Proceedings of
25th IEEE International Parallel and Distributed Processing Symposium

IPDPS 2011 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

IEEE International Parallel & Distributed Processing Symposium

IPDPS 2011

Session 1
Resource Management

Power-aware replica placement and update strategies in tree networks

Anne Benoit, Paul Renaud-Goud, Yves Robert
LIP, Ecole Normale Supérieure de Lyon
46 Allée d'Italie, 69364 Lyon Cedex 07, France
Email: {Anne.Benoit|Paul.Renaud-Goud|Yves.Robert}@ens-lyon.fr

Abstract

This paper deals with optimal strategies to place replicas in tree networks, with the double objective to minimize the total cost of the servers, and/or to optimize power consumption. The client requests are known beforehand, and some servers are assumed to pre-exist in the tree. Without power consumption constraints, the total cost is an arbitrary function of the number of existing servers that are reused, and of the number of new servers. Whenever creating and operating a new server has higher cost than reusing an existing one (which is a very natural assumption), cost optimal strategies have to trade-off between reusing resources and load-balancing requests on new servers. We provide an optimal dynamic programming algorithm that returns the optimal cost, thereby extending known results without pre-existing servers. With power consumption constraints, we assume that servers operate under a set of M different modes depending upon the number of requests that they have to process. In practice M is a small number, typically 2 or 3, depending upon the number of allowed voltages. Power consumption includes a static part, proportional to the total number of servers, and a dynamic part, proportional to a constant exponent of the server mode, which depends upon the model for power. The cost function becomes a more complicated function that takes into account reuse and creation as before, but also upgrading or downgrading an existing server from one mode to another. We show that with an arbitrary number of modes, the power minimization problem is NP-complete, even without cost constraint, and without static power. Still, we provide an optimal dynamic programming algorithm that returns the minimal power, given a threshold value on the total cost; it has exponential complexity in the number of modes M, and its practical usefulness is limited to small values of M. Still, experiments conducted with this algorithm show that it can process large trees in reasonable time, despite its worst-case complexity.

Minimum Cost Resource Allocation for Meeting Job Requirements

Venkatesan T. Chakaravarthy, Gyana R Parija, Sambuddha Roy, Yogish Sabharwal
IBM Research India, New Delhi, India 110070
Email:vechakra, gyana.parija, sambuddha, ysabharwal@in.ibm.com
Amit Kumar
IIT Delhi, New Delhi, India 110016
Email:amitk@cse.iitd.ac.in

Abstract

We consider the problem of allocating resources for completing a collection of jobs. Each resource is specified by a start-time, finish-time and the capacity of resource available and has an associated cost; and each job is specified by a start-time, finish-time and the amount of the resource required (demand) during this interval. A feasible solution is a multiset of resources (i.e., multiple units of each resource may be picked) such that at any point of time, the sum of the capacities offered by the resources is at least the total demand of the jobs active at that point of time. The cost of the solution is the sum of the costs of the resources included in the solution (taking into account the units of the resources). The goal is to find a feasible solution of minimum cost. This problem arises naturally in many scenarios. For example, given a set of jobs, we would like to allocate some resource such as machines, memory or bandwidth in order to complete all the jobs.

This problem generalizes a covering version of the knapsack problem which is known to be NP-hard. We present a constant factor approximation algorithm for this problem based on a Primal-Dual approach.

Power and Performance Management in Priority-type Cluster Computing Systems

Kaiqi Xiong
Department of Computer Science
Texas A&M University
Commerce, TX USA

Abstract

Cluster computing not only improves performance but also increase power consumption. It is a challenge to increase the performance of a cluster computing system and reduce its power consumption simultaneously. In this paper, we consider a collection of cluster computing resources owned by a service provider to host an enterprise application for multiple class business customers where customer requests are distinguished, with different request characteristics and service requirements. We start with a development of computing an average end-to-end delay and an average energy consumption for multiple class customers in such an application. Then, we present approaches for optimizing the average end-to-end delay subject to the constraint of an average energy consumption and optimizing the average end-to-end energy consumption subject to the constraints of an average end-to-end delay for all class and each class customer requests respectively. Moreover, a service provider processes the service requests of customers according to a service level agreement (SLA), which is a contract agreed between a customer and a service provider. It becomes important and commonplace to prioritize multiple customer services in favor of customers who are willing to pay higher fees. We propose an approach for minimizing the total cost of cluster computing resources allocated to ensure multiple priority customer service guarantees by the service provider. It is demonstrated through our simulation that the proposed approaches are efficient and accurate for power management and performance guarantees in priority-type cluster computing systems.

Willow: A Control System for Energy and Thermal Adaptive Computing

Krishna Kant
Intel Corporation
Email: krishna.kant@intel.com

Muthukumar Murugan
University Of Minnesota
Minneapolis, USA-55414
Email: murugan@cs.umn.edu

David H.C.Du
University Of Minnesota
Minneapolis, USA-55414
Email: du@cs.umn.edu

Abstract

The increasing energy demand coupled with emerging sustainability concerns requires a re-examination of power/thermal issues in data centers from the perspective of short term energy deficiencies. Such energy deficient scenarios arise for a variety of reasons including variable energy supply from renewable sources and inadequate power, thermal and cooling capacities. In this paper we propose a hierarchical control scheme to adapt assignments of tasks to servers in a way that can cope with the varying energy limitations and still provide necessary QoS . The rescheduling of tasks on different servers has direct (migration related) and indirect (changed traffic patterns) network energy impacts that we also consider. We show the stability of our scheme and evaluate its performance via detailed simulations and experiments.

Session 2
Communication and I/O Optimization

Communication-Avoiding QR Decomposition for GPUs

Michael Anderson, Grey Ballard, James Demmel and Kurt Keutzer
UC Berkeley: Department of Electrical Engineering and Computer Science
Berkeley, CA USA
{mjanders,ballard,demmel,keutzer}@cs.berkeley.edu

Abstract

We describe an implementation of the Communication-Avoiding QR (CAQR) factorization that runs entirely on a single graphics processor (GPU). We show that the reduction in memory traffic provided by CAQR allows us to outperform existing parallel GPU implementations of QR for a large class of tall-skinny matrices. Other GPU implementations of QR handle panel factorizations by either sending the work to a general-purpose processor or using entirely bandwidth-bound operations, incurring data transfer overheads. In contrast, our QR is done entirely on the GPU using compute-bound kernels, meaning performance is good regardless of the width of the matrix. As a result, we outperform CULA, a parallel linear algebra library for GPUs by up to 17x for tall-skinny matrices and Intel's Math Kernel Library (MKL) by up to 12x.

We also discuss stationary video background subtraction as a motivating application. We apply a recent statistical approach, which requires many iterations of computing the singular value decomposition of a tall-skinny matrix. Using CAQR as a first step to getting the singular value decomposition, we are able to get the answer 3x faster than if we use a traditional bandwidth- bound GPU QR factorization tuned specifically for that matrix size, and 30x faster than if we use Intel's Math Kernel Library (MKL) singular value decomposition routine on a multicore CPU.

Overlapping Computation and Communication for Advection on Hybrid Parallel Computers

JB White III
Climate and Global Dynamics
National Center for Atmospheric Research
Boulder, Colorado
trey@ucar.edu
JJ Dongarra
Department of Electrical Engineering and Computer Science
University of Tennessee
Knoxville, Tennessee
dongarra@cs.utk.edu

Abstract

We describe computational experiments exploring the performance improvements from overlapping computation and communication on hybrid parallel computers. Our test case is explicit time integration of linear advection with constant uniform velocity in a three-dimensional periodic domain. The test systems include a Cray XT5, a Cray XE6, and two multicore Infiniband clusters with different generations of NVIDIA graphics processing units (GPUs). We describe results for Fortran implementations using various combinations of MPI, OpenMP, and CUDA, with and without overlap of computation and communication. We find that overlapping CPU computation, GPU computation, parallel communication, and CPU-GPU communication can provide performance improvements of more than a factor of two.

VisIO: Enabling Interactive Visualization of Ultra-Scale, Time Series Data via High-Bandwidth Distributed I/O Systems

Christopher Mitchell* , James Ahrens† , and Jun Wang*
*Department of Electrical Engineering & Computer Science,
University of Central Florida,
Orlando, Florida 32816-2450
†Computer, Computational, & Statistical Sciences,
Los Alamos National Laboratory, Los Alamos, New Mexico 87545
Email: {mitchell, jwang}@eecs.ucf.edu* , ahrens@lanl.gov†

Abstract

Petascale simulations compute at resolutions ranging into billions of cells and write terabytes of data for visualization and analysis. Interactive visualization of this time series is a desired step before starting a new run. The I/O subsystem and associated network often are a significant impediment to interactive visualization of time-varying data; as they are not configured or provisioned to provide necessary I/O read rates.

In this paper, we propose a new I/O library for visualization applications: VisIO. Visualization applications commonly use N-to-N reads within their parallel enabled readers which provides an incentive for a shared-nothing approach to I/O, similar to other data-intensive approaches such as Hadoop. However, unlike other data-intensive applications, visualization requires: (1) interactive performance for large data volumes, (2) compatibility with MPI and POSIX file system semantics for compatibility with existing infrastructure, and (3) use of existing file formats and their stipulated data partitioning rules.

VisIO, provides a mechanism for using a non-POSIX distributed file system to provide linear scaling of I/O bandwidth. In addition, we introduce a novel scheduling algorithm that helps to co-locate visualization processes on nodes with the requested data. Testing using VisIO integrated into ParaView was conducted using the Hadoop Distributed File System (HDFS) on TACC's Longhorn cluster. A representative dataset, VPIC, across 128 nodes showed a 64.4% read performance improvement compared to the provided Lustre installation. Also tested, was a dataset representing a global ocean salinity simulation that showed a 51.4% improvement in read performance over Lustre when using our VisIO system. VisIO, provides powerful high-performance I/O services to visualization applications, allowing for interactive performance with ultra-scale, time-series data.

Architectural constraints to attain 1 Exaflop/s for three scientific application classes

Abhinav Bhatele, Pritish Jetley, Hormozd Gahvari, Lukasz Wesolowski,
William D. Gropp, Laxmikant V. Kalé
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA
E-mail: {bhatele, pjetley2, gahvari, wesolwsk, wgropp, kale}@illinois.edu

Abstract

The first Teraflop/s computer, the ASCI Red, became operational in 1997, and it took more than 11 years for a Petaflop/s performance machine, the IBM Roadrunner, to appear on the Top500 list. Efforts have begun to study the hardware and software challenges for building an exascale machine. It is important to understand and meet these challenges in order to attain Exaflop/s performance. This paper presents a feasibility study of three important application classes to formulate the constraints that these classes will impose on the machine architecture for achieving a sustained performance of 1 Exaflop/s.

The application classes being considered in this paper are – classical molecular dynamics, cosmological simulations and unstructured grid computations (finite element solvers). We analyze the problem sizes required for representative algorithms in each class to achieve 1 Exaflop/s and the hardware requirements in terms of the network and memory. Based on the analysis for achieving an Exaflop/s, we also discuss the performance of these algorithms for much smaller problem sizes.

Session 3
Hardware-Software Interaction

A Novel Power management for CMP Systems in Data-intensive Environment

Pengju Shang, Jun Wang
Computer Engineering and Computer Science Dept.
University of Central Florida
Orlando, USA.
Email: shang@eecs.ucf.edu, jwang@eecs.ucf.edu

Abstract

The emerging data-intensive applications of today are comprised of non-uniform CPU and I/O intensive workloads, thus imposing a requirement to consider both CPU and I/O effects in the power management strategies. Only scaling down the processor's frequency based on its busy/idle ratio cannot fully exploit opportunities of saving power. Our experiments show that besides the busy and idle status, each processor may also have I/O wait phases waiting for I/O operations to complete. During this period, the completion time is decided by the I/O subsystem rather than the CPU thus scaling the processor to a lower frequency will not affect the performance but save more power. In addition, the CPU's reaction to the I/O operations may be significantly affected by several factors, such as I/O type (sync or unsync), instruction/job level parallelism; it cannot be accurately modeled via physics laws like mechanical or chemical systems.

In this paper, we propose a novel power management scheme called MAR (modeless, adaptive, rule-based) in multiprocessor systems to minimize the CPU power consumption under performance constraints. By using richer feedback factors, e.g. the I/O wait, MAR is able to accurately describe the relationships among core frequencies, performance and power consumption.We adopt a modeless control model to reduce the complexity of system modeling. MAR is designed for CMP (Chip Multi Processor) systems by employing multi-input/multi-output (MIMO) theory and percore level DVFS (Dynamic Voltage and Frequency Scaling). Our extensive experiments on a physical test bed demonstrate that, for the SPEC benchmark and data-intensive (TPC-C) benchmark, the efficiency of MAR is 93.6-96.2% accurate to the ideal power saving strategy calculated off-line. Compared with baseline solutions, MAR could save 22.5-32.5% more power while keeping the comparable performance loss of about 1.8-2.9%. In addition, simulation results show the efficiency of our design for various CMP configurations.

Characterization of System Services and Their Performance Impact in Multi-core Nodes

Seetharami Seelam
Liana Fong
IBM T. J. Watson Research Center,
{sseelam,llfong}@us.ibm.com

John Lewars, John Divirgilio
Brian F. Veale, Kevin Gildea
IBM Systems and Technology Group,
{jlewars,jdivirgi,bfveale,gildeak}@us.ibm.com

Abstract

The performance of parallel applications on large scale systems is shown to disproportionately degrade due to interference from system services. This interference from system services is also known as jitter. However, there is limited understanding of sources and patterns of jitter on multi-core systems. In this paper, we identify and characterize jitter sources in terms of their amplitude and execution interval distributions on multi-core IBM Power systems with UNIX-based general purpose operating systems: AIX and Linux. Our analysis shows that there are various kinds of jitter sources and their execution varies drastically between different cores and between hardware threads within each core for practical reasons.

This in-depth knowledge of jitter events is leveraged to devise effective approaches to mitigate the jitter impact on application performance in large scale systems. Moreover, such knowledge would provide useful insights to a new generation of operating system designs such as multikernel or satellite kernel for multi-core systems.

Automatic Recognition of Performance Idioms in Scientific Applications

Jiahua He* , Allan E. Snavely* , Rob F. Van der Wijngaart† and Michael A. Frumkin‡
* San Diego Supercomputer Center (SDSC), University of California, San Diego
† Software Solutions Group, Intel Corporation
‡ Google Corporation
Emails: jiahua@gmail.com, allans@sdsc.edu, rob.f.van.der.wijngaart@intel.com, frumkin@google.com

Abstract

Basic data flow patterns that we call **performance idioms**, such as stream, transpose, reduction, random access and stencil, are common in scientific numerical applications. We hypothesize that a small number of idioms can cover most programming constructs that dominate the execution time of scientific codes and can be used to approximate the application performance. To check these hypotheses, we proposed an automatic idioms recognition method and implemented the method, based on the open source compiler Open64. With the NAS Parallel Benchmark (NPB) as a case study, the prototype system is about 90% accurate compared with idiom classification by a human expert. Our results showed that the above five idioms suffice to cover 100% of the six NPB codes (MG, CG, FT, BT, SP and LU). We also compared the performance of our idiom benchmarks with their corresponding instances in the NPB codes on two different platforms with different methods. The approximation accuracy is up to 96.6%. The contribution is to show that a small set of idioms can cover more complex codes, that idioms can be recognized automatically, and that suitably defined idioms may approximate application performance.

Iso-energy-efficiency: An approach to power-constrained parallel computation

Shuaiwen Song[1], Chun-Yi Su[1], Rong Ge[2], Abhinav Vishnu[3], and Kirk W. Cameron[1]
[1]SCAPE Laboratory, Virginia Tech [2]Marquette University [3]Pacific Northwest National Laboratory
s562673@vt.edu, sonicat@vt.edu, rong.ge@marquette.edu, Abhinav.Vishnu@pnl.gov, cameron@vt.edu

Abstract

Future large scale high performance supercomputer systems require high energy efficiency to achieve exaflops computational power and beyond. Despite the need to understand energy efficiency in high-performance systems, there are few techniques to evaluate energy efficiency at scale. In this paper, we propose a system-level iso-energy-efficiency model to analyze, evaluate and predict energy-performance of data intensive parallel applications with various execution patterns running on large scale power-aware clusters. Our analytical model can help users explore the effects of machine and application dependent characteristics on system energy efficiency and isolate efficient ways to scale system parameters (e.g. processor count, CPU power/frequency, workload size and network bandwidth) to balance energy use and performance. We derive our iso-energy-efficiency model and apply it to the NAS Parallel Benchmarks on two power-aware clusters. Our results indicate that the model accurately predicts total system energy consumption within 5% error on average for parallel applications with various execution and communication patterns. We demonstrate effective use of the model for various application contexts and in scalability decision-making.

Session 4
Runtime Systems

A Study of Speculative Distributed Scheduling on the Cell/B.E.

Pieter Bellens
Barcelona Supercomputing Center, Spain

Josep M. Perez
Barcelona Supercomputing Center, Spain

Rosa M. Badia
Barcelona Supercomputing Center, Spain
Intelligence Research Institute (IIIA), Spain
Spanish National Research Council (CSIC), Spain

Jesus Labarta
Barcelona Supercomputing Center, Spain
Universitat Politecnica de Catalunya, Spain

Abstract

Star Superscalar's (StarSs) programming model converts a sequential application in C or Fortran into an efficient parallel program. The resulting parallel code is highly dynamic in the sense that data analysis and task scheduling occur at runtime, while the application executes. In this paper we compare this approach to the strategy adopted by other multi-core programming environments. The prize to pay for dynamic scheduling and dependence tracking is higher runtime overhead. We propose a distributed scheduler for Task Dependence Graphs (TDGs) to attenuate the scheduling cost in heterogeneous multi-core architectures. This scheduler allows the cores to speculatively select tasks from a conservative estimate of the TDG.In case of conflicts or lack of tasks a lightweight centralized scheduler services the faulting core after which the latter resumes its participation in the distributed scheme. Experiments with Cell Superscalar (CellSs) on a representative set of benchmarks demonstrate the reduction in runtime overhead achieved by the distributed scheduler. This reduction in runtime overhead carries over directly to a performance improvement for a large fraction of the benchmarks.

Exploiting Data Similarity to Reduce Memory Footprints

Susmit Biswas, Bronis R. de Supinski, Martin Schulz
Lawrence Livermore National Laboratory
Livermore, CA - 94550, USA
Email: {biswas3, bronis, schulzm}@llnl.gov
Diana Franklin, Timothy Sherwood, Frederic T. Chong
Department of Computer Science
University of California, Santa Barbara, USA
Email: {franklin, sherwood, chong}@cs.ucsb.edu

Abstract

Memory size has long limited large-scale applications on high-performance computing (HPC) systems. Since compute nodes frequently do not have swap space, physical memory often limits problem sizes. Increasing core counts per chip and power density constraints, which limit the number of DIMMs per node, have exacerbated this problem. Further, DRAM constitutes a significant portion of overall HPC system cost. Therefore, instead of adding more DRAM to the nodes, mechanisms to manage memory usage more efficiently —preferably transparently— could increase effective DRAM capacity and thus the benefit of multicore nodes for HPC systems.

MPI application processes often exhibit significant data similarity. These data regions occupy multiple physical locations across the individual rank processes within a multicore node and thus offer a potential savings in memory capacity. These regions, primarily residing in heap, are dynamic, which makes them difficult to manage statically.

