Proceedings of the 24^{th} IEEE International Parallel & Distributed Processing Symposium, Workshops and Phd Forum

IPDPS 2010 Advance Program Abstracts

Abstracts for 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

Workshop 1
Heterogeneity in Computing Workshop
HCW 2010

Characterizing Heterogeneous Computing Environments using Singular Value Decomposition

Abdulla M. Al-Qawasmeh[1], Anthony A. Maciejewski[1] and Howard Jay Siegel[1,2]
[1]Department of Electrical and Computer Engineering
[2]Department of Computer Science
Colorado State University
Fort Collins, Colorado, USA
{Abdulla.Al-Qawasmeh, aam, hj}@colostate.edu

Abstract

We consider a heterogeneous computing environment that consists of a collection of machines and task types. The machines vary in capabilities and different task types are better suited to specific machine architectures. We describe some of the difficulties with the current measures that are used to characterize heterogeneous computing environments and propose two new measures. These measures relate to the aggregate machine performance (relative to the given task types) and the degree of affinity that specific task types have to different machines. The latter measure of task-machine affinity is quantified using singular value decomposition. One motivation for using these new measures is to be able to represent a wider range of heterogeneous environments than is possible with previous techniques. An important application of studying the heterogeneity of heterogeneous systems is predicting the performance of different computing hardware for a given task type mix.

Statistical Predictors of Computing Power in Heterogeneous Clusters

Ron C. Chiang[1], Anthony A. Maciejewski[1], Arnold L. Rosenberg[1,2] and Howard Jay Siegel[1,2]
[1]Electrical and Computer Engineering Department
[2]Computer Science Department
Colorado State University
Fort Collins, CO 80523, USA
{ron.chiang,aam,rsnbrg,hj}@colostate.edu

Abstract

If cluster C_1 consists of computers with a faster mean speed than the computers in cluster C_2, does this imply that cluster C_1 is more productive than cluster C_2? What if the computers in cluster C_1 have the same mean speed as the computers in cluster C_2: is the one with computers that have a higher variance in speed more productive? Simulation experiments are performed to explore the above questions within a formal framework for measuring the performance of a cluster. Simulation results show that both mean speed and variance in speed (when mean speeds are equal) are typically correlated with the performance of a cluster, but not always; these statements are quantified statistically for our simulation environments. In addition, simulation results also show that: (1) If the mean speed of computers in cluster C_1 is faster by at least a threshold amount than the mean speed of computers in cluster C_2, then C_1 is more productive than C_2. (2) If the computers in clusters C_1 and C_2 have the same mean speed, then C_1 is more productive than C_2 when the variance in speed of computers in cluster C_1 is higher by at least a threshold amount than the variance in speed of computers in cluster C_2.

A First Step to the Evaluation of SimGrid in the Context of a Real Application

Abdou Guermouche
LaBRI and Univ. Bordeaux1
Bordeaux, France
Email: Abdou.Guermouche@labri.fr

Hélène Renard
I3S and Univ. Nice/Sophia-Antipolis
Sophia-Antipolis, France
Email: Helene.Renard@polytech.unice.fr

Abstract

Simulation is a "simple" method to experimentally evaluate the behavior of algorithms designed for parallel and distributed platforms. Moreover, the reliability of the evaluation strongly depends on the models used inside the simulator. This paper is devoted to the study and the evaluation of the behavior of the SimGrid simulator through a comparison between a simulated and a real execution of the same target application (i.e. heat propagation). Our target platforms are heterogeneous and their characteristics may dynamically vary during the execution. The obtained results (on a set of various platforms) show that the behavior observed when using the simulated platform is very close to the one obtained on the real one.

Dynamic Adaptation of DAGs with Uncertain Execution Times in Heterogeneous Computing Systems

Qin Zheng
Advanced Computing Programme, Institute of High Performance Computing
Fusionopolis, 1 Fusionopolis Way, 16-16 Connexis, Singapore 138632

Abstract

In this paper, we consider the problem of schedule DAGs with uncertainties in task execution times. Firstly, given an offline planned schedule based on the estimated task execution times, we consider when the schedule should be adapted during runtime based on the current information about the start and completion times of its tasks. The objective is to limit the number of runtime adaptations upon task overruns and underruns and minimize the response time of the DAG. We then consider the case without offline planned schedules and discuss dynamic planning and adaptation of tasks. We conduct extensive simulation experiments to quantify the performance of the proposed algorithms.

Unibus: Aspects of Heterogeneity and Fault Tolerance in Cloud Computing

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

Abstract

The paper describes our on-going project, termed Unibus, in the context of facilitating fault-tolerant executions of MPI applications on computing chunks in the cloud. In general, Unibus focuses on resource access virtualization and automatic, user-transparent resource provisioning that simplify use of heterogeneous resources available to users. In this work, we present the key Unibus concepts (the Capability Model, composite operations, mediators, soft and successive conditionings, metaapplications), and demonstrate how to employ Unibus to orchestrate resources provided by a commercial cloud provider into a fault-tolerant platform, capable of executing message passing applications. In order to support fault tolerance we use DMTCP [1] (Distributed MultiThreaded CheckPointing) that enables checkpointing at the user's level. To demonstrate that the Unibus-created, FT-enabled platform allows to execute MPI applications we ran NAS Parallel Benchmarks and measured the overhead introduced by FT.

Robust Resource Allocation of DAGs in a Heterogeneous Multicore System

Luis Diego Briceño[1], Jay Smith[1,3], Howard Jay Siegel[1,2],
Anthony A. Maciejewski[1], Paul Maxwell[1,4], Russ Wakefield[2],
Abdulla Al-Qawasmeh[1], Ron C. Chiang[1] and Jiayin Li[1]
Colorado State University
[1] Department of Electrical and Computer Engineering
[2] Department of Computer Science
Fort Collins,CO 80523
[3]DigitalGlobe, Longmont,CO,80503
[4]United States Army

Abstract

In this study, we consider an environment composed of a heterogeneous cluster of multicore-based machines used to analyze satellite images. The workload involves large data sets, and is typically subject to deadline constraints. Multiple applications, each represented by a directed acyclic graph (DAG), are allocated to a dedicated heterogeneous distributed computing system. Each vertex in the DAG represents a task that needs to be executed and task execution times vary substantially across machines. The goal of this research is to assign applications to multicore-based parallel system in such a way that all applications complete before a common deadline, and their completion times are robust against uncertainties in execution times. We define a measure that quantifies robustness in this environment. We design, compare, and evaluate two resource allocation heuristics that attempt to maximize robustness.

Decentralized Dynamic Scheduling across Heterogeneous Multi-core Desktop Grids

Jaehwan Lee, Pete Keleher and Alan Sussman
UMIACS and Department of Computer Science
University of Maryland
College Park, MD, USA
{jhlee, keleher, als}@cs.umd.edu

Abstract

The recent advent of multi-core computing environments increases both the heterogeneity and complexity of managing desktop grid resources, making efficient load balancing challenging even for a centralized manager. Even with good initial job assignments, dynamic scheduling is still needed to adapt to dynamic environments, as well as for applications whose running times are not known a priori.

In this paper, we propose new decentralized scheduling schemes that backfill jobs locally and dynamically migrate waiting jobs across nodes to leverage residual resources, while guaranteeing bounded waiting times for all jobs. The methods attempt to maximize total throughput while balancing load across available grid resources. Experimental results via simulation show that our scheduling scheme has performance competitive with an online centralized scheduler.

Custom Built Heterogeneous Multi-Core Architectures (CUBEMACH): Breaking the Conventions

Nagarajan Venkateswaran, Karthikeyan Palavedu Saravanan, Nachiappan Chidambaram Nachiappan
Aravind Vasudevan, Balaji Subramaniam and Ravindhiran Mukundarajan
Waran Research Foundation Chennai, India
www.warftindia.org

Abstract

Increasing computational demand has stirred node architectures to move towards SuperComputer-On-Chips(SCOCs), where computational efficiency is emphasized over peak performance. Suitability of the architecture to a wider class of applications is becoming an pre-eminent design constraint for future HPC systems. This paper explores a novel design paradigm, the Custom Built Heterogeneous Multi-core Architectures (CUBEMACH) for realizing future generation node architectures. The CUBEMACH design flavoring a set of applications offers the possibility of increased resource utilization, which is exploited by running traces of multiple independent applications within a node without time or space sharing. A wide variety of complex Algorithm Level Functional units (ALFUs) besides scalars are used to meet high performance requirements of the grand challenge applications. To cater to the high communication bandwidth requirements across heterogeneous cores comprising these Algorithm Level Functional Units, a novel hierarchical communication backbone structure referred to as the On-Node-Network (ONNET) is used. The demand for high instruction issue rate due to the presence of a large number Algorithm Level Functional Units is catered by an Hardware based Compiler- On-Silicon(COS). The cost-effectiveness is achieved due to the the fact that the CUBEMACH design paradigm helps create an architecture for a single user for executing multiple independent applications without space time sharing.The cost effectiveness of implementing the CUBEMACH Design Paradigm is also achieved by developing SCOC IP cores for Higher Level Functional Us, Compiler-On-Silicon and On-Node-Network.

Improving MapReduce Performance through Data Placement in Heterogeneous Hadoop Clusters

Jiong Xie, Shu Yin, Xiaojun Ruan, Zhiyang Ding, Yun Tian,
James Majors, Adam Manzanares and Xiao Qin
Department of Computer Science and Software Engineering
Auburn University, Auburn, AL 36849-5347
Email: {jzx0009, szy0004, xzr0001, dingzhi, tianyun, majorjh, acm0008}@eng.auburn.edu,
xqin@auburn.edu http://www.eng.auburn.edu/~xqin

Abstract

MapReduce has become an important distributed processing model for large-scale data-intensive applications like data mining and web indexing. Hadoop–an open-source implementation of MapReduce is widely used for short jobs requiring low response time. The current Hadoop implementation assumes that computing nodes in a cluster are homogeneous in nature. Data locality has not been taken into account for launching speculative map tasks, because it is assumed that most maps are data-local. Unfortunately, both the homogeneity and data locality assumptions are not satisfied in virtualized data centers. We show that ignoring the data-locality issue in heterogeneous environments can noticeably reduce the MapReduce performance. In this paper, we address the problem of how to place data across nodes in a way that each node has a balanced data processing load. Given a data-intensive application running on a Hadoop MapReduce cluster, our data placement scheme adaptively balances the amount of data stored in each node to achieve improved data-processing performance. Experimental results on two real data-intensive applications show that our data placement strategy can always improve the MapReduce performance by rebalancing data across nodes before performing a data-intensive application in a heterogeneous Hadoop cluster.

An Empirical Study of a Scalable Byzantine Agreement Algorithm

Olumuyiwa Oluwasanmi and Jared Saia
Department of Computer Science,
University of New Mexico,
Albuquerque, NM 87131-1386.
Email: {muyiwa,saia}@cs.unm.edu

Valerie King
Department of Computer Science, University of Victoria,
P.O. Box 3055, Victoria, BC, Canada V8W 3P6.
Email:val@cs.uvic.ca

Abstract

A recent theoretical result by King and Saia shows that it is possible to solve the Byzantine agreement, leader election and universe reduction problems in the full information model with $\tilde{O}(n^{3/2})$ total bits sent. However, this result, while theoretically interesting, is not practical due to large hidden constants. In this paper, we design a new practical algorithm, based on this theoretical result. For networks containing more than about $1,000$ processors, our new algorithm sends significantly fewer bits than a well-known algorithm due to Cachin, Kursawe and Shoup. To obtain our practical algorithm, we relax the fault model compared to the model of King and Saia by (1) allowing the adversary to control only a $1/8$, and not a $1/3$ fraction of the processors; and (2) assuming the existence of a cryptographic bit commitment primitive. Our algorithm assumes a partially synchronous communication model, where any message sent from one honest player to another honest player needs at most Δ time steps to be received and processed by the recipient for some fixed Δ, and we assume that the clock speeds of the honest players are roughly the same. However, the clocks do not have to be synchronized (i.e., show the same time)

Workshop 2
Reconfigurable Architectures Workshop
RAW 2010

A Configurable-Hardware Document-Similarity Classifier to Detect Web Attacks

Craig Ulmer[1] and Maya Gokhale[2]
[1]Sandia National Laboratories, CA
[2]Lawrence Livermore National Laboratory
cdulmer@sandia.gov, maya@llnl.gov

Abstract

This paper describes our approach to adapting a text document similarity classifier based on the Term Frequency Inverse Document Frequency (TFIDF) metric to reconfigurable hardware. The TFIDF classifier is used to detect web attacks in HTTP data. In our reconfigurable hardware approach, we design a streaming, real-time classifier by simplifying an existing sequential algorithm and manipulating the classifier's model to allow decision information to be represented compactly. We have developed a set of software tools to help automate the process of converting training data to synthesizable hardware and to provide a means of trading off between accuracy and resource utilization. The Xilinx Virtex 5-LX implementation requires two orders of magnitude less memory than the original algorithm. At 166MB/s (80X the software) the hardware implementation is able to achieve Gigabit network throughput at the same accuracy as the original algorithm.

A Configurable High-Throughput Linear Sorter System

Jorge Ortiz
Information and Telecommunication Technology Center
2335 Irving Hill Road
Lawrence, KS
jorgeo@ku.edu

David Andrews
Computer Science and Computer Engineering
The University of Arkansas
504 J. B. Hunt Building, Fayetteville, AR
dandrews@uark.edu

Abstract

Popular sorting algorithms do not translate well into hardware implementations. Instead, hardware-based solutions like sorting networks and linear sorters exploit parallelism to increase sorting efficiency. Linear sorters, built from identical nodes with simple control, have less area and latency than sorting networks, but they are limited in their throughput. We present a system composed of multiple linear sorters acting in parallel in order to increase throughput. Interleaving is used to increase bandwidth and allow sorting of multiple values per clock cycle, and the amount of interleaving and depth of the linear sorters can be adapted to suit specific applications. Implementation of this system into a Field Programmable Gate Array (FPGA) results in a speedup of 68 compared to quicksort running in a MicroBlaze processor.

Hardware Implementation for Scalable Lookahead Regular Expression Detection

Masanori Bando, N. Sertac Artan, Nishit Mehta, Yi Guan and H. Jonathan Chao
Department of Electrical and Computer Engineering
Polytechnic Institute of NYU, Brooklyn, NY

Abstract

Regular Expressions (RegExes) are widely used in various applications to identify strings of text. Their flexibility, however, increases the complexity of the detection system and often limits the detection speed as well as the total number of RegExes that can be detected using limited resources. The two classical detection methods, Deterministic Finite Automaton (DFA) and Non-Deterministic Finite Automaton (NFA), have the potential problems of prohibitively large memory requirements and a large number of concurrent operations, respectively. Although recent schemes addressing these problems to improve DFA and NFA are promising, they are inherently limited by their scalability, since they follow the state transition model in DFA and NFA, where the state transitions occur per each character of the input. We recently proposed a scalable RegEx detection system called Lookahead Finite Automata (LaFA) to solve these problems with three novel ideas: 1. Provide specialized and optimized detection modules to increase resource utilizations. 2. Systematically reordering the RegEx detection sequence to reduce number of concurrent operations. 3. Sharing states among automata for different RegExes to reduce resource requirements. In this paper, we propose an efficient hardware architecture and prototype design implementation based on LaFA. Our proof-of-concept prototype design is built on a fraction of a single commodity Field Programmable Gate Array (FPGA) chip and can accommodate up to twenty-five thousand (25k) RegExes. Using only 7% of the logic area and 25% of the memory on a Xilinx Virtex-4 FX100, the prototype design can achieve 2-Gbps (gigabits-per-second) detection throughput with only one detection engine. We estimate that 34-Gbps detection throughput can be achieved if the entire resources of a state-of-the-art FPGA chip are used to implement multiple detection engines.

A GPU-Inspired Soft Processor for High-Throughput Acceleration

Jeffrey Kingyens and J. Gregory Steffan
Department of Electrical and Computer Engineering, University of Toronto
{kingyen,steffan}@eecg.toronto.edu

Abstract

There is building interest in using FPGAs as accelerators for high-performance computing, but existing systems for programming them are so far inadequate. In this paper we propose a soft processor programming model and architecture inspired by graphics processing units (GPUs) that are well-matched to the strengths of FPGAs, namely highly-parallel and pipelinable computation. In particular, our soft processor architecture exploits multithreading and vector operations to supply a floating-point pipeline of 64 stages via hardware support for up to 256 concurrent thread contexts. The key new contributions of our architecture are mechanisms for managing threads and register files that maximize data-level and instruction-level parallelism while overcoming the challenges of port limitations of FPGA block memories, as well as memory and pipeline latency. Through simulation of a system that (i) supports AMD's CTM r5xx GPU ISA [1], and (ii) is realizable on an XtremeData XD1000 FPGA-based accelerator system, we demonstrate that our soft processor can achieve 100% utilization of the deeply-pipelined floating-point datapath.

A Reconfigurable Architecture for Multicore Systems

Annie Avakian, Jon Nafziger, Amayika Panda and Ranga Vemuri
Department of Electrical and Computer Engineering
University of Cincinnati
Cincinnati, Ohio, USA
{avakiaam,nafzigjw,pandaaa}@mail.uc.edu, ranga.vemuri@uc.edu

Abstract

Various studies concluded that bus-based multiprocessor architectures outperform Network-on-Chip (NoC) architectures when the number of processors is relatively small. On the other hand, NoC architectures offer distinct performance advantages when the number of processors is large. This led to recent proposals for hybrid architectures where each node in a mesh-style packet-switched NoC architecture contains a bus-based subsystem with a small number of processors. Experimental results using select benchmarks demonstrated that these hybrid architectures offer superior performance when compared with purely bus based or purely NoC style architectures. Our studies indicate that while a hybrid architecture is preferable, the optimal number of processors on each bus subsystem varies based on the application. This number appears to vary between 1 and 8 depending on the communication requirements of the application. Further, various applications simultaneously executing on the same system require differing numbers of processors on each bus-based subsystem to minimize the overall throughput time. In this paper, we present a new reconfigurable NoC architecture which allows scalable bus-based multiprocessor subsystems on each node in the NoC. Following configuration, the system provides a multi-bus execution environment where each processor is connected to a bus and the bus-based subsystems communicate via routers connected in a mesh-style configuration. The system can be reconfigured to vary the number of bus subsystems and the number of processors on each subsystem. Each processor contains a Level 1 (L1) cache and each bus, connected to a router, has access to a Level 2 (L2) cache. The L2 caches distributed across the network together form a large virtual L2 that can shared by all the processors in the system via the router network. We present the architecture in detail, discuss a configuration algorithm, and discuss experimental results (using the NS2 and SIMICS simulators) on standard and synthetic benchmarks indicating the performance advantages of the proposed architecture.

A Shared Reconfigurable VLIW Multiprocessor System

Fakhar Anjam, Stephan Wong and Faisal Nadeem
Computer Engineering Laboratory
Delft University of Technology
Delft, The Netherlands
E-mail: {F.Anjam, J.S.S.M.Wong, M.F.Nadeem}@tudelft.nl

Abstract

In this paper, we present the design and implementation of an open-source reconfigurable very long instruction word (VLIW) multiprocessor system. This processor is implemented as a softcore on a field-programmable gate arrays (FPGA) and its instruction set architecture (ISA) is based on the Lx/ST200 ISA. This multiprocessor design is based on our earlier ρ-VEX processor design. Since the ρ-VEX processor is a parameterized processor, our multiprocessor design is also parameterized. By utilizing a freely available compiler and simulator in our development framework, we are able to optimize our design and map any application written in C to our multiprocessor system. This VLIW multiprocessor can exploit data level as well as instruction level parallelism inherent in an application and make its execution faster. More importantly, we achieve our results by saving expensive FPGA area through the sharing of resources. The results show that we can achieve two times better performance for our dual-processor system (with shared resources) compared to a uni-processor system or a 2-cluster processor system for applications having data level and instruction level parallelism.

TLP and ILP exploitation through a Reconfigurable Multiprocessor System

Mateus B. Rutzig[1], Felipe Madruga[1], Marco A. Alves[1], Henrique Cota[1], Antonio C.S.Beck[2], Nicolas Maillard[1], Philippe O. A. Navaux[1], Luigi Carro[1]
[1]Universidade Federal do Rio Grande do Sul, Instituto de Informática – Porto Alegre/Brazil
[2]Universidade Federal de Santa Maria, Departamento de Eletrônica e Computação – Santa Maria/Brazil
{mbrutzig, flmadruga, mazalves, hcfreitas, caco, nicolas, navaux, carro}@inf.ufrgs.br

Abstract

Limits of instruction level parallelism and the higher transistor density sustain the increasing need for multiprocessor systems: they are rapidly taking over both general purpose and embedded processor domains. Nowadays, since these processors must handle a wide range of different application classes, there is no consensus over which are the best hardware solutions to exploit the best of ILP and TLP together. Current multiprocessing systems are composed either of many homogeneous and simple cores, or of complex superscalar SMT processing elements. In this work, we have expanded a reconfigurable architecture to be used in a multiprocessing scenario, showing the need for an adaptable ILP exploitation even in TLP architectures. We have successfully coupled a dynamic reconfigurable system to a SPARC-based multiprocessor, and obtained performance gains of up to 40% even for applications that show a great level of parallelism at thread level, demonstrating the need for an adaptable ILP exploration.

CAP-OS: Operating System for Runtime Scheduling, Task Mapping and Resource Management on Reconfigurable Multiprocessor Architectures

Diana Göhringer[1], Michael Hübner[2], Etienne Nguepi Zeutebouo[1] and Jürgen Becker[2]
[1]Fraunhofer IOSB, Germany
[2]ITIV, Karlsruhe Institute of Technology (KIT) Germany
e-mail: {dgoehringer, zeutebouo}@fom.fgan.de[1], {michael.huebner, becker}@kit.edu[2]

Abstract

Operating systems traditionally handle the task scheduling of one or more application instances on a processor like hardware architecture. Novel runtime adaptive hardware exploits the dynamic reconfiguration on FPGAs, where hardware blocks are generated, started and terminated. This is similar to software tasks in well established operating system approaches. The hardware counterparts to the software tasks have to be transferred to the reconfigurable hardware via a configuration access port. This port enables the allocation of hardware blocks on the FPGA. Current reconfigurable hardware, like e.g. Xilinx Virtex 5 provide two internal configuration access ports (ICAPs), where only one of these ports can be accessed at one point of time. In e.g. a multiprocessor system on an FPGA, it can happen that multiple instances try to access these ports simultaneously. To prevent conflicts, the access to these ports as well as the hardware resource management needs to be controlled by a special purpose operating system running on an embedded processor. This special purpose operating system, called CAPOS (Configuration Access Port-Operating System), which will be presented in this paper, supports the clients using the configuration port with the service of priority-based access scheduling, hardware task mapping and resource management.

PATIS: Using Partial Configuration to Improve Static FPGA Design Productivity

T. Frangieh[1], A. Chandrasekharan[1], S. Rajagopalan[1], Y. Iskander[2], S. Craven[3], C. Patterson[1]
[1]Configurable Computing Lab
ECE Department, Virginia Tech, Blacksburg, VA 24061
{tannous,athira,sureshr,cdp}@vt.edu
[2]Secure Computing & Communications
Luna Innovations Incorporated, 1 Riverside Circle, Suite 400, Roanoke, VA 24016
iskandery@lunainnovations.com
[3]EE Department, University of Tennessee at Chattanooga, Chattanooga, TN 37403
stephen-craven@utc.edu

Abstract

Reconfigurable hardware development and debugging tools aspire to provide software-like productivity. A major impediment, however, is the lack of a module linkage capability permitting hardware blocks to be compiled concurrently, limiting the effective use of multi-core and multiprocessor platforms. Although modular and incremental design flows can reuse the layouts of unmodified blocks, non-local changes to the logical hierarchy or physical layout, or addition of debug circuitry, generally force complete re-implementation. We describe the PATIS dynamic floorplanner, targeting development environments in which some circuit speed and area optimization may be sacrificed for improved implementation and debug turnaround. The floorplan consists of partial modules with structured physical interfaces observable through configuration readback rather than synthesized logic analysis circuitry, allowing module ports to be passively probed without disturbing the layout. Although PATIS supports incremental design, complete re-implementation is still rapid because the partial bitstream for each block is generated by independent and concurrent invocations of the standard Xilinx tools running on separate cores or hosts. A continuous background task proactively generates floorplan variants to accelerate global layout changes. The partial reconfiguration design flow is easier to automate in PATIS because run-time module swapping is not required, suggesting that partial reconfiguration may serve a useful role in large-scale static design.

Wirelength driven floorplacement for FPGA-based partial reconfigurable systems

A. Montone[2], M. D. Santambrogio[1,2], D Sciuto[2]
[1]Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
santambr@mit.edu
[2]Dipartimento di Elettronica e Informazione
Politecnico di Milano
alessio.montone@dresd.org, {santambr, sciuto}@elet.polimi.it

Abstract

The proposed work aims at identifying groups of Reconfigurable Functional Units that are likely to be configured in the same chip area, identifying these areas based on resource requirements, device capabilities and wirelength. The proposed floorplacement framework, tailored for Xilinx Virtex 4 and 5 FPGAs, uses an objective function based on external wirelength, i.e., the estimated length of the nets connecting each Reconfigurable Functional Unit to the corresponding required chip Input Output Blocks. The proposed approach results, as also demonstrated in the experimental results section, in a shorter external wirelength (an average reduction of 50%) with respect to purely area-driven approaches and a highly increased probability of re-use of existing links (90% reduction can be obtained in the best case).

Fast dynamic and partial reconfiguration Data Path with low Hardware overhead on Xilinx FPGAs

Michael Hübner, Diana Göhringer, Juanjo Noguera, Jürgen Becker
Fraunhofer FOM, Germany
Karlsruhe Institute of Technology - KIT, Germany
Xilinx Research Labs, Dublin Ireland
{michael.huebner, becker}@kit.edu, dgoehringer@fom.fgan.de
Juanjo.noguera@xilinx.com

Abstract

Dynamic and partial reconfiguration of Xilinx FPGAs is a well known technique in runtime adaptive system design. With this technique, parts of a configuration can be substituted while other parts stay operative without any disturbance. The advantage is the fact, that the spatial and temporal partitioning can be exploited with the goal to increase performance and to reduce power consumption due to the re-use of chip area. This paper shows a novel methodology for the inclusion of the configuration access port into the data path of a processor core in order to adapt the internal architecture and to re-use this access port as data- sink and source. It is obvious that the chip area, which is utilized by the hardware drivers for the internal configuration access port (ICAP), has to be as small as possible in comparison to the application functionality. Therefore, a hardware design with a small footprint, but with an adequate performance in terms of data throughput, is necessary. This paper presents a fast data path for dynamic and partial reconfiguration data with the advantage of a small footprint on the hardware resources.

High-Level Synthesis Techniques for In-Circuit Assertion-Based Verification

John Curreri, Greg Stitt and Alan D. George
NSF Center for High-Performance Reconfigurable Computing (CHREC)
ECE Department, University of Florida

Abstract

Field-Programmable Gate Arrays (FPGAs) are increasingly employed in both high-performance computing and embedded systems due to performance and power advantages compared to microprocessors. However, widespread usage of FPGAs has been limited by increased design complexity. High-level synthesis has reduced this complexity but often relies on inaccurate software simulation or lengthy register-transfer-level simulations for verification and debugging, which is unattractive to software developers. In this paper, we present high-level synthesis techniques that allow application designers to efficiently synthesize ANSI-C assertions into FPGA circuits, enabling real-time verification and debugging of circuits generated from high-level languages, while executing in the actual FPGA environment. Although not appropriate for all systems (e.g., safety-critical systems), the proposed techniques enable software developers to rapidly verify and debug FPGA applications, while reducing frequency by less than 3% and increasing FPGA resource utilization by less than 0.13% for several application case studies on an Altera Stratix-II EP2S180 using Impulse-C. The presented techniques reduced area overhead by as much as 3x and improved assertion performance by as much as 100% compared to unoptimized in-circuit assertions.

Support of Cross Calls between a Microprocessor and FPGA in CPU-FPGA Coupling Architecture

Giang Nguyen thi Huong
School of Electrical Engineering
Korea University
Seoul 136-701, Republic of Korea
redriver@korea.ac.kr

Seon Wook Kim
School of Electrical Engineering
Korea University
Seoul 136-701, Republic of Korea
seon@korea.ac.kr

Abstract

The coupling architecture containing an FPGA device and a microprocessor has been widely used to accelerate microprocessor execution. Therefore, there have been intensive researches about synthesizing high-level programming languages (HLL) such as C and C++ into HW in the high-level synthesis community in order to make the work of reconfiguring the FPGA easier. However, the difference in a calling method in terms of semantics between HDLs and HLLs makes their interface implementation very difficult. This paper presents a novel communication framework between a microprocessor and FPGA, which allows the full implementation of cross calls between SW and HW and even recursive calls in HW without any limitation. We show that our proposed calling overhead is very small. With our communication framework, hardware components inside the FPGA are no longer isolated accelerators, and they can work as other master components in a system configuration.

An Architectural Space Exploration Tool for Domain Specific Reconfigurable Computing

Gayatri Mehta
Department of Electrical Engineering
University of North Texas
Denton, Texas 76203
Email: gayatri.mehta@unt.edu

Alex K. Jones
Department of Electrical and Computer Engineering
University of Pittsburgh
Pittsburgh, PA 15261
Email: akjones@ece.pitt.edu

Abstract

In this paper, we describe a design space exploration (DSE) tool for domain specific reconfigurable computing where the needs of the applications drive the construction of the device architecture. The tool has been developed to automate the design space case studies which allows application developers to explore architectural tradeoffs efficiently and reach solutions quickly. We selected some of the core signal processing benchmarks from the MediaBench benchmark suite and some of the edge-detection benchmarks from the image processing domain for our case studies. We compare the energy consumption of the architecture selected from manual design space case studies with the architectural solution selected by the design space exploration tool. The architecture selected by the DSE tool consumes approximately 9% less energy on an average as compared to the best candidate from the manual design space case studies. The fabric architecture selected from the manual design case studies and the one selected by the tool were synthesized on 130 nm cell-based ASIC fabrication process from IBM. We compare the energy of the benchmarks implemented onto the fabric with other hardware and software implementations. Both fabric architectures (manual and tool) yield energy within 3X of a direct ASIC implementation, 330X better than a Virtex-II Pro FPGA and 2016X better than an Intel XScale processor.

Memory Architecture Template for Fast Block Matching Algorithms on FPGAs

Shant Chandrakar, Abraham Clements, Arvind Sudarsanam and Aravind Dasu
Electrical and Computer Engineering Department
Utah State University, Logan UT-84322, USA
Email: fshant.chandrakar@aggiemail, aclements@cc,
arvind.sudarsanam@aggiemail, dasu@engineeringg.usu.edu

Abstract

Fast Block Matching (FBM) algorithms for video compression are well suited for acceleration using parallel data-path architecture on Field Programmable Gate Arrays (FPGAs). However, designing an efficient on-chip memory subsystem to provide the required throughput to this parallel data-path architecture is a complex problem. This paper proposes a memory architecture template that is explored using a Bounded Set algorithm to design efficient on-chip memory subsystems for FBM algorithms. The resulting memory subsystems are compared with three existing memory subsystems. Results show that our memory subsystems can provide full parallelism in majority of test cases and can process integer pixels of a 1080p video sequence up to a rate of 275 frames per second.

A Low-Energy Approach for Context Memory in Reconfigurable Systems

Thiago Berticelli Ló, Antonio Carlos S. Beck, Mateus Beck Rutzig and Luigi Carro
Instituto de Informática
Universidade Federal do Rio Grande do Sul
Porto Alegre/RS - Brazil
{tblo, caco, mbrutzig, carro}@inf.ufrgs.br

Abstract

In most of the works concerning reconfigurable computing, the main objective is system optimization by taking into account the known requirements of a project, such as speedup, energy or area. However, as it will be shown in this paper, although very significant, the impact of the context memory is often ignored. Since the context memory is responsible for keeping configurations of the reconfigurable unit, the word size and hence the number of output bits is orders of magnitude larger than the regular memories, considerably increasing the energy consumption and area occupation. Therefore, in this article we propose a technique to handle these issues, while maintaining system performance. Using as case study a coarsegrain architecture tightly coupled to the MIPS R3000 processor, we show that the context memory can represent up to 63% of the total system energy and, by using the proposed approach, it is possible to save 59% of this amount, without any performance penalties.

Efficient Floating-Point Logarithm Unit for FPGAs

Nikolaos Alachiotis and Alexandros Stamatakis
The Exelixis Lab
Dept. of Computer Science
Technische Universitat Munchen
email: {alachiot,stamatak}@in.tum.de

Abstract

As FPGAs become larger, new fabrics, in particular DSPs, allow for a wider range of applications, specifically floating-point intensive codes, to be efficiently executed.

The logarithm is a widely used function in many scientific applications. We present the design of an efficient and sufficiently accurate Logarithm Approximation Unit (LAU) that uses a Look-Up Table (LUT) based approximation, in reconfigurable logic. The LAU has been verified through post place and route simulations, tested on actual FPGA, and is freely available for download. An important property of the LAU architecture is, that it only requires 2% of overall hardware resources on a medium-size FPGA (Xilinx V5SX95T) and thereby allows for easy integration with more complex architectures. Under single precision (SP) the LAU is 11 and 1.6 times faster than the GNU and Intel Math Kernel Library (MKL) implementations and up to 1.44 times faster than the FloPoCo reconfigurable logarithm unit, while occupying slightly less resources. Under double precision (DP) the LAU is 18 and 2.5 times faster than the GNU and Intel MKL implementations and up to 1.66 times faster than the FloPoCo logarithm while occupying significantly less resources.

The LUT-based approximation is sufficiently accurate for our target application and provides a flexible mechanism to adapt the LAU to specific accuracy requirements.

Flexible IP cores for the k-NN classification problem and their FPGA implementation

Elias S. Manolakos and Ioannis Stamoulias
Department of Informatics and Telecommunications, University of Athens
Panepistimioupolis, Ilisia, 15784, Athens, Greece
eliasm@di.uoa.gr

Abstract

The k-nearest neighbor (k-NN) is a popular nonparametric benchmark classification algorithm to which new classifiers are usually compared. It is used in numerous applications, some of which may involve thousands of data vectors in a possibly very high dimensional feature space. For real-time classification a hardware implementation of the algorithm can deliver high performance gains by exploiting parallel processing and block pipelining. We present two different linear array architectures that have been described as soft parameterized IP cores in VHDL. The IP cores are used to synthesize and evaluate a variety of array architectures for a different k-NN problem instances and Xilinx FPGAs. It is shown that we can solve efficiently, using a medium size FPGA device, very large size classification problems, with thousands of reference data vectors or vector dimensions, while achieving very high throughput. To the best of our knowledge, this is the first effort to design flexible IP cores for the FPGA implementation of the widely used k-NN classifier.

Automatic Mapping of Control-Intensive Kernels onto Coarse-Grained Reconfigurable Array Architecture with Speculative Execution

Ganghee Lee, Kyungwook Chang and Kiyoung Choi
Department of Electrical Engineering and Computer Science
Seoul National University
Seoul, South Korea
berean97@snu.ac.kr, kyungwook222@poppy.snu.ac.kr, kchoi@snu.ac.kr

Abstract

Coarse-grained reconfigurable array architectures have drawn increasing attention due to their good performance and flexibility. In general, they show high performance for compute-intensive kernel code, but cannot handle control-intensive parts efficiently, thereby degrading the overall performance. In this paper, we present automatic mapping of control-intensive kernels onto coarse-grained reconfigurable array architecture by using kernel-level speculative execution. Experimental results show that our automatic mapping tool successfully handles control-intensive kernels for coarse-grained reconfigurable array architecture. In particular, it improves the performance of the H.264 deblocking filters for luma and chroma over 26 and 16 times respectively compared to conventional software implementation. Compared to the approach using predicated execution, the proposed approach achieves 2.27 times performance enhancement.

Virtual Area Management: Multitasking on Dynamically Partially Reconfigurable Devices

Josef Angermeier[1], Sándor P. Fekete[2], Tom Kamphans[2], Nils Schweer[2] and Jürgen Teich[1]
[1] Department of Computer Science 12, University of Erlangen-Nuremberg
Erlangen, Germany
[2] Department of Computer Science, Braunschweig University of Technology
Braunschweig, Germany

Abstract

Every year the computing resources available on dynamically partially reconfigurable devices increase enormously. In the near future, we expect many applications to run on a single reconfigurable device. In this paper, we present a concept for multitasking on dynamically partially reconfigurable systems called *virtual area management*. We explain its advantages, show its challenges, and discuss possible solutions. Furthermore, we investigate one problem in more detail: Packing modules with time-varying resource requests. This problem from the reconfigurable computing field results in a completely new optimization problem not tackled before. ILP-based and heuristic approaches are compared in an experimental study and the drawbacks and benefits discussed.

Self-Configurable Architecture for Reusable Systems with Accelerated Relocation Circuit (SCARS-ARC)

Adarsha Sreeramareddy
Department of ECE
University of Arizona
Tucson, AZ
adarshs@email.arizona.edu

Ramachandra Kallam
Department of ECE
Utah State University
Logan, UT
kallam @engineering.usu.edu

Aravind R. Dasu
Department of ECE
Utah State University
Logan, UT
dasu@engineering.usu.edu

Ali Akoglu
Department of ECE
University of Arizona
Tucson, AZ
akoglu@ece.arizona.edu

Abstract

Field Programmable Gate Arrays (FPGAs), with partial reconfiguration (PR) technology present an attractive option for creating reliable platforms that adapt to changes in user objectives over time and respond to hardware/software anomalies automatically with self-healing action. Conventional solutions for partial reconfiguration based self-configurable architectures experience severe hardware limitations on ability to move any partially reconfigurable module to any available region of the reconfigurable fabric and ability to relocate the module quickly. In this study we adopt the hardware-based partial bitstream relocation technique, Accelerated Relocation Circuit (ARC), into the FPGA based wirelessly networked self-configurable architecture that employs traditional module based partial reconfiguration strategy. We show that the integrated architecture allows flexibility for module relocation, reduces the off-chip communication overhead, and observes up to 17x speedup for module relocation over the traditional Xilinx hardware internal configuration access port wrapper (HWICAP) based implementation.

Reconfiguration-aware Spectrum Sharing for FPGA based Software Defined Radio

Hessam Kooti, Elaheh Bozorgzadeh, Shenghui Liao, Lichun Bao
Computer Science Department
University of California, Irvine, CA, USA
{hkooti,eli,shenghul,lbao}@ics.uci.edu

Abstract

This paper focuses on reconfigurable systems for software defined radio applications in which the underlying hardware is dynamically reconfigured for packet processing of multiple protocols. However, due to non-negligible reconfiguration delay overhead, the physical layer may not be able to respond to all the packets scheduled by MAC layer. In this paper, we present a reconfiguration-aware spectrum sharing and spectrum access scheduling at MAC layer. We considered four protocols similar to WiFi, WiMax, GPRS, and WCDMA on a FPGA-based system. Our results show that the optimal solution outperforms the adopted existing heuristic for time slot scheduling by 13.29%.

Implementation of the Compression Function for Selected SHA-3 Candidates on FPGA

A. H. Namin and M. A. Hasan
Department of Electrical and Computer Engineering, University of Waterloo
Waterloo, Ontario N2L 3G1 Canada
Email: {anamin,ahasan}@uwaterloo.ca

Abstract

Implementation of the main building block (compression function) for five different SHA-3 candidates on reconfigurable hardware is presented. The five candidates, namely Blue Midnight Wish, Luffa, Skein, Shabal, and Blake have been considered since they present faster software implementation results compared to the rest of the SHA-3 proposals. The results allow an easy comparison for hardware performance of the candidates.

Improving application performance with hardware data structures

Ravikesh Chandra and Oliver Sinnen
Department of Electrical and Computer Engineering
The University of Auckland, New Zealand
{r.chandra, o.sinnen}@auckland.ac.nz

Abstract

Contemporary processors are becoming wider and more parallel. Thus developers must work hard to extract performance gains. An alternative computing paradigm is to use FPGA technology in a reconfigurable computing environment–where both software and hardware can be specified. This has the potential to realise substantial performance gains in a variety of applications, however it is a daunting task as hardware development is required to harness the benefits. In this research the acceleration of common data structuresCwith the priority queue (PQ) as a case studyChas been explored in the context of such a reconfigurable computing environment. A Java-based hybrid hardware/software PQ has been developed that is a 'drop-in' replacement for a software implementation; achieved by strictly adhering to the same programming interface. The accelerated PQ has demonstrated up to 3x speedup when performing a minimum spanning tree graph computation. Taking this further, a suite of accelerated data structures represents an attractive way for developers to harness the potential of reconfigurable computing in the future across a wide gamut of application domains.

Adaptive Traffic Scheduling Techniques for Mixed Real-Time and Streaming Applications on Reconfigurable Hardware

Tobias Ziermann and Juergen Teich
Hardware/Software Co-Design, Department of Computer Science, University of Erlangen-Nuremberg

Abstract

With the ongoing development of new FPGA generations, the reconfiguration time decreases and therefore the benefit of runtime reconfiguration increases. In this paper, we describe how to use runtime reconfiguration to improve the efficiency of transmitting streaming data on a communication channel shared with real-time applications. This means, the bandwidth that the streaming data has available is dynamically changing. To use the bandwidth effectively, different modules can be loaded on the reconfigurable hardware. These modules have a tradeoff between bandwidth and area requirements. The target now is to find an optimal reconfiguration schedule that minimizes an objective function consisting of two conflicting objectives: reducing the average area needed and providing a certain quality of transmission. In this paper, a model for this scheduling problem is presented and an Integer Linear Programming (ILP) formulation is introduced to calculate an optimal offline solution for benchmarking.

In addition, an online scheduling system is presented. It uses the current delay of the streaming application to calculate the schedule. Extensive simulations have been made to show the benefits of the proposed solution.

Reconfigurable Architecture for Mathematical Morphology Using Genetic Programming and FPGAs

Emerson Carlos Pedrino and Osmar Ogashawara
Department of Computer Science
Federal University of São Carlos
Rod. Washington Luís, km235, SC, SP, Brazil
CEP: 13565905, Caixa Postal: 676
Tel: 55 16 33518232, Fax: 55 16 33518233
email: emerson@dc.ufscar.br,
osmaroga@ufscar.br

Valentin Obac Roda
Department of Electrical Engineering
University of São Paulo
Av. Trabalhador São-carlense, 400, SC, SP,
Brazil
CEP: 13566590
Tel/Fax: 55 16 33739335
email: valentin@sel.eesc.usp.br

Abstract

The task of designing manually morphological operators for a given application is not always a trivial one. Genetic programming is a branch of evolutionary computing and it is consolidating as a promising method for applications of digital image processing. The main objective of genetic programming is to discover how computers can learn to solve problems without being programmed for that. In the literature little has been found about the automatic morphological construction of operators using genetic programming. In this paper, it's presented an original architecture implemented in a FPGA for classical mathematical morphological (binary and gray level) operations that are generated automatically by a genetic programming approach. The possible applications for the system are: pattern recognition and emulation of simple filters, to name just a few. Practical examples using the developed system are presented.

MU-Decoders: A Class of Fast and Efficient Configurable Decoders

Matthew C. Jordan
Quadrus Corporation
Huntsville, Alabama, USA
mjordan@quadruscorp.com

Ramachandran Vaidyanathan
Department of Electrical & Computer Engineering
Louisiana State University, Baton Rouge, Louisiana, USA
vaidy@ece.lsu.edu

Abstract

A decoder is a hardware module that expands an x-bit input into an n-bit output, where $x \ll n$. It can be viewed as producing a set $\mathcal{P}$ of subsets of an n-element set Z_n. If this set $\mathcal{P}$ can be altered by the user, the decoder is said to be configurable. In this paper we propose a class of configurable decoders (called "mapping-unit" based decoders or simply MU-decoders) that facilitate efficient selection of elements in an FPGA(in general, in any chip). Current solutions for this selection use either (a) a fixed (non-reconfigurable) decoder that lacks the flexibility to generate many subsets quickly, or (b) a large look-up table (LUT) which is flexible, but too expensive. The proposed class of MU-decoders offers a range of trade-offs between flexibility of subset generation and cost. Specifically, we show that for any fixed order of gate cost, the MU-decoder can produce any set of subsets that the LUT decoder can; in addition, the MU-decoder can exploit any available structure in the application at hand to produce many more subsets than the LUT decoder. We illustrate this ability in the context of totally ordered sets of subsets.

Analysis and validation of partially dynamically reconfigurable architecture based on Xilinx FPGAs

M. D. Santambrogio[1,2], P. R. Grassi[2], D. Candiloro[2] and Donatella Sciuto[2]
[1]Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
santambr@mit.edu
[2]Dipartimento di Elettronica e Informazione
Politecnico di Milano
santambr, sciuto@elet.polimi.it, paolo.grassi, davide.candiloro@dresd.org

Abstract

Partial dynamic reconfiguration of FPGAs is a methodology that allows the efficient use of an FPGAs resources and an improved degree of flexibility with respect to static hardware when designing an architecture on FPGA. However, due to the number of technology dependent constraints to be satisfied and to the lack of a complete automatize design flow, which implies that several manual operations still need to be done in the respect of those constraints, the design of a partial dynamic reconfigurable system can be seen as a complex task to be accomplished. We introduce a methodology for the low level configuration analysis and for the debugging and validation of the bitstream files of architectures exploiting dynamic partial reconfiguration on Xilinx FPGAs. The proposed methodology has been validated using different Xilinx FPGAs Spartan 3, Virtex II Pro and Virtex 4. The methodology has subsequently yielded a framework in which the validation and debugging techniques have been implemented. The framework, called Rebit, supports the designer in validating a partial dynamic reconfigurable architecture, debugging the user-defined constraints for the reconfigurable regions of the design and validating partial bitstream occupation data against those defined reconfigurable regions.

Stack Protection Unit as a step towards securing MPSoCs

Slobodan Lukovic, Paolo Pezzino, Leandro Fiorin
University of Lugano
Lugano, Switzerland
{lukovics, pezzinop, fiorin}@alari.ch

Abstract

Reconfigurable technologies are getting popular as an instrument not only for verification and prototyping but also for commercial implementation of Multi-Processor System-on-Chip (MPSoC) architectures. These systems, in particular Network-on-Chip (NoC) based ones, have emerged as a design strategy to cope with increased requirements and complexity of modern applications. However, the increasing heterogeneity, coupled with possibility of reconfiguration, makes security become one of major concerns in MPSoC design. In this work, we show a solution for FPGA based designs against one of the most widespread types of attacks - code injection. Our response to tackle this challenge is given in form of Stack Protection Unit (SPU) embedded into processing cores. MicroBlaze soft-core processor serves as a case study for verification of the proposed solution in FPGA technology.

Fast Smith-Waterman hardware implementation

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

H. Ekin Sümbül
Electronic Engineering Program
Sabancı University
Turkey
ekinsumbul@su.sabanciuniv.edu

Abstract

The Smith-Waterman (SW) algorithm is one of the widely used algorithms for sequence alignment in computational biology. With the growing size of the sequence database, there is always a need for even faster implementation of SW. In this paper, we have implemented two Recursive Variable Expansion (RVE) based techniques, which are proved to give better speedup than any best dataflow approach at the cost of extra area. Compared to dataflow approach, our HW implementation is 2.29 times faster at the expense of 2.82 times more area.

Workshop 3

Workshop on High-Level Parallel Programming Models & Supportive Environments

HIPS 2010

The Gozer Workflow System

Jason Madden[1], Nicolas G. Grounds[1], Jay Sachs[1] and John K. Antonio[2]
[1]RiskMetrics Group, 201 David L. Boren Blvd, Suite 300, Norman, OK, USA
[2]School of Computer Science, University of Oklahoma, Norman, OK, USA

Abstract

The Gozer workflow system is a production workflow authoring and execution platform that was developed at RiskMetrics Group. It provides a high-level language and supporting libraries for implementing local and distributed parallel processes. Gozer was developed with an emphasis on distributed processing environments in which workflows may execute for hours or even days. Key features of Gozer include: implicit parallelization that exploits both local and distributed parallel resources; survivability of system faults/shutdowns without losing state; automatic distributed process migration; and implicit resource management and control. The Gozer language is a dialect of Lisp, and the Gozer system is implemented on a service-oriented architecture.

Static Macro Data Flow: Compiling Global Control into Local Control

Pritish Jetley and Laxmikant V. Kalé,
Department of Computer Science,
University of Illinois at Urbana-Champaign, USA
{pjetley2, kale}@illinois.edu

Abstract

The expression of parallel codes through abstract, high-level specifications of global control and data flow can greatly simplify the task of creating large parallel programs. We discuss the challenges of compiling such global flows into the behavioral descriptions of individual component objects in an SPMD environment. We present our work in the context of Charisma, a language that describes global data and control flow through a simple script-like language. Inter-object interactions are realized through the *production* and *consumption* of data. The compiler infers communication patterns between objects and generates appropriate messaging code. We discuss the productivity and performance benefits of compiling such global specifications into local descriptions of control flow embodied by a language called Structured Dagger (SDAG).

False Conflict Reduction in the Swiss Transactional Memory (SwissTM) System

Aravind Natarajan
Department of Computer Engineering
The University of Texas at Dallas
Richardson, TX - 75080, USA
Email: aravindn@utdallas.edu

Neeraj Mittal
Department of Computer Science
The University of Texas at Dallas
Richardson, TX - 75080, USA
Email: neerajm@utdallas.edu

Abstract

Software Transactional Memory (STM) is a programming paradigm that allows a programmer to write parallel programs, without having to deal with the intricacies of synchronization. That burden is instead borne by the underlying STM system. SwissTM is a lock-based STM, developed at EPFL, Switzerland. Memory locations map to entries in a lock table to detect conflicts. Increasing the number of locations that map to a lock reduces the number of locks to be acquired and improves throughput, while also increasing the possibility of false conflicts. False conflicts occur when a transaction that updates a location mapping to a lock, causes validation failure of another transaction, that reads a different location mapping to the same lock. In this paper, we present a solution for the false conflict problem and suggest an adaptive version of the same algorithm, to improve performance. Our algorithms produce significant throughput improvement in benchmarks with false conflicts.

Transforming Linear Algebra Libraries: From Abstraction to Parallelism

Ernie Chan, Robert van de Geijn and Field G. Van Zee
Department of Computer Sciences
The University of Texas at Austin
Austin, TX 78712
{echan,rvdg,field}@cs.utexas.edu

Jim Nagle
LabVIEW Math and Signal Processing
National Instruments
Austin, TX 78759
jim.nagle@ni.com

Abstract

We have built a body of evidence which shows that, given a mathematical specification of a dense linear algebra operation to be implemented, it is possible to mechanically derive families of algorithms and subsequently to mechanically translate these algorithms into high-performing code. In this paper, we add to this evidence by showing that the algorithms can be statically analyzed and translated into directed acyclic graphs (DAGs) of coarse-grained operations that are to be performed. DAGs naturally express parallelism, which we illustrate by representing the DAGs with the G graphical programming language used by LabVIEW. The LabVIEW compiler and runtime execution system then exploit parallelism from the resulting code. Respectable speedup on a sixteen core architecture is reported.

AUTO-GC: Automatic Translation of Data Mining Applications to GPU Clusters

Wenjing Ma andGagan Agrawal
Department of Computer Science and Engineering
The Ohio State University
Columbus, OH 43210
{mawe,agrawal}@cse.ohio-state.edu

Abstract

Because of the very favorable price to performance ratio of the GPUs, a popular parallel programming configuration today is a cluster of GPUs. However, extracting performance on such a configuration would typically require programming in both MPI and CUDA, thus requiring a high degree of expertise and effort. It is clearly desirable to be able to support higher-level programming of this emerging high-performance computing platform.

This paper reports on a code generation system that can translate data mining applications on a GPU cluster. Our work is driven by the observation that a common processing structure, that of generalized reductions, fits a large number of popular data mining algorithms. In our solution, the programmers simply need to specify the sequential reduction loop(s) with some additional information about the parameters. We use program analysis and code generation to automatically map the applications to the API of FREERIDE, which is a middleware for parallel data mining. We also automatically generate CUDA code for using the GPU on each node of the cluster.

We have evaluated our system using two popular data mining applications, k-means clustering and Principal Component Analysis (PCA). We observed good scalability over the number of computing nodes, and the automatically generated version did not have any noticeable overheads compared to hand written codes. The speedup obtained by using GPU over using only the CPU on each node of a cluster is between 3 and 21.

Dense Linear Algebra Solvers for Multicore with GPU Accelerators

Stanimire Tomov, Rajib Nath, Hatem Ltaief and Jack Dongarra
Department of Electrical Engineering and Computer Science,
University of Tennessee, Knoxville
tomov, rnath1, ltaief, dongarra@eecs.utk.edu

Abstract

Solving dense linear systems of equations is a fundamental problem in scientific computing. Numerical simulations involving complex systems represented in terms of unknown variables and relations between them often lead to linear systems of equations that must be solved as fast as possible. We describe current efforts toward the development of these critical solvers in the area of dense linear algebra (DLA) for multicore with GPU accelerators. We describe how to code/develop solvers to effectively use the high computing power available in these new and emerging hybrid architectures. The approach taken is based on *hybridization techniques* in the context of Cholesky, LU, and QR factorizations. We use a high-level parallel programming model and leverage existing software infrastructure, e.g. optimized BLAS for CPU and GPU, and LAPACK for sequential CPU processing. Included also are architecture and algorithm-specific optimizations for standard solvers as well as mixed-precision iterative refinement solvers. The new algorithms, depending on the hardware configuration and routine parameters, can lead to orders of magnitude acceleration when compared to the same algorithms on standard multicore architectures that do not contain GPU accelerators. The newly developed DLA solvers are integrated and freely available through the MAGMA library.

Experiences of Using a Dependence Profiler to Assist Parallelization for Multi-cores

Dibyendu Das
IBM India
dibyendu.das@in.ibm.com

Peng Wu
IBM T. J. Watson Research Center
pengwu@us.ibm.com

Abstract

In this work we show how to use a data-dependence profiling tool called DProf, which can be utilized to assist parallelization for multi-core systems. DProf is based on an optimizing compiler and uses reference runs to emit information on runtime dependences between various memory accesses within a loop. The profiler not only marks the dependent statements and the accesses but also emits details regarding the percentage of time the dependences are encountered – the percentage being taken over the loop iteration size. Though DProf has been primarily built to capture opportunities for speculative thread-level parallelism, it has been found that the report generated by the DProf can be utilized very effectively in detecting and parallelizing complex code. To demonstrate this, we have taken two complex benchmarks – 435.gromacs and 437.leslie3d from the SPECfp CPU2006 suite and show how they can be parallelized effectively using DProf as an assist. To the best of our knowledge none of the existing parallelizing compilers can detect and parallelize all the instances reported in this work. We find that by using DProf we are able to parallelize these benchmarks very effectively for IBM P5+ and P6 multi-core systems. We have parallelized the benchmarks using OpenMP leading to speedups up to 2.6x. Also, due to the detailed reporting by DProf, we could cut down on our parallelization development effort significantly by concentrating on portions of the code that require attention. DProf can also be used to identify applications, where applying parallelization may lead to regression in performance. This allows the developers to discard applications or parts of it quickly, which may not lead to performance improvements when deployed on multi-cores. Thus, a data dependence profiler like DProf can act as an excellent assist mechanism to move applications to multi-cores in an effective way.

Integrating Parallel Application Development with Performance Analysis in Periscope

Ventsislav Petkov
Department of Informatics, I10
Technische Universität München
85748 Garching, Germany
petkovve@in.tum.de

Michael Gerndt
Department of Informatics, I10
Technische Universität München
85748 Garching, Germany
gerndt@in.tum.de

Abstract

High-performance computing (HPC) is making its way in every field of science and engineering by providing advanced methods for getting deeper comprehension of different processes and phenomena. However, due to the increased complexity of computer architectures and their multi-level parallelism, the development of efficient highly parallel applications is considerably complicated. This process has to be inevitably augmented with continuous performance analysis in order for one to be successful in optimizing applications and squeezing the potential out of today's supercomputers.

Periscope is a distributed performance analysis tool capable of collecting and processing measurement data from large scale application runs. In comparison to other similar tools, Periscope provides high-level performance bottlenecks and not the low-level values of hardware counters. This paper presents an enhanced and powerful graphical user interface that was recently developed for Periscope. It was successfully integrated in the Eclipse development platform as a plug-in and takes advantage of one of its extensions – the Parallel Tools Platform (PTP). This approach combines some of platform's advanced programming features with those of the Periscope performance measurement toolkit. As a result, a convenient software development and performance analysis environment was produced that aims at increasing the productivity of developers during the creation of highly efficient HPC applications.

Handling Errors in Parallel Programs Based on *Happens Before* Relations

Nicholas D. Matsakis
Laboratory for Software Technology
ETH Zurich
Zurich, Switzerland
nmatsaki@inf.ethz.ch

Thomas R. Gross
Laboratory for Software Technology
ETH Zurich
Zurich, Switzerland
trg@inf.ethz.ch

Abstract

Intervals are a new model for parallel programming based on an explicit *happens before* relation. Intervals permit fine-grained but high-level control of the program scheduler, and they dynamically detect and prevent deadlocking schedules. In this paper, we discuss the design decisions that led to the intervals model, focusing on error detection and handling. Our error propagation scheme makes use of the *happens before* relation to detect and abort dependent tasks that occur between the point where a failure occurs and where the failure is handled.

Workshop 4

Workshop on Nature Inspired Distributed Computing

NIDISC 2010

Evolving Hybrid Time-Shuffled Behavior of Agents

Patrick Ediger and Rolf Hoffmann
Technische Universität Darmstadt
FB Informatik, FG Rechnerarchitektur
Hochschulstraße 10, 64289 Darmstadt, Germany
{ediger, hoffmann}@ra.informatik.tu-darmstadt.de

Abstract

We searched for methods to evolve the hybrid behavior of moving agents for the All-to-All Communication task. The multi-agent system is modeled in two-dimensional Cellular Automata. An agent is controlled by one or more finite state machines. We use a time-shuffling method to join the state machines into one hybrid "algorithm". We propose a method to directly evolve a hybrid behavior consisting of multiple state machines including their time-shuffling periods. We compared the evolved hybrid algorithms to other evolved non-hybrid algorithms (consisting of only one finite state machine) and to hybrid algorithms that were composed of separately evolved non-hybrid algorithms. The performance of the directly evolved hybrid algorithms was significantly better, and the computation time for the evolution was roughly the same.

Diagnosing Permanent Faults in Distributed and Parallel Computing Systems Using Artificial Neural Networks

Mourad Elhadef
College of Engineering and Computer Science
Abu Dhabi University
Abu Dhabi, UAE
Email: mourad.elhadef@adu.ac.ae

Abstract

This paper deals with the problem of identifying faulty nodes (or units) in diagnosable distributed and parallel systems under the PMC model. In this model, each unit is tested by a subset of the other units, and it is assumed that, at most, a bounded subset of these units is permanently faulty. When performing testing, faulty units can incorrectly claim that fault-free units are faulty or that faulty units are fault-free. Since the introduction of the PMC model, significant progress has been made in both theory and practice associated with the original model and its offshoots. Nevertheless, this problem of efficiently identifying the set of faulty units of a diagnosable system remained an outstanding research issue. In this paper, we describe a new neural-network-based diagnosis algorithm, which exploits the off-line learning phase of artificial neural network to speed up the diagnosis algorithm. The novel approach has been implemented and evaluated using randomly generated diagnosable systems. The simulation results showed that the new neural-network-based fault identification approach constitutes an addition to existing diagnosis algorithms. Extreme faulty situations, where the number of faults is around the bound t, and large diagnosable systems have been also experimented to show the efficiency of the new neural-network-based diagnosis algorithm.

Particle Swarm Optimization under Fuzzy Logic Controller for solving a Hybrid Reentrant Flow Shop problem

Naim Yalaoui, Lionel Amodeo, Farouk Yalaoui and Halim Mahdi
ICD-LOSI (FRE CNRS 2848), University of Technology of Troyes
12 rue Marie Curie, 10010 Troyes, France
Email: [naim.yalaoui, farouk.yalaoui, lionel.amodeo]@utt.fr
Caillau Company
28, Rue Ernest Renan, 92130, Issy les Moulineaux. France
Email: [nyalaoui, hmahdi]@caillau.com

Abstract

Our study aims to solve a hybrid flow shop scheduling problem. This one has a specification which is: it contains different stages in series, each one is composed of some identical parallel machines, each order is composed of several batches and some ones are re-entrant. The objective function of this problem is to minimize the total tardiness. A new method is developed to solve the problem based on the Nature which is the particle swarm optimization method under fuzzy logic controller (FLCPSO). This one is compared to another modified Particle Swarm Optimization developed by us in previous works. The results are very interesting.

pALS: An Object-Oriented Framework for Developing Parallel Cooperative Metaheuristics

Andrés Bernal
Departamento de Ingeniería de Sistemas y Computación
Comunicación y Tecnología de Información, COMIT
Bogotá, Colombia
andre-be@uniandes.edu.co

Harold Castro
Departamento de Ingeniería de Sistemas y Computación
Comunicación y Tecnología de Información, COMIT
Bogotá, Colombia
hcastro@uniandes.edu.co

Abstract

pALS acronym for parallel Adaptive Learning Search is a computational object oriented framework for the development of parallel and cooperative metaheuristics for solving complex optimization problems. The library exploits the paralellization allowing the deployment of mainly two models: the parallel execution of operators and the execution of separate instances or multi-start models. pALS also allows to include in the design of the problem's solution cooperation strategies such as the islands model for genetic algorithms or the parallel exploration of neighborhoods in metaheuristics derived from local searches, including a broad set of topologies associated with these models. pALS has been successfully used in different optimization problems and has proven to be a flexible, extensible and commanding library to promptly develop prototypes offering a collection of ready to use operators that encompass the nucleus of many metaheuristics including hybrid metaheuristics.

A New Parallel Asynchronous Cellular Genetic Algorithm for Scheduling in Grids

Frédéric Pinel, Bernabé Dorronsoro and Pascal Bouvry
Faculty of Science, Technology, and Communications
University of Luxembourg
{frederic.pinel, bernabe.dorronsoro, pascal.bouvry}@uni.lu

Abstract

We propose a new parallel asynchronous cellular genetic algorithm for multi-core processors. The algorithm is applied to the scheduling of independent tasks in a grid. Finding such optimal schedules is in general an NP-hard problem, to which evolutionary algorithms can find near-optimal solutions. We analyze the parallelism of the algorithm, as well as different recombination and new local search operators. The proposed algorithm improves previous schedules on benchmark problems. The parallelism of this algorithm suits it to bigger problem instances.

CA-based Generator of S-boxes for Cryptography Use

Miroslaw Szaban
Institute of Computer Science
University of Podlasie
Sienkiewicza 51, 08-110 Siedlce, Poland
mszaban@ap.siedlce.pl

Franciszek Seredynski
Institute of Computer Science
Polish Academy of Sciences
Ordona 21, 01-237 Warsaw, Poland
sered@ipipan.waw.pl
Polish-Japanese Institute of Information Technology
Koszykowa 86, 02-008 Warsaw, Poland
sered@pjwstk.edu.pl

Abstract

Substitution boxes (S-boxes) are important elements of many block ciphers, which serve as tools of nonlinear transformation of information in the cipher process. Classical S-boxes are usually represented by specially designed tables, which are used today in current cryptographic standards, such as Data Encryption Standard (DES) or Advanced Encryption Standard (AES), but in the result of developing methods of cryptanalysis they do not ensure enough safety of ciphers. Therefore, the open research issue now is to design new more sophisticated classes of S-boxes, in particular dynamic ones. In this paper we propose a methodology to design dynamic cellular automata (CA)-based S-boxes, which can be considered as generator od CA-based S-boxes. We provide an exhaustive experimental analysis of the proposed CA-based S-boxes in terms of non-linearity, autocorrelation, balance and strict avalanche criterion. We show that the proposed S-boxes have high quality cryptographic properties (high nonlinearity and balance, also low autocorrelation and distance to fulfill strict avalanche criterion). The interesting feature of the proposed S-boxes is a dynamic flexible structure, fully functionally realized by CA, while the classical S-boxes are represented by predefined unchangeable table structures.

Modeling memory resources distribution on multicore processors using games on cellular automata lattices

Michail-Antisthenis I. Tsompanas, Georgios Ch. Sirakoulis and Ioannis Karafyllidis
Department of Electrical and Computer Engineering
Democritus University of Thrace, DUTH
67100, Xanthi, Greece
{mtsompan, gsirak, ykar}@ee.duth.gr

Abstract

Nowadays, there is an increasingly recognized need for more computing power, which has led to multicore processors. However, this evolution is still restrained by the poor efficiency of memory chips. As a possible solution to the problem, this paper examines a model of re-distributing the memory resources assigned to the processor, especially the on-chip memory, in order to achieve higher performance. The proposed model uses the basic concepts of game theory applied to cellular automata lattices and the iterated spatial prisoner's dilemma game. A simulation was established in order to evaluate the performance of this model under different circumstances. Moreover, a corresponding FPGA logic circuit was designed as a part of an embedded, real-time co-circuit, aiming at memory resources fair distribution. The proposed FPGA implementation proved advantageous in terms of low-cost, high-speed, compactness and portability features. Finally, a significant improvement on the performance of the memory resources was ascertained from simulation results.

A Survey On Bee Colony Algorithms

Salim Bitam[1]
[1]Computer science department
Mohamed Khider University
Biskra, Algeria
salimbitam@gmail.com

Mohamed Batouche[2]
[2]COEIA - CCIS
King Saud University
Riyadh, Saudi Arabia
batouche@ccis.edu.sa

El-ghazali Talbi[2,3]
[3]INRIA
University of Lille
Lille, France
el-ghazali.talbi@lifl.fr

Abstract

This paper presents a survey of current research activities inspired by bee life. This work is intended to provide a broad and comprehensive view of the various principles and applications of these bio-inspired systems. We propose to classify them into two major models. The first one is based on the foraging behavior in the bee quotidian life and the second is inspired by the marriage principle. Different original studies are described and classified along with their applications, comparisons against other approaches and results. We then summarize a review of their derived algorithms and research efforts.

Particle Swarm Optimization to solve the Vehicle Routing Problem with Heterogeneous fleet, Mixed Backhauls, and Time Windows

Farah Belmecheri, Christian Prins, Farouk Yalaoui and Lionel Amodeo
ICD-LOSI (FRE CNRS 2848)
University of Technology of Troyes, 12 rue Marie Curie
10010 Troyes, France
Email: farah.belmecheri, christian.prins, farouk.yalaoui, lionel.amodeo@utt.fr
TCP Distribution
115, route d'Auxerre, Saint-Andre les vergers 10120 Troyes, France
Email: farah.belmecheri@tcp-troyes.com

Abstract

Many distribution companies must deliver and pick up goods to satisfy customers. This problem is called the Vehicle Routing Problem with Mixed linehauls and Backhauls (VRPMB) which considers that some goods must be delivered from a depot to linehaul customers, while others must be picked up at backhaul customers to be brought to the depot. This paper studies an enriched version called Heterogeneous fleet VRPMB with Time Windows or HVRPMBTW which has not much been studied in the literature. A Particle Swarm Optimization heuristic (PSO) is proposed to solve this problem. This approach uses and models the social behavior of bird flocking, fish schooling. The adaptation and implementation of PSO search strategy to HVRPMBTW is explained, then the results are compared to previous works (Ant Colony Optimization) and compared also to the high quality solutions obtained by an exact method (solver CPLEX). Good promising results are reported and have shown the effectiveness of the method.

Heterogeneous Parallel Algorithms to Solve Epistatic Problems

Carolina Salto
Universidad Nacional de La Pampa
Calle 100 esq. 9, General Pico (6360)
La Pampa, Argentina
Email: saltoc@ing.unlpam.edu.ar

Enrique Alba
Universidad de Málaga
E.T.S.I. Informática, 29071
Málaga, España
Email: eat@lcc.uma.es

Abstract

In this paper, we propose parallel heterogeneous metaheuristics (PHM) to solve a kind of epistatic problem (NK-Landscape). The main feature of our heterogeneous algorithms is the utilization of multiple search threads using different configurations to guide the search process. We propose an operator-based PHM, where each search thread uses a different combination of recombination and mutation operators. We compare the performance of our heterogeneous proposal against an homogeneous algorithm (multiple threads with the same parameter configuration) in a numerical and real time ways. Our experiments show that the heterogeneity could help to design powerful and robust optimization algorithms on high dimensional landscapes with an additional reduction in execution times.

A Bio-Inspired Coverage-Aware Scheduling Scheme for Wireless Sensor Networks

Xin Fei
PARADISE Research Laboratory
SITE, University of ottawa, Canada
xfei@site.uottawa.ca

Samer Samarah
PARADISE Research Laboratory
SITE, University of ottawa, Canada
ssamarah@site.uottawa.ca

Azzedine Boukerche
PARADISE Research Laboratory
SITE, University of ottawa, Canada
boukerch@site.uottawa.ca

Abstract

In the densely deployed wireless sensor networks, sensors are scheduled over time in order to maintain the coverage while saving energy of networks. In this article, we investigate the coverage-aware scheduling problem using genetic algorithms. Sensors are optimally scheduled in different time slots to maximize the overall coverage under the given k-cover requirement and lifetime of networks. A set of simulation is carried out. The simulation result shows that, using the optimal schedule generated by genetic algorithm, our algorithm can optimize the coverage performance of wireless sensor network in terms of overall coverage rate and number of active sensors.

Workshop 5

Workshop on High Performance Computational Biology

HiCOMB 2010

GPU-Accelerated Multi-scoring Functions Protein Loop Structure Sampling

Yaohang Li
Department of Computer Science
North Carolina A&T State University
Greensboro, NC 27411, USA
e-mail: yaohang@ncat.edu

Weihang Zhu
Department of Industrial Engineering
Lamar University
Beaumont, TX 77710, USA
e-mail: Weihang.Zhu@lamar.edu

Abstract

Accurate protein loop structure models are important to understand functions of many proteins. One of the main problems in correctly modeling protein loop structures is sampling the large loop backbone conformation space, particularly when the loop is long. In this paper, we present a GPU-accelerated loop backbone structure modeling approach by sampling multiple scoring functions based on pair-wise atom distance, torsion angles of triplet residues, or soft-sphere van der Waals potential. The sampling program implemented on a heterogeneous CPU-GPU platform has observed a speedup of ~40 in sampling long loops, which enables the sampling process to carry out computation with large population size. The GPU-accelerated multi-scoring functions loop structure sampling allows fast generation of decoy sets composed of structurally-diversified backbone decoys with various compromises of multiple scoring functions. In the 53 long loop benchmark targets we tested, our computational results show that in more than 90% of the targets, the decoy sets we generated include decoys within 1.5A RMSD (Root Mean Square Deviation) from native while in 77% of the targets, decoys in 1.0A RMSD are reached.

Acceleration of Spiking Neural Networks in Emerging Multi-core and GPU Architectures

Mohammad A. Bhuiyan, Vivek K. Pallipuram and Melissa C. Smith
Department of Electrical and Computer Engineering,
Clemson University, Clemson, SC 29634, USA
{mbhuiya, kpallip, smithmc}@clemson.edu

Abstract

Recently, there has been strong interest in large-scale simulations of biological spiking neural networks (SNN) to model the human brain mechanisms and capture its inference capabilities. Among various spiking neuron models, the Hodgkin-Huxley model is the oldest and most compute intensive, whereas the more recent Izhikevich model is very compute efficient. Some of the recent multi-core processors and accelerators including Graphical Processing Units, IBM's Cell Broadband Engine, AMD Opteron, and Intel Xeon can take advantage of task and thread level parallelism, making them good candidates for large-scale SNN simulations. In this paper we implement and analyze two character recognition networks based on these spiking neuron models. We investigate the performance improvement and optimization techniques for SNNs on these accelerators over an equivalent software implementation on a 2.66 GHz Intel Core 2 Quad. We report significant speedups of the two SNNs on these architectures. It has been observed that given proper application of optimization techniques, the commodity X86 processors are viable options for those applications that require a nominal amount of flops/byte. But for applications with a significant number of flops/byte, specialized architectures such as GPUs and cell processors can provide better performance. Our results show that a proper match of architecture with algorithm complexity provides the best performance.

A Tile-based Parallel Viterbi Algorithm for Biological Sequence Alignment on GPU with CUDA

Zhihui Du[1+], Zhaoming Yin[2] and David A. Bader[3]
[1]Tsinghua National Laboratory for Information Science and Technology
Department of Computer Science and Technology, Tsinghua University, 100084, Beijing, China
[+]Corresponding Author's Email: duzh@tsinghua.edu.cn
[2]School of Software and Microelectronics, Peking University, 100871, China.
Email zhaoming_leon@pku.edu.cn
[3]College of Computing, Georgia Institute of Technology, Atlanta, GA, 30332, USA.

Abstract

The Viterbi algorithm is the compute-intensive kernel in Hidden Markov Model (HMM) based sequence alignment applications. In this paper, we investigate extending several parallel methods, such as the wave-front and streaming methods for the Smith-Waterman algorithm, to achieve a significant speed-up on a GPU. The wave-front method can take advantage of the computing power of the GPU but it cannot handle long sequences because of the physical GPU memory limit. On the other hand, the streaming method can process long sequences but with increased overhead due to the increased data transmission between CPU and GPU. To further improve the performance on GPU, we propose a new tile-based parallel algorithm. We take advantage of the homological segments to divide long sequences into many short pieces and each piece pair (tile) can be fully held in the GPU's memory. By reorganizing the computational kernel of the Viterbi algorithm, the basic computing unit can be divided into two parts: independent and dependent parts. All of the independent parts are executed with a balanced load in an optimized coalesced memory-accessing manner, which significantly improves the Viterbi algorithm's performance on GPU. The experimental results show that our new tile-based parallel Viterbi algorithm can outperform the wave-front and the streaming methods. Especially for the long sequence alignment problem, the best performance of tile-based algorithm is on average about an order magnitude faster than the serial Viterbi algorithm.

Fast Binding Site Mapping using GPUs and CUDA

Bharat Sukhwani and Martin C. Herbordt
Computer Architecture and Automated Design Laboratory
Department of Electrical and Computer Engineering
Boston University, Boston, MA 02215

Abstract

Binding site mapping refers to the computational prediction of the regions on a protein surface that are likely to bind a small molecule with high affinity. The process involves flexibly docking a variety of small molecule probes and finding a consensus site that binds most of those probes. Due to the computational complexity of flexible docking, the process is often split into two steps: the first performs rigid docking between the protein and the probe; the second models the side chain flexibility by energy-minimizing the (few thousand) top scoring protein-probe complexes generated by the first step. Both these steps are computationally very expensive, requiring many hours of runtime per probe on a serial CPU. In the current article, we accelerate a production mapping software program using NVIDIA GPUs. We accelerate both the rigid-docking and the energy minimization steps of the program. The result is a 30x speedup on rigid docking and 12x on energy minimization, resulting in a 13x overall speedup over the current single core implementation.

Hybrid MPI/Pthreads Parallelization of the RAxML Phylogenetics Code

Wayne Pfeiffer
San Diego Supercomputer Center
University of California, San Diego
La Jolla, California, USA
pfeiffer@sdsc.edu

Alexandros Stamatakis
Department of Computer Science
Technische Universität München
Munich, Germany
stamatak@cs.tum.edu

Abstract

A hybrid MPI/Pthreads parallelization was implemented in the RAxML phylogenetics code. New MPI code was added to the existing Pthreads production code to exploit parallelism at two algorithmic levels simultaneously: coarse-grained with MPI and fine-grained with Pthreads. This hybrid, multi-grained approach is well suited for current high-performance computers, which typically are clusters of multi-core, shared-memory nodes.

The hybrid version of RAxML is especially useful for a comprehensive phylogenetic analysis, i.e., execution of many rapid bootstraps followed by a full maximum likelihood search. Multiple multi-core nodes can be used in a single run to speed up the computation and, hence, reduce the turnaround time. The hybrid code also allows more efficient utilization of a given number of processor cores. Moreover, it often returns a better solution than the stand-alone Pthreads code, because additional maximum likelihood searches are conducted in parallel using MPI.

The comprehensive analysis algorithm involves four stages, in which coarse-grained parallelism continually decreases from stage to stage. The first three stages speed up well with MPI, while the last stage speeds up only with Pthreads. This leads to a tradeoff in effectiveness between MPI and Pthreads parallelization.

The useful number of MPI processes increases with the number of bootstraps performed, but typically is limited to 10 or 20 by the parameters of the algorithm. The optimal number of Pthreads increases with the number of distinct patterns in the columns of the multiple sequence alignment, but is limited to the number of cores per node of the computer being used.

For a benchmark problem with 218 taxa, 1,846 patterns, and 100 bootstraps run on the Dash computer at SDSC, the speedup of the hybrid code on 10 nodes (80 cores) was 6.5 compared to the Pthreads-only code on one node (8 cores) and 35 compared to the serial code. This run used 10 MPI processes with 8 Pthreads each. For another problem with 125 taxa, 19,436 patterns, and 100 bootstraps, the speedup on the Triton PDAF computer at SDSC was 38 on two nodes (64 cores) compared to the serial code. This run used 2 MPI processes with 32 Pthreads each.

Measuring Properties of Molecular Surfaces Using Ray Casting

Mike Phillips[1,2], Iliyan Georgiev[2], Anna Katharina Dehof[3], Stefan Nickels[3,4],
Lukas Marsalek[2], Hans-Peter Lenhof[3], Andreas Hildebrandt[3] and Philipp Slusallek[1,2,4]
[1]DFKI Saarbrücken
[2]Computer Graphics Lab, Saarland University
[3]Center for Bioinformatics, Saarland University
[4]Intel Visual Computing Institute, Saarland University
66123 Saarbrücken, Germany

Abstract

Molecular geometric properties, such as volume, exposed surface area, and occurrence of internal cavities, are important inputs to many applications in molecular modeling. In this work we describe a very general and highly efficient approach for the accurate computation of such properties, which is applicable to arbitrary molecular surface models. The technique relies on a high performance ray casting framework that can be easily adapted to the computation of further quantities of interest at interactive speed, even for huge models.

On the Parallelisation of MCMC-based Image Processing

Jonathan M. R. Byrd
Department of Computer Science
University of Warwick
Coventry, CV4 7AL, UK
Email: jbyrd@dcs.warwick.ac.uk

Stephen A. Jarvis
Department of Computer Science
University of Warwick
Coventry, CV4 7AL, UK
Email: saj@dcs.warwick.ac.uk

Abhir H. Bhalerao
Department of Computer Science
University of Warwick
Coventry, CV4 7AL, UK
Email: abhir@dcs.warwick.ac.uk

Abstract

The increasing availability of multi-core and multiprocessor architectures provides new opportunities for improving the performance of many computer simulations. Markov Chain Monte Carlo (MCMC) simulations are widely used for approximate counting problems, Bayesian inference and as a means for estimating very high-dimensional integrals. As such MCMC has found a wide variety of applications in fields including computational biology and physics, financial econometrics, machine learning and image processing.

Whilst divide and conquer is an obvious means to simplify image processing tasks, "naively" dividing an image into smaller images to be processed separately results in anomalies and breaks the statistical validity of the MCMC algorithm. We present a method of grouping the spatially local moves and temporarily partitioning the image to allow those moves to be processed in parallel, reducing the runtime whilst conserving the properties of the MCMC method. We calculate the theoretical reduction in runtime achievable by this method, and test its effectiveness on a number of different architectures. Experiments are presented that show reductions in runtime of 38% using a dual-core dual-processor machine.

In circumstances where absolute statistical validity are not required, an algorithm based upon, but not strictly adhering to, MCMC may suffice. For such situation two additional algorithms are presented for partitioning of images to which MCMC will be applied. Assuming preconditions are met, these methods may be applied with minimal risk of anomalous results. Although the extent of the runtime reduction will be data dependent, the experiments performed showed the runtime reduced to 27% of its original value.

Exploring Parallelism in Short Sequence Mapping Using Burrows-Wheeler Transform

Doruk Bozdağ[1], Ayat Hatem[1,2] and Umit V. Catalyurek[1,2]
[1]Department of Biomedical Informatics
[2]Department of Electrical and Computer Engineering
The Ohio State University, Columbus, OH 43210
Email: {bozdagd,dayat,umit}@bmi.osu.edu

Abstract

Next-generation high throughput sequencing instruments are capable of generating hundreds of millions of reads in a single run. Mapping those reads to a reference genome is an extremely compute-intensive process that takes more than a day on a modern computer even when the accuracy of the results is traded off to speed up the execution. In this work, we explore various data distribution strategies for parallel execution of three state-of-the-art mapping tools, namely Bowtie, BWA and SOAP2, that are based on the Burrows-Wheeler Transformation. We report on the performance of these strategies and show that the best strategy depends on the input scenario as well as the relative efficiency of the tools in the indexing and matching steps of the mapping process. The parallelization strategies investigated in this paper are general and can easily be applied to different mapping algorithms. With the availability of parallel execution methods, it will be possible to carry out more intensive computations that cannot be accomplished in a reasonable time using sequential tools, including mapping with larger mismatch tolerance.

pFANGS: Parallel High Speed Sequence Mapping for Next Generation 454-Roche Sequencing Reads

Sanchit Misra[1], Ramanathan Narayanan[2], Wei-keng Liao[3] and Alok Choudhary[4]
Electrical Engineering and Computer Science
Northwestern University
Evanston, IL, USA 60208
Email: [1]smi539@eecs.northwestern.edu,
[2]ran310@eecs.northwestern.edu,
[3]wkliao@eecs.northwestern.edu,
[4]choudhar@eecs.northwestern.edu
Simon Lin
Robert H. Lurie Comprehensive Cancer Center
Northwestern University
Chicago, IL, USA 60611
Email: S-Lin2@northwestern.edu

Abstract

Millions of DNA sequences (reads) are generated by Next Generation Sequencing machines everyday. There is a need for high performance algorithms to map these sequences to the reference genome to identify single nucleotide polymorphisms or rare transcripts to fulfill the dream of personalized medicine. In this paper, we present a high-throughput parallel sequence mapping program pFANGS. pFANGS is designed to find all the matches of a query sequence in the reference genome tolerating a large number of mismatches or insertions/deletions. pFANGS partitions the computational workload and data among all the processes and employs load-balancing mechanisms to ensure better process efficiency. Our experiments show that, with 512 processors, we are able to map approximately 31 million 454/Roche queries of length 500 each to a reference human genome per hour allowing 5 mismatches or insertion/deletions at full sensitivity. We also report and compare the performance results of two alternative parallel implementations of pFANGS: a shared memory OpenMP implementation and a MPI-OpenMP hybrid implementation.

Efficient and scalable parallel reconstruction of sibling relationships from genetic data in wild populations

Saad Sheikh, Ashfaq Khokhar, Tanya Berger-Wolf
Department of Computer Science, University of Illinois at Chicago, Chicago, IL 60607

Abstract

Wild populations of organism are often difficult to study in their natural settings. Often, it is possible to infer mating information about these species by genotyping the offspring and using the genetic information to infer sibling, and other kinship, relationships. While sibling reconstruction has been studied for a long time, none of the existing approaches have targeted scalability. In this paper, we introduce the first parallel approach to reconstructing sibling relationships from microsatellite markers. We use both functional and data domain decomposition to break down the problem and argue that this approach can be applied to other problems where columns are independent and simple constraint-based enumeration is required. We discuss algorithmic and implementation choices and their effects on results. We show that our approach is highly efficient and scalable.

Workshop 6

Advances in Parallel and Distributed Computing Models

APDCM 2010

Throughput optimization for micro-factories subject to task and machine failures

Anne Benoit
ENS Lyon, Université de Lyon, France
LIP laboratory (ENS, CNRS, INRIA, UCBL)
Anne.Benoit@ens-lyon.fr

Alexandru Dobrila, Jean-Marc Nicod and Laurent Philippe
Laboratoire d'Informatique de Franche-Comté,
Université de Franche-Comté, France
adobrila,jmnicod,lphilippe@lifc.univ-fcomte.fr

Abstract

In this paper, we study the problem of optimizing the throughput for micro-factories subject to failures. The challenge consists in mapping several tasks of different types onto a set of machines. The originality of our approach is the failure model for such applications in which not only the machines are subject to failures but the reliability of a task may depend on its type. The failure rate is unrelated: a probability of failure is associated to each couple (task type, machine). We consider different kind of mappings: in one-to-one mappings, each machine can process only a single task, while several tasks of the same type can be processed by the same machine in specialized mappings. Finally, general mappings have no constraints. The optimal one-to-one mapping can be found in polynomial time for particular problem instances, but the problem is NP-hard in most of the cases. For the most realistic case of specialized mappings, which turns out to be NP-hard, we design several polynomial time heuristics and a linear program allows us to find the optimal solution (in exponential time) for small problem instances. Experimental results show that the best heuristics obtain a good throughput, much better than the throughput achieved with a random mapping. Moreover, we obtain a throughput close to the optimal solution in the particular cases where the optimal throughput can be computed.

An Efficient GPU Implementation of the Revised Simplex Method

Jakob Bieling, Patrick Peschlow and Peter Martini
Institute of Computer Science 4, University of Bonn
Römerstr. 164, 53117 Bonn, Germany
{bieling, peschlow, martini}@cs.uni-bonn.de

Abstract

The computational power provided by the massive parallelism of modern graphics processing units (GPUs) has moved increasingly into focus over the past few years. In particular, general purpose computing on GPUs (GPGPU) is attracting attention among researchers and practitioners alike. Yet GPGPU research is still in its infancy, and a major challenge is to rearrange existing algorithms so as to obtain a significant performance gain from the execution on a GPU. In this paper, we address this challenge by presenting an efficient GPU implementation of a very popular algorithm for linear programming, the revised simplex method. We describe how to carry out the steps of the revised simplex method to take full advantage of the parallel processing capabilities of a GPU. Our experiments demonstrate considerable speedup over a widely used CPU implementation, thus underlining the tremendous potential of GPGPU.

OpenCL – An effective programming model for data parallel computations at the Cell Broadband Engine

Jens Breitbart and Claudia Fohry
Research Group Programming Languages / Methodologies
Universität Kassel, Germany
Email:{jbreitbart, fohry}@uni-kassel.de

Abstract

Current processor architectures are diverse and heterogeneous. Examples include multicore chips, GPUs and the Cell Broadband Engine (CBE). The recent Open Compute Language (OpenCL) standard aims at efficiency and portability. This paper explores its efficiency when implemented on the CBE, without using CBE-specific features such as explicit asynchronous memory transfers. We based our experiments on two applications: matrix multiplication, and the client side of the Einstein@Home distributed computing project. Both were programmed in OpenCL, and then translated to the CBE. For matrix multiplication, we deployed different levels of OpenCL performance optimization, and observed that they pay off on the CBE. For the Einstein@Home application, our translated OpenCL version achieves almost the same speed as a native CBE version.

Another main contribution of the paper is a proposal for an additional memory level in OpenCL, called static local memory. With little programming expense, it can lead to significant speedups such as factor seven for reduction. Finally, we studied two versions of the OpenCL to CBE mapping, in which the PPE component of the CBE does or does not take the role of a compute unit.

A random walk based clustering with local recomputations for mobile ad hoc networks

Alain BUI
PRiSM (UMR CNRS 8144)
Université de Versailles
St-Quentin-en-Yvelines
alain.bui@prism.uvsq.fr

Abdurusul KUDIRETI
CReSTIC-SysCom
Université de Reims
Champagne-Ardenne
kudireti.abdurusul@univ-reims.fr

Devan SOHIER
PRiSM (UMR CNRS 8144)
Université de Versailles
St-Quentin-en-Yvelines
devan.sohier@prism.uvsq.fr

Abstract

In this paper, we present a clustering algorithm for MANETs. A cluster is divided in two parts: core nodes, that are recruited by a randomly moving token, and ordinary nodes, that are adjacent to core nodes. The clustering built by this algorithm is adaptive to topological changes. When a topological change occurs, nodes that are not in the cluster in which it occurs, or in adjacent clusters, are not affected. This property allows the scalability of this algorithm, by ensuring that only a bounded number of nodes have to recompute their cluster.

Stability of a localized and greedy routing algorithm

Christelle Caillouet
Drakkar, LIG Lab
Grenoble, France

Florian Huc
TCS-SENSOR Lab,
University of Geneva, Switzerland

Nicolas Nisse and Stéphane Pérennes
CNRS-INRIA-UNS Mascotte
Sophia Antipolis, France

Hervé Rivano
CITI, INRIA-INSA Lyon
University of Lyon, France

Abstract

In this work, we study the problem of routing packets between undifferentiated sources and sinks in a network modeled by a multigraph. We consider a distributed and local algorithm that transmits packets hop by hop in the network and study its behavior. At each step, a node transmits its queued packets to its neighbors in order to optimize a local gradient. This protocol is greedy since it does not require to record the history about the past actions, and localized since nodes only need information about their neighborhood.

A transmission protocol is *stable* if the number of packets in the network does not diverge. To prove the stability, it is sufficient to prove that the number of packets stored in the network remains bounded as soon as the sources inject a flow that another method could have exhausted. The localized and greedy protocol considered has been shown to be stable in some specific cases related to the arrival rate of the packets. We investigate its stability in a more general context and therefore reinforce results from the literature that worked for differentiated suboptimal flows.

We show that, to prove the stability of this protocol, it is sufficient to prove the intuitive following conjecture: roughly, if the protocol is stable when all sources inject the maximum number of packets at each turn and no packets are lost, then the protocol is stable whatever be the behavior of the network (i.e., when less packets are injected and some of them may be lost).

Detecting and Using Critical Paths at Runtime in Message Driven Parallel Programs

Isaac Dooley
Department of Computer Science
University of Illinois
idooley2@illinois.edu

Laxmikant V. Kale
Department of Computer Science
University of Illinois
kale@illinois.edu

Abstract

Detecting critical paths in traditional message passing parallel programs can be useful for post-mortem performance analysis. This paper presents an efficient online algorithm for detecting critical paths for message-driven parallel programs. Initial implementations of the algorithm have been created in three message-driven parallel languages: Charm++, Charisma, and Structured Dagger. Not only does this work describe a novel implementation of critical path detection for the message-driven programs, but also the resulting critical paths are successfully used as the program runs for automatic performance tuning. The actionable information provided by the critical path is shown to be useful for online performance tuning within the context of the message driven parallel model, whereas it has never been used for online purposes within the traditional message passing model.

Randomized Self-Stabilizing Leader Election in Preference-Based Anonymous Trees

Daniel Fajardo-Delgado[1], José Alberto Fernández-Zepeda[1] and Anu G. Bourgeois[2]
[1]Department of Computer Science
CICESE
Ensenada, B.C., Mexico
{dfajardo, fernan}@cicese.mx

[2]Department of Computer Science
Georgia State University
Atlanta, GA, USA
abourgeois@cs.gsu.edu

Abstract

The performance of processors in a distributed system can be measured by parameters such as bandwidth, storage capacity, work capability, reliability, manufacture technology, years of usage, among others. An algorithm using a preference-based approach uses these parameters to make decisions. In this paper we introduce a randomized self-stabilizing leader election algorithm for preference-based anonymous trees. Our algorithm uses the preference of the processors as criteria to select a leader under symmetric or non-symmetric configurations. It is partially inspired on Xu and Srimani's algorithm [1], but we use a distributed daemon and randomization to break symmetry. We prove that our algorithm has an optimal average complexity time and performed simulations to verify our results.

A PRAM-NUMA Model of Computation for Addressing Low-TLP Workloads

Martti Forsell
VTT Technical Research Center of Finland
Platform Architectures Team
Box 1100, FI-90571 Oulu, Finland
Martti.Forsell@VTT.Fi

Abstract

It is possible to implement the parallel random access machine (PRAM) on a chip multiprocessor (CMP) efficiently with an emulated shared memory (ESM) architecture to gain easy parallel programmability crucial to wider penetration of CMPs to general purpose computing. This implementation relies on exploitation of the slack of parallel applications to hide the latency of the memory system instead of caches, sufficient bisection bandwidth to guarantee high throughput, and hashing to avoid hot spots in intercommunication. Unfortunately this solution can not handle workloads with low thread-level parallelism (TLP) efficiently because then there is not enough parallel slackness available for hiding the latency. In this paper we show that integrating nonuniform memory access (NUMA) support to the PRAM implementation architecture can solve this problem. The obtained PRAM-NUMA hybrid model is described and architectural implementation of it is outlined on our Eclipse ESM CMP framework.

Self-Stabilizing Master–Slave Token Circulation and Efficient Size-Computation in a Unidirectional Ring of Arbitrary Size

Wayne Goddard and Pradip K Srimani
School of Computing
Clemson University
Clemson, SC 29634–0974
{goddard, srimani}@cs.clemson.edu

Abstract

Self-stabilizing algorithms represent an extension of distributed algorithms in which nodes of the network have neither coordination, synchronization, nor initialization. We consider the model where there is one designated master node and all other nodes are anonymous and have constant space. Recently, Lee et al. obtained such an algorithm for determining the size of a unidirectional ring. We provide a new algorithm that converges much quicker. This algorithm exploits a token-circulation idea due to Afek and Brown. Disregarding the time for stabilization, our algorithm computes the size of the ring at the master node in $O(n \log n)$ time compared to $O(n^3)$ steps used in the algorithm by Lee et al. using the same computing paradigm. It seems likely that one should be able to obtain master–slave algorithms for other problems in networks.

Distributed Tree Decomposition of Graphs and Applications to Verification

Stéphane Grumbach
INRIA-LIAMA,
stephane.grumbach@inria.fr

Zhilin Wu
LaBRI, Université de Bordeaux,
zlwu@labri.fr

Abstract

The tree decomposition of graphs is a fundamental algorithmic tool. It has been shown that difficult problems, such as some NP-complete ones, can be solved efficiently over classes of graphs of bounded tree-width. We consider in this paper the distributed construction of the tree decompositions of network topology graphs. We propose algorithms to distributively construct the tree-decomposition of respectively (i) planar networks of bounded diameter and (ii) networks of bounded degree and bounded tree-length. Both algorithms are very efficient, requiring only a constant number of messages sent over each link. We then use these algorithms to distributively verify properties of graphs expressible in Monadic Second Order Logic, MSO.

Collaborative Execution Environment for Heterogeneous Parallel Systems

Aleksandar Ilić and Leonel Sousa
INESC-ID, IST, TU Lisbon
Rua Alves Redol 9
1000-029 Lisbon, Portugal
Email: {Aleksandar.Ilic, Leonel.Sousa}@inesc-id.pt

Abstract

Nowadays, commodity computers are complex heterogenous systems that provide a huge amount of computational power. However, to take advantage of this power we have to orchestrate the use of processing units with different characteristics. Such distributed memory systems make use of relatively slow interconnection networks, such as system buses. Therefore, most of the time we only individually take advantage of the central processing unit (CPU) or processing accelerators, which are simpler homogeneous subsystems. In this paper we propose a collaborative execution environment for exploiting data parallelism in a heterogeneous system. It is shown that this environment can be applied to program both CPU and graphics processing units (GPUs) to collaboratively compute matrix multiplication and fast Fourier transform (FFT). Experimental results show that significant performance benefits are achieved when both CPU and GPU are used.

Efficient Exhaustive Verification of the Collatz Conjecture using DSP48E blocks of Xilinx Virtex-5 FPGAs

Yasuaki Ito and Koji Nakano
Department of Information Engineering
Hiroshima University
Kagamiyama 1-4-1, Higashi-Hiroshima, 739-8527, JAPAN
{yasuaki, nakano}@cs.hiroshima-u.ac.jp

Abstract

Consider the following operation on an arbitrary positive number: if the number is even, divide it by two, and if the number is odd, triple it and add one. The Collatz conjecture asserts that, starting from any positive number m, repeated iteration of the operations eventually produces the value 1. The main contribution of this paper is to present an efficient implementation of a coprocessor that performs the exhaustive search to verify the Collatz conjecture using a DSP48E Xilinx Virtex-5 blocks, each of which contains one multiplier and one adder. The experimental results show that, our coprocessor can verify 3.88×10^8 64-bit numbers per second.

Modeling and Analysis of Real-Time Systems with Mutex Components

Guoqiang Li
BASICS, School of Software
Shanghai Jiao Tong University
Shanghai, China
li.g@sjtu.edu.cn

Xiaojuan Cai
BASICS, Department of Computer Science
Shanghai Jiao Tong University
Shanghai, China
cxj@sjtu.edu.cn

Shoji Yuen
Graduate School of Information Science
Nagoya University
Nagoya, Japan
yuen@is.nagoya-u.ac.jp

Abstract

Timed automata are popular for formally analyzing real-time systems. However, it is difficult to depict real-time systems with compositional components that interact with each other in a synchronization way or a mutex way. Synchronized components are modeled using parallel composition of timed automata by Larsen et al. [1]. This paper proposes controller automata to represent real-time systems with mutex components. In a controller automaton each state corresponds to a timed automaton with a built-in mechanism of relations, e.g., preemptions, in which every such timed automaton models a component of the real-time system. It is shown that given a strict partial order over states, an ordered controller automaton can be translated into a timed automaton. Various analyses are thus performed by checking the reachability to an error state.

Performance Analysis and Evaluation of Random Walk Algorithms on Wireless Networks

Keqin Li
Department of Computer Science
State University of New York
New Paltz, New York 12561, USA
Email: lik@newpaltz.edu

Abstract

We propose a model of dynamically evolving random networks and give an analytical result of the cover time of the simple random walk algorithm on a dynamic random symmetric planar point graph. Our dynamic network model considers random node distribution and random node mobility. We analyze the cover time of the parallel random walk algorithm on a complete network and show by numerical data that k parallel random walks reduce the cover time by almost a factor of k. We present simulation results for four random walk algorithms on random asymmetric planar point graphs. These algorithms include the simple random walk algorithm, the intelligent random walk algorithm, the parallel random walk algorithm, and the parallel intelligent random walk algorithm. Our random network model considers random node distribution and random battery transmission power.

Polylogarithmic Time Simulation of Reconfigurable Row/Column Buses by Static Buses

Susumu Matsumae
Department of Information Science
Saga University
Saga, Japan
Email: s.matsumae@computer.org

Abstract

This paper studies the difference in computational power between the mesh-connected parallel computers equipped with dynamically reconfigurable bus systems and those with static ones. The mesh with separable buses (MSB) is the mesh-connected computer with dynamically reconfigurable row/column buses. The broadcasting buses of the MSB can be dynamically sectioned into smaller bus segments by program control. We examine the impact of reconfigurable capability on the computational power of the MSB model, and investigate how computing power of the MSB decreases when we deprive the MSB of its reconfigurability. We show that any single step of the MSB of size $n \times n$ can be simulated in $O(\log n)$ time by the MSB without its reconfigurable function, which means that the MSB of size $n \times n$ can work with $O(\log n)$ step slowdown even if its dynamic reconfigurable function is disabled.

Parallel external sorting for CUDA-enabled GPUs with load balancing and low transfer overhead

Hagen Peters, Ole Schulz-Hildebrandt and Norbert Luttenberger
Research Group for Communication Systems
Department of Computer Science
Christian-Albrechts-University Kiel, Germany
{hap,osh,nl}@informatik.uni-kiel.de

Abstract

Sorting is a well-investigated topic in Computer Science in general and by now many efficient sorting algorithms for CPUs and GPUs have been developed. There is no swapping, paging, etc. available on GPUs to provide more virtual memory than physically available, thus if one wants to sort sequences that exceed GPU memory using the GPU the problem of external sorting arises.

In this contribution we present a novel merge-based external sorting algorithm for one or more CUDA-enabled GPUs. We reduce the performance impact of memory transfers to and from the GPU by using an approach similar to regular samplesort and by overlapping memory transfers with GPU computation. We achieve a good utilization of GPUs and load balancing among them by carefully choosing the samples and the amount of GPU memory used for computation.

We demonstrate the performance of our algorithm by extended testing. Using two GTX280 the implementation outperforms the fastest CPU sorting algorithms known to the authors.

Efficient Traffic Simulation Using the GCA Model

Christian Schäck, Rolf Hoffmann and Wolfgang Heenes
TU Darmstadt
FB Informatik, FG Rechnerarchitektur
Hochschulstraße 10, 64289 Darmstadt, Germany
{schaeck, hoffmann, heenes}@ra.informatik.tu-darmstadt.de

Abstract

The GCA (Global Cellular Automata) model consists of a collection of cells which change their states synchronously depending on the states of their neighbors like in the classical CA (Cellular Automata) model. In contrast to the CA model the neighbors can be freely and dynamically selected at runtime. The GCA model is applicable to a wide range of parallel algorithms. We present a mapping of the the well known Nagel-Schreckenberg algorithm for traffic simulation onto the GCA model using agents. The vehicles are considered as agents that are modeled as GCA cells with a certain state. The proposed GCA algorithm uses multiple data and link fields per cell to interconnect the relevant cells. An agent is connected to its agent in front, and an empty cell is connected to its agent behind. In the current generation t the position of an agent is already computed for the generation t+2 (one generation in advance). Thereby the agents movements and all cell updates can directly be calculated as defined by the cell rule. No searching of specific cells during the computation is necessary. The complexity is of $O(N)$ when simulating the N cells of an GCA sequentially. Compared to an optimized CA algorithm (with searching for agents) the GCA algorithm executes significantly faster, especially for low traffic densities and high vehicle speeds. Simulating 2048 cells and 204 agents on a multiprocessor system resulted in a speed-up (measured in clock ticks) of 14.75 for a system with 16 NIOS II processors configured on an FPGA.

Parallel Discrete Wavelet Transform using the Open Computing Language: a performance and portability study

Bharatkumar Sharma and Naga Vydyanathan
Siemens Corporate Technology
Bangalore, India
{bharatkumar.sharma, nagavijayalakshmi.vydyanathan}@siemens.com

Abstract

The discrete wavelet transform (DWT) is a powerful signal processing technique used in the JPEG 2000 image compression standard. The multi-resolution sub-band encoding provided by DWT allows for higher compression ratios, avoids blocking artifacts and enables progressive transmission of images. However, these advantages come at the expense of additional computational complexity. Achieving real-time or interactive compression/de-compression speeds, therefore, requires a fast implementation of DWT that leverages emerging parallel hardware systems. In this paper, we develop an optimized parallel implementation of the lifting-based DWT algorithm using the recently proposed Open Computing Language (OpenCL). OpenCL is a standard for cross-platform parallel programming of heterogeneous systems comprising of multi-core CPUs, GPUs and other accelerators. We explore the potential of OpenCL in accelerating the DWT computation and analyze the programmability, portability and performance aspects of this language. Our experimental analysis is done using NVIDIA's and AMD's drivers that support OpenCL.

Accelerating Mutual-Information-Based Registration on Multi-Core Systems

Jian Shen[1,2], Yurong Chen[2] He Li[1,2], Yimin Zhang[2] and Yinlong Xu[1]
[1]Dept. of Computer Science, Univ. of Science and Technology of China
[2]Intel Labs China, Intel Corporation
{yurong.chen}@intel.com

Abstract

Mutual-Information-Based Registration (MIBR) is an image registration method that maps points from one image to another. It has been widely used in medical image processing applications. However, MIBR is a very compute-intensive task, and fast processing speed is often required in medical diagnosis. Nowadays, with the multi-core processor becoming the mainstream, MIBR can be accelerated by fully utilizing the computing power of available multi-core processors. In this paper, we propose a parallel MIBR algorithm and present some optimization techniques to improve the implementation's performance. The result shows our optimized implementation can register a pair of 512x512x30 3D images in one second on an 8-core system, which meets the real-time processing requirement. We also conduct a detailed scalability and memory performance analysis on the multi-core system. The analysis helps us to identify the causes of bottlenecks, and make suggestion for future improvement on large-scale multi-core systems.

Cross Layer Design of Heterogeneous Virtual MIMO Radio Networks with Multi-Optimization

Wei Chen, Heh Miao, Liang Hong, Jim Savage and Husam Adas
College of Engineering, Technology & Computer Science
Tennessee State University
Nashville, TN 37209, USA
(wchen, hmiao, lhong)@tnstate.edu; jim.savage@nashville.gov; adas@comcast.net

Abstract

In virtual MIMO technology, distributed single- antenna radio systems cooperate on information transmission and reception as a multiple- antenna MIMO radio system. In this paper, a new virtual MIMO communication scheme and routing protocol are cross layered designed for a wireless sensor network (WSN) to jointly achieve reliability, energy efficiency and delay reduction. The proposed communication scheme minimizes the intra communication in virtual MIMO nodes. It saves energy and reduces latency simultaneously at transmission links even when the diameters of virtual nodes are large. Given a required reliability, the proposed routing protocol simultaneously optimizes energy and latency along the route. Virtual MIMO nodes/links are allowed to be heterogeneous in order to apply the visual MIMO technology to a general WSN. A virtual MIMO radio network can be formed for any underlying WSN. The network is reconfigurable with low cost. The performance evaluation shows that the proposed design can fully realize the potential of the virtual MIMO technology and largely improve network throughput, network lifetime, and reliability in a WSN.

Workshop 7
Communication Architecture for Clusters
CAC 2010

Optimizing MPI Communication Within Large Multicore Nodes with Kernel Assistance

Stéphanie Moreaud, Brice Goglin and Raymond Namyst
INRIA, LaBRI, Université of Bordeaux
351, cours de la Lib?ation
F-33405 Talence – France
Email: {smoreaud,goglin,namyst}@labri.fr

David Goodell
Mathematics and Computer Science Division
Argonne National Laboratory
Argonne, IL 60439, USA
Email: goodell@mcs.anl.gov

Abstract

As the number of cores per node increases in modern clusters, intra-node communication efficiency becomes critical to application performance. We present a study of the traditional double-copy model in MPICH2 and a kernel-assisted single-copy strategy with KNEM on different shared-memory hosts with up to 96 cores.

We show that KNEM suffers less from process placement on these complex architectures. It improves throughput up to a factor of 2 for large messages for both point-to-point and collective operations, and significantly improves NPB execution time. We detail when to switch from one strategy to the other depending on the communication pattern and we show that I/OAT copy offload only appears to be an interesting solution for older architectures.

Acceleration for MPI Derived Datatypes Using an Enhancer of Memory and Network

Noboru Tanabe
Corporate Research and Development Center
Toshiba Corporation
Kawasaki, Japan
noboru.tanabe@toshiba.co.jp

Hironori Nakajo
Division of Systems and Information Technology,
Institute of Symbiotic Science and Technology,
Tokyo University of Agriculture and Technology
Koganei, Japan
nakajo@cc.tuat.ac.jp

Abstract

This paper presents a support function for MPI derived datatypes on an enhancer of memory and network named DIMMnet-3. It is a network interface with vector access functions and multi-banked extended memory, which is under development. Semi-hardwired derived datatype communication based on RDMA with hardwired scatter and gather is proposed. This mechanism and MPI using it are implemented and validated on DIMMnet-2 which is a former prototype operating on DDR DIMM slot. The performance of scatter and gather transfer of 8byte elements with large interval by using vector commands of DIMMnet-2 is 6.8 compared with software on a host. Proprietary benchmark of MPI derived datatype communication for transferring a submatrix corresponding to a narrow HALO area is executed. Observed bandwidth on DIMMnet-2 is far higher than that for similar condition with VAPI based MPI implementation on InfniBand, even though very old generation FPGA, poorer CPU and motherboard are used. This function will avoid cache pollution and save CPU time for processing with local data which can be overlapped with communication. A new commercial machine with vector scatter/gather functions in NIC named SGI Altix UV is launched recently. It may be able to adopt our proposed concept partially, even though the capacity and fine grain access throughput of main memory attached with CPU are not enhanced on it.

Efficient Hardware Support for the Partitioned Global Address Space

Holger Fröning and Heiner Litz
Computer Architecture Group, Institute for Computer Engineering
University of Heidelberg
Mannheim, Germany
{holger.froening, heiner.litz}@ziti.uni-heidelberg.de

Abstract

We present a novel architecture of a communication engine for non-coherent distributed shared memory systems. The shared memory is composed by a set of nodes exporting their memory. Remote memory access is possible by forwarding local load or store transactions to remote nodes. No software layers are involved in a remote access, neither on origin or target side: a user level process can directly access remote locations without any kind of software involvement. We have implemented the architecture as an FPGA-based prototype in order to demonstrate the functionality of the complete system. This prototype also allows real world measurements in order to show the performance potential of this architecture, in particular for fine grain memory accesses like they are typically used for synchronization tasks.

Overlapping Computation and Communication: Barrier Algorithms and ConnectX-2 CORE-*Direct* Capabilities

Richard L. Graham[1], Steve Poole[1], Pavel Shamis[2], Gil Bloch[2], Noam Bloch[2],
Hillel Chapman[2], Michael Kagan[2], Ariel Shahar[2], Ishai Rabinovitz[2], Gilad Shainer[2]
[1]Oak Ridge National Laboratory (ORNL), Oak Ridge, TN, USA
Email: {rlgraham,spoole}@ornl.gov
[2]Mellanox Technologies, Inc.
Email: {pasha,gil,noam,hillel,michael,ariels,ishai}@mellanox.co.il, shainer@mellanox.com

Abstract

This paper explores the computation and communication overlap capabilities enabled by the new CORE-*Direct* hardware capabilities introduced in the InfiniBand (IB) Host Channel Adapter (HCA) ConnectX-2. These capabilities enable the progression and completion of data-dependent communications sequences to progress and complete at the network level without any Central Processing Unit (CPU) involvement. We use the latency dominated nonblocking barrier algorithm in this study, and find that at 64 process count, a contiguous time slot of about 80 percent of the nonblocking barrier time is available for computation. This time slot increases as the number of processes participating increases. In contrast, CPU based implementations provide a time slot of up to 30 percent of the nonblocking barrier time. This bodes well for the scalability of simulations employing offloaded collective operations. These capabilities can be used to reduce the effects of system noise, and when using nonblocking collective operations have the potential to hide the effects of application load imbalance.

Designing Topology-Aware Collective Communication Algorithms for Large Scale InfiniBand : Case Studies with Scatter and Gather

Krishna Kandalla[1], Hari Subramoni[1], Abhinav Vishnu[2] and Dhabaleswar K. (DK) Panda[1]

[1]Department of Computer Science and Engineering,
The Ohio State University
{kandalla, subramon, panda}@cse.ohio-state.edu

[2]High Performance Computing Group,
Pacific Northwest National Laboratory
{abhinav.vishnu}@pnl.gov

Abstract

Modern high performance computing systems are being increasingly deployed in a hierarchical fashion with multi-core computing platforms forming the base of the hierarchy. These systems are usually comprised of multiple racks, with each rack consisting of a finite number of chassis, and each chassis having multiple compute nodes or blades, based on multi-core architectures. The networks are also hierarchical with multiple levels of switches. Message exchange operations between processes that belong to different racks involve multiple hops across different switches and this directly affects the performance of collective operations. In this paper, we take on the challenges involved in detecting the topology of large scale InfiniBand clusters and leveraging this knowledge to design efficient topology-aware algorithms for collective operations. We also propose a communication model to analyze the communication costs involved in collective operations on large scale supercomputing systems. We have analyzed the performance characteristics of two collectives, MPI_Gather and MPI_Scatter, on such systems and we have proposed topology-aware algorithms for these operations. Our experimental results have shown that the proposed algorithms can improve the performance of these collective operations by almost 54% at the micro-benchmark level.

Designing High-Performance and Resilient Message Passing on InfiniBand

Matthew J. Koop[1], Pavel Shamis[2], Ishai Rabinovitz[2] and Dhabaleswar K. (DK) Panda[3]
[1]High Performance Technologies, Inc (HPTi), mkoop@hpti.com
[2]Mellanox Technologies, {pasha, ishai}@mellanox.co.il
[3]Dept. of Computer Science and Engineering, The Ohio State University, panda@cse.ohio-state.edu

Abstract

Clusters featuring the InfiniBand interconnect are continuing to scale. As an example, the Ranger system at the Texas Advanced Computing Center (TACC) includes over 60,000 cores with nearly 4,000 InfiniBand ports. The latest Top500 list shows 30% of systems and over 50% of the top 100 are now using InfiniBand as the compute node interconnect. As these systems continue to scale, the Mean-Time-Between-Failure (MTBF) is reducing and additional resiliency must be provided to the important components of HPC systems, including the MPI library.

In this paper we present a design that leverages the reliability semantics of InfiniBand, but provides a higher-level of resiliency. We are able to avoid aborting jobs in the case of network failures as well as failures on the endpoints in the InfiniBand Host Channel Adapters (HCA). We propose reliability designs for rendezvous designs using both Remote DMA (RDMA) read and write operations. We implement a prototype of our design and show that performance is near-identical to that of a non-resilient design. This shows that we can have both the performance and the network reliability needed for large-scale systems.

Index Tuning for Adaptive Multi-Route Data Stream Systems

Karen Works
Worcester Polytechnic Institute
Worcester, MA USA
Email: kworks@cs.wpi.edu

Elke A. Rundensteiner
Worcester Polytechnic Institute
Worcester, MA USA
Email: rundenst@cs.wpi.edu

Emmanuel Agu
Worcester Polytechnic Institute
Worcester, MA USA
Email: emmanuel@cs.wpi.edu

Abstract

Adaptive multi-route query processing (AMR) is an emerging paradigm for processing stream queries in highly fluctuating environments. AMR dynamically routes batches of tuples to operators in the query network based on routing criteria and up-to-date system statistics. In the context of AMR systems, indexing, a core technology for efficient stream processing, has received little attention. Indexing in AMR systems is demanding as indices must adapt to serve continuously evolving query paths while maintaining index content under high volumes of data. Our Adaptive Multi-Route Index (AMRI) employs a bitmap design. Our AMRI design is both versatile in serving a diverse ever changing workload of multiple query access patterns as well as lightweight in terms of maintenance and storage requirements. In addition, our AMRI index tuner exploits the hierarchical interrelationships between query access patterns to compress the statistics collected for assessment. Our experimental study using synthetic data streams has demonstrated that AMRI strikes a balance between supporting effective query processing in dynamic stream environments while keeping the overhead to a minimum.

Towards Execution Guarantees for Stream Queries

Rafael J. Fernández-Moctezuma, David Maier and Kristin A. Tufte
Department of Computer Science, Portland State University
1900 SW 4th Ave. Portland, OR 97207
Email: {rfernand, maier, tufte}@cs.pdx.edu
Telephone: (503) 725C2406
Fax: (503) 725C3211

Abstract

The unbounded nature of data streams and the low-latency requirements of stream processing present interesting challenges in Data Stream Management System (DSMS) design. Streaming query operators are typically designed to produce results with low latency, as well as to efficiently manage their state. Stream-progress delimitation techniques, such as punctuation, can help query operators achieve these goals. In this work, we look at deriving execution guarantees with respect to result production and state management for complete queries over punctuated streams. These guarantees are derived before query execution. We formalize notions of successful stream processing at an operator level, and extend these definitions to stream queries as a whole. We introduce a framework, punctuation contracts, for analyzing data processing and punctuation propagation from input to output on individual operators. We then use our framework to analyze complete queries and determine, prior to execution, if every valid input is eventually emitted, and no item remains in operator state indefinitely. Finally, we discuss extensions needed to bound query memory requirements; we describe four stream properties that can be used to help understand and quantify memory and CPU usage.

Exploiting Constraints to Build a Flexible and Extensible Data Stream Processing Middleware

Nazario Cipriani, Carlos Lubbe and Alexander Moosbrugger
Universität Stuttgart, Institute of Parallel and Distributed Systems,
Universitätsstraße 38, 70569 Stuttgart, Germany
{cipriani | luebbe | moosbrar}@ipvs.uni-stuttgart.de

Abstract

A wide range of real-time applications process stream-based data. To process this stream-based data in an application-independent manner, many stream processing systems have been built. However, none of them reached a huge domain of applications, such as databases did. This is due to the fact that they do not consider the specific needs of real-time applications. For instance, an application which visualizes stream-based data has stringent timing constraints, or may even need a specific hardware environment to smoothly process the data. Furthermore, users may even add additional constraints. E.g., for security reasons they may want to restrict the set of nodes that participates in processing. Thus, constraints naturally arise on different levels of query processing.

In this work we classify constraints that occur on different levels of query processing. Furthermore we propose a scheme to classify the constraints and show how these can be integrated into the query processing of the distributed data stream middleware NexusDS.

Distributed Monitoring of Conditional Entropy for Anomaly Detection in Streams

Chrisil Arackaparambil, Sergey Bratus, Joshua Brody and Anna Shubina
Dept. of Computer Science, Dartmouth College
Hanover, NH 03755, USA
{cja, sergey, jbrody, ashubina}@cs.dartmouth.edu

Abstract

In this work we consider the problem of monitoring information streams for anomalies in a scalable and efficient manner. We study the problem in the context of network streams where the problem has received significant attention.

Monitoring the empirical Shannon entropy of a feature in a network packet stream has previously been shown to be useful in detecting anomalies in the network traffic. Entropy is an information-theoretic statistic that measures the variability of the feature under consideration. Anomalous activity in network traffic can be captured by detecting changes in this variability.

There are several challenges, however, in monitoring this statistic. Computing the statistic efficiently is non-trivial. Further, when monitoring multiple features, the streaming algorithms proposed previously would likely fail to keep up with the ever-increasing channel bandwidth of network traffic streams. There is also the concern that an adversary could attempt to mask the effect of his attacks on variability by a mimicry attack disguising his traffic to mimic the distribution of normal traffic in the network, thus avoiding detection by an entropy monitoring sensor. Also, the high rate of false positives is a big problem with Intrusion Detection Systems, and the case of entropy monitoring is no different.

In this work we propose a way to address the above challenges. First, we leverage recent progress in sketching algorithms to develop a *distributed* approach for computing entropic statistics accurately, at reasonable memory costs. Secondly, we propose monitoring not only regular entropy, but the related statistic of conditional entropy, as a more reliable measure in detecting anomalies. We implement our approach and evaluate it with real data collected at the link layer of an 802.11 wireless network.

Workshop 8
High-Performance, Power-Aware Computing
HPPAC 2010

VMeter: Power Modelling for Virtualized Clouds

Ata E Husain Bohra and Vipin Chaudhary
Department of Computer Science and Engineering,
University at Buffalo, State University of New York, New York, U.S.A.
{aehusain, vipin}@buffalo.edu

Abstract

Datacenters are seeing unprecedented growth in recent years. The energy requirements to operate these large scale facilities are increasing significantly both in terms of operation cost as well as their indirect impact on ecology due to high carbon emissions. There are several ongoing research efforts towards the development of an integrated cloud management system to provide comprehensive online monitoring of resource utilization along with the implementation of power-aware policies to reduce the total energy consumption. However, most of these techniques provide online power monitoring based on the power consumption of a physical node running one or more Virtual Machines (VM). They lack a fine-grained mechanism to profile the power of an individual hosted VM. In this work we present a novel power modelling technique, VMeter, based on online monitoring of system-resources having high correlation with the total power consumption. The monitored system sub-components include: CPU, cache, disk, and DRAM. The proposed model predicts instantaneous power consumption of an individual VM hosted on a physical node besides the full system power consumption. Our model is validated using computationally diverse and industry standard benchmark programs. Our evaluation results show that our model is able to predict instantaneous power with an average mean and median accuracy of 93% and 94%, respectively, against the actual measured power using an externally attached power meter.

Characterizing Energy Efficiency of I/O Intensive Parallel Applications on Power-Aware Clusters

Rong Ge and Xizhou Feng
Dept. of Mathematics, Statistics, and Computer Science
Marquette University
Milwaukee, WI 53201, USA
{rong.ge,xizhou.feng}@marquette.edu

Sindhu Subramanya and Xian-he Sun
Dept. of Computer Science
Illinois Institute of Technology
Chicago, IL 60616, USA
sindhu.subramanya@gmail.com, sun@iit.edu

Abstract

Energy efficiency and parallel I/O performance have become two critical measures in high performance computing (HPC). However, there is little empirical data that characterize the energy-performance behaviors of parallel I/O workload. In this paper, we present a methodology to profile the performance, energy, and energy efficiency of parallel I/O access patterns and report our findings on the impacting factors of parallel I/O energy efficiency. Our study shows that choosing the right buffer size can change the energy-performance efficiency by up to 30 times. High spatial and temporal spacing can also lead to significant improvement in energy-performance efficiency (about 2X). We observe CPU frequency has a more complex impact, depending on the IO operations, spatial and temporal, and memory buffer size. The presented methodology and findings are useful for evaluating the energy efficiency of I/O intensive applications and for providing a guideline to develop energy efficient parallel I/O technology.

The Green500 List: Year Two

Wu-chun Feng and Heshan Lin
{feng,hlin2}@cs.vt.edu
Department of Computer Science
Virginia Tech
Blacksburg, VA 24060

Abstract

The Green500 turned two years old this past November at the ACM/IEEE SC|09 Conference. As part of the grassroots movement of the Green500, this paper takes a look back and reflects on how the Green500 has evolved in its second year as well as since its inception. Specifically, it analyzes trends in the Green500 and reports on the implications of these trends. In addition, based on significant feedback from the high-end computing (HEC) community, the Green500 announced three exploratory sub-lists: the Little Green500, the Open Green500, and the HPCC Green500, which are each discussed in this paper.

Reducing Grid Energy Consumption through Choice of Resource Allocation Method

Timothy M. Lynar, Ric D. Herbert, Simon and William J. Chivers
School of Design, Communication, and Information Technology
The University of Newcastle
Ourimbah, NSW, Australia
timothy.lynar@newcastle.edu.au

Abstract

Energy consumption is an increasingly important consideration in computing. High-performance computing environments consume substantial amounts of energy, at an increasing financial and environmental cost. We explore the possibility of reducing the energy consumption of a grid of heterogeneous computers through appropriate resource allocation strategies. We examine a number of possible grid workload scenarios and analyse the impact of different resource allocation mechanisms on energy consumption. We perform this analysis first on a cluster of heterogeneous nodes, then on a grid of several clusters. Our results show that different resource allocation mechanisms perform better under different scenarios, and that selection of an appropriate resource allocation mechanism can significantly reduce the total grid energy consumption.

BSLD Threshold Driven Power Management Policy for HPC Centers

Maja Etinski[1] maja.etinski@bsc.es
Julita Corbalan[1,2] julita.corbalan@bsc.es
Jesus Labarta[1,2] jesus.labarta@bsc.es
Mateo Valero[1,2] mateo.valero@bsc.es

[1]Barcelona Supercomputing Center
Jordi Girona 31, 08034 Barcelona, Spain
[2]Department of Computer Architecture
Technical University of Catalonia

Abstract

In this paper, we propose a power-aware parallel job scheduler assuming DVFS enabled clusters. A CPU frequency assignment algorithm is integrated into the well established EASY backfilling job scheduling policy. Running a job at lower frequency results in a reduction in power dissipation and accordingly in energy consumption. However, lower frequencies introduce a penalty in performance. Our frequency assignment algorithm has two adjustable parameters in order to enable fine grain energy-performance trade-off control. Furthermore, we have done an analysis of HPC system dimension. This paper investigates whether having more DVFS enabled processors for same load can lead to better energy efficiency and performance. Five workload traces from systems in production use with up to 9216 processors are simulated to evaluate the proposed algorithm and the dimensioning problem. Our approach decreases CPU energy by 7% - 18% on average depending on allowed job performance penalty. Using the power-aware job scheduling for 20% larger system, CPU energy needed to execute same load can be decreased by almost 30% while having same or better job performance.

Scheduling Parallel Tasks on Multiprocessor Computers with Efficient Power Management

Keqin Li
Department of Computer Science
State University of New York
New Paltz, New York 12561, USA
Email: lik@newpaltz.edu

Abstract

In this paper, scheduling parallel tasks on multiprocessor computers with dynamically variable voltage and speed is addressed as combinatorial optimization problems. Our scheduling problems are defined such that the energy-delay product is optimized by fixing one factor and minimizing the other. It is noticed that power-aware scheduling of parallel tasks has rarely been discussed before. Our investigation in this paper makes some initial attempt to energy efficient scheduling of parallel tasks on multiprocessor computers with dynamic voltage and speed. Our scheduling problems contain three nontrivial subproblems, namely, system partitioning, task scheduling, and power supplying. The harmonic system partitioning and processor allocation scheme is used, which divides a multiprocessor computer into clusters of equal sizes and schedules tasks of similar sizes together to increase processor utilization. A three-level energy/time/power allocation scheme is adopted for a given schedule, such that the schedule length is minimized by consuming given amount of energy or the energy consumed is minimized without missing a given deadline. The performance of our heuristic algorithms is analyzed and accurate performance bounds are derived. Simulation data which validate our analytical results are also presented. It is found that our analytical results provide very accurate estimation of the expected normalized schedule length and the expected normalized energy consumption, and that our heuristic algorithms are able to produce solutions very close to optimum.

Performance Evaluation of a Green Scheduling Algorithm for Energy Savings in Cloud Computing

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

Abstract

With energy shortages and global climate change leading our concerns these days, the power consumption of datacenters has become a key issue. Obviously, a substantial reduction in energy consumption can be made by powering down servers when they are not in use. This paper aims at designing, implementing and evaluating a Green Scheduling Algorithm integrating a neural network predictor for optimizing server power consumption in Cloud computing. We employ the predictor to predict future load demand based on historical demand. According to the prediction, the algorithm turns off unused servers and restarts them to minimize the number of running servers, thus minimizing the energy use at the points of consumption to benefit all other levels. For evaluation, we perform simulations with two load traces. The results show that the PP20 mode can save up to 46.3

T-NUCA - A Novel Approach to Non-Uniform Access Latency Cache Architectures for 3D CMPs

Konrad Malkowski, Padma Raghavan, Mahmut Kandemir and Mary Jane Irwin
{malkowsk, raghavan, kandemir, mji}@cse.psu.edu
Department of Computer Science and Engineering
The Pennsylvania State University
University Park, PA 16802, USA

Abstract

We consider a non-uniform access latency cache architecture (NUCA) design for 3D chip multi-processors (CMPs) where cache structures are divided into small banks interconnected by a network-on-chip (NoC). In earlier NUCA designs, data is placed in banks either statically (S-NUCA) or dynamically (D-NUCA). In both S-NUCA and D-NUCA designs, scaling to hundreds of cores can pose several challenges. Thus, we propose a new NUCA architecture with an inclusive, octal tree-based, hierarchical directory (T-NUCA-8), with the potential to scale to hundreds of cores with performance comparable to D-NUCA at a fraction of the energy cost. Our evaluations indicate that relative to D-NUCA, our T-NUCA-8 reduces network usage by 92%, energy by 87%, and EDP by 87%, at performance cost of 10%.

Integrated Energy-Aware Cyclic and Acyclic Scheduling for Clustered VLIW Processors

Jimmy Bahuleyan, Rahul Nagpal and Y. N. Srikant
Department of Computer Science and Automation
Indian Institute of Science
Bangalore, India
{jimmybahuleyan,rahul,srikant}@csa.iisc.ernet.in

Abstract

The technological trend towards smaller feature size and related implications of reduced threshold voltage and increased number of transistors pose a power management challenge. Leakage power dominates total processor power in smaller technologies. In VLIW and clustered VLIW architectures, the large number of functional units with relatively simpler issue logic contribute significantly to the overall power consumption. The underutilization of functional units (because of inherent limitations in ILP among others) causes a significant amount of this power to be consumed in the form of leakage power.

Architectural schemes proposed in the past suffer from a limited program view thereby compromising a good amount of energy savings in the form of extra transitions to/from low leakage mode. Energy aware scheduling in the context of VLIW architectures has mostly focused on acyclic scheduling. In this paper, we propose a simple and integrated compiler directed scheme that obtains significant energy savings in functional units for VLIW and clustered VLIW and works equally well for both cyclic as well as acyclic scheduled regions.

Our compiler directed scheme increases energy savings in functional units up to 9% and 26% in VLIW architecture and clustered VLIW architecture respectively. We provide a preliminary experimental analysis of our algorithms using the Trimaran 4.0 compiler infrastructure.

Dynamic Core Partitioning for Energy Efficiency

Yang Ding, Mahmut Kandemir, Mary Jane Irwin and Padma Raghavan
CSE Department, Pennsylvania State University
{yding, kandemir, mji, raghavan}@cse.psu.edu

Abstract

Chip multiprocessors (CMPs) are expected to dominate the landscape of computer architecture in the near future. The potential performance gains that can be achieved by the use of CMPs depend, to a large extent, on how much parallelism can be extracted from applications. One effective way of utilizing CMP architectures is to execute multiple (potentially multi-threaded) applications at the same time. In this work, we propose and evaluate a dynamic (runtime) core partitioning scheme for CMPs that exploits application level information. Focusing on an optimization metric called the weighted energy-delay product gain (W-EDPG), we dynamically partition available cores across competing applications during the course of execution. This dynamic partitioning uses input from a curve fitting model to predict the best operating points for an application at runtime. It can generate nonuniform core allocations across applications (i.e., some applications may have more or fewer cores than others) if doing so increases the value of the W-EDPG metric. We compare this approach against several alternative schemes (including equal partitioning of cores and standard operating system based scheduling). Our experiments indicate that the proposed core partitioning scheme improves the WEDPG metric significantly (e.g., 14.0% on average over the equal partitioning scheme on a 16-core CMP when four multi-threaded applications are executing concurrently).

Workshop 9
High Performance Grid Computing
HPGC 2010

An Interoperable & Optimal Data Grid Solution for Heterogeneous and SOA based Grid- GARUDA

Payal Saluja, Prahlada Rao BB., ShashidharV, Paventhan A., Neetu Sharma
System Software Development Group
C-DAC Knowledge Park,
#1, Old Madras Road, Bangalore-560038, INDIA
{payals, prahladab, ShashidharV, neetus}@cdacb.ernet.in

Abstract

Storage plays an important role in sufficing the requirements of data intensive applications in a Grid computing environment. Current Scientific applications perform complex computational analysis, and consume/produce hundreds of terabytes of data. The authors in this paper have surveyed available data grid solutions, viz., Storage Resource Broker (SRB), Grid File System (GFS), Storage Resource Manager (SRM), iRODS and WS-DAI and presented their operational experiences in Service Oriented Architecture (SOA) based GARUDA grid. SOA introduces more challenges to achieve: availability, security, scalability and performance to the storage system. Based on the survey, the authors proposed GARUDA-Storage Resource Manager (GSRM) that adheres to SRM specifications. GSRM is a disk based SRM implementation based on DPM (Disk Pool manager) architecture. It addresses the various aspects like virtualization, security, latency, performance, and data availability. We discussed how GSRM architecture can leverage CDAC's Parallel File System (C-PFS).

Improvements of Common Open Grid Standards to Increase High Throughput and High Performance Computing Effectiveness on Large-scale Grid and e-Science Infrastructures

M. Riedel, M.S. Memon, A.S. Memon,
A. Streit, F. Wolf, Th. Lippert
Jülich Supercomputing Centre
Forschungszentrum Jülich
Jülich, Germany
m.riedel@fz-juelich.de

A. Konstaninov
Vilnius University
Vilnius, Lithuania

M. Marzolla
University of Bologna
Bologna, Italy

B. Konya, O. Smirnova
Lund University
Lund, Sweden

L. Zangrando
INFN
Padova, Italy

J. Watzl, D.Kranzlmüller
Ludwig Maximillians University Munich
Munich, Germany

Abstract

Grid and e-science infrastructure interoperability is an increasing demand for Grid applications but interoperability based on common open standards adopted by Grid middle-wares are only starting to emerge on Grid infrastructures and are not broadly provided today. In earlier work we have shown how open standards can be improved by lessons learned from cross-Grid applications that require access to both, High Throughput Computing (HTC) resources as well as High Performance Computing (HPC) resources. This paper provides more insights in several concepts with a particular focus on effectively describing Grid job descriptions in order to satisfy the demands of e-scientists and their cross- Grid applications. Based on lessons learned over years gained with interoperability setups between production Grids such as EGEE, DEISA, and NorduGrid, we illustrate how common open Grid standards (i.e. JSDL and GLUE2) can take cross- Grid application experience into account.

A Distributed Diffusive Heuristic for Clustering a Virtual P2P Supercomputer

Joachim Gehweiler and Henning Meyerhenke
Department of Computer Science
University of Paderborn
Fürstenallee 11, D-33102 Paderborn, Germany
Email: {joge,henningm}@upb.de

Abstract

For the management of a virtual P2P supercomputer one is interested in subgroups of processors that can communicate with each other efficiently. The task of finding these subgroups can be formulated as a graph clustering problem, where clusters are vertex subsets that are densely connected within themselves, but sparsely connected to each other. Due to resource constraints, clustering using global knowledge (i. e., knowing (nearly) the whole input graph) might not be permissible in a P2P scenario, e. g., because collecting the data is not possible or would consume a high amount of resources. That is why we present a distributed heuristic using only limited local knowledge for clustering static and dynamic graphs.

Based on disturbed diffusion, our algorithm DIDIC implicitly optimizes cut-related quality measures such as modularity. It thus settles between distributed clustering algorithms for other quality measures (e. g., energy efficiency in the field of ad-hoc-networking) and graph clustering algorithms optimizing cut-related measures with global knowledge. Our experiments with graphs resembling a virtual P2P supercomputer show the promising potential of the new approach: Although each node starts with a random cluster number, may communicate only with its direct neighbors within the graph, and requires only a small amount of additional memory space, the solutions computed by DIDIC converge to clusterings that are comparable in quality to those computed by the established non-distributed graph clustering library mcl, whose main algorithm uses global knowledge.

How Algorithm Definition Language (ADL) Improves the Performance of SmartGridSolve Applications

Michele Guidolin
School of Engineering
University of Exeter
Exeter, EX4 4QF, UK
m.guidolin@exeter.ac.uk

Thomas Brady and Alexey Lastovetsky
School of Computer Science and Informatics
University College Dublin
Belfield, Dublin 4, Ireland
{thomasbrady, alexey.lastovetsky}@ucd.ie

Abstract

In this paper, we study the importance of languages for the specification of algorithms in high performance Grid computing. We present one such language, the Algorithm Definition Language (ADL), designed and implemented for the use in conjunction with SmartGridSolve. We demonstrate that the use of this type of language can significantly improve the performance of Grid applications. We discuss how ADL can be used to improve the execution of some typical algorithms that use conditional statements, iterative computations and adaptive methods. We present experimental results demonstrating significant performance gains due to the use of ADL.

GridP2P: Resource Usage in Grids and Peer-to-Peer Systems

Sérgio Esteves Luís Veiga Paulo Ferreira
sesteves@gsd.inesc-id.pt luis.veiga@inesc-id.pt paulo.ferreira@inesc-id.pt
INESC-ID/IST, Distributed Systems Group, Rua Alves Redol, 9, 1000-029 Lisboa, Portugal

Abstract

The last few years have witnessed huge growth in computer technology and available resources throughout the Internet. These resources can be used to run CPU-intensive applications requiring long periods of processing time.

Grid systems allow us to take advantage of available resources lying over a network. However, these systems impose several difficulties to their usage (e.g. heavy authentication and configuration management); in order to overcome them, Peer-to-Peer systems provide open access making the Grid available to any user.

Our solution consists of a platform for distributed cycle sharing which attempts to combine Grid and Peer-to-Peer models. A major goal is to allow any ordinary user to use remote idle cycles in order to speedup commodity applications. On the other hand, users can also provide spare cycles of their machines when they are not using them.

Our solution encompasses the following functionalities: application management, job creation and scheduling, resource discovery, security policies, and overlay network management. The simple and modular organization of this system allows that components can be changed at minimum cost. In addition, the use of history-based policies provides powerful usage semantics concerning the resource management.

A Grid Simulation Framework to Study Advance Scheduling Strategies for Complex Workflow Applications

Adan Hirales-Carbajal and Andrei Tchernykh
Computer Science Department
CICESE Research Center
Ensenada, Baja California, Mexico
e-mail: ahirales@uabc.mx,
chernykh@cicese.mx

Thomas Röblitz, Ramin Yahyapour
IT und Medien Centrum & Fakultät für Informatik
Technische Universität Dortmund
Dortmund, Germany
e-mail: thomas.roeblitz@udo.edu,
ramin.yahyapour@.udo.edu

Abstract

Workflow scheduling in Grids becomes an important area as it allows users to process large scale problems in an atomic way. However, validating the performance of workflow scheduling strategies in real production environment cannot be feasibly carried out. The complexity of production systems, dynamicity of Grid execution environments, and the difficulty to reproduce experiments, make workflow scheduling production systems a complex research environment. Instead, this work is based on a trace driven simulator.

This work presents workflow scheduling support as an extension to the Teikoku Grid Scheduling Framework (tGSF). tGSF was developed as a response for a standard compliant and trace based Grid scheduling simulation environment. Workflow scheduling is provided via a second layer of Grid scheduler, extensible to new workflow and parallel scheduling strategies. This work also includes a usage case scenario, which illustrates how this extension can be used for quantitative experimental study.

Meta-Scheduling in Advance using Red-Black Trees in Heterogeneous Grids

Luis Tomás, Carmen Carrión and Blanca Caminero
Dept. of Computing Systems
The University of Castilla La Mancha.
Albacete, Spain
{luistb, carmen, blanca}@dsi.uclm.es

Agustín Caminero
Dept. of Communication and Control Systems
The National University of Distance Education.
Madrid, Spain
accaminero@scc.uned.es

Abstract

The provision of Quality of Service in Grid environments is still an open issue that needs attention from the research community. One way of contributing to the provision QoS in Grids is by performing meta-scheduling of jobs in advance, that is, jobs are scheduled some time before they are actually executed. In this way, the apropprate resources will be available to run the job when needed, so that QoS requirements (i.e., deadline) are met.

This paper presents two new techniques, implemented over the red-black tree data structure, to manage the idle/busy periods of resources. One of them takes into account the heterogeneity of resources when estimating the execution times of jobs. A performance evaluation using a real testbed is presented that illustrates the efficiency of this approach to meet the QoS requirements of users.

SPSE: A Flexible QoS-based Service Scheduling Algorithm for Service-Oriented Grid

Laiping Zhao[1], Yizhi Ren[1,2], Mingchu Li[2], and Kouichi Sakurai[3] [1]Department of Informatics
Kyushu University, Fukuoka, Japan
Email: {zlp,ren}@itslab.csce.kyushu-u.ac.jp
[2]School of Software
Dalian University of Technology, Dalian, China
Email: li_mingchu@yahoo.com
[3]Department of Informatics
Kyushu University, Fukuoka, Japan
Email: sakurai@inf.kyushu-u.ac.jp

Abstract

With the development of the Grid computing, increased attention is paid to services and user personalization. How to search and schedule the most suitable service for an end user direct affects the popularization use of service oriented Grid. Inspired from the mode of web search engine, such as Yahoo, Google, this paper proposes an innovative service searching and scheduling algorithm (SPSE: Service Providers Search Engine) for the Grid. The SPSE sorts all services from Internet and returns the most appropriate ones to the end user. Compared with the existing scheduling algorithms, our method is much more flexible in meeting user's QoS requirements, especially supporting the multiobjective and user personalization. The related simulation experiments show that our method performs well in scalability, and can capture user's preferences value precisely and automatically.

Fault-Tolerance for PastryGrid Middleware

Heithem Abbes[1,2] Christophe Cérin[1,2] Mohamed Jemni[2] Yazid Missaoui[2]
[1,2]LIPN CNRS UMR 7030 — Université Paris 13,
99, avenue Jean-Baptiste Clément, 93430 Villetaneuse, FRANCE
[2]Research Unit UTIC, ESSTT, University of Tunis,
5, Av. Taha Hussein, B.P. 56, Bab Mnara, Tunis, TUNISIA
{Heithem.Abbes,Christophe.Cerin}@lipn.univ-paris13.fr
{Mohamed.Jemni,Yasid.Missaoui}@esstt.rnu.tn

Abstract

This paper analyses the performance of a decentralized and fault-tolerant software layer for Desktop Grid resources management. The grid middleware under concern is named PastryGrid. Its main design principle is to eliminate the need for a centralized server, therefore to remove the single point of failure and bottleneck of existing Desktop Grids. PastryGrid (based on Pastry) supports the execution of distributed application with precedence between tasks in a decentralized way. Indeed, each node can play alternatively the role of client or server. Our main contribution is to propose a fault tolerant mechanism for PastryGrid middleware. Since the management of PastryGrid is distributed over the participants without central manager, its control becomes a challenging problem, especially when dealing with faults. The experimental results on the Grid'5000 testbed demonstrate that our decentralized fault-tolerant system is robust because it supports high fault rates.

Workshop 10

Workshop on System Management Techniques, Processes, and Services

SMTPS 2010

Desktop Workload Study with Implications for Desktop Cloud Resource Optimization

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

Abstract

Desktop cloud is a new delivery model in which end users connect to virtual desktops running in remote data centers. This paradigm offers multiple benefits both in terms of manageability as well as efficiency improvements. However, realizing this potential requires better understanding of desktop workload and its implications for desktop consolidation. We analyze CPU and memory usage on a sample of 35 desktops using a fine-grained 10 second averaging interval. Results provide insights into achievable efficiency improvements from desktop consolidation as well as detailed autocorrelation and variability behavior as a function of number of aggregated desktops. We also propose an interactivity classification method leading to functional form suitable for estimating residual durations of interactivity states. This finding can be leveraged in on-line proactive management algorithms for desktop cloud optimization.

Automation and Management of Scientific Workflows in Distributed Network Environments

Qishi Wu[1], Mengxia Zhu[2], Xukang Lu[1], Patrick Brown[2], Yunyue Lin[1], Yi Gu[1],
Fei Cao[2] and Michael A. Reuter[3]
[1]Department of Computer Science, University of Memphis, Memphis, TN 38152, USA
Email: {qishiwu,xlv,ylin1,yigu}@memphis.edu
[2]Department of Computer Science, Southern Illinois University, Carbondale, IL 62901, USA
Email: {mzhu,patiek,fcao}@cs.siu.edu
[3]Neutron Science Scattering Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA
Email: reuterma@ornl.gov

Abstract

Large-scale computation-intensive applications in various science fields feature complex DAG-structured workflows comprised of distributed computing modules with intricate inter-module dependencies. Supporting such workflows in heterogeneous network environments and optimizing their endto- end performance are crucial to the success of large-scale collaborative scientific applications. We design and develop a generic Scientific Workflow Automation and Management Platform (SWAMP), which contains a set of easy-to-use computing and networking toolkits for application scientists to conveniently assemble, execute, monitor, and control complex computing workflows in distributed network environments. The current version of SWAMP integrates the graphical user interface of Kepler to compose abstract workflows and employs Condor DAGMan for workflow dispatch and execution. SWAMP provides a web-based user interface to automate and manage workflow executions and uses a special workflow mapper to optimize the end-to-end workflow performance. A case study of the workflow for Spallation Neutron Source datasets in real networks is presented to show the efficacy of the proposed platform.

Simplifying solution deployment on a Cloud through composite appliances

Trieu Chieu, Alexei Karve, Ajay Mohindra and Alla Segal
IBM T.J. Watson Research Center
19 Skyline Drive
Hawthorne, NY
{tchieu,karve,ajaym,segal}@us.ibm.com

Abstract

Containing runaway IT costs is one of the top priorities for enterprises. Cloud Computing, with its on-demand provisioning capability on shared resources, has emerged as a new paradigm for managing IT costs. In this paper, we describe a framework to simplify deployment of complex solutions on a Cloud infrastructure. We discuss the concept of a composite appliance and show how it can be used to reduce management costs. We illustrate the benefits of our approach with a complex three-tiered solution that can be deployed and configured on a set of virtual machines instances without any manual intervention.

Formulating the Real Cost of DSM-Inherent Dependent Parameters in HPC Clusters

Mohsen Sharifi
School of Computer Engineering
Iran University of Science and Technology
Tehran, Iran
msharifi@iust.ac.ir

Alfredo Tirado-Ramos
Wallace H. Coulter Department of Biomedical Engineering
Georgia Institute of Technology and Emory University
School of Medicine
Atlanta, GA, USA
atirado@emory.edu

Ehsan Mousavi Khaneghah
School of Computer Engineering
Iran University of Science and Technology
Tehran, Iran
emousavi@comp.iust.ac.ir

Seyedeh Leili Mirtaheri
School of Computer Engineering
Iran University of Science and Technology
Tehran, Iran
mirtaheri@comp.iust.ac.ir

Abstract

The choice of an appropriate interprocess communication (IPC) mechanism is critical to the performance of distributed systems and parallel programs. There is however a trade-off between the accrued system performance and the imposed cost of the deployed IPC mechanism. Message passing (MP) and distributed shared memory (DSM) mechanisms have been extensively studied and compared, but the state of the art work on formulating the cost of DSM is not comprehensive. In this paper we distinguish between DSM-inherent and application-specific parameters that contribute most to the real cost of DSM, and focus on DSM-inherent parameters in high performance computing (HPC) clusters. The weights of each parameter's effectiveness on the cost, as well as the weights of their influence on each other are determined. The derived formula for the real cost of DSM is then calibrated by a clustering coefficient that varies with different sizes of clusters. We have used the principles of management and accounting sciences in calculating the real cost of DSM.

Combining Virtualization, Resource Characterization, and Resource Management to Enable Efficient High Performance Compute Platforms Through Intelligent Dynamic Resource Allocation

J. Brandt[1], F. Chen[2], V. De Sapio[1], A. Gentile[2], J. Mayo[1], P. Pebay[1], D. Roe[1], D. Thompson[1] and M. Wong[2]
Sandia National Laboratories MS [1]9159 / [2]9152 P.O. Box 969, Livermore, CA 94551 U.S.A.
{{brandt,fxchen,vdesap,gentile,jmayo,pppebay,dcroe,dcthomp,mhwong}|{ovis}}@sandia.gov

Abstract

Improved resource utilization and fault tolerance of large-scale HPC systems can be achieved through fine-grained, intelligent, and dynamic resource (re)allocation. We explore components and enabling technologies applicable to creating a system to provide this capability: specifically 1) Scalable fine-grained monitoring and analysis to inform resource allocation decisions, 2) Virtualization to enable dynamic reconfiguration, 3) Resource management for the combined physical and virtual resources and 4) Orchestration of the allocation, evaluation, and balancing of resources in a dynamic environment. We discuss both general and HPC-centric issues that impact the design of such a system. Finally, we present our prototype system, giving both design details and examples of its application in real-world scenarios.

ROME: Road Monitoring and Alert System through Geocache

Bin Zan, Tingting Sun, Marco Gruteser and Yanyong Zhang
WINLAB, Rutgers University
Technology Center of New Jersey
671 Route 1 South
North Brunswick, NJ 08902-3390
{zanb,sunting,gruteser,yyzhang}@winlab.rutgers.edu

Abstract

We present a road monitoring and alert system implemented using a novel Geocache concept to