Bowen Zhang
Delft University of Technology
Mekelweg 4, 2628 CD
Delft, the Netherlands
bowen.zhang@tudelft.nl

Cornelis W.Oosterlee
Centrum Wiskunde & Informatica
Amsterdam, the Netherlands
c.w.oosterlee@cwi.nl

Abstract

In this paper, acceleration on the GPU for option pricing by the COS method is demonstrated. In particular, both European and Bermudan options will be discussed in detail. For Bermudan options, we consider both the Black-Scholes model and Lévy processes of infinite activity. Moreover, the influence of the number of terms in the Fourier-cosine expansion, N, as well as the number of exercise dates, M, on the acceleration factor of the GPU is explored. We also give a comparison between different ways of GPU and CPU implementation. For instance, we have optimized the GPU implementation for maximum performance which is compared to a hybrid CPU/GPU version which outperforms the pure GPU or CPU versions for European options. Furthermore, for each process and each option type that is covered by this paper, we give a discussion on the precision of the GPU.

Calculation of Default Probability (PD) solving Merton Model PDEs on Sparse Grids

Philipp Schroeder
Goethe-Center for Scientific Computing, G-CSC
Goethe-University
Frankfurt am Main, Germany
philipp.schroeder@gcsc.uni-frankfurt.de
Prof. Dr. Gabriel Wittum
Goethe-Center for Scientific Computing, G-CSC
Goethe-University
Frankfurt am Main, Germany
wittum@gcsc.uni-frankfurt.de

Abstract

Actual developements of the sub-prime crisis of 2008 have put a strong focus on the importance of credit default models. The Merton Model is one of these models, using partial differential equations to calculate the probability of default (PD) for a correlated credit portfolio. The resulting equations are discretized on structured sparse grids through the method of Finite-Differences and numerically solved using the software package SG2. Parallel Computing is used to speed up the calculations.

An Aggregated Ant Colony Optimization Approach for Pricing Options

Yeshwanth Udayshankar, Sameer Kumar, Girish K. Jha and Ruppa K. Thulasiram
and Parimala Thulasiraman
Department of Computer Science
University of Manitoba
Winnipeg, Manitoba, Canada
maniyesh, sameer, girish08, tulsi, thulasir@cs.umanitoba.ca

Abstract

Estimating the current cost of an option by predicting the underlying asset prices is the most common methodology for pricing options. Pricing options has been a challenging problem for a long time due to unpredictability in market which gives rise to unpredictability in the option prices. Also the time when the options have to be exercised has to be determined to maximize the profits. This paper proposes an algorithm for predicting the time and price when the option can be exercised to gain expected profits.

The proposed method is based on Nature inspired algorithm i.e. Ant Colony Optimization (ACO) which is used extensively in combinatorial optimization problems and dynamic applications such as mobile ad-hoc networks where the objective is to find the shortest path. In option pricing, the primary objective is to find the best node in terms of price and time that would bring expected profit to the investor. Ants traverse the solution space (asset price movements) in the market to identify a profitable node. We have designed and implemented an Aggregated ACO algorithm to price options which is distributed and robust. The initial results are encouraging and we are continuing this work further.

A Novel Application of Option Pricing to Distributed Resources Management

David Allenotor
Dept. of Computer Science
University of Manitoba
Winnipeg, R3T 2N2, Canada
dallen@cs.umanitoba.ca

Ruppa Thulasiram
Dept. of Computer Science
University of Manitoba
Winnipeg, R3T 2N2, Canada
tulsi@cs.umanitoba.ca

Parimala Thulasiraman
Dept. of Computer Science
University of Manitoba
Winnipeg, R3T 2N2, Canada
thulasir@cs.umanitoba.ca

Abstract

In this paper, we address a novel application of financial option pricing theory to the management of distributed computing resources. To achieve the set objective, first, we highlight the importance of finance models for the given problem and explain how option theory fits well to price the distributed grid compute resources. Second, we design and develop a pricing model and generate pricing results based on the trace data drawn from two real grids: one commercial grid Auvergrid and one experimental platform grid LCG. We evaluate our proposed model using various grid compute resources (such as memory, storage, software, and compute cycles) as individual commodities. By carrying out several experiments, a justification of the pricing model is obtained by comparing real behavior to a simulated system based on the spot price for the resources. We further enhanced our model to achieve a desirable balance between Quality of Service (QoS) and profitability from the perspectives of the users and resource operators respectively.

Workshop 20

Workshop on Large-Scale Parallel Processing

LSPP 2009

The world's fastest CPU and SMP node: Some performance results from the NEC SX-9

Thomas Zeiser, Georg Hager and Gerhard Wellein
Erlangen Regional Computing Center
University of Erlangen-Nuremberg, Germany
thomas.zeiser@rrze.uni-erlangen.de

Abstract

Classic vector systems have all but vanished from recent TOP500 lists. Looking at the newly introduced NEC SX-9 series, we benchmark its memory subsystem using the low level vector triad and employ an advanced lattice Boltzmann flow solver kernel to demonstrate that classic vectors still combine excellent performance with a well-established optimization approach. Results for commodity x86-based systems are provided for reference.

GPU Acceleration of Zernike Moments for Large-scale Images

Manuel Ujaldon
Computer Architecture Department
University of Malaga
Malaga, Spain
ujaldon@uma.es

Abstract

Zernike moments are trascendental digital image descriptors used in many application areas like biomedical image processing and computer vision due to their good properties of orthogonality and rotation invariance. However, their computation is too expensive and limits its application in practice, overall when real-time constraints are imposed. This work introduces a novel approach to the high-performance computation of Zernike moments using CUDA on graphics processors. The proposed method is applicable to the computation of an individual Zernike moment as well as a set of Zernike moments of a given order, and it is compared against three of the fastest implementations performed on CPUs over the last decade. Our experimental results on a commodity PC reveal up to 5x faster execution times on a GeForce 8800 GTX against the best existing implementation on a Pentium 4 CPU.

Harnessing the Power of idle GPUs for Acceleration of Biological Sequence Alignment

Fumihiko Ino, Yuki Kotani and Kenichi Hagihara
Graduate School of Information Science and Technology, Osaka University
1-3 Machikaneyama, Toyonaka, Osaka 560-8531, Japan
ino@ist.osaka-u.ac.jp

Abstract

This paper presents a parallel system capable of accelerating biological sequence alignment on the graphics processing unit (GPU) grid. The GPU grid in this paper is a desktop grid system that utilizes idle GPUs and CPUs in the office and home. Our parallel implementation employs a master-worker paradigm to accelerate Liu's OpenGL-based algorithm that runs on a single GPU. We integrate this implementation into a screensaver-based grid system that detects idle resources on which the alignment code can run. We also show some experimental results comparing our implementation with three different implementations running on a single GPU, a single CPU, or multiple CPUs. As a result, we find that a single non-dedicated GPU can provide us almost the same throughput as two dedicated CPUs in our laboratory environment, where GPU-equipped machines are ordinarily used to develop GPU applications.

Application Profiling on Cell-based Clusters

Hikmet Dursun[1,2], Kevin J. Barker[1], Darren J. Kerbyson[1] and Scott Pakin[1]
[1]Performance and Architecture Laboratory (PAL), Computer Science for HPC (CCS-1)
Los Alamos National Laboratory, NM 87545, USA
[2]Collaboratory for Advanced Computing and Simulations, Department of Computer Science
University of Southern California, CA 90089, USA
{hdursun, kjbarker, djk, pakin}@lanl.gov

Abstract

In this paper, we present a methodology for profiling parallel applications executing on the IBM PowerXCell 8i (commonly referred to as the "Cell" processor). Specifically, we examine Cell-centric MPI programs on hybrid clusters containing multiple Opteron and Cell processors per node such as those used in the petascale Roadrunner system. Our implementation incurs less than 3.2 μs of overhead per profile call while efficiently utilizing the limited local store of the Cell's SPE cores. We demonstrate the use of our profiler on a cluster of hybrid nodes running a suite of scientific applications. Our analyses of inter-SPE communication (across the entire cluster) and function call patterns provide valuable information that can be used to optimize application performance.

Non-Uniform Fat-Meshes for Chip Multiprocessors

Yu Zhang and Alex K. Jones
University of Pittsburgh
Pittsburgh, PA 15261 USA
yuz28@pitt.edu, akjones@ece.pitt.edu

Abstract

This paper studies the traffic hot spots of mesh networks in the context of chip multiprocessors. To mitigate these effects, this paper describes a non-uniform fat-mesh extension to mesh networks, which are popular for chip multiprocessors. The fat-mesh is inspired by the fat-tree and dedicates additional links for connections with heavy traffic (e.g. near the center) with fewer links for lighter traffic (e.g. near the periphery). Two fat-mesh schemes are studied based on the traffic requirements of chip multiprocessors using dimensional ordered XY routing and a randomized XY-YX routing algorithms, respectively. Analytical fat-mesh models are constructed by theoretically presenting the expressions for the traffic requirements of personalized all-to-all traffic for both the raw message numbers and their normalized equivalents. We demonstrate how traffic scales for a traditional mesh compared to a non-uniform fat mesh.

An Evaluative Study on the Effect of Contention on Message Latencies in Large Supercomputers

Abhinav Bhatele and Laxmikant V. Kale
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA
{bhatele, kale}@illinois.edu

Abstract

Significant theoretical research was done on interconnect topologies and topology aware mapping for parallel computers in the 80s. With the deployment of virtual cut-through, wormhole routing and faster interconnects, message latencies reduced and research in the area died down. This paper presents a study showing that with the emergence of very large supercomputers, typically connected as a 3D torus or mesh, topology effects have become important again. It presents an evaluative study on the effect of contention on message latencies on torus and mesh networks.

The paper uses three MPI benchmarks to evaluate the effect of hops (links) traversed by messages, on their latencies. The benchmarks demonstrate that when multiple messages compete for network resources, link occupancy or contention can increase message latencies by up to a factor of 8 times. In other words, contention leads to increased message latencies and reduces effective available bandwidth for each message. This suggests that application developers should consider interconnect topologies when mapping tasks to processors in order to obtain the best performance. Results are shown for two parallel machines – ANL's Blue Gene/P and PSC's XT3.

The Impact of Network Noise at Large-Scale Communication Performance

Torsten Hoefler, Timo Schneider and Andrew Lumsdaine
Open Systems Laboratory
Indiana University
Bloomington IN 47405, USA
{htor,timoschn,lums}@cs.indiana.edu

Abstract

The impact of operating system noise on the performance of large-scale applications is a growing concern and ameliorating the effects of OS noise is a subject of active research. A related problem is that of network noise, which arises from shared use of an interconnection network by parallel processes. To characterize the impact of network noise on parallel applications we conducted a series of simulations and experiments using a newly-developed benchmark. Experiment results show a decrease in the communication performance of a parallel reduction operation by a factor of two on 246 nodes. In addition, simulations show that influence of network noise grows with the system size. Although network noise is not as well-studied as OS noise, our results clearly show that it is an important factor that must be considered when running large-scale applications.

Large Scale Experiment and Optimization of a Distributed Stochastic Control Algorithm. Application to Energy Management Problems

Pascal. Vezolle[1], Stephane Vialle[2,3] and Xavier Warin[4]
[1]IBM Deep Computing Europe, 34060 Montpellier, FRANCE
[2]SUPELEC, IMS group, 2 rue Edouard Belin, 57070 Metz, France
[3]AlGorille INRIA Project Team, 615, rue du Jardin Botanique 54600 Villers-les-Nancy France, France
[4]EDF - R&D, OSIRIS group, 92141 Clamart, France

Abstract

Asset management for the electricity industry leads to very large stochastic optimization problem. We explain in this article how to efficiently distribute the Bellman algorithm used, re-distributing data and computations at each time step, and we examine the parallelization of a simulation algorithm usually used after this optimization part. We focus on distributed architectures with shared memory multi-core nodes, and we design a multiparadigm parallel algorithm, implemented with both MPI and multithreading mechanisms. Then we lay emphasis on the serial optimizations carried out to achieve high performances both on a dual-core PC cluster and a Blue Gene/P IBM supercomputer with quadcore nodes.

Finally, we introduce experimental results achieved on two large testbeds, running a 7-stocks and 10-state-variables benchmark, and we show the impact of multithreading and serial optimizations on our distributed application.

Performance Analysis and Projections for Petascale Applications on Cray XT Series Systems

Sadaf R. Alam, Richard F. Barrett, Jeffery A. Kuehn and Steve. W. Poole
Oak Ridge National Laboratory
Oak Ridge, USA
alamsr,rbarrett,kuehn,spoole@ornl.gov

Abstract

The Petascale Cray XT5 system at the Oak Ridge National Laboratory (ORNL) Leadership Computing Facility (LCF) shares a number of system and software features with its predecessor, the Cray XT4 system including the quad-core AMD processor and a multi-core aware MPI library. We analyze performance of scalable scientific applications on the quad-core Cray XT4 system as part of the early system access using a combination of micro-benchmarks and Petascale ready applications. Particularly, we evaluate impact of key changes that occurred during the dual-core to quad-core processor upgrade on applications behavior and provide projections for the next-generation massively-parallel platforms with multicore processors, specifically for proposed Petascale Cray XT5 system. We compare and contrast the quad-core XT4 system features with the upcoming XT5 system and discuss strategies for improving scaling and performance for our target applications.

Performance Modeling in Action: Performance Prediction of a Cray XT4 System during Upgrade

Kevin J. Barker, Kei Davis and Darren J. Kerbyson
Performance and Architecture Lab (PAL), Los Alamos National Laboratory
{kjbarker,kei.davis,djk}@lanl.gov

Abstract

We present predictive performance models of two of the petascale applications, S3D and GTC, from the DOE Office of Science workload. We outline the development of these models and demonstrate their validation on an Opteron/Infiniband cluster and the pre-upgrade ORNL Jaguar system (Cray XT3/XT4). Given the high accuracy of the full application models, we predict the performance of the Jaguar system after the upgrade of its nodes, and subsequently compare this to the actual performance of the upgraded system. We then analyze the performance of the system based on the models to quantify bottlenecks and potential optimizations. Finally, the models are used to quantify the benefits of alternative node allocation strategies.