Our novel memory allocation library, *SBLLmallocShort*, automatically identifies identical memory blocks and merges them into a single copy. Our implementation is transparent to the application and does not require any kernel modifications. Overall, we demonstrate that *SBLLmalloc* reduces the memory footprint of a range of MPI applications by 32.03% on average and up to 60.87%. Further, *SBLLmalloc* supports problem sizes for IRS over 21.36% larger than using standard memory management techniques, thus significantly increasing effective system size. Similarly, *SBLLmalloc* requires 43.75% fewer nodes than standard memory management techniques to solve an AMG problem.

The Evaluation of an Effective Out-of-core Run-Time System in the Context of Parallel Mesh Generation

Andriy Kot
Computer Science Department
The College of William and Mary
Williamsburg, VA
kot@cs.wm.edu

Andrey N. Chernikov, Nikos P. Chrisochoides
Computer Science Department
Old Dominion University
Norfolk, VA
{achernik, nikos}@cs.odu.edu

Abstract

We present an out-of-core run-time system that supports effective parallel computation of large irregular and adaptive problems, in particular unstructured mesh generation (PUMG). PUMG is a highly challenging application due to intensive memory accesses, unpredictable communication patterns, and variable and irregular data dependencies reflecting the unstructured spatial connectivity of mesh elements.

Our runtime system allows to transform the footprint of parallel applications from wide and shallow into narrow and deep by extending the memory utilization to the out-of-core level. It simplifies and streamlines the development of otherwise highly time consuming out-of-core applications as well as the converting of existing applications. It utilizes disk, network and memory hierarchy to achieve high utilization of computing resources without sacrificing performance with PUMG. The runtime system combines different programming paradigms: multi-threading within the nodes using industrial strength software framework, one-sided active messages among the nodes, and an out-of-core subsystem for managing large datasets.

We performed an evaluation on traditional parallel platforms to stress test all layers of the run-time system using three different PUMG methods with significantly varying communication and synchronization patterns. We demonstrated high overlap in computation, communication, and disk I/O which results in good performance when computing large out-of-core problems. The runtime system adds very small overhead (up to 18% on most configurations) when computing in-core which means performance is not compromised.

Enriching 3-D Video Games on Multicores

Romain Cledat* , Tushar Kumar † , Jaswanth Sreeram* and Santosh Pande*
* School of Computer Science, College of Computing † School of Electrical and Computer Engineering
Georgia Institute of Technology, Atlanta, GA
romain@gatech.edu, tusharkumar@gatech.edu, jaswanth@cc.gatech.edu, santosh@cc.gatech.edu

Abstract

The introduction of multicore processors on desktops and other personal computing platforms has given rise to multiple interesting end-user application possibilities. One important trend is the increased presence of resource hungry applications like gaming and multimedia applications. One of the key distinguishing factors of these applications is that they are amenable to variable semantics (ie, multiple possibilities of results) unlike traditional applications wherein a fixed, unique answer is expected. For example, varying degrees of image processing improves picture quality; different model complexities used in game physics allow different degrees of realism during game play, and so on. The goal of this paper is to demonstrate that scalable semantics in applications such as video games can be enriched with optional tasks that can be launched and thus adapt to the amount of available resources at runtime. We propose a C/C++ API that allows the programmer to define how the current semantics of a program can be opportunistically enriched, as well as the underlying runtime system that orchestrates the different computations We show how this infrastructure can be used to enrich a well known game called Quake 3. Our results show that it is possible to perform significant enrichment without degrading the application's performance by utilizing additional cores.

Session 5
Routing and Communication

On Nonblocking Folded-Clos Networks in Computer Communication Environments

Xin Yuan
Department of Computer Science, Florida State University
Tallahassee, FL 32306, USA
xyuan@cs.fsu.edu

Abstract

Folded-Clos networks, also referred to as fat-trees, have been widely used as interconnects in large scale high performance computing clusters. The switching capability of such interconnects in computer communication environments, however, is not well understood. In particular, the concept of nonblocking interconnects, which is often used by system vendors, has only been studied in the telephone communication environment with the assumption of a centralized controller. Such "nonblocking"networks do not support nonblocking communications in computer communication environments where the network control is distributed. This paper theoretically analyzes the conditions for folded-Clos networks to achieve nonblocking communications in computer communication environments with various routing schemes including deterministic routing and adaptive routing, and establishes nonblocking conditions.

vFtree - A Fat-tree Routing Algorithm using Virtual Lanes to Alleviate Congestion

Wei Lin Guay, Bartosz Bogdanski, Sven-Arne Reinemo, Olav Lysne, Tor Skeie
Simula Research Laboratory
P.O. Box 134,
NO-1325, Lysaker, Norway
E-mail:{weilin, bartoszb, svenar, olavly, tskeie}@simula.no

Abstract

It is a well known fact that multiple virtual lanes can improve performance in interconnection networks, but this knowledge has had little impact on real clusters. Currently, a large number of clusters using InfiniBand is based on fat-tree topologies that can be routed deadlock-free using only one virtual lane. Consequently, all the remaining virtual lanes are left unused.

In this paper we suggest an enhancement to the fat-tree algorithm that utilizes virtual lanes to improve performance when hot-spots are present. Even though the bisection bandwidth in a fat-tree is constant, hot-spots are still possible and they will degrade performance for flows not contributing to them due to head-of-line blocking. Such a situation may be alleviated through adaptive routing or congestion control, however, these methods are not yet readily available in InfiniBand technology. To remedy this problem, we have implemented an enhanced fat-tree algorithm in OpenSM that distributes traffic across all available virtual lanes without any configuration needed. We evaluated the performance of the algorithm on a small cluster and did a large-scale evaluation through simulations. In a congested environment, results show that we are able to achieve throughput increases up to 38% on a small cluster and from 221% to 757% depending on the hot-spot scenario for a 648-port simulated cluster.

Measuring Temporal Lags in Delay-Tolerant Networks

Arnaud Casteigts*, Paola Flocchini*, Bernard Mans† and Nicola Santoro‡
*University of Ottawa, Canada,
{casteig,flocchin}@site.uottawa.ca
†Macquarie University, Sydney, Australia,
bernard.mans@mq.edu.au
‡Carleton University, Ottawa, Canada,
santoro@scs.carleton.ca

Abstract

Delay-tolerant networks (DTNs) are characterized by a possible absence of end-to-end communication routes at any instant. In most cases, however, a form of connectivity can be established over time and space. This particularity leads to consider the relevance of a given route not only in terms of hops (topological length), but also in terms of time (temporal length). The problem of measuring temporal distances between individuals in a social network was recently addressed, based on a posteriori analysis of interaction traces. This paper focuses on the distributed version of this problem, asking whether every node in a network can know precisely and in real time how out-of-date it is with respect to every other. Answering affirmatively is simple when contacts between the nodes are punctual, using the temporal adaptation of vector clocks provided in (Kossinets et al., 2008). It becomes more difficult when contacts have a duration and can overlap in time with each other. We demonstrate that the problem remains solvable with arbitrarily long contacts and non-instantaneous (though invariant and known) propagation delays on edges. This is done constructively by extending the temporal adaptation of vector clocks to non-punctual causality. The second part of the paper discusses how the knowledge of temporal lags could be used as a building block to solve more concrete problems, such as the construction of foremost broadcast trees or network backbones in periodically-varying DTNs.

Session 6
Self Stabilization and Security

A Lightweight Method for Automated Design of Convergence

Ali Ebnenasir
Computer Science Department
Michigan Technological University
Houghton MI 49931, USA
Email: aebnenas@mtu.edu

Aly Farahat
Computer Science Department
Michigan Technological University
Houghton MI 49931, USA
Email: anfaraha@mtu.edu

Abstract

Design and verification of Self-Stabilizing (SS) network protocols are difficult tasks in part because of the requirement that a SS protocol must recover to a set of legitimate states from *any* state in its state space (when perturbed by transient faults). Moreover, distribution issues exacerbate the design complexity of SS protocols as processes should take local actions that result in global recovery/convergence of a network protocol. As such, most existing design techniques focus on protocols that are locally-correctable. To facilitate the design of finite-state SS protocols (that may not necessarily be locally-correctable), this paper presents a lightweight formal method supported by a software tool that automatically adds convergence to non-stabilizing protocols. We have used our method/tool to automatically generate several SS protocols with up to 40 processes (and 3^{40} states) in a few minutes on a regular PC. Surprisingly, our tool has automatically synthesized both protocols that are the same as their manually-designed versions as well as new solutions for well-known problems in the literature (e.g., Dijkstra's token ring [**?**]). Moreover, the proposed method has helped us reveal flaws in a manually designed SS protocol.

Snap-Stabilizing Committee Coordination

Borzoo Bonakdarpour* , Stéphane Devismes† , Franck Petit‡
* Department of Electrical and Computer Engineering
University of Waterloo
200 University Avenue West
Waterloo, Ontario, N2L 3G1, Canada
Email: borzoo@ece.uwaterloo.ca
† VERIMAG UMR 5104 Grenoble 1, France
Email: stephane.devismes@imag.fr
‡ LIP6, UPMC Paris 6, France
Email: franck.petit@lip6.fr

Abstract

In this paper, we propose two *snap-stabilizing* distributed algorithms for the *committee coordination problem*. In this problem, a committee consists of a set of processes and committee meetings are synchronized, so that each process participates in at most one committee meeting at a time. Snap-stabilization is a versatile technique allowing to design algorithms that efficiently tolerate transient faults. Indeed, after a finite number of such faults (*e.g.* memory corruptions, message losses, etc), a snap-stabilizing algorithm immediately operates correctly, without any external intervention.

We design snap-stabilizing committee coordination algorithms enriched with some desirable properties related to *concurrency*, *(weak) fairness*, and a stronger synchronization mechanism called *2-Phase Discussion Time*. From previous papers, we know that (1) in the general case, (weak) fairness cannot be achieved in the committee coordination, and (2) it becomes feasible provided that each process waits for meetings infinitely often. Nevertheless, we show that even under this latter assumption, it is impossible to implement a fair solution that allows *maximal concurrency*. Hence, we propose two orthogonal snap-stabilizing algorithms, each satisfying 2-phase discussion time, and either maximal concurrency or fairness. The algorithm implementing fairness requires that every process waits for meetings infinitely often. Moreover, for this algorithm, we introduce and evaluate a new efficiency criterion called the *degree of fair concurrency*. This criterion shows that even if it does not satisfy maximal concurrency, our snap-stabilizing fair algorithm still allows a high level of concurrency.

SC-OA: a Secure and Efficient Scheme for Origin Authentication of Interdomain Routing in Cloud Computing Networks

Zhongjian Le[*†] , Naixue Xiong[‡*¶], Bo Yang[*] and Yuezhi Zhou[§]
[*]School of Information Technology, Jiangxi University of Finance and Economics, Nanchang, China 330013
[†]School of Management, Huazhong University of Science and Technology, Wuhan, China 430074
[‡]Department of Computer Science, Georgia State University, Atlanta, GA, USA
[§] Department of Computer Sci. and Tech., Tsinghua University, Beijing 100084, China
[¶] Naixue Xiong is the corresponding author in this paper. Email: nxiong@gsu.edu

Abstract

IP prefix hijacking is one of the top threats in the cloud computing Internets. Based on cryptography, many schemes for preventing prefix hijacks have been proposed. Securing binding between IP prefix and its owner underlies these schemes. We believe that a scheme for securing this binding should try to satisfy these seven critical requirements: no key escrow, no other secure channel, defending against Malicious Key Issuer (MKI) in the phase of prefix announcement, defending against MKI in the phase of key issuing, no certificate, in-band delegation attestation, and in-band public key witness. In this paper, we propose a new scheme, Origin Authentication based on Self-Certified public keys (SC-OA), using self-certified public keys to authenticate origin autonomous systems. To the best of our knowledge, it is the first work for securing prefix ownership using self-certified public keys to achieve an efficient and secure scheme that satisfies all seven requirements. The analyses show that SC-OA can defend against regular prefix, subprefix, unassigned prefix, interception-based, and MKI hijacking, and improve performance in many aspects. It will be pushed ahead to practical deployment for preventing prefix hijacks.

Session 7
Numerical Algorithms

Automatic Library Generation for BLAS3 on GPUs

Huimin Cui*†, Lei Wang*†, Jingling Xue‡ , Yang Yang*† and Xiaobing Feng*
* Institute of Computing Technology, Chinese Academy of Sciences, China
† Graduate University, Chinese Academy of Sciences, China
‡ School of Computer Science and Engineering, University of New South Wales, Australia
{cuihm,wlei,yangyang,fxb}@ict.ac.cn jingling@cse.unsw.edu.au

Abstract

High-performance libraries, the performance-critical building blocks for high-level applications, will assume greater importance on modern processors as they become more complex and diverse. However, automatic library generators are still immature, forcing library developers to manually tune library to meet their performance objectives.

We are developing a new script-controlled compilation framework to help domain experts reduce much of the tedious and error-prone nature of manual tuning, by enabling them to leverage their expertise and reuse past optimization experiences. We focus on demonstrating improved performance and productivity obtained through using our framework to tune BLAS3 routines on three GPU platforms: up to 5.4x speedups over the CUBLAS achieved on NVIDIA GeForce 9800, 2.8x on GTX285, and 3.4x on Fermi Tesla C2050. Our results highlight the potential benefits of exploiting domain expertise and the relations between different routines (in terms of their algorithms and data structures).

Redesign of Higher-level Matrix Algorithms for Multicore and Distributed Architectures and Applications in Quantum Monte Carlo Simulation

Che-Rung Lee
Department of Computer Science
National TsingHua University
Hsinchu, Taiwan 30013
cherung@cs.nthu.edu.tw

Zhaojun Bai
Department of Computer Science
University of California
Davis, CA 95616, USA
bai@cs.ucdavis.edu

Abstract

A matrix operation is referred to as a hard-to-parallel matrix operation (HPMO) if it has serial bottlenecks that are hardly parallelizable. Otherwise, it is referred to as an easy-to-parallel matrix operation (EPMO). Empirical evidences showed the performance scalability of an HPMO is significantly poorer than an EPMO on multicore and distributed architectures. As the result, the design of higher-level algorithms for applications, for the performance considerations on multicore and distributed architectures, should avoid the use of HPMOs as the computational kernels.

In this paper, as a case study, we present an HPMO-avoiding algorithm for the Green's function calculation in quantum Monte Carlo simulation. The original algorithm utilizes the QR-decomposition with column pivoting (QRP) as its computational kernel. QRP is an HPMO. The redesigned algorithm maintains the same simulation stability but employs the standard QR decomposition without pivoting (QR), which is an EPMO. Different implementations of the redesigned algorithm on multicore and distributed architectures are investigated. Although some implementations of the redesigned method use about a factor of three more floating-point operations than the original algorithm, they are about 20% faster on a quadcore system and 2.5 times faster on a 1024-CPU massively parallel processing system. The broader impact of the redesign of higher-level matrix algorithms to avoid HPMOs in other computational science applications is also discussed.

Challenges of Scaling Algebraic Multigrid across Modern Multicore Architectures

Allison H. Baker, Todd Gamblin, Martin Schulz, and Ulrike Meier Yang
Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, 7000 East Avenue, Livermore CA 94550
{abaker,tgamblin,schulzm,umyang}@llnl.gov

Abstract

Algebraic multigrid (AMG) is a popular solver for large-scale scientific computing and an essential component of many simulation codes. AMG has shown to be extremely efficient on distributed-memory architectures. However, when executed on modern multicore architectures, we face new challenges that can significantly deteriorate AMG's performance. We examine its performance and scalability on three disparate multicore architectures: a cluster with four AMD Opteron Quad-core processors per node (Hera), a Cray XT5 with two AMD Opteron Hex-core processors per node (Jaguar), and an IBM BlueGene/P system with a single Quad-core processor (Intrepid). We discuss our experiences on these platforms and present results using both an MPI-only and a hybrid MPI/OpenMP model. We also discuss a set of techniques that helped to overcome the associated problems, including thread and process pinning and correct memory associations.

Session 8
Reliability and Security

Hauberk: Lightweight Silent Data Corruption Error Detector for GPGPU

Keun Soo Yim, Cuong Pham, Mushfiq Saleheen, Zbigniew Kalbarczyk, and Ravishankar Iyer
Center for Reliable and High Performance Computing, Coordinated Science Laboratory
University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
{yim6, pham9, msalehe2, kalbarcz, rkiyer}@illinois.edu

Abstract

High performance and relatively low cost of GPU-based platforms provide an attractive alternative for general purpose high performance computing (HPC). However, the emerging HPC applications have usually stricter output correctness requirements than typical GPU applications (i.e., 3D graphics). This paper first analyzes the error resiliency of GPGPU platforms using a fault injection tool we have developed for commodity GPU devices. On average, 16-33% of injected faults cause silent data corruption (SDC) errors in the HPC programs executing on GPU. This SDC ratio is signifi-cantly higher than that measured in CPU programs (<2.3%). In order to tolerate SDC errors, customized error detectors are strategically placed in the source code of target GPU programs so as to minimize performance impact and error propagation and maximize recoverability. The presented Hauberk technique is deployed in seven HPC benchmark programs and evaluated using a fault injection. The results show a high average error detection coverage (~87%) with a small performance overhead (~15%).

A Performance and Area Efficient Architecture for Intrusion Detection Systems

Govind Sreekar Shenoy*, Jordi Tubella* and Antonio González*†
*Department of Computer Architecture, Universitat Politècnica de Catalunya, Barcelona, Spain.
† Intel Barcelona Research Center, Barcelona, Spain.
Email: {govind,jordit}@ac.upc.edu, antonio.gonzalez@intel.com

Abstract

Intrusion Detection Systems (IDS) have emerged as one of the most promising ways to secure systems in network. An IDS operates by scanning packet-data for known signatures and accordingly takes requisite action. However, scanning bytes in the packet payload and checking for more than 20,000 signatures becomes a computationally intensive task. Additionally, with signatures doubling almost every 30 months, this complexity will aggravate further.

IDS commonly uses the Aho-Corasick state machine based search to scan packets for signatures. However, the huge size of the state machine negatively impacts the performance and area efficiency of the underlying hardware. In this work, we propose novel mechanisms to compactly store the state machine thereby improving the area efficiency. We observe over 2X reduction in area for storing the state machine in comparison to BS-FSM [19]. We investigate various approaches to improve the performance efficiency. We pipeline the processing of consecutive bytes accessing the upper-most level, the frequently accessed level, of the state machine. In order to further enhance the performance efficiency, we use a dedicated hardware unit specifically tuned for traversal using our proposed storage mechanism. We observe that our proposed architecture outperforms BS-FSM based approaches [13, 14, 19].

Time-Ordered Event Traces: A New Debugging Primitive for Concurrency Bugs

Martin Dimitrov
Software and Services Group
Intel Corporation
Chandler, AZ
martin.p.dimitrov@intel.com

Huiyang Zhou
Department of Electrical and Computer Engineering
North Carolina State University
Raleigh, NC
hzhou@ncsu.edu

Abstract

Non-determinism makes concurrent bugs extremely difficult to reproduce and to debug. In this work, we propose a new debugging primitive to facilitate the debugging process by exposing this non-deterministic behavior to the programmer. The key idea is to generate a time-ordered trace of events such as function calls/returns and memory accesses across different threads. The architectural support for this primitive is lightweight, including a high-precision, frequency-invariant time-stamp counter and an event trace buffer in each processor core. The proposed primitive continuously records and timestamps the last N function calls/returns per core by default, and can also be configured to watch specific memory addresses or code regions through a flexible software interface. To examine the effectiveness of the proposed primitive, we studied a variety of concurrent bugs in large commercial software and our results show that exposing the time-ordered information, function calls/returns in particular, to the programmer is highly beneficial for diagnosing the root causes of these bugs.

Session 9
Wireless and Sensor Networks

Singlehop Collaborative Feedback Primitives for Threshold Querying in Wireless Sensor Networks

Murat Demirbas, Serafettin Tasci, Hanifi Gunes, Atri Rudra
Department of Computer Science and Engineering,
University at Buffalo, SUNY
Buffalo, NY, 14260 { demirbas | serafett | hanifigu | atri }@buffalo.edu

Abstract

In wireless sensor network (WSN) deployments, Receiver-side Collision Detection (RCD) has been proposed for speeding up collaborative feedback collection from a singlehop neighborhood. Using RCD, an initiator node can query the existence of a predicate P in its neighborhood in constant time by making all P-positive nodes answer simultaneously. Despite the collisions, the initiator is still able to infer useful information from a broadcast using RCD: an activity in the network means the predicate P holds for at least one node while silence indicates that P does not hold at any queried node in the network. In this study we investigate the threshold querying problem, where the initiator has to learn whether P holds in the network for at least threshold t number of nodes in singlehop of the initiator. To answer the threshold queries in an efficient fashion, we present a number of adaptive RCD-based querying mechanisms that dynamically re-groups the queried nodes in the network. We evaluate our algorithms on a real sensor network implementation and also carry out several simulations to contrast our approach with the traditional techniques. The experiments reveal that our algorithms achieve significant time improvements in threshold queries over traditional techniques.

Completely Distributed Particle Filters for Target Tracking in Sensor Networks

Bo Jiang, Binoy Ravindran
Department of Electrical and Computer Engineering
Virginia Tech, Blacksburg, VA 24061
Email: {bjiang,binoy}@vt.edu

Abstract

Particle filters (or PFs) are widely used for the tracking problem in dynamic systems. Despite their remarkable tracking performance and flexibility, PFs require intensive computation and communication, which are strictly constrained in wireless sensor networks (or WSNs). Thus, distributed particle filters (or DPFs) have been studied to distribute the computational workload onto multiple nodes while minimizing the communication among them. However, weight normalization and re-sampling in generic PFs cause significant challenges in the distributed implementation. Few existing efforts on DPF could be implemented in a completely distributed manner. In this paper, we design a completely distributed particle filter (or CDPF) for target tracking in sensor networks, and further improve it with neighborhood estimation toward minimizing the communication cost. First, we describe the particle maintenance and propagation mechanism, by which particles are maintained on different sensor nodes and propagated along the target trajectory. Then, we design the CDPF algorithm by adjusting the order of PFs' four steps and leveraging the data aggregation during particle propagation. Finally, we develop a neighborhood estimation method to replace the measurement broadcasting and the calculation of likelihood functions. With this approximate estimation, the communication cost of DPFs can be minimized. Our experimental evaluations show that although CDPF incurs about 50% more estimation error than semi-distributed particle filter (or SDPF), its communication cost is lower than that of SDPF by as much as 90%.

Connectivity Trade-offs in 3D Wireless Sensor Networks using Directional Antennae

Evangelos Kranakis
School of Computer Science Carleton University
Ottawa, Canada
Email: kranakis@scs.carleton.ca
Danny Krizanc
Department of Mathematics Wesleyan University
Middletown, USA
Email: dkrizanc@wesleyan.edu

Ashish Modi
Department of Electronics and Electrical Communication Engineering
IIT Kharagpur, India
Email: ashishmodi.kgp@gmail.com
Oscar Morales-Ponce
School of Computer Science Carleton University
Ottawa, Canada
Email: oscarmponce@gmail.com

Abstract

We consider a 3D antenna orientation problem for maintaining connectivity of a wireless network in 3D space using only directional antennae. Sensors are located at points in 3D space and are equipped with directional antennae. The *strong connectivity antenna orientation problem* is concerned with deciding whether or not for given solid angle Ω and range r it is possible to orient the antennae so as to ensure that the sensor network resulting from the induced transmissions is strongly connected. In this paper we 1) present an algorithm ensuring optimal antenna range for the case when $\Omega \geq \frac{18\pi}{5}$, 2) show that determining whether or not there exists a strong orientation of directional sensors of solid angle $\Omega < \pi - \epsilon$ having optimal range is NP-complete, for any $\epsilon > 0$, and 3) provide an algorithm for approximating the antennae range so as to ensure strong connectivity of the resulting graph, provided the solid angle of the antennae is $2\pi \leq \Omega < \frac{18\pi}{5}$. In addition, we study the effect of replacing omnidirectional antennae with directional antennae on the hop stretch factor of the resulting network of directional antennae and present some simulation results on the variation of hop stretch factor with different network sizes and solid angles of directional antennae. This is the first paper concerning the strong connectivity antennae orientation problem in 3D space.

Distributed Fine-grained Access Control in Wireless Sensor Networks

Sushmita Ruj, Amiya Nayak and Ivan Stojmenovic
SITE, University of Ottawa,
Ottawa K1N 6N5, Canada
Email: {sruj, anayak, ivan}@site.uottawa.ca

Abstract

In mission-critical activities, each user is allowed to access some specific, but not all, data gathered by wireless sensor networks. Yu et al [**?**] recently proposed a centralized fine grained data access control mechanism for sensor networks, which exploits a cryptographic primitive called attribute based encryption (ABE). There is only one trusted authority to distribute keys to the sensor nodes and the users. Compromising the single authority can undermine the whole network. We propose a fully distributed access control method, which has several authorities instead of one. Each sensor has a set of attributes and each user has an access structure of attributes. A message from a sensor is encrypted such that only a user with matching set of attributes can decrypt. Compared to [**?**], our schemes need simpler access structure which make secret key distribution more computation efficient, when user rights are modified. We prove that our scheme can tolerate compromising all but one distribution centers, which independently distribute their contributions to a single user key. Our scheme do not increase the computation and communication costs of the sensors, making it highly desirable for fine grained access control.

Session 10
GPU Acceleration

Design of MILC lattice QCD application for GPU clusters

Guochun Shi[†], Steven Gottlieb[‡,†], Aaron Torok[‡], Volodymyr Kindratenko[†]
[†]National Center for Supercomputing Applications (NCSA), University of Illinois, Urbana, IL, USA
[‡]Department of Physics, Indiana University, Bloomington, IN, USA
gshi@ncsa.illinois.edu, sg@indiana.edu, amtorok@indiana.edu, kindr@ncsa.illinois.edu

Abstract

We present an implementation of the improved staggered quark action lattice QCD computation designed for execution on a GPU cluster. The parallelization strategy is based on dividing the space-time lattice along the time dimension and distributing the sub-lattices among the GPU cluster nodes. We provide a mixed-precision floating-point GPU implementation of the multi-mass conjugate gradient solver. Our single GPU implementation of the conjugate gradient solver achieves a 9x performance improvement over the highly optimized code executed on a state-of-the-art eight-core CPU node. The overall application executes almost six times faster on a GPU-enabled cluster vs. a conventional multi-core cluster. The developed code is currently used for running production QCD calculations with electromagnetic corrections.

Multifrontal Factorization of Sparse SPD Matrices on GPUs

Thomas George[*], Vaibhav Saxena[*], Anshul Gupta[†], Amik Singh[‡] and Anamitra R. Choudhury[*]
[*]High Performance Computing Group
IBM Research India, New Delhi, India 110070
Email: thomasgeorge, vaibhavsaxena, anamchou@in.ibm.com
[†]Department of Mathematical Sciences
IBM T.J. Watson Research Center, Yorktown Heights, USA
Email: anshul@us.ibm.com
[‡]Department of Electronics and Computer Engineering, IIT Roorkee, India
Email: amiksuec@iitr.ernet.in

Abstract

Solving large sparse linear systems is often the most computationally intensive component of many scientific computing applications. In the past, sparse multifrontal direct factorization has been shown to scale to thousands of processors on dedicated supercomputers resulting in a substantial reduction in computational time. In recent years, an alternative computing paradigm based on GPUs has gained prominence, primarily due to its affordability, power-efficiency, and the potential to achieve significant speedup relative to desktop performance on regular and structured parallel applications. However, sparse matrix factorization on GPUs has not been explored sufficiently due to the complexity involved in an efficient implementation and concerns of low GPU utilization.

In this paper, we present an adaptive hybrid approach for accelerating sparse multifrontal factorization based on a judicious exploitation of the processing power of the host CPU and GPU. We present four different policies for distributing and scheduling the workload between the host CPU and the GPU, and propose a mechanism for a runtime selection of the appropriate policy for each step of sparse Cholesky factorization. This mechanism relies on auto-tuning based on modeling the best policy predictor as a parametric classifier. We estimate the classifier parameters from the available empirical computation time data such that the expected computation time is minimized. This approach is readily adaptable for using the current or an extended set of policies for different CPU-GPU combinations as well as for different combinations of dense kernels for both the CPU and the GPU.

Large-Scale Semantic Concept Detection on Manycore Platforms for Multimedia Mining

Mamadou Diao
School of ECE
Georgia Institute of Technology
Atlanta, GA, USA
diao@gatech.edu

Chrysostomos Nicopoulos
Department of ECE
University of Cyprus
Nicosia, CYPRUS
nicopoulos@ucy.ac.cy

Jongman Kim
School of ECE
Georgia Institute of Technology
Atlanta, GA, USA
jkim@ece.gatech.edu

Abstract

Media mining, the extraction of meaningful knowledge from multimedia content has become a major application and poses significant computational challenges in today's platforms. Media mining applications contain many sophisticated algorithms that include data-intensive analysis, classification, and learning. This paper explores the use of Graphics Processing Units (GPU) in media mining. We are particularly focused on large-scale semantic concept detection, a state-of-the-art approach that maps media content to hight-level semantic concepts, and a building block in many Media mining applications. We present a fast, parallel, large-scale, high-level semantic concept detector that leverages the GPU for image/video retrieval and content analysis. Through efficient data partitioning and movement, we parallelize feature extraction routines. By interleaving feature extraction routines of different types, we increase the computational intensity and mitigate the negative effects of histogram-like reduction operations. To cope with the very large number of semantic concepts, we propose a data layout of concept models on a multi-GPU hybrid architecture for high throughput semantic concept detection. We achieve one to two orders of magnitude speedups compared to serial implementations and our experiments show that we can detect 374 semantic concepts at a rate of over 100 frames/sec. This is over 100 times faster than a LibSVM-based semantic concept detection.

Efficient GPU implementation for Particle in Cell algorithm

Rejith George Joseph, Girish Ravunnikutty, and Sanjay Ranka
UFL
rjoseph, girishr, sranka@ufl.edu
Eduardo D'Azevedo and Scott Klasky
ORNL
dazevedoef, klasky@ornl.gov

Abstract

Particle in cell (PIC) algorithm is a widely used method in plasma physics to study the trajectories of charged particles under electromagnetic fields. The PIC algorithm is computationally intensive and its time requirements are proportional to the number of charged particles involved in the simulation. The focus of the paper is to parallelize the PIC algorithm on Graphics Processing Unit (GPU). We present several performance trade-offs related to small shared memory and atomic operations on the GPU to achieve high performance.

Session 11
Multiprocessing and Concurrency

Hardware-based Job Queue Management for Manycore Architectures and OpenMP Environments

Junghee Lee*, Chrysostomos Nicopoulos†, Yongjae Lee*, Hyung Gyu Lee* and Jongman Kim*
*School of Electrical and Computer Engineering, Georgia Institute of Technology,
Email: {junghee.lee, ylee308, hyunggyu, jkim}@gatech.edu
†Department of Electrical and Computer Engineering, University of Cyprus,
Email: nicopoulos@ucy.ac.cy

Abstract

The seemingly interminable dwindle of technology feature sizes well into the nano-scale regime has afforded computer architects with an abundance of computational resources on a single chip. The Chip Multi-Processor (CMP) paradigm is now seen as the de facto architecture for years to come. However, in order to efficiently exploit the increasing number of on-chip processing cores, it is imperative to achieve and maintain efficient utilization of the resources at run time. Uneven and skewed distribution of workloads misuses the CMP resources and may even lead to such undesired effects as traffic and temperature hotspots. While existing techniques rely mostly on software for the undertaking of load balancing duties and exploit hardware mainly for synchronization, we will demonstrate that there are wider opportunities for hardware support of load balancing in CMP systems. Based on this fact, this paper proposes IsoNet, a conflict-free dynamic load distribution engine that exploits hardware aggressively to reinforce massively parallel computation in manycore settings. Moreover, the proposed architecture provides extensive fault-tolerance against both CPU faults and intra-IsoNet faults. The hardware takes charge of both (1) the management of the list of jobs to be executed, and (2) the transfer of jobs between processing elements to maintain load balance. Experimental results show that, unlike the existing popular techniques of blocking and job stealing, IsoNet is scalable with as many as 1024 processing cores.

HK-NUCA: Boosting Data Searches in Dynamic Non-Uniform Cache Architectures for Chip Multiprocessors

Javier Lira
Dept. of Computer Architecture
Universitat Politècnica de Catalunya
08034 Barcelona, Spain
javier.lira@ac.upc.edu

Carlos Molina
Dept. of Computer Engineerging
Universitat Rovira i Virgili
43007 Tarragona, Spain
carlos.molina@urv.net

Antonio González
Intel Barcelona Research Center
Intel Labs - UPC
08034 Barcelona, Spain
antonio.gonzalez@intel.com

Abstract

The exponential increase in the cache sizes of multicore processors (CMPs) accompanied by growing on-chip wire delays make it difficult to implement traditional caches with single and uniform access latencies. Non-Uniform Cache Architecture (NUCA) designs have been proposed to address this problem. NUCA divides the whole cache memory into smaller banks and allows nearer cache banks to have lower access latencies than farther banks, thus mitigating the effects of the cache's internal wires. Traditionally, NUCA organizations have been classified as static (S-NUCA) and dynamic (D-NUCA). While in S-NUCA a data block is mapped to a unique bank in the NUCA cache, D-NUCA allows a data block to be mapped in multiple banks. Besides, D-NUCA designs are dynamic in the sense that data blocks may migrate towards the cores that access them most frequently. Recent works consider D-NUCA as a promising design, however, in order to obtain significant performance benefits, they used a non-affordable access scheme mechanism to find data in the NUCA cache. In this paper, we propose a novel and implementable data search algorithm for D-NUCA designs in CMP architectures, which is called HK-NUCA (*Home Knows where to find data within the NUCA cache*). It exploits migration features by providing fast and power efficient accesses to data which is located close to the requesting core. Moreover, HK-NUCA implements an efficient and cost-effective search mechanism to reduce miss latency and on-chip network contention. We show that using HK-NUCA as data search mechanism in a D-NUCA design reduces about 40% energy consumed per each memory request, and achieves an average performance improvement of 6%.

Power Token Balancing: Adapting CMPs to Power Constraints for Parallel Multithreaded Workloads

Juan M. Cebrián, Juan L. Aragón
Dept. of Computer Engineering,
University of Murcia, Murcia, Spain
{jcebrian,jlaragon}@ditec.um.es

Stefanos Kaxiras
Dept. of Information Technology,
University of Uppsala, Uppsala, Sweden
stefanos.kaxiras@it.uu.se

Abstract

In the recent years virtually all processor architectures employ multiple cores per chip (CMPs). It is possible to use legacy (i.e., single-core) power saving techniques in CMPs which run either sequential applications or independent multithreaded workloads. However, new challenges arise when running parallel shared-memory applications. In the later case, sacrificing some performance in a single core (thread) in order to be more energy-efficient might unintentionally delay the rest of cores (threads) due to synchronization points (locks/barriers), therefore, harming the performance of the whole application.

CMPs increasingly face thermal and power-related problems during their typical use. Such problems can be solved by setting a power budget to the processor/core. This paper initially studies the behavior of different techniques to match a predefined power budget in a CMP processor. While legacy techniques properly work for thread independent/multi-programmed workloads, parallel workloads exhibit the problem of independently adapting the power of each core in a thread dependent scenario. In order to solve this problem we propose a novel mechanism, Power Token Balancing (PTB), aimed at accurately matching an external power constraint by balancing the power consumed among the different cores using a power token-based approach while optimizing the energy efficiency. We can use power (seen as tokens or coupons) from non-critical threads for the benefit of critical threads. PTB runs transparent for thread independent / multiprogrammed workloads and can be also used as a spinlock detector based on power patterns. Results show that PTB matches more accurately a predefined power budget (total energy consumed over the budget is reduced to 8% for a 16-core CMP) than DVFS with only a 3% energy increase. Finally, we can trade accuracy on matching the power budget for energy-efficiency reducing the energy a 4% with a 20% of accuracy.

A Very Fast Simulator for Exploring the Many-Core Future

Olivier Certner
ST Microelectronics & INRIA Saclay
France

Zheng Li, Olivier Temam
INRIA Saclay
Orsay, France

Arun Raman
Princeton University
New Jersey, USA

Abstract

Although multi-core architectures with a large number of cores ("many-cores") are considered the future of computing systems, there are currently few practical tools to quickly explore both their design and general program scalability. In this paper, we present *SiMany*, a discrete-event-based many-core simulator able to support more than a thousand cores while being orders of magnitude faster than existing flexible approaches.

One of the difficult challenges for a reasonably realistic many-core simulation is to model faithfully the potentially high concurrency a program can exhibit. SiMany uses a novel virtual time synchronization technique, called *spatial synchronization*, to achieve this goal in a completely *local* and *distributed* fashion, which diminishes interactions and preserves locality. Compared to previous simulators, it raises the level of *abstraction* by focusing on modeling concurrent interactions between cores, which enables fast coarse comparisons of high-level architecture design choices and parallel programs performance. Sequential pieces of code are executed *natively* for maximal speed.

We exercise the simulator with a set of dwarf-like task-based benchmarks with dynamic control flow and irregular data structures. Scalability results are validated through comparison with a cycle-level simulator up to 64 cores. They are also shown consistent with well-known benchmark characteristics. We finally demonstrate how SiMany can be used to efficiently compare the benchmarks' behavior over a wide range of architectural organizations, such as polymorphic architectures and network of clusters.

Session 12
Compilers

Variable Granularity Access Tracking Scheme for Improving the Performance of Software Transactional Memory

Sandya S.Mannarswamy
CSA, IISc and Hewlett Packard
Bangalore,India
sandya@hp.com

Ramaswamy Govindarajan
SERC, Indian Institute of Science,
Bangalore, India
govind@serc.iisc.ernet.in

Abstract

Software transactional memory (STM) has been proposed as a promising programming paradigm for shared memory multi-threaded programs as an alternative to conventional lock based synchronization primitives. Typical STM implementations employ a conflict detection scheme, which works with uniform access granularity, tracking shared data accesses either at word/cache line or at object level. It is well known that a single fixed access tracking granularity cannot meet the conflicting goals of reducing false conflicts without impacting concurrency adversely. A fine grained granularity while improving concurrency can have an adverse impact on performance due to lock aliasing, lock validation overheads, and additional cache pressure. On the other hand, a coarse grained granularity can impact performance due to reduced concurrency. Thus, in general, a fixed or uniform granularity access tracking (UGAT) scheme is application-unaware and rarely matches the access patterns of individual application or parts of an application, leading to sub-optimal performance for different parts of the application(s). In order to mitigate the disadvantages associated with UGAT scheme, we propose a Variable Granularity Access Tracking (VGAT) scheme in this paper. We propose a compiler based approach wherein the compiler uses inter-procedural whole program static analysis to select the access tracking granularity for different shared data structures of the application based on the application's data access pattern. We describe our prototype VGAT scheme, using TL2 as our STM implementation. Our experimental results reveal that VGAT-STM scheme can improve the application performance of STAMP benchmarks from 1.87% to up to 21.2%.

Automated architecture-aware mapping of streaming applications onto GPUs

Andrei Hagiescu* Huynh Phung Huynh† Weng-Fai Wong* Rick Siow Mong Goh†
*School of Computing, National University of Singapore
† A*STAR Institute of High Performance Computing, Singapore
{hagiescu,wongwf}@comp.nus.edu.sg, {huynhph,gohsm}@ihpc.a-star.edu.sg

Abstract

Graphic Processing Units (GPUs) are made up of many streaming multiprocessors, each consisting of processing cores that interleave the execution of a large number of threads. Groups of threads - called *warps* and *wavefronts*, respectively, in nVidia and AMD literature - are selected by the hardware scheduler and executed in lockstep on the available cores. If threads in such a group access the slow off-chip global memory, the entire group has to be stalled, and another group is scheduled instead. The utilization of a given multiprocessor will remain high if there is a sufficient number of alternative thread groups to select from. Many parallel general purpose applications have been efficiently mapped to GPUs. Unfortunately, many stream processing applications exhibit unfavorable data movement patterns and low computation-to-communication ratio that may lead to poor performance. In this paper, we describe an automated compilation flow that maps most stream processing applications onto GPUs by taking into consideration two important architectural features of nVidia GPUs, namely interleaved execution as well as the small amount of shared memory available in each streaming multiprocessors. In particular, we show that using a small number of compute threads such that the memory footprint is reduced, we can achieve high utilization of the GPU cores. Our scheme goes against the conventional wisdom of GPU programming which is to use a large number of homogeneous threads. Instead, it uses a mix of *compute* and *memory access* threads, together with a carefully crafted schedule that exploits parallelism in the streaming application, while maximizing the effectiveness of the unique memory hierarchy. % small on-chip memory located within each streaming multiprocessor. We have implemented our scheme in the compiler of the StreamIt programming language, and our results show a significant speedup compared to the state-of-the-art solutions.

Automatic Loop Tiling for Direct Memory Access

Haibo Lin*, Tao Liu*, Lakshminarayanan Renganarayana‡, Huoding Li†,
Tong Chen‡, Kevin O'Brien‡, Ling Shao*
* IBM Research - China, {linhb, liuttao, shaol}@cn.ibm.com
† IBM GCG Systems and Technology Lab, lihuod@cn.ibm.com
‡ IBM T. J. Watson Research Center, {lrengan, chentong, caomhin}@us.ibm.com

Abstract

In heterogeneous multi-core systems, such as the Cell BE processor, each accelerator core has its own fast local memory without hardware supported coherence and the software is responsible to dynamically transfer data between the fast local and slow global memory. The data can be transferred through either a software controlled cache or a direct buffer. The software controlled cache maintains correctness for arbitrary access patterns, but introduces the extra overhead of cache lookup. Direct buffer is efficient for regular accesses, while requiring precise analysis, detailed modeling of execution, and significant code generation. In this paper we present the design and implementation of *DMATiler* which combines compiler analysis and runtime management to optimize local memory performance via automatic loop tiling and buffer optimization techniques.

The DMATiler chooses a data transfer friendly loop order and using a empirically validated DMA performance model, it formulates and solves a convex optimization problem to determine globally optimal tile sizes. Further, the DMATiler applies optimization techniques such compressed data transfers and DMA commands to achieve the best DMA performance for a given loop nest. We have implemented the DMATiler in the IBM XL Single Source Compiler (SSC), and have conducted experiments with a set of loop nest benchmarks. The results show that the DMATiler is much more efficient than software controlled cache (average speedup of 9.8x) and single level loop blocking (average speedup of 6.2x) on the Cell BE processor.

Tolerant Value Speculation in Coarse-Grain Streaming Computations

Nathaniel Azuelos, Idit Keidar
Technion - Israel Institute of Technology
Electrical Engineering Faculty
Technion City, Haifa, Israel
Email: nazuel@tx.technion.ac.il, idish@ee.technion.ac.il
Ayal Zaks
IBM Research Haifa, and
Technion - Israel Institute of Technology
Haifa, Israel
Email: zaks@il.ibm.com

Abstract

Streaming applications are the subject of growing interest, as the need for fast access to data continues to grow. In this work, we present the design requirements and implementation of coarse-grain value speculation in streaming applications. We explain how this technique can be useful in cases where serial parts of applications constitute bottlenecks, and when slower I/O favors using available prefixes of the data. Contrary to previous work, we show how allowing some tolerance can justify early predictions on a scale of a large window of values. We suggest a methodology for runtime support of speculation, along with the mechanisms required for rollback. We present resource management issues consequent to our technique. We study how validation and speculation frequencies impact the performance of the program. Finally, we present our implementation in the context of the Huffman encoder benchmark, running it in different configurations and on different architectures.

Session 13
Distributed Algorithms and Models

Adding a referee to an interconnection network: What can(not) be computed in one round.

Florent Becker[1], Martin Matamala[2,3], Nicolas Nisse[4], Ivan Rapaport[2,3], Karol Suchan[5,6] and Ioan Todinca[1]
[1] LIFO, Université d'Orléans, France
[2] Departamento de Ingeniería Matemática, Universidad de Chile, Santiago, Chile
[3] Centro de Modelamiento Matemático (UMI 2807 CNRS), Univ. de Chile, Santiago, Chile
[4] MASCOTTE, INRIA, I3S (CNRS/Univ. Nice Sophia Antipolis) France
[5] Facultad de Ingeniería y Ciencias, Universidad Adolfo Ibáñez, Santiago, Chile
[6] Faculty of Applied Mathematics, AGH - Univ. of Science and Technology, Cracow, Poland

Abstract

In this paper we ask which properties of a distributed network can be computed from a few amount of local information provided by its nodes. The distributed model we consider is a restriction of the classical $\mathcal{CONGEST}$ (distributed) model and it is close to the simultaneous messages (communication complexity) model defined by Babai, Kimmel and Lokam. More precisely, each of these n nodes -which only knows its own ID and the IDs of its neighbors- is allowed to send a message of $O(\log n)$ bits to some central entity, called the referee. Is it possible for the referee to decide some basic structural properties of the network topology G? We show that simple questions like, "does G contain a square?", "does G contain a triangle?" or "Is the diameter of G at most 3?" cannot be solved in general. On the other hand, the referee can decode the messages in order to have full knowledge of G when G belongs to many graph classes such as planar graphs, bounded treewidth graphs and, more generally, bounded degeneracy graphs. We leave open questions related to the connectivity of arbitrary graphs.

Improved Algorithms for the Distributed Trigger Counting Problem

Venkatesan T. Chakaravarthy, Anamitra R. Choudhury and Yogish Sabharwal
IBM Research India, New Delhi, India 110070
Email: vechakra, anamchou, ysabharwal@in.ibm.com

Abstract

Consider a distributed system with n processors, in which each processor receives some triggers from an external source. The distributed trigger counting (DTC) problem is to raise an alert and report to a user when the number of triggers received by the system reaches w, where w is a user-specified input. The problem has applications in monitoring, global snapshots, synchronizers and other distributed settings.

In this paper, we present two decentralized and randomized algorithms for the DTC problem.

The first algorithm has message complexity O(n log w) and no processor receives more than O(log w) messages with high probability. It does not provide any bound on the messages sent per processor. This algorithm assumes complete connectivity between the processors.

The second algorithm has message complexity O(n log n log w) and no processor exchanges more than O(log n log w) messages with high probability. However, there is a negligible failure probability in raising the alert on receiving w triggers. This algorithm only requires that a constant degree tree be embeddable in the underlying communication graph.

The Weighted Byzantine Agreement Problem

Vijay K. Garg and John Bridgman
Department of Electrical and Computer Engineering
The University of Texas at Austin
Austin, TX 78712-1084, USA
Email: garg@ece.utexas.edu, johnfb@mail.utexas.edu

Abstract

This paper presents a weighted version of the Byzantine Agreement Problem and its solution under various conditions. In this version, each machine is assigned a weight depending on the application. Instead of assuming that at most f out of N machines fail, the algorithm assumes that the total weight of the machines that fail is at most f/N. When each machine has weight 1/N, this problem reduces to the standard Byzantine Generals Agreement Problem. By choosing weights appropriately, the weighted Byzantine Agreement Problem can be applied to situations where a subset of processes are more trusted. By using weights, the system can reach consensus in the presence of Byzantine failures, even when more than N/3 processes fail, so long as the total weight of the failed processes is less than 1/3. Also, a method to update the weights of the processes after execution of the weighted Byzantine Agreement is given. The update method guarantees that the weight of any correct process is never reduced and the weight of any faulty process, suspected by correct processes whose total weight is at least 1/4, is reduced to 0 for future instances. A short discussion of some weight assignment strategies is also given.

Leveraging Social Networks to Combat Collusion in Reputation Systems for Peer-to-Peer Networks

Ze Li, Haiying Shen and Karan Sapra
Department of Electrical and Computer Engineering
Clemson University, Clemson, SC, 29631
Email: {zel, shenh, ksapra}@clemson.edu

Abstract

In peer-to-peer networks (P2Ps), many autonomous peers without preexisting trust relationships share resources with each other. Due to their open environment, the P2Ps usually employ reputation systems to provide guidance in selecting trustworthy resource providers for high reliability and security. However, node collusion impairs the effectiveness of reputation systems in trustworthy node selection. Although some reputation systems have certain mechanisms to counter collusion, the effectiveness of the mechanisms is not sufficiently high. In this paper, we leverage social networks to enhance the capability of reputation systems in combating collusion. We first analyzed real trace of the reputation system in the Overstock online auction platform which incorporates a social network. The analysis reveals the important impact of the social network on user purchasing and reputation rating patterns. We thus identified suspicious collusion behavior patterns and propose a social network based mechanism, namely SocialTrust, to counter collusion. SocialTrust adaptively adjusts the weight of ratings based on the social distance and interest relationship between peers. Experimental results show that SocialTrust can significantly strengthen the capability of current reputation systems in combating collusion.

Session 14
Parallel Graph and Particle Algorithms

Computing Strongly Connected Components in Parallel on CUDA

Jiří Barnat, Petr Bauch, Luboš Brim, and Milan Češka
Faculty of Informatics, Masaryk University, Botanická 68a, 60200 Brno, Czech Republic

Abstract

The problem of decomposing a directed graph into its strongly connected components is a fundamental graph problem inherently present in many scientific and commercial applications. In this paper we show how some of the existing parallel algorithms can be reformulated in order to be accelerated by NVIDIA CUDA technology. In particular, we design a new CUDA-aware procedure for pivot selection and we adapt selected parallel algorithms for CUDA accelerated computation. We also experimentally demonstrate that with a single GTX 480 GPU card we can easily outperform the optimal serial CPU implementation by an order of magnitude in most cases, 40 times on some sufficiently big instances. This is an interesting result as unlike the serial CPU case, the asymptotic complexity of the parallel algorithms is not optimal.

On optimal tree traversals for sparse matrix factorization

Mathias Jacquelin, Loris Marchal, Yves Robert and Bora Ucar
LIP, Ecole Normale Supérieure de Lyon
46 Allée d'Italie, 69364 Lyon Cedex 07, France
Email: {Mathias.Jacquelin|Loris.Marchal|Yves.Robert|Bora.Ucar}@ens-lyon.fr

Abstract

We study the complexity of traversing tree-shaped workflows whose tasks require large I/O files. Such workflows typically arise in the multifrontal method of sparse matrix factorization. We target a classical two-level memory system, where the main memory is faster but smaller than the secondary memory. A task in the workflow can be processed if all its predecessors have been processed, and if its input and output files fit in the currently available main memory. The amount of available memory at a given time depends upon the ordering in which the tasks are executed. What is the minimum amount of main memory, over all postorder schemes, or over all possible traversals, that is needed for an in-core execution? We establish several complexity results that answer these questions. We propose a new, polynomial time, exact algorithm which runs faster than a reference algorithm. Next, we address the setting where the required memory renders a pure in-core solution unfeasible. In this setting, we ask the following question: what is the minimum amount of I/O that must be performed between the main memory and the secondary memory? We show that this latter problem is NP-hard, and propose efficient heuristics. All algorithms and heuristics are thoroughly evaluated on assembly trees arising in the context of sparse matrix factorizations.

Fast Community Detection Algorithm With GPUs and Multicore Architectures

Jyothish Soman
IBM Research India
New Delhi, India
jyosoman@in.ibm.com

Ankur Narang
IBM Research India
New Delhi, India
annarang@in.ibm.com

Abstract

In this paper, we present the design of a novel scalable parallel algorithm for community detection optimized for multi-core and GPU architectures. Our algorithm is based on label propagation, which works solely on local information, thus giving it the scalability advantage over conventional approaches. We also show that weighted label propagation can overcome typical quality issues in communities detected with label propagation. Experimental results on well known massive scale graphs such as Wikipedia (100M edges) and also on RMAT graphs with 10M - 40M edges, demonstrate the superior performance and scalability of our algorithm compared to the well known approaches for community detection. On the *hep-th* graph (352K edges) and the *wikipedia* graph (100M edges), using Power 6 architecture with 32 cores, our algorithm achieves one to two orders of magnitude better performance compared to the best known prior results on parallel architectures with similar number of CPUs. Further, our GPGPU based algorithm achieves 8× improvement over the Power 6 performance on 40M edge R-MAT graph. Alongside, we achieve high quality (modularity) of communities detected, with experimental evidence from well-known graphs such as Zachary karate club, Dolphin network and Football club, where we achieve modularity that is close to the best known alternatives. To the best of our knowledge these are best known results for community detection on massive graphs (100M edges) in terms of performance and also quality vs. performance trade-off. This is also a unique work on community detection on GPGPUs with scalable performance.

A Study of Parallel Particle Tracing for Steady-State and Time-Varying Flow Fields

Tom Peterka
Robert Ross
Argonne National Laboratory
Argonne, IL, USA
tpeterka@mcs.anl.gov
rross@mcs.anl.gov

Boonthanome Nouanesengsy
Teng-Yok Lee
Han-Wei Shen
The Ohio State University
Columbus, OH, USA
nouanese@cse.ohio-state.edu
leeten@cse.ohio-state.edu
hwshen@cse.ohio-state.edu

Wesley Kendall
Jian Huang
University of Tennessee at Knoxville
Knoxville, TN, USA
kendall@eecs.utk.edu
huangj@eecs.utk.edu

Abstract

Particle tracing for streamline and pathline generation is a common method of visualizing vector fields in scientific data, but it is difficult to parallelize efficiently because of demanding and widely varying computational and communication loads. In this paper we scale parallel particle tracing for visualizing steady and unsteady flow fields well beyond previously published results. We configure the 4D domain decomposition into spatial and temporal blocks that combine in-core and out-of-core execution in a flexible way that favors faster run time or smaller memory. We also compare static and dynamic partitioning approaches. Strong and weak scaling curves are presented for tests conducted on an IBM Blue Gene/P machine at up to 32 K processes using a parallel flow visualization library that we are developing. Datasets are derived from computational fluid dynamics simulations of thermal hydraulics, liquid mixing, and combustion.

Session 15
Distributed Systems and Networks

Critical Bubble Scheme: An Efficient Implementation of Globally-aware Network Flow Control

Lizhong Chen, Ruisheng Wang, Timothy M. Pinkston
Ming Hsieh Department of Electrical Engineering
University of Southern California
Los Angeles, California, USA
{lizhongc, ruishenw, tpink}@usc.edu

Abstract

Network flow control mechanisms that are aware of global conditions potentially can achieve higher performance than flow control mechanisms that are only locally aware. Owing to high implementation overhead, globally-aware flow control mechanisms in their purest form are seldom adopted in practice, leading to less efficient simplified implementations. In this paper, we propose an efficient implementation of a globally-aware flow control mechanism, called Critical Bubble Scheme, and apply it successfully to k-ary n-cube networks for the general class of buffer occupancy-based network flow control techniques. Simulation results show that the proposed scheme can reduce the buffer access portion of packet latency by as much as 77%, leading to much lower average packet latency at medium and high network loads while sustaining 11% throughput improvement after network saturation.

A Scalable Reverse Lookup Scheme using Group-based Shifted Declustering Layout

Junyao Zhang, Pengju Shang, and Jun Wang
Department of Electrical Engineering
and Computer Science
University of Central Florida
Orlando, Florida 32826
junyao, shang, jwang@eecs.ucf.edu

Abstract

Recent years have witnessed an increasing demand for super data clusters. The super data clusters have reached the petabyte-scale that can consist of thousands or tens of thousands storage nodes at a single site. For this architecture, reliability is becoming a great concern. In order to achieve a high reliability, data recovery and node reconstruction is a must. Although extensive research works have investigated how to sustain high performance and high reliability in case of node failures at large scale, a reverse lookup problem, namely finding the objects list for the failed node remains open. This is especially true for storage systems with high requirement of data integrity and availability, such as scientific research data clusters and etc. Existing solutions are either time consuming or expensive. Meanwhile, replication based block placement can be used to realize fast reverse lookup. However, they are designed for centralized, small-scale storage architectures. In this paper, we propose a fast and efficient reverse lookup scheme named Group- based Shifted Declustering (G-SD) layout that is able to locate the whole content of the failed node. G-SD extends our previous shifted declustering layout and applies to large-scale file systems. Our mathematical proofs and real-life experiments show that G- SD is a scalable reverse lookup scheme that is up to one order of magnitude faster than existing schemes.

Deadlock-Free Oblivious Routing for Arbitrary Topologies

Jens Domke
Center for Information Services and
High Performance Computing
Technische Universität Dresden
Dresden, Germany
jens.domke@zih.tu-dresden.de

Torsten Hoefler
Blue Waters Directorate
National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA
htor@illinois.edu

Wolfgang E. Nagel
Center for Information Services and
High Performance Computing
Technische Universität Dresden
Dresden, Germany
wolfgang.nagel@tu-dresden.de

Abstract

Efficient deadlock-free routing strategies are crucial to the performance of large-scale computing systems. There are many methods but it remains a challenge to achieve lowest latency and highest bandwidth for irregular or unstructured high-performance networks. We investigate a novel routing strategy based on the single-source-shortest-path routing algorithm and extend it to use virtual channels to guarantee deadlock-freedom. We show that this algorithm achieves minimal latency and high bandwidth with only a low number of virtual channels and can be implemented in practice. We demonstrate that the problem of finding the minimal number of virtual channels needed to route a general network deadlock-free is NP-complete and we propose different heuristics to solve the problem. We implement all proposed algorithms in the Open Subnet Manager of InfiniBand and compare the number of needed virtual channels and the bandwidths of multiple real and artificial network topologies which are established in practice. Our approach allows to use the existing virtual channels more effectively to guarantee deadlock-freedom and increase the effective bandwidth of up to a factor of two. Application benchmarks show an improvement of up to 95%. Our routing scheme is not limited to InfiniBand but can be deployed on existing InfiniBand installations to increase network performance transparently without modifications to the user applications.

RDMA Capable iWARP over Datagrams

Ryan E. Grant, Mohammad J. Rashti, Ahmad Afsahi
Electrical and Computer Engineering,
Queen's University,
Kingston, ON, Canada
{ryan.grant, mohammad.rashti, ahmad.afsahi}@queensu.ca
Pavan Balaji
Mathematics and Computer Science,
Argonne National Laboratory,
Argonne, IL, USA
balaji@mcs.anl.gov

Abstract

iWARP is a state of the art high-speed connection-based RDMA networking technology for Ethernet networks to provide InfiniBand-like zero-copy and one-sided communication capabilities over Ethernet. Despite the benefits offered by iWARP, many datacenter and web-based applications, such as stock-market trading and media-streaming applications, that rely on datagram-based semantics (mostly through UDP/IP) cannot take advantage of it because the iWARP standard is only defined over reliable, connection-oriented transports. This paper presents an RDMA model that functions over reliable and unreliable datagrams. The ability to use datagrams significantly expands the application space serviced by iWARP and can bring the scalability advantages of a connectionless transport to iWARP. In our previous work, we had developed an iWARP datagram solution using send/receive semantics showing excellent memory scalability and performance benefits over the current TCP-based iWARP. In this paper, we demonstrate an improved iWARP design that provides true RDMA semantics over datagrams. Specifically, because traditional RDMA semantics do not map well to unreliable communication, we propose RDMA Write-Record, the first and the only method capable of supporting RDMA Write over both unreliable and reliable datagrams. We demonstrate through a proof-of-concept software implementation that datagram-iWARP is feasible for real-world applications. Our proposed RDMA Write-Record method has been designed with data loss in mind and can provide superior performance under conditions of packet loss. It is shown through micro-benchmarks that by using RDMA capable datagram-iWARP a maximum of 256% increase in large message bandwidth and a maximum of 24.4% improvement in small message latency can be achieved over traditional iWARP. For application results we focus on streaming applications, showing a 24% improvement in memory usage and up to a 74% improvement in performance, although the proposed approach is also applicable to the HPC domain.

Session 16
Programming Environments and Tools

Reconciling Sampling and Direct Instrumentation for Unintrusive Call-Path Profiling of MPI Programs

Zoltán Szebenyi[*‡], Todd Gamblin[§], Martin Schulz[§], Bronis R. de Supinski[§], Felix Wolf[*†‡], Brian J. N. Wylie[*]
[*] Jülich Supercomputing Centre, Forschungszentrum Jülich, 52425 Jülich, Germany
[†] German Research School for Simulation Sciences, 52062 Aachen, Germany
[‡] RWTH Aachen University, 52056 Aachen, Germany
[§] Lawrence Livermore National Laboratory, CA 94550 Livermore, USA
{z.szebenyi, f.wolf, b.wylie}@fz-juelich.de
{tgamblin, schulzm, bronis}@llnl.gov

Abstract

We can profile the performance behavior of parallel programs at the level of individual call paths through sampling or direct instrumentation. While we can easily control measurement dilation by adjusting the sampling frequency, the statistical nature of sampling and the difficulty of accessing the parameters of sampled events make it unsuitable for obtaining certain communication metrics, such as the size of message payloads. Alternatively, direct instrumentation, which is preferable for capturing message-passing events, can excessively dilate measurements, particularly for C++ programs, which often have many short but frequently called class member functions. Thus, we combine these techniques in a unified framework that exploits the strengths of each approach while avoiding their weaknesses: We use direct instrumentation to intercept MPI routines while we record the execution of the remaining code through low-overhead sampling. One of the main technical hurdles mastered was the inexpensive and portable determination of call-path information during the invocation of MPI routines. We show that the overhead of our implementation is sufficiently low to support substantial performance improvement of a C++ fluid-dynamics code.

A Practical Approach for Performance Analysis of Shared-Memory Programs

Bogdan Marius Tudor and Yong Meng Teo
Department of Computer Science
National University of Singapore
Computing 1, 13 Computing Drive, Singapore 117417
[bogdanma,teoym]@comp.nus.edu.sg

Abstract

Parallel programming has transcended from HPC into mainstream, enabled by a growing number of programming models, languages and methodologies, as well as the availability of multicore systems. However, performance analysis of parallel programs is still difficult, especially for large and complex programs, or applications developed using different programming models. This paper proposes a simple analytical model for studying the speedup of shared-memory programs on multicore systems. The proposed model derives the speedup and speedup loss from data dependency and memory overhead for various configurations of threads, cores and memory access policies in UMA and NUMA systems. The model is practical because it uses only generally available and non-intrusive inputs derived from the trace of the operating system run-queue and hardware events counters. Using six OpenMP HPC dwarfs from the NPB benchmark, our model differs from measurement results on average by 9% for UMA and 11% on NUMA. Our analysis shows that speedup loss is dominated by memory contention, especially for larger problem sizes. For the worst performing structured grid dwarf on UMA, memory contention accounts for up to 99% of the speedup loss. Based on this insight, we apply our model to determine the optimal number of cores that alleviates memory contention, maximizing speedup and reducing execution time.

Single Node On-Line Simulation of MPI Applications with SMPI

Pierre-Nicolas Clauss*, Mark Stillwell†, Stéphane Genaud‡, Frédéric Suter§, Henri Casanova†, Martin Quinson*
* Nancy University, LORIA, INRIA, Nancy, France
† Department of Information and Computer Sciences, University of Hawai'i at Mānoa, Honolulu, U.S.A.
‡ University of Strasbourg, ICPS-LSIIT, Illkirch, France
§ IN2P3 Computing Center, CNRS, IN2P3, Lyon-Villeurbanne, France

Abstract

Simulation is a popular approach for predicting the performance of MPI applications for platforms that are not at one's disposal. It is also a way to teach the principles of parallel programming and high-performance computing to students without access to a parallel computer. In this work we present SMPI, a simulator for MPI applications that uses on-line simulation, i.e., the application is executed but part of the execution takes place within a simulation component. SMPI simulations account for network contention in a fast and scalable manner. SMPI also implements an original and validated piece- wise linear model for data transfer times between cluster nodes. Finally SMPI simulations of large-scale applications on large-scale platforms can be executed on a single node thanks to techniques to reduce the simulation's compute time and memory footprint. These contributions are validated via a large set of experiments in which SMPI is compared to popular MPI implementations with a view to assess its accuracy, scalability, and speed.

PATUS: A Code Generation and Autotuning Framework For Parallel Iterative Stencil Computations on Modern Microarchitectures

Matthias Christen, Olaf Schenk, Helmar Burkhart
Department of Mathematics and Computer Science
University of Basel, Switzerland
{ m.christen | olaf.schenk | helmar.burkhart } @unibas.ch

Abstract

Stencil calculations comprise an important class of kernels in many scientific computing applications ranging from simple PDE solvers to constituent kernels in multigrid methods as well as image processing applications. In such types of solvers, stencil kernels are often the dominant part of the computation, and an efficient parallel implementation of the kernel is therefore crucial in order to reduce the time to solution. However, in the current complex hardware microarchitectures, meticulous architecture-specific tuning is required to elicit the machine's full compute power. We present a code generation and auto-tuning framework PATUS for stencil computations targeted at multi- and manycore processors, such as multicore CPUs and graphics processing units, which makes it possible to generate compute kernels from a specification of the stencil operation and a parallelization and optimization strategy, and leverages the autotuning methodology to optimize strategy-dependent parameters for the given hardware architecture.

Session 17
Parallel Algorithms

A New Data Layout For Set Intersection on GPUs

Rasmus Resen Amossen and Rasmus Pagh
IT University of Copenhagen, Denmark
Email: {resen,pagh}@itu.dk

Abstract

Set intersection is the core in a variety of problems, e.g. frequent itemset mining and sparse boolean matrix multiplication. It is well-known that large speed gains can, for some computational problems, be obtained by using a graphics processing unit (GPU) as a massively parallel computing device. However, GPUs require highly regular control flow and memory access patterns, and for this reason previous GPU methods for intersecting sets have used a simple bitmap representation. This representation requires excessive space on sparse data sets. In this paper we present a novel data layout, "BatMap", that is particularly well suited for parallel processing, and is compact even for sparse data.

Frequent itemset mining is one of the most important applications of set intersection. As a case-study on the potential of BatMaps we focus on frequent pair mining, which is a core special case of frequent itemset mining. The main finding is that our method is able to achieve speedups over both Apriori and FP-growth when the number of distinct items is large, and the density of the problem instance is above 0.01. Previous implementations of frequent itemset mining on GPU have not been able to show speedups over the best single-threaded implementations.

Partitioning Spatially Located Computations using Rectangles

Erik Saule*, Erdeniz Ö. Baş*† and Ümit V. Çatalyürek*‡
* Department of Biomedical Informatics
† Department of Computer Science and Engineering
‡Department of Electrical and Computer Engineering
The Ohio State University
Email: {esaule,erdeniz,umit}@bmi.osu.edu

Abstract

The ideal distribution of spatially located heterogeneous workloads is an important problem to address in parallel scientific computing. We investigate the problem of partitioning such workloads (represented as a matrix of positive integers) into rectangles, such that the load of the most loaded rectangle (processor) is minimized. Since finding the optimal arbitrary rectangle-based partition is an NP-hard problem, we investigate particular classes of solutions, namely, rectilinear partitions, jagged partitions and hierarchical partitions. We present a new class of solutions called m-way jagged partitions, propose new optimal algorithms for m-way jagged partitions and hierarchical partitions, propose new heuristic algorithms, and provide worst case performance analyses for some existing and new heuristics. Moreover, the algorithms are tested in simulation on a wide set of instances. Results show that two of the algorithms we introduce lead to a much better load balance than the state-of-the-art algorithms.

Reduced-Bandwidth Multithreaded Algorithms for Sparse-Matrix Vector Multiplication

Aydın Buluç* Samuel Williams*
abuluc@lbl.gov swwilliams@lbl.gov
* Computational Research Division
Lawrence Berkeley National Laboratory
Berkeley, CA 94720

Leonid Oliker* James Demmel*†
loliker@lbl.gov demmel@cs.berkeley.edu
† Mathematics Department and CS Division
University of California
Berkeley, CA 94720

Abstract

On multicore architectures, the ratio of peak memory bandwidth to peak floating-point performance (byte:flop ratio) is decreasing as core counts increase, further limiting the performance of bandwidth limited applications. Multiplying a sparse matrix (as well as its transpose in the unsymmetric case) with a dense vector is the core of sparse iterative methods. In this paper, we present a new multithreaded algorithm for the symmetric case which potentially cuts the bandwidth requirements in half while exposing lots of parallelism in practice. We also give a new data structure transformation, called bitmasked register blocks, which promises significant reductions on bandwidth requirements by reducing the number of indexing elements without introducing additional fill-in zeros. Our work shows how to incorporate this transformation into existing parallel algorithms (both symmetric and unsymmetric) without limiting their parallel scalability. Experimental results indicate that the combined benefits of bitmasked register blocks and the new symmetric algorithm can be as high as a factor of 3.5x in multicore performance over an already scalable parallel approach. We also provide a model that accurately predicts the performance of the new methods, showing that even larger performance gains are expected in future multicore systems as current trends (decreasing byte:flop ratio and larger sparse matrices) continue.

Session 18
Distributed Systems

GRAL: A Grouping Algorithm to Optimize Application Placement in Wireless Embedded Systems

N.Tziritas[†⋆], T.Loukopoulos[†‡], S.Lalis[†⋆] and P.Lampsas [†‡]
[†]Center for Research and Technology Thessaly
Volos, Greece
[⋆]Computer & Communications Engineering Dept. University of Thessaly
Volos, Greece
[‡]Dept. of Informatics and Computer Technology
TEI of Lamia
Lamia, Greece
Email: ⋆{ nitzirit, lalis }@inf.uth.gr , ‡ { luke, plam }@teilam.gr

Abstract

Recent embedded middleware initiatives enable the structuring of an application as a set of collaborating agents deployed in the various sensing/actuating entities of the system. Of particular importance is the incurred cost due to agent communication which in terms depends on agent positions in the system. In this paper we present GRAL a grouping algorithm that migrates groups of agents with the aim of minimizing communication. The algorithm works in a distributed fashion based on knowledge available locally at each node and can be used both for one-shot initial application deployment and for the continuous updating of agent placement. Through simulation experiments under various scenarios we evaluate the algorithm, comparing the solution quality reached against the optimal obtained from exhaustive search.

Vitis: A Gossip-based Hybrid Overlay for Internet-scale Publish/Subscribe

Enabling Rendezvous Routing in Unstructured Overlay Networks

Fatemeh Rahimian[†‡], Sarunas Girdzijauskas[†], Amir H. Payberah[†‡], Seif Haridi[†‡]
[†]Swedish Institute of Computer Science (SICS), Stockholm, Sweden
[‡]Royal Institute of Technology (KTH), Stockholm, Sweden
Email: {fatemeh, sarunas, amir, seif}@sics.se

Abstract

Peer-to-peer overlay networks are attractive solutions for building Internet-scale publish/subscribe systems. However, scalability comes with a cost: a message published on a certain topic often needs to traverse a large number of uninterested (unsubscribed) nodes before reaching all its subscribers. This might sharply increase resource consumption for such relay nodes (in terms of bandwidth transmission cost, CPU, etc) and could ultimately lead to rapid deterioration of the system's performance once the relay nodes start dropping the messages or choose to permanently abandon the system. In this paper, we introduce *Vitis*, a gossip-based publish/subscribe system that significantly decreases the number of relay messages, and scales to an unbounded number of nodes and topics. This is achieved by the novel approach of enabling rendezvous routing on unstructured overlays. We construct a hybrid system by injecting structure into an otherwise unstructured network. The resulting structure resembles a navigable small-world network, which spans along clusters of nodes that have similar subscriptions. The properties of such an overlay make it an ideal platform for efficient data dissemination in large-scale systems. We perform extensive simulations and evaluate Vitis by comparing its performance against two base-line publish/subscribe systems: one that is oblivious to node subscriptions, and another that exploits the subscription similarities. Our measurements show that Vitis significantly outperforms the base-line solutions on various subscription and churn scenarios, from both synthetic models and real-world traces.

CABdedupe: A Causality-based Deduplication Performance Booster for Cloud Backup Services

Yujuan Tan[1], Hong Jiang[2], Dan Feng[1,*], Lei Tian[1,2], Zhichao Yan[1]
[1]School of Computer Science & Technology, Huazhong University of Science & Technology
[2]Department of Computer Science & Engineering, University of Nebraska-Lincoln
[*]Corresponding author: dfeng@hust.edu.cn
{tanyujuan, jarod2046}@gmail.com, {jiang}@cse.unl.edu, {dfeng,ltian}@hust.edu.cn

Abstract

Due to the relatively low bandwidth of WAN (Wide Area Network) that supports cloud backup services, both the backup time and restore time in the cloud backup environment are in desperate need for reduction to make cloud backup a practical and affordable service for small businesses and telecommuters alike. Existing solutions that employ the deduplication technology for cloud backup services only focus on removing redundant data from transmission during backup operations to reduce the backup time, while paying little attention to the restore time that we argue is an important aspect and affects the overall quality of service of the cloud backup services. In this paper, we propose a *CAusality-Based deduplication performance booster for both cloud backup and restore operations*, called CABdedupe, which captures the causal relationship among chronological versions of datasets that are processed in multiple backups/restores, to remove the unmodified data from transmission during not only backup operations but also restore operations, thus to improve both the backup and restore performances. CABdedupe is a middleware that is orthogonal to and can be integrated into any existing backup system. Our extensive experiments, where we integrate CABdedupe into two existing backup systems and feed real world datasets, show that both the backup time and restore time are significantly reduced, with a reduction ratio of up to 103 : 1.

High performance scalable and expressive modeling environment to study mobile malware in large dynamic networks

Karthik Channakeshava[*†] , Keith Bisset[†] , V. S. Anil Kumar[†‡] , Madhav Marathe[†‡] and Shrirang Yardi[§†]
[*] Ericsson, San Jose CA 95134,
[†] Virginia Bioinformatics Institute, [‡] Department of Computer Science, Virginia Tech, Blacksburg VA 24061
[§] NVIDIA Corporation, Santa Clara CA 95050

Abstract

Large scale realistic simulations of malware on mobile wireless networks have recently become an increasingly important application of high-performance computing. We propose EpiCure - an individual-based, scalable high performance computing oriented modeling environment to study malware propagation over realistic mobile networks. It is designed specifically to work on commodity cluster architectures.

EpiCure runs extremely fast for realistic instances that involve: (i) large time-varying networks consisting of millions of heterogeneous individuals with time varying interaction neighborhoods, (ii) dynamic interactions between disease propagation, device behavior, and the exogenous interventions, and (iii) large number of replicated runs necessary for statistically sound estimates about the stochastic epidemic evolution. We find that EpiCure runs several orders of magnitude faster than another comparable simulation tool while delivering similar results. Beyond simple compute speed, EpiCure has been designed so that analysts can easily represent a range of interventions leading to improved human productivity and ease of use. This is an increasingly important metric in high performance computing. We illustrate EpiCure using three case studies that bring out the novel features of EpiCure.

Session 19
Storage Systems and Memory

H-Code: A Hybrid MDS Array Code to Optimize Partial Stripe Writes in RAID-6

Chentao Wu[1*], Shenggang Wan[2†], Xubin He[1*], Qiang Cao[2‡], and Changsheng Xie[2‡]
[1]Department of Electrical & Computer Engineering,
Virginia Commonwealth University, Richmond, VA 23284, USA
[2]School of Computer Science & Technology,
Huazhong University of Science & Technology, Wuhan, China 430074
* {wuc4,xhe2}@vcu.edu, † wanshenggang@gmail.com, ‡{caoqiang, cs_xie}@hust.edu.cn

Abstract

RAID-6 is widely used to tolerate concurrent failures of any two disks to provide a higher level of reliability with the support of erasure codes. Among many implementations, one class of codes called **M**aximum **D**istance **S**eparable (**MDS**) codes aims to offer data protection against disk failures with optimal storage efficiency. Typical MDS codes contain horizontal and vertical codes. Due to the horizontal parity, in the case of *partial stripe write* (refers to I/O operations that write new data or update data to a subset of disks in an array) in a row, horizontal codes may get less I/O operations in most cases, but suffer from unbalanced I/O distribution. They also have limitation on high single write complexity. Vertical codes improve single write complexity compared to horizontal codes, while they still suffer from poor performance in partial stripe writes. In this paper, we propose a new XOR-based MDS array code, named Hybrid Code (H-Code), which optimizes partial stripe writes for RAID-6 by taking advantages of both horizontal and vertical codes. H-Code is a solution for an array of $(p + 1)$ disks, where p is a prime number. Unlike other codes taking a dedicated anti-diagonal parity strip, H-Code uses a special anti-diagonal parity layout and distributes the anti-diagonal parity elements among disks in the array, which achieves a more balanced I/O distribution. On the other hand, the horizontal parity of H-Code ensures a partial stripe write to continuous data elements in a row share the same row parity chain, which can achieve optimal partial stripe write performance. Not only within a row but also within a stripe, H-Code offers optimal partial stripe write complexity to two continuous data elements and optimal partial stripe write performance among all MDS codes to the best of our knowledge. Specifically, compared to RDP and EVENODD codes, H-Code reduces I/O cost by up to 15.54% and 22.17%. Overall, H-code has optimal storage efficiency, optimal encoding/decoding computational complexity, optimal complexity of both single write and partial stripe write.

LACIO: A New Collective I/O Strategy for Parallel I/O Systems

Yong Chen [§] , Xian-He Sun [†] , Rajeev Thakur [‡] , Philip C. Roth [♮], and William D. Gropp [♯]
[§] Department of Computer Science, Texas Tech University, Lubbock, Texas, USA
[†] Department of Computer Science, Illinois Institute of Technology, Chicago, Illinois, USA
[‡] Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois, USA
[♮] Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA
[♯] Department of Computer Science, University of Illinois Urbana-Champaign, Urbana, Illinois, USA
Email: {yong.chen@ttu.edu, sun@iit.edu, thakur@mcs.anl.gov, rothpc@ornl.gov, wgropp@illinois.edu}

Abstract

Parallel applications benefit considerably from the rapid advance of processor architectures and the available massive computational capability, but their performance suffers from large latency of I/O accesses. The poor I/O performance has been attributed as a critical cause of the low sustained performance of parallel systems. Collective I/O is widely considered a critical solution that exploits the correlation among I/O accesses from multiple processes of a parallel application and optimizes the I/O performance. However, the conventional collective I/O strategy makes the optimization decision based on the logical file layout to avoid multiple file system calls and does not take the physical data layout into consideration. On the other hand, the physical data layout in fact decides the actual I/O access locality and concurrency. In this study, we propose a new collective I/O strategy that is aware of the underlying physical data layout. We confirm that the new Layout-Aware Collective I/O (LACIO) improves the performance of current parallel I/O systems effectively with the help of noncontiguous file system calls. It holds promise in improving the I/O performance for parallel systems.

Using Shared Memory to Accelerate MapReduce on Graphics Processing Units

Feng Ji
Department of Computer Science
North Carolina State University
Email: fji@ncsu.edu

Xiaosong Ma
Department of Computer Science
North Carolina State University
&Computer Science and Mathematics Division
Oak Ridge National Laboratory
Email: ma@cs.ncsu.edu

Abstract

Modern General Purpose Graphics Processing Units (GPGPUs) provide high degrees of parallelism in computation and memory access, making them suitable for data parallel applications such as those using the elastic MapReduce model. Yet designing a MapReduce framework for GPUs faces significant challenges brought by their multi-level memory hierarchy. Due to the absence of atomic operations in the earlier generations of GPUs, existing GPU MapReduce frameworks have problems in handling input/output data with varied or unpredictable sizes. Also, existing frameworks utilize mostly a single level of memory, *i.e.*, the relatively spacious yet slow global memory.

In this work, we attempt to explore the potential benefit of enabling a GPU MapReduce framework to use multiple levels of the GPU memory hierarchy. We propose a novel GPU data staging scheme for MapReduce workloads, tailored toward the GPU memory hierarchy. Centering around the efficient utilization of the fast but very small shared memory, we designed and implemented a GPU MapReduce framework, whose key techniques include (1) shared memory staging area management, (2) thread-role partitioning, and (3) intra-block thread synchronization. We carried out evaluation with five popular MapReduce workloads and studied their performance under different GPU memory usage choices. Our results reveal that exploiting GPU shared memory is highly promising for the Map phase (with an average 2.85x speedup over using global memory only), while in the Reduce phase the benefit of using shared memory is much less pronounced, due to the high input-to-output ratio. In addition, when compared to Mars, an existing GPU MapReduce framework, our system is shown to bring a significant speedup in Map/Reduce phases.

Unified Signatures for Improving Performance in Transactional Memory

Woojin Choi and Jeff Draper
Information Sciences Institute
University of Southern California
Marina del Rey, CA 90292, USA
{woojinch, draper}@ISI.EDU

Abstract

Transactional Memory (TM) promises to increase programmer productivity by making it easier to write correct parallel programs. In fulfilling this goal, a TM system should maximize its performance with limited hardware resources. Conflict detection is an essential element for maintaining correctness among concurrent transactions in a TM system. Hardware signatures have been proposed as an area-efficient method for detecting conflicts. However, signatures can degrade TM performance by falsely declaring conflicts. Hence, increasing the quality of signatures within a given hardware budget is a crucial issue for TM to be adopted as a mainstream programming model. In this paper, we propose a simple and effective signature design, unified signature. Instead of using separate read- and write-signatures, as is often done in TM systems, we implement a single signature to track all read- and write-accesses. By merging read- and write-signatures, a unified signature can effectively enlarge the signature size without additional overhead. Within the constraints of a given hardware budget, a TM system with a unified signature outperforms a baseline system with the same hardware budget by reducing the number of falsely detected conflicts. Even though the unified signature scheme incurs read-after-read dependencies, we show that these false dependencies do not negate the benefit of unified signatures for practical signature sizes. A TM system with 2K-bit unified signatures achieves average speedups of 22% over baseline TM systems.

Session 20

Operating Systems and Resource Management

Reducing Fragmentation on Torus-Connected Supercomputers

Wei Tang,* Zhiling Lan,* Narayan Desai,† Daniel Buettner,‡ Yongen Yu*
* Department of Computer Science, Illinois Institute of Technology
Chicago, IL 60616, USA
{wtang6, lan, yyu22}@iit.edu
† Mathematics and Computer Science Division
‡ Argonne Leadership Computing Facility
Argonne National Laboratory, Argonne, IL 60439, USA
† desai@mcs.anl.gov
‡ buettner@alcf.anl.gov

Abstract

Torus-based networks are prevalent on leadership-class petascale systems, providing a good balance between network cost and performance. The major disadvantage of this network architecture is its susceptibility to fragmentation. Many studies have attempted to reduce resource fragmentation in this architecture. Although the approaches suggested can make good allocation decisions reducing fragmentation at job start time, none of them considers a job's walltime, which can cause resource fragmentation when neighboring jobs do not complete closely. In this paper, we propose a walltime-aware job allocation strategy, which adjacently packs jobs that finish around the same time, in order to minimize resource fragmentation caused by job length, discrepancy. Event-driven simulations using real job traces from a production Blue Gene/P system at Argonne National Laboratory demonstrate that our walltime-aware strategy can effectively reduce system fragmentation and improve overall system performance.

Co-analysis of RAS Log and Job Log on Blue Gene/P

Ziming Zheng, Li Yu, Wei Tang, Zhiling Lan
Department of Computer Science
Illinois Institute of Technology
{zzheng11, lyu17, wtang6, lan}@iit.edu
Rinku Gupta,* Narayan Desai,* Susan Coghlan,† Daniel Buettner†
*Mathematics and Computer Science Division
†Leadership Computing Facility
Argonne National Laboratory
{rgupta, desai}@mcs.anl.gov, {smc, buettner}@alcf.anl.gov

Abstract

With the growth of system size and complexity, reliability has become of paramount importance for petascale systems. Reliability, Availability, and Serviceability (RAS) logs have been commonly used for failure analysis. However, analysis based on just the RAS logs has proved to be insufficient in understanding failures and system behaviors. To overcome the limitation of this existing methodologies, we analyze the Blue Gene/P RAS logs and the Blue Gene/P job logs in a cooperative manner. From our co-analysis effort, we have identified a dozen important observations about failure characteristics and job interruption characteristics on the Blue Gene/P systems. These observations can significantly facilitate the research in fault resilience of large-scale systems.

A Quantitative Analysis of OS Noise

Alessandro Morari* , Roberto Gioiosa* , Robert W. Wisniewski§ , Francisco J. Cazorla*‡ , Mateo Valero*‡

* Computer Science Division
Barcelona Supercomputing Center, ES

† Computer Architecture Department
Universitat Politecnica de Catalunya, ES

‡ IIA-CSIC(Spanish National Research Council)

§ IBM T. J. Watson Research Center
{alessandro.morari,roberto.gioiosa,francisco.cazorla}@bsc.es, bobww@watson.ibm.com, mateo@ac.upc.edu

Abstract

Operating system noise is a well-known problem that may limit application scalability on large-scale machines, significantly reducing their performance. Though the problem is well studied, much of the previous work has been qualitative.

We have developed a technique to provide a *quantitative* descriptive analysis for each OS event that contributes to OS noise. The mechanism allows us to detail all sources of OS noise through precise kernel instrumentation and provides frequency and duration analysis for each event. Such a description gives OS developers better guidance for reducing OS noise. We integrated this data with a trace visualizer allowing quicker and more intuitive understanding of the data.

Specifically, the contributions of this paper are three-fold. First, we describe a methodology whereby detailed quantitative information may be obtained for each OS noise event. Though not the thrust of the paper, we show how we implemented that methodology by augmenting LTTng. We validate our approach by comparing it to other well-known standard techniques to analyze OS noise. Second, we provide a case study in which we use our methodology to analyze the OS noise when running benchmarks from the LLNL Sequoia applications. Our experiments enrich and expand previous results with our quantitative characterization. Third, we describe how a detailed characterization permits to disambiguate noise signatures of qualitatively similar events, allowing developers to address the true cause of each noise event.

Decal: Transparent Checkpointing and Process Migration of OpenCL Applications

Hiroyuki Takizawa[*‡] , Kentaro Koyama[*] , Katsuto Sato[*] , Kazuhiko Komatsu[*] , and Hiroaki Kobayashi[*‡]
[*] Tohoku University
Sendai, Miyagi 980-8578, Japan
E-mail: {tacky@, kentaro@sc., katuto@sc., komatsu@sc., koba@}isc.tohoku.ac.jp
[‡] Japan Science and Technology Agency, CREST

Abstract

In this paper, we propose a new transparent checkpoint/restart (CPR) tool, named CheCL, for high-performance and dependable GPU computing. CheCL can perform CPR on an OpenCL application program without any modification and recompilation of its code. A conventional checkpointing system fails to checkpoint a process if the process uses OpenCL. Therefore, in CheCL, every API call is forwarded to another process called an API proxy, and the API proxy invokes the API function; two processes, an application process and an API proxy, are launched for an OpenCL application. In this case, as the application process is not an OpenCL process but a standard process, it can be safely checkpointed. While CheCL intercepts all API calls, it records the information necessary for restoring OpenCL objects. The application process does not hold any OpenCL handles, but CheCL handles to keep such information. Those handles are automatically converted to OpenCL handles and then passed to API functions. Upon restart, OpenCL objects are automatically restored based on the recorded information. This paper demonstrates the feasibility of transparent checkpointing of OpenCL programs including MPI applications, and quantitatively evaluates the runtime overheads. It is also discussed that CheCL can enable process migration of OpenCL applications among distinct nodes, and among different kinds of compute devices such as a CPU and a GPU.

Plenary Session
Best Papers

Online Adaptive Code Generation and Tuning

Ananta Tiwari
Department of Computer Science,
University of Maryland,
College Park, MD, 20742
Email: tiwari@cs.umd.edu

Jeffrey K. Hollingsworth
Department of Computer Science,
University of Maryland,
College Park, MD, 20742
Email: hollings@cs.umd.edu

Abstract

In this paper, we present a runtime compilation and tuning framework for parallel programs. We extend our prior work on our auto-tuner, Active Harmony, for tunable parameters that require code generation (for example, different unroll factors). For such parameters, our auto-tuner generates and compiles new code on-the-fly. Effectively, we merge traditional feedback directed optimization and just-in-time compilation. We show that our system can leverage available parallelism in today's HPC platforms by evaluating different code-variants on different nodes simultaneously. We evaluate our system on two parallel applications and show that our system can improve runtime execution by up to 46% compared to the original version of the program.

GLocks: Efficient Support for Highly-Contended Locks in Many-Core CMPs

José L. Abellán, Juan Fernández and Manuel E. Acacio
Dept. de Ingeniería y Tecnología de Computadores, Universidad de Murcia
Facultad de Informática, Campus de Espinardo s/n, 30100 Murcia, Spain
email: {jl.abellan, juanf, meacacio}@ditec.um.es

Abstract

Synchronization is of paramount importance to exploit thread-level parallelism on many-core CMPs. In these architectures, synchronization mechanisms usually rely on shared variables to coordinate multithreaded access to shared data structures thus avoiding data dependency conflicts. Lock synchronization is known to be a key limitation to performance and scalability. On the one hand, lock acquisition through busy waiting on shared variables generates additional coherence activity which interferes with applications. On the other hand, lock contention causes serialization which results in performance degradation. This paper proposes and evaluates *GLocks*, a hardware-supported implementation for highly-contended locks in the context of many-core CMPs. *GLocks* use a token-based message-passing protocol over a dedicated network built on state-of-the-art technology. This approach skips the memory hierarchy to provide a non-intrusive, extremely efficient and fair lock implementation with negligible impact on energy consumption or die area.

A comprehensive comparison against the most efficient shared-memory-based lock implementation for a set of microbenchmarks and real applications quantifies the goodness of *GLocks*. Performance results show an average reduction of 42% and 14% in execution time, an average reduction of 76% and 23% in network traffic, and also an average reduction of 78% and 28% in energy-delay^2 product (ED^{2}P) metric for the full CMP for the microbenchmarks and the real applications, respectively. In light of our performance results, we can conclude that *GLocks* satisfy our initial working hypothesis. *GLocks* minimize cache-coherence network traffic due to lock synchronization which translates into reduced power consumption and execution time.

Profiling Heterogeneous Multi-GPU Systems to Accelerate Cortically Inspired Learning Algorithms

Andrew Nere, Atif Hashmi, and Mikko Lipasti
Department of Electrical and Computer Engineering
University of Wisconsin-Madison
Madison, WI 53706, USA
{nere, ahashmi}@wisc.edu, {mikko}@engr.wisc.edu

Abstract

Recent advances in neuroscientific understanding make parallel computing devices modeled after the human neocortex a plausible, attractive, fault-tolerant, and energy-efficient possibility. Such attributes have once again sparked an interest in creating learning algorithms that aspire to reverse-engineer many of the abilities of the brain.

In this paper we describe a GPGPU-accelerated extension to an intelligent learning model inspired by the structural and functional properties of the mammalian neocortex. Our cortical network, like the brain, exhibits massive amounts of processing parallelism, making today's GPGPUs a highly attractive and readily-available hardware accelerator for such a model.

Furthermore, we consider two inefficiencies inherent to our initial design: multiple kernel-launch overhead and poor utilization of GPGPU resources. We propose optimizations such as a software work-queue structure and pipelining the hierarchical layers of the cortical network to mitigate such problems. Our analysis provides important insight into the GPU architecture details including the number of cores, the memory system, and the global thread scheduler. Additionally, we create a runtime profiling tool for our parallel learning algorithm which proportionally distributes the cortical network across the host CPU as well as multiple GPUs, whether homogeneous or heterogeneous, that may be available to the system. Using the profiling tool with these optimizations on Nvidia's CUDA framework, we achieve up to 60x speedup over a single-threaded CPU implementation of the model.

PHAST: Hardware-Accelerated Shortest Path Trees

Daniel Delling, Andrew V. Goldberg, Andreas Nowatzyk, and Renato F. Werneck
Microsoft Research Silicon Valley
Mountain View, CA, 94043, USA
Email: {dadellin,goldberg,andnow,renatow}@microsoft.com

Abstract

We present a novel algorithm to solve the nonnegative single-source shortest path problem on road networks and other graphs with low highway dimension. After a quick preprocessing phase, we can compute all distances from a given source in the graph with essentially a linear sweep over all vertices. Because this sweep is independent of the source, we are able to reorder vertices in advance to exploit locality. Moreover, our algorithm takes advantage of features of modern CPU architectures, such as SSE and multi-core. Compared to Dijkstra's algorithm, our method needs fewer operations, has better locality, and is better able to exploit parallelism at multi-core and instruction levels. We gain additional speedup when implementing our algorithm on a GPU, where our algorithm is up to three orders of magnitude faster than Dijkstra's algorithm on a high-end CPU. This makes applications based on all-pairs shortest-paths practical for continental-sized road networks. Several algorithms, such as computing the graph diameter, exact arc flags, or centrality measures (exact reaches or betweenness), can be greatly accelerated by our method.

Session 21
Numerical Algorithms

QR Factorization on a Multicore Node Enhanced with Multiple GPU Accelerators

Emmanuel Agullo*, Cédric Augonnet*, Jack Dongarra† , Mathieu Faverge†,
Hatem Ltaief†, Samuel Thibault* and Stanimire Tomov†
* INRIA, LaBRI, University of Bordeaux, France
Email: Emmanuel.Agullo,Cedric.Augonnet,Samuel.Thibault@inria.fr
† Innovative Computing Laboratory, University of Tennessee, Knoxville, TN 37996, USA
Email: Dongarra,Faverge,Ltaief,Tomov@eecs.utk.edu

Abstract

One of the major trends in the design of exascale architectures is the use of multicore nodes enhanced with GPU accelerators. Exploiting all resources of a hybrid accelerators-based node at their maximum potential is thus a fundamental step towards exascale computing. In this article, we present the design of a highly efficient QR factorization for such a node. Our method is in three steps. The first step consists of expressing the QR factorization as a sequence of tasks of well chosen granularity that will aim at being executed on a CPU core or a GPU. We show that we can efficiently adapt high-level algorithms from the literature that were initially designed for homogeneous multicore architectures. The second step consists of designing the kernels that implement each individual task. We use CPU kernels from previous work and present new kernels for GPUs that complement kernels already available in the MAGMA library. We show the impact on performance of these GPU kernels. In particular, we present the benefits of new hybrid CPU/GPU kernels. The last step consists of scheduling these tasks on the computational units. We present two alternative approaches, respectively based on static and dynamic scheduling. In the case of static scheduling, we exploit the a priori knowledge of the schedule to perform successive optimizations leading to very high performance. We, however, highlight the lack of portability of this approach and its limitations to relatively simple algorithms on relatively homogeneous nodes. Alternatively, by relying on an efficient runtime system, StarPU, in charge of ensuring data availability and coherency, we can schedule more complex algorithms on complex heterogeneous nodes with much higher productivity. In this latter case, we show that we can achieve high performance in a portable way thanks to a fine interaction between the application and the runtime system. We demonstrate that the obtained performance is very close to the theoretical upper bounds that we obtained using Linear Programming.

Two-Stage Tridiagonal Reduction for Dense Symmetric Matrices using Tile Algorithms on Multicore Architectures

Piotr Luszczek*, Hatem Ltaief* and Jack Dongarra*
* Innovative Computing Laboratory, University of Tennessee, Knoxville, TN 37996, USA
Email: luszczek,ltaief,dongarra@eecs.utk.edu

Abstract

While successful implementations have already been written for one-sided transformations (e.g., QR, LU and Cholesky factorizations) on multicore architecture, getting high performance for two-sided reductions (e.g., Hessenberg, tridiagonal and bidiagonal reductions) is still an open and difficult research problem due to expensive memory-bound operations occurring during the panel factorization. The processor-memory speed gap continues to widen, which has even further exacerbated the problem. This paper focuses on an efficient implementation of the tridiagonal reduction, which is the first algorithmic step toward computing the spectral decomposition of a dense symmetric matrix. The original matrix is translated into a *tile* layout i.e., a high performance data representation, which substantially enhances data locality. Following a two-stage approach, the tile matrix is then transformed into band tridiagonal form using compute intensive kernels. The band form is further reduced to the required tridiagonal form using a *left-looking* bulge chasing technique to reduce memory traffic and memory contention. A dependence translation layer associated with a dynamic runtime system allows for scheduling and overlapping tasks generated from both stages. The obtained tile tridiagonal reduction significantly outperforms the state-of-the-art numerical libraries (10X against multithreaded LAPACK with optimized MKL BLAS and 2.5X against the commercial numerical software Intel MKL) from medium to large matrix sizes.

An Auto-tuned Method for Solving Large Tridiagonal Systems on the GPU

Andrew Davidson
University of California, Davis
aaldavidson@ucdavis.edu

Yao Zhang
University of California, Davis
yaozhang@ucdavis.edu

John D. Owens
University of California, Davis
jowens@ece.ucdavis.edu

Abstract

We present a multi-stage method for solving large tridiagonal systems on the GPU. Previously large tridiagonal systems cannot be efficiently solved due to the limitation of on-chip shared memory size. We tackle this problem by splitting the systems into smaller ones and then solving them on-chip. The multi-stage characteristic of our method, together with various workloads and GPUs of different capabilities, obligates an auto-tuning strategy to carefully select the switch points between computation stages. In particular, we show two ways to effectively prune the tuning space and thus avoid an impractical exhaustive search: (1) apply algorithmic knowledge to decouple tuning parameters, and (2) estimate search starting points based on GPU architecture parameters. We demonstrate that auto-tuning is a powerful tool that improves the performance by up to 5x, saves 17% and 32% of execution time on average respectively over static and dynamic tuning, and enables our multi-stage solver to outperform the Intel MKL tridiagonal solver on many parallel tridiagonal systems by 6-11x.

A communication-avoiding, hybrid-parallel, rank-revealing orthogonalization method

Mark Hoemmen
Scalable Algorithms Department, Sandia National Laboratories
P.O. Box 5800, MS 1320, Albuquerque, NM 87185-1320, USA
Email: mhoemme@sandia.gov

Abstract

Orthogonalization consumes much of the run time of many iterative methods for solving sparse linear systems and eigenvalue problems. Commonly used algorithms, such as variants of Gram-Schmidt or Householder QR, have performance dominated by communication. Here, "communication" includes both data movement between the CPU and memory, and messages between processors in parallel. Our Tall Skinny QR (TSQR) family of algorithms requires asymptotically fewer messages between processors and data movement between CPU and memory than typical orthogonalization methods, yet achieves the same accuracy as Householder QR factorization. Furthermore, in block orthogonalizations, TSQR is faster and more accurate than existing approaches for orthogonalizing the vectors within each block ("normalization"). TSQR's rank-revealing capability also makes it useful for detecting deflation in block iterative methods, for which existing approaches sacrifice performance, accuracy, or both.

We have implemented a version of TSQR that exploits both distributed-memory and shared-memory parallelism, and supports real and complex arithmetic. Our implementation is optimized for the case of orthogonalizing a small number (5–20) of very long vectors. The shared-memory parallel component uses Intel's Threading Building Blocks, though its modular design supports other shared-memory programming models as well, including computation on the GPU. Our implementation achieves speedups of 2 times or more over competing orthogonalizations. It is available now in the development branch of the Trilinos software package, and will be included in the 10.8 release.

Session 22
Fault Tolerance

Flease - Lease Coordination without a Lock Server

Björn Kolbeck, Mikael Högqvist, Jan Stender
Zuse Institute Berlin
Berlin, Germany
{kolbeck,stender,hoegqvist}@zib.de

Felix Hupfeld
Google Switzerland GmbH
Zurich, Switzerland
hupfeld@google.com

Abstract

Large-scale distributed systems often require scalable and fault-tolerant mechanisms to coordinate exclusive access to shared resources such as files, replicas or the primary role. The best known algorithms to implement distributed mutual exclusion with leases, such as Multipaxos, are complex, difficult to implement, and rely on stable storage to persist lease information.

In this paper we present **F**LEASE, an algorithm for fault-tolerant lease coordination in distributed systems that is simpler than Multipaxos and does not rely on stable storage. The evaluation shows that **F**LEASE can be used to implement scalable, decentralized lease coordination that outperforms a central lock service implementation by an order of magnitude.

Uncoordinated Checkpointing Without Domino Effect for Send-Deterministic MPI Applications

Amina Guermouche [‡*], Thomas Ropars [*], Elisabeth Brunet [*], Marc Snir[†], Franck Cappello [*†]
[*] INRIA Saclay-Île de France, F-91893, Orsay, France
[‡] Université Paris Sud, F-91405 Orsay, France
[†] University of Illinois at Urbana-Champaign - College of Engineering, Urbana, IL, USA
guermou@lri.fr, tropars@inria.fr, brunet@it-sudparis.eu, snir@illinois.edu, fci@lri.fr

Abstract

As reported by many recent studies, the mean time between failures of future post-petascale supercomputers is likely to reduce, compared to the current situation. The most popular fault tolerance approach for MPI applications on HPC Platforms relies on coordinated checkpointing which raises two major issues: a) global restart wastes energy since all processes are forced to rollback even in the case of a single failure; b) checkpoint coordination may slow down the application execution because of congestions on I/O resources. Alternative approaches based on uncoordinated checkpointing and message logging require logging all messages, imposing a high memory/storage occupation and a significant overhead on communications. It has recently been observed that many MPI HPC applications are *send-deterministic*, allowing to design new fault tolerance protocols. In this paper, we propose an uncoordinated checkpointing protocol for send-deterministic MPI HPC applications that (i) logs only a subset of the application messages and (ii) does not require to restart systematically all processes when a failure occurs. We first describe our protocol and prove its correctness. Through experimental evaluations, we show that its implementation in MPICH2 has a negligible overhead on application performance. Then we perform a quantitative evaluation of the properties of our protocol using the NAS Benchmarks. Using a clustering approach, we demonstrate that this protocol actually succeeds to combine the two expected properties: a) it logs only a small fraction of the messages and b) it reduces by a factor approaching 2 the average number of processes to rollback compared to coordinated checkpointing.

Minimal Obstructions for the Coordinated Attack Problem and Beyond

Tristan Fevat[1] and Emmanuel Godard[1,2]
[1] Laboratoire Informatique Fondamentale
Aix-Marseille Université - CNRS (UMR 6166)
[2]CNRS (UMI3069) - Pacific Institute for the Mathematical Sciences
(tristan.fevat|emmanuel.godard)@lif.univ-mrs.fr

Abstract

We consider the well known Coordinated Attack Problem, where two generals have to decide on a common attack, when their messengers can be captured by the enemy. Informally, this problem represents the difficulties to agree in the present of communication faults. We consider here only omission faults (loss of message), but contrary to previous studies, we do not to restrict the way messages can be lost, ie. we use no specific failure metric. Our contribution is threefold. First, we introduce the study of arbitrary patterns of failure ("omission schemes"), proposing notions and notations that revealed very convenient to handle. In the large subclass of omission schemes where the double simultaneous omission can never happen, we characterize which one are obstructions for the Coordinated Attack Problem. We present then some interesting applications. We show for the first time that the well studied omission scheme, where at most one message can be lost at each round, is a kind of least worst case environment for the Coordinated Attack Problem. We also extend our study to networks of arbitrary size. In particular, we address an open question of Santoro and Widmayer about the Consensus Problem in communication networks with omission faults.

Scheduling Parallel Iterative Applications on Volatile Resources

Henri Casanova
University of Hawaíi at Mānoa, USA
henric@hawaii.edu

Fanny Dufossé, Yves Robert, Frédéric Vivien
Ecole Normale Supérieure de Lyon, France
Fanny.Dufosse,Yves.Robert,Frederic.Vivien@ens-lyon.fr

Abstract

In this paper we study the execution of iterative applications on volatile processors such as those found on desktop grids. We develop master-worker scheduling schemes that attempt to achieve good trade-offs between worker speed and worker availability. A key feature of our approach is that we consider a communication model where the bandwidth capacity of the master for sending application data to workers is limited. This limitation makes the scheduling problem more difficult both in a theoretical sense and in a practical sense. Furthermore, we consider that a processor can be in one of three states: available, down, or temporarily preempted by its owner. This preempted state also complicates the scheduling problem. In practical settings, e.g., desktop grids, master bandwidth is limited and processors are temporarily reclaimed. Consequently, addressing the aforementioned difficulties is necessary for successfully deploying master-worker applications on volatile platforms.

Our first contribution is to determine the complexity of the scheduling problem in its off-line version, i.e., when processor availability behaviors are known in advance. Even with this knowledge, the problem is NP-hard, and cannot be approximated within a factor $8/7$. Our second contribution is a closed-form formula for the expectation of the time needed by a worker to complete a set of tasks. This formula relies on a Markovian assumption for the temporal availability of processors, and is at the heart of some heuristics that aim at favoring "reliable" processors in a sensible manner. Our third contribution is a set of heuristics, which we evaluate in simulation. Our results provide guidance to selecting the best strategy as a function of processor state availability versus average task duration.

Session 23
Resource Utilization

Shared Resource Monitoring and Throughput Optimization in Cloud-Computing Datacenters

Jaideep Moses (Intel), Ravi Iyer (Intel), Ramesh Illikkal (Intel), Sadagopan Srinivasan (Intel)
Konstantinos Aisopos (Princeton University)

Abstract

Many datacenters employ server consolidation to maximize the efficiency of platform resource usage. As a result, multiple virtual machines (VMs) simultaneously run on each datacenter platform. Contention for shared resources between these virtual machines has an undesirable and non-deterministic impact on their performance behavior in such platforms. This paper proposes the use of shared resource monitoring to (a) understand the resource usage of each virtual machine on each platform, (b) collect resource usage and performance across different platforms to correlate implications of usage to performance, and (c) migrate VMs that are resource-constrained to improve overall datacenter throughput and improve Quality of Service (QoS). We focus our efforts on monitoring and addressing shared cache contention and propose a new optimization metric that captures the priority of the VM and the overall weighted throughput of the datacenter. We conduct detailed experiments emulating datacenter scenarios including on-line transaction processing workloads (based on TPC-C) middle-tier workloads (based on SPECjbb and SPECjAppServer) and financial workloads (based on PARSEC). We show that monitoring shared resource contention (such as shared cache) is highly beneficial to better manage throughput and QoS in a cloud-computing datacenter environment.

The Impact of Soft Resource Allocation on n-Tier Application Scalability

Qingyang Wang, Simon Malkowski, Deepal Jayasinghe
Pengcheng Xiong, Calton Pu
College of Computing
Georgia Institute of Technology
Atlanta, USA
{qywang, zmon, deepal, pxiong3, calton}@cc.gatech.edu
Yasuhiko Kanemasa, Motoyuki Kawaba, Lilian Harada
Research Center for Cloud Computing
Fujitsu Laboratories Ltd
Kawasaki, Japan
{kanemasa, kawaba, harada.lilian}@jp.fujitsu.com

Abstract

Good performance and efficiency, in terms of high quality of service and resource utilization for example, are important goals in a cloud environment. Through extensive measurements of an n-tier application benchmark (RUBBoS), we show that overall system performance is surprisingly sensitive to appropriate allocation of soft resources (e.g., server thread pool size). Inappropriate soft resource allocation can quickly degrade overall application performance significantly. Concretely, both under-allocation and over-allocation of thread pool can lead to bottlenecks in other resources because of non-trivial dependencies. We have observed some non-obvious phenomena due to these correlated bottlenecks. For instance, the number of threads in the Apache web server can limit the total useful throughput, causing the CPU utilization of the C-JDBC clustering middleware to decrease as the workload increases. We provide a practical iterative solution approach to this challenge through an algorithmic combination of operational queuing laws and measurement data. Our results show that soft resource allocation plays a central role in the performance scalability of complex systems such as n-tier applications in cloud environments.

Profiling Directed NUMA Optimization on Linux Systems: A Case Study of the Gaussian Computational Chemistry Code

Rui Yang*, Joseph Antony†, Alistair Rendell‡, Danny Robson‡ and Peter Strazdins‡
* Information Technology Services
University of Wollongong
Wollongong, NSW 2522 Australia
‡School of Computer Science
†ANU Supercomputer Facility
Australian National University
Canberra, ACT 0200 Australia
Email: ruiy@uow.edu.au; joseph.antony@anu.edu.au; alistair.rendell@anu.edu.au;
drobson@cs.anu.edu.au; peter.strazdins@cs.anu.edu.au

Abstract

The parallel performance of applications running on Non-Uniform Memory Access (NUMA) platforms is strongly influenced by the relative placement of memory pages to the threads that access them. As a consequence there are Linux application programmer interfaces (APIs) to control this. For large parallel codes it can, however, be difficult to determine how and when to use these APIs. In this paper we introduce the `NUMAgrind` profiling tool which can be used to simplify this process. It extends the `Valgrind` binary translation framework to include a model which incorporates cache coherency, memory locality domains and interconnect traffic for arbitrary NUMA topologies. Using `NUMAgrind`, cache misses can be mapped to memory locality domains, page access modes determined, and pages that are referenced by multiple threads quickly determined. We show how the `NUMAgrind` tool can be used to guide the use of Linux memory and thread placement APIs in the Gaussian computational chemistry code. The performance of the code before and after use of these APIs is also presented for three different commodity NUMA platforms.

Model-Driven SIMD Code Generation for a Multi-Resolution Tensor Kernel

Kevin Stock, Tom Henretty, Iyyappa Murugandi, P. Sadayappan
The Ohio State University
{stockk,henretty,murugand,saday}@cse.ohio-state.edu

Robert Harrison
Oak Ridge National Laboratoty
rjharrison@ornl.gov

Abstract

In this paper, we describe a model-driven compile-time code generator that transforms a class of tensor contraction expressions into highly optimized short-vector SIMD code. We use as a case study a multi-resolution tensor kernel from the MADNESS quantum chemistry application. Performance of a C-based implementation is low, and because the dimensions of the tensors are small, performance using vendor optimized BLAS libraries is also suboptimal. We develop a model-driven code generator that determines the optimal loop permutation and placement of vector load/store, transpose, and splat operations in the generated code, enabling portable performance on short-vector SIMD architectures. Experimental results on an SSE-based platform demonstrate the efficiency of the vector-code synthesizer.

Session 24
Parallel Programming Models and Languages

Multi-GPU MapReduce on GPU Clusters

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

We present GPMR, our stand-alone MapReduce library that leverages the power of GPU clusters for large-scale computing. To better utilize the GPU, we modify MapReduce by combining large amounts of map and reduce items into chunks and using partial reductions and accumulation. We use persistent map and reduce tasks and stress aspects of GPMR with a set of standard MapReduce benchmarks. We run these benchmarks on a GPU cluster and achieve desirable speedup and efficiency for all benchmarks. We compare our implementation to the current-best GPU-MapReduce library (runs only on a solo GPU) and a highly-optimized multi-core MapReduce to show the power of GPMR. We demonstrate how typical MapReduce tasks are easily modified to fit into GPMR and leverage a GPU cluster. We highlight how total and relative amounts of communication affect GPMR. We conclude with an exposition on the types of MapReduce tasks well-suited to GPMR, and why some tasks need more modifications than others to work well with GPMR.

X10 as a parallel language for scientific computation: practice and experience

Josh Milthorpe
V. Ganesh
Alistair P. Rendell
School of Computer Science
Australian National University
{josh.milthorpe,ganesh.venkateshwara,alistair.rendell}@anu.edu.au
David Grove
Thomas J. Watson Research Center
IBM
groved@us.ibm.com

Abstract

X10 is an emerging Partitioned Global Address Space (PGAS) language intended to increase significantly the productivity of developing scalable HPC applications. The language has now matured to a point where it is meaningful to consider writing large scale scientific application codes in X10. This paper reports our experiences writing three codes from the chemistry/material science domain: Fast Multipole Method (FMM), Particle Mesh Ewald (PME) and Hartree-Fock (HF), entirely in X10. Performance results are presented for up to 256 places on a Blue Gene/P system.

During the course of this work our experiences have been shared with the X10 development team, so that application requirements could inform language design discussions as the language capabilities influenced algorithm design. This resulted in improvements in the language implementation and standard class libraries, including the design of the array API and support for complex math.

Data constructs in X10 such as *places* and *distributed arrays*, and parallel constructs such as *finish* and *async*, simplify implementation of the applications in comparison with MPI. However, current implementation limitations in X10 2.1.2 make it difficult to achieve scalable performance using the most natural expressions of the algorithms. The most serious limitation is the use of point-to-point communication patterns, rather than collectives, to implement parallel constructs and array operations. This issue will be addressed in future releases of X10.

Implementation and Performance Evaluation of the HPC Challenge Benchmarks in Coarray Fortran 2.0

Guohua Jin, Laksono Adhianto, John Mellor-Crummey, William N. Scherer III, Chaoran Yang
Rice University, Department of Computer Science
{jin, laksono, johnmc, scherer, chaoran}@rice.edu

Abstract

Today's largest supercomputers have over two hundred thousand CPU cores and even larger systems are under development. Typically, these systems are programmed using message passing. Over the past decade, there has been considerable interest in developing simpler and more expressive programming models for them. Partitioned global address space (PGAS) languages are viewed as perhaps the most promising alternative.

In this paper, we report on our experience developing a set of PGAS extensions to Fortran that we call Coarray Fortran 2.0 (CAF 2.0). Our design for CAF 2.0 goes well beyond the original 1998 design of Coarray Fortran (CAF) by Numrich and Reid. CAF 2.0 includes language support for many features including teams, collective communication, asynchronous communication, function shipping, and synchronization. We describe the implementation of these features and our experiences using them to implement the High Performance Computing Challenge (HPCC) benchmarks, including High Performance Linpack (HPL), RandomAccess, Fast Fourier Transform (FFT), and STREAM triad. On 4096 CPU cores of a Cray XT with 2.3 GHz single socket quad-core Opteron processors, we achieved 18.3 TFLOP/s with HPL, 2.01 GUP/s with RandomAccess, 125 GFLOP/s with FFT, and a bandwidth of 8.73 TByte/s with STREAM triad.

Communication Optimizations for Distributed-Memory X10 Programs

Rajkishore Barik
Intel Corporation, Santa Clara, CA
rajkishore.barik@intel.com

Jisheng Zhao
Rice University, Houston, TX
jisheng.zhao@rice.edu

David Grove
IBM T.J Watson Research Center, NY
groved@us.ibm.com

Igor Peshansky
IBM T.J Watson Research Center, NY
igorp@us.ibm.com

Zoran Budimlić
Rice University, Houston, TX
zoran@rice.edu

Vivek Sarkar
Rice University, Houston, TX
vsarkar@rice.edu

Abstract

X10 is a new object-oriented PGAS (Partitioned Global Address Space) programming language with support for distributed asynchronous dynamic parallelism that goes beyond past SPMD message-passing models such as MPI and SPMD PGAS models such as UPC and Co-Array Fortran. The concurrency constructs in X10 make it possible to express complex computation and communication structures with higher productivity than other distributed-memory programming models. However, this productivity often comes at the cost of high performance overhead when the language is used in its full generality.

This paper introduces high-level compiler optimizations and transformations to reduce communication and synchronization overheads in distributed-memory implementations of X10 programs. Specifically, we focus on locality optimizations such as scalar replacement and task localization, combined with supporting transformations such as loop distribution, scalar expansion, loop tiling, and loop splitting. We have completed a prototype implementation of these high-level optimizations, and performed a performance evaluation that shows significant improvements in performance, scalability, communication volume and number of tasks. We evaluated the communication optimizations on three platforms: a 128-node BlueGene/P cluster, a 32-node Nehalem cluster, and a 16-node Power7 cluster. On the BlueGene/P cluster, we observed a maximum performance improvement of 31.46x relative to the unoptimized case (for the MolDyn benchmark). On the Nehalem cluster, we observed a maximum performance improvement of 3.01x (for the NQueens benchmark) and on the Power7 cluster, we observed a maximum performance improvement of 2.73x (for the MolDyn benchmark). In addition, there was no case in which the optimized code was slower than the unoptimized case. We also believe that the optimizations presented in this paper will be necessary for any high-productivity PGAS language based on modern object-oriented principles, that is designed for execution on future Extreme Scale systems that place a high premium on locality improvement for performance and energy efficiency.

Session 25
Algorithms for Distributed Computing

I/O-Optimal Distribution Sweeping on Private-Cache Chip Multiprocessors

Deepak Ajwani
Centre for Unified Computing
University College Cork
Cork, Ireland
d.ajwani@cs.ucc.ie

Nodari Sitchinava
MADALGO
Department of Computer Science
University of Aarhus
Aarhus, Denmark
nodari@madalgo.au.dk

Norbert Zeh
Faculty of Computer Science
Dalhousie University
Halifax, Canada
nzeh@cs.dal.ca

Abstract

The parallel external memory (PEM) model has been used as a basis for the design and analysis of a wide range of algorithms for private-cache multi-core architectures. As a tool for developing geometric algorithms in this model, a parallel version of the I/O-efficient distribution sweeping framework was introduced recently, and a number of algorithms for problems on axis-aligned objects were obtained using this framework. The obtained algorithms were efficient but not optimal. In this paper, we improve the framework to obtain algorithms with the optimal I/O complexity of $O(sort_P(N) + K/PB)$ for a number of problems on axis-aligned objects; P denotes the number of cores/processors, B denotes the number of elements that fit in a cache line, N and K denote the sizes of the input and output, respectively, and $sort_P(N)$ denotes the I/O complexity of sorting N items using P processors in the PEM model.

To obtain the above improvement, we present a new one-dimensional batched range counting algorithm on a sorted list of ranges and points that achieves an I/O complexity of $O((N + K)/PB)$, where K is the sum of the counts of all the ranges. The key to achieving efficient load balancing among the processors in this algorithm is a new method to count the output without enumerating it, which might be of independent interest.

A Fast Algorithm for Constructing Inverted Files on Heterogeneous Platforms

Zheng Wei and Joseph JaJa
Department of Electrical and Computer Engineering
Institute for Advanced Computer Studies, University of Maryland
College Park, U. S. A
{zwei, joseph}@umiacs.umd.edu

Abstract

Given a collection of documents residing on a disk, we develop a new strategy for processing these documents and building the inverted files extremely fast. Our approach is tailored for a heterogeneous platform consisting of a multicore CPU and a highly multithreaded GPU. Our algorithm is based on a number of novel techniques including: (i) a high-throughput pipelined strategy that produces parallel parsed streams that are consumed at the same rate by parallel indexers; (ii) a hybrid trie and B-tree dictionary data structure in which the trie is represented by a table for fast look-up and each B-tree node contains string caches; (iii) allocation of parsed streams with frequent terms to CPU threads and the rest to GPU threads so as to match the throughput of parsed streams; and (iv) optimized CUDA indexer implementation that ensures coalesced memory accesses and effective use of shared memory. We have performed extensive tests of our algorithm on a single node (two Intel Xeon X5560 Quad-core) with two NVIDIA Tesla C1060 attached to it, and were able to achieve a throughput of more than 262 MB/s on the ClueWeb09 dataset. Similar results were obtained for widely different datasets. The throughput of our algorithm is superior to the best known algorithms reported in the literature even when compared to those run on large clusters.

Graph Partitioning with Natural Cuts

Daniel Delling*, Andrew V. Goldberg* , Ilya Razenshteyn*§ , and Renato F. Werneck*
* Microsoft Research Silicon Valley
Mountain View, CA, 94043, USA
Email: {dadellin,goldberg,renatow}@microsoft.com
§ Mathematics Department, Logic and Algorithms Theory Division
Lomonosov Moscow State University, Russia
Email: ilyaraz@gmail.com

Abstract

We present a novel approach to graph partitioning based on the notion of *natural cuts*. Our algorithm, called PUNCH, has two phases. The first phase performs a series of minimum-cut computations to identify and contract dense regions of the graph. This reduces the graph size, but preserves its general structure. The second phase uses a combination of greedy and local search heuristics to assemble the final partition. The algorithm performs especially well on road networks, which have an abundance of natural cuts (such as bridges, mountain passes, and ferries). In a few minutes, it obtains the best known partitions for continental-sized networks, significantly improving on previous results.

Reader Activation Scheduling in Multi-Reader RFID Systems: A Study of General Case

Shaojie Tang* , Cheng Wang†‡ , Xiang-Yang Li* , Changjun Jiang†‡
Department of Computer Science, Illinois Institute of Technology, Chicago, IL, 60616
† Department of Computer Science, Tongji University, Shanghai, China
‡ Key Laboratory of Embedded System and Service Computing, Ministry of Education, Shanghai, China

Abstract

Radio frequency identification (RFID) is a technology where a reader device can "sense" the presence of a close by object by reading a tag device attached to the object. To guarantee the coverage quality, multiple RFID readers can be deployed in the given region. In this paper, we consider the problem of activation schedule for readers in a multi-reader environment. In particular, we try to design a schedule for readers to maximize the number of served tags per time-slot while avoiding various interferences. We first develop a centralized algorithm under the assumption that different readers may have different interference and interrogation radius. Next, we propose a novel algorithm which does not need any location information of the readers. Finally, we extend the previous algorithm in distributed manner in order to suit the case where no central entity exists. We conduct extensive simulations to study the performances of our proposed algorithm. And our evaluation results corroborate our theoretical analysis.

Session 26
Scheduling

Efficient Parallel Scheduling of Malleable Tasks

Peter Sanders
Department of Informatics
Karlsruher Institut für Technologie
Karlsruhe, Germany
sanders@kit.edu

Jochen Speck
Department of Informatics
Karlsruher Institut für Technologie
Karlsruhe, Germany
speck@kit.edu

Abstract

We give an $O(n + \min\{n, m\} \log m)$ work algorithm for scheduling n tasks with flexible amount of parallelism on m processors, provided the speedup functions of the tasks are concave. We give efficient parallelizations of the algorithm that run in polylogarithmic time. Previous algorithms were sequential and required quadratic work. This is in some sense a best-possible result since the problem is NP-hard for more general speedup functions.

Offline Scheduling of Multi-Threaded Request Streams on a Caching Server

Veronika Rehn-Sonigo
Franche-Comté University
Besançon, France
Email: veronika.sonigo@lifc.univ-fcomte.fr

Denis Trystram
Institut Universitaire de France
Grenoble University, France
Email: denis.trystram@imag.fr

Frédéric Wagner
Grenoble University
France
Email: frederic.wagner@imag.fr

Haifeng Xu, Guochuan Zhang
Zhejiang University
Hanghzou, China
Email: {xuhaifeng, zgc}@zju.edu.cn

Abstract

In this work, we are interested in the problem of satisfying multiple concurrent requests submitted to a computing server. Informally, there are users each sending a sequence of requests to the server. The requests consist of tasks linked by precedence constraints. Tasks may occur several times in the same sequence as well as in a request sequence of another user. The computing server has to execute tasks with variable processing times. The server owns a cache of limited size where intermediate results of the processing may be stored. If an intermediate result for a task is stored into the cache, no processing cost has to be paid and the result can directly be fetched from the cache.

The goal of this work is to determine a schedule of the tasks such that an optimization function is minimized (the only objective studied up to now is the makespan). This problem is a variant of caching which considers only one sequence of requests. We then extend the study to the minimization of the mean completion time of the request sequences. Two models are considered. In the first model, caching is forced whereas in the second model caching is optional and one can choose whether an intermediate result is stored in the cache or not. All combinations turn out to be NP-hard for fixed cache sizes and we provide a formulation as dynamic program as well as bounds for inapproximation. We propose polynomial time approximation algorithms for some variants and analyze their approximation ratios. Finally, we also devise some heuristics and present experimental results.

Tight Analysis of Relaxed Multi-Organization Scheduling Algorithms

Daniel Cordeiro[*‡] , Pierre-François Dutot[‡] , Grégory Mounié[‡] , and Denis Trystram[†‡]
[*] Instituto de Matemática e Estatística
Universidade de São Paulo
São Paulo, Brazil
[†] Institut universitaire de France
Paris, France
[‡] LIG, Grenoble University
Montbonnot Saint-Martin, France
{cordeiro,pfdutot,mounie,trystram}@imag.fr

Abstract

The goal of this paper is to study how limited cooperation can impact the quality of the schedule obtained by multiple independent organizations in a typical grid computing platform. This relaxed version of the problem known as the Multi-Organization Scheduling Problem (MOSP) models an environment where organizations providing both resources and jobs tolerate a bounded degradation on the makespan of their own jobs in order to minimize the makespan over the entire platform.

More precisely, the technical contributions are the following. First, we improve the existing inapproximation bounds for this problem proving that what was previously though as not polynomially approximable (*unless P = NP*) is actually not approximable at all. We achieve this using two families of instances whose Pareto optimal solutions are on par with the previous inaproximability bounds.

Then, we present two algorithms that solve the problem with approximation ratios of (2; 3/2) and (3; 4/3) respectively. This means that when using the first (second) algorithm, if an organization tolerates that the completion time of its last job cannot exceed twice (three times) the time it would have obtained by itself, then the algorithm provides a solution that is a 3/2-approximation (4/3-approximation) for the optimal global makespan. Both algorithms are efficient since their performance ratio correspond to the Pareto optimal solutions of the previously defined instances.

Scheduling Functionally Heterogeneous Systems with Utilization Balancing

Yuxiong He, Jie Liu
Microsoft Research
yuxhe;liuj@microsoft.com

Hongyang Sun
Nanyang Technological University
sunh0007@ntu.edu.sg

Abstract

Heterogeneous systems become popular in both client and cloud. A parallel program can incur operations on multiple processing resources such as CPU, GPU, and vector processor units. This paper investigates scheduling problems on functionally heterogeneous systems with the objective of minimizing the completion time of parallel jobs. We first present performance bounds of online scheduling and show that any online algorithm is at best around (K + 1)-competitive with respect to job completion time, where K is the total number of resource types. There exist "bad" jobs that prevent any online algorithms from obtaining good interleaving of heterogeneous tasks. This lower bound suggests that the relative performance of online algorithms versus an offline optimal could degrade linearly as types of heterogeneous resources increase. The limitation of online scheduling motivates our study of how additional offline or lookahead information can help improve scheduling performance. We propose a Multi-Queue Balancing algorithm (MQB) that effectively transforms the problem of minimizing completion time to one of maximizing utilization of heterogeneous resources. It promotes interleaving of heterogeneous tasks through balancing the task queues of different types. Our simulation results suggest that MQB reduces the execution time of online greedy algorithms up to 40% over various workloads and outperforms other offline schemes in most cases. Furthermore, MQB can use limited and approximated offline information to improve scheduling decisions.

Session 27
Computational Biology and Simulations

Smith-Waterman Alignment of Huge Sequences with GPU in Linear Space

Edans Flavius de Oliveira Sandes, Alba Cristina Magalhaes Alves de Melo
Department of Computer Science
University of Brasilia (UnB)
Brasilia, Brazil
{edans,albamm}@cic.unb.br

Abstract

Cross-species chromosome alignments can reveal ancestral relationships and may be used to identify the peculiarities of the species. It is thus an important problem in Bioinformatics. So far, aligning huge sequences, such as whole chromosomes, with exact methods has been regarded as unfeasible, due to huge computing and memory requirements. However, high performance computing platforms such as GPUs are being able to change this scenario, making it possible to obtain the exact result for huge sequences in reasonable time. In this paper, we propose and evaluate a parallel algorithm that uses GPU to align huge sequences, executing the Smith-Waterman algorithm combined with Myers-Miller, with linear space complexity. In order to achieve that, we propose optimizations that are able to reduce significantly the amount of data processed and that enforce full parallelism most of the time. Using the GTX 285 Board, our algorithm was able to produce the optimal alignment between sequences composed of 33 Millions of Base Pairs (MBP) and 47 MBP in 18.5 hours.

Accelerating Protein Sequence Search in a Heterogeneous Computing System

Shucai Xiao* , Heshan Lin† , and Wu-chun Feng*†

* Department of Electrical and Computer Engineering
† Department of Computer Science
Virginia Tech
Blacksburg, Virginia 24061
Email: {shucai, hlin2, wfeng}@vt.edu

Abstract

The "Basic Local Alignment Search Tool" (BLAST) is arguably the most widely used computational tool in bioinformatics. However, the computational power required for routine BLAST analysis has been outstripping Moore's Law due to the exponential growth in the size of the genomic sequence databases that BLAST searches on.

To address the above issue, we propose the design and optimization of the BLAST algorithm for searching protein sequences (i.e., BLASTP) in a heterogeneous computing system. The end result is a BLASTP implementation that delivers a seven-fold speedup over the sequential BLASTP for the most computationally intensive phase (i.e., hit detection and ungapped extension) on a NVIDIA Fermi C2050 GPU. In addition, when pipelining the processing on a dual-core CPU and the NVIDIA Fermi GPU, our implementation can achieve a six-fold speedup for the overall program execution.

Parallel Metagenomic Sequence Clustering via Sketching and Maximal Quasi-clique Enumeration on Map-reduce Clouds

Xiao Yang[1], Jaroslaw Zola[1] and Srinivas Aluru[1,2]
[1]Department of Electrical and Computer Engineering
Iowa State University, Ames, IA, USA
[2]Department of Computer Science and Engineering
Indian Institute of Technology Bombay, Mumbai, India
Email: {xyang,zola,aluru}@iastate.edu

Abstract

Taxonomic clustering of species is an important and frequently arising problem in metagenomics. High-throughput next generation sequencing is facilitating the creation of large metagenomic samples, while at the same time making the clustering problem harder due to the short sequence length supported and unknown species sampled. In this paper, we present a parallel algorithm for hierarchical taxonomic clustering of large metagenomic samples with support for overlapping clusters. We adapt the sketching techniques originally developed for web document clustering to deduce significant similarities between pairs of sequences without resorting to expensive all vs. all alignments. We formulate the metagenomics classification problem as that of maximal quasi-clique enumeration in the resulting similarity graph, at multiple levels of the hierarchy as prescribed by different similarity thresholds. We cast execution of the underlying algorithmic steps as applications of the map-reduce framework to achieve a cloud based implementation. Apart from solving an important problem in metagenomics, this work demonstrates the applicability of map-reduce framework in relatively complicated algorithmic settings.

Large-scale lattice gas Monte Carlo simulations for the generalized Ising model

Tobias C. Kerscher and Stefan Müller
Institut für Keramische Hochleistungswerkstoffe
Technische Universität Hamburg-Harburg
21073 Hamburg,Germany

Quinn O. Snell and Gus L. W. Hart
Brigham Young University
Provo, Utah 84602

Abstract

We present an efficient parallel algorithm for lattice gas Monte Carlo simulations in the framework of an Ising model that allows arbitrary interaction on any lattice, a model often called a cluster expansion. Thermodynamic Monte Carlo simulations strive for the equilibrium properties of a system by exchanging atoms over a long range, while preserving detailed balance. This long-range exchange of atoms renders other frequent parallelization techniques, like domain decomposition, unfavorable due to excessive communication cost. Our ansatz, based on the Metropolis algorithm, minimizes communication between parallel processes. We present this new "partial sequence preserving" (PSP) algorithm, as well as benchmark data for a physical alloy system (NiAl) comprised of one billion atoms.

Session 28
Cloud Computing

CATCH: A Cloud-based Adaptive Data Transfer Service for HPC

Henry M. Monti, Ali R. Butt
Dept. of Computer Science
Virginia Tech
Email: {hmonti, butta}@cs.vt.edu

Sudharshan S. Vazhkudai
Computer Science and Mathematics Division
Oak Ridge National Laboratory
vazhkudaiss@ornl.gov

Abstract

Modern High Performance Computing (HPC) applications process very large amounts of data. A critical research challenge lies in transporting input data to the HPC center from a number of distributed sources, e.g., scientific experiments and web repositories, etc., and offloading the result data to geographically distributed, intermittently available end-users, often over under-provisioned connections. Such end-user data services are typically performed using point-to-point transfers that are designed for well-endowed sites and are unable to reconcile the center's resource usage and users' delivery deadlines, unable to adapt to changing dynamics in the end-to-end data path and are not fault-tolerant. To overcome these inefficiencies, decentralized HPC data services are emerging as viable alternatives. In this paper, we develop and enhance such distributed data services by designing CATCH, a Cloud-based Adaptive data Transfer serviCe for HPC. CATCH leverages a bevy of cloud storage resources to orchestrate a decentralized data transport with fail-over capabilities. Our results demonstrate that CATCH is a feasible approach, and can help improve the data transfer times at the HPC center by as much as 81.1% for typical HPC workloads.

A Scalable and Elastic Publish/Subscribe Service

Ming Li, Fan Ye, Minkyong Kim, Han Chen, Hui Lei
IBM T.J. Watson Research Center,
Hawthorne NY 10532, USA
Email: {liming, fanye, minkyong, chenhan, hlei}@us.ibm.com

Abstract

The rapid growth of sense-and-respond applications and the emerging cloud computing model present a new challenge: providing publish/subscribe as a scalable and elastic cloud service. This paper presents the BlueDove attribute based publish/subscribe service that seeks to address such a challenge. BlueDove uses a gossip-based one-hop overlay to organize servers into a scalable cluster. It proactively exploits skewness in data distribution to achieve high performance. By assigning each subscription to multiple servers through a multidimensional subscription space partitioning technique, it provides multiple candidate servers for each publication message. A message can be matched on any of its candidate servers with one hop forwarding. The performance-aware forwarding in BlueDove ensures that the message is sent to the least loaded candidate server for processing, leading to low latency and high throughput. The evaluation shows that BlueDove has a linear capacity increase as the system scales up, adapts to sudden workload changes within tens of seconds, and achieves multifold higher throughput than the techniques used in the existing enterprise and peer-to-peer pub/sub systems.

Moving the Code to the Data - Dynamic Code Deployment using ActiveSpaces

Ciprian Docan and Manish Parashar
Rutgers University
Piscataway, New Jersey, USA
E-mail: {docan,parashar}@cac.rutgers.edu

Julian Cummings
California Institute of Technology
Pasadena, California, USA
E-mail: cummings@cacr.caltech.edu

Scott Klasky
Oak Ridge National Laboratory
Oak Ridge, Tennessee, USA
E-mail: sklasky@ornl.gov

Abstract

Managing the large volumes of data produced by emerging scientific and engineering simulations running on leadership-class resources has become a critical challenge. The data has to be extracted off the computing nodes and transported to consumer nodes so that it can be processed, analyzed, visualized, archived, etc. Several recent research efforts have addressed data-related challenges at different levels. One attractive approach is to offload expensive I/O operations to a smaller set of dedicated computing nodes known as a staging area. However, even using this approach, the data still has to be moved from the staging area to consumer nodes for processing, which continues to be a bottleneck. In this paper, we investigate an alternate approach, namely moving the data-processing code to the staging area rather than moving the data. Specifically, we present the ActiveSpaces framework, which provides (1) programming support for defining the data-processing routines to be downloaded to the staging area, and (2) run-time mechanisms for transporting binary codes associated with these routines to the staging area, executing the routines on the nodes of the staging area, and returning the results. We also present an experimen- tal performance evaluation of ActiveSpaces using applications running on the Cray XT5 at Oak Ridge National Laboratory. Finally, we use a coupled fusion application workflow to explore the trade-offs between transporting data and transporting the code required for data processing during coupling, and we characterize the sweet spots for each option.

DryadOpt: Branch-and-Bound on Distributed Data-Parallel Execution Engines

Mihai Budiu, Daniel Delling, and Renato F. Werneck
Microsoft Research Silicon Valley
Mountain View, CA, USA
Email: {mbudiu,dadellin,renatow}@microsoft.com

Abstract

We introduce DryadOpt, a library that enables massively parallel and distributed execution of optimization algorithms for solving hard problems. DryadOpt performs an exhaustive search of the solution space using branch-and-bound, by recursively splitting the original problem into many simpler subproblems. It uses both parallelism (at the core level) and distributed execution (at the machine level). DryadOpt provides a simple yet powerful interface to its users, who only need to implement sequential code to process individual subproblems (either by solving them in full or generating new subproblems). The parallelism and distribution are handled automatically by DryadOpt, and are invisible to the user. The distinctive feature of our system is that it is implemented on top of DryadLINQ, a distributed data-parallel execution engine similar to Hadoop and Map-Reduce. Despite the fact that these engines offer a constrained application model, with restricted communication patterns, our experiments show that careful design choices allow DryadOpt to scale linearly with the number of machines, with very little overhead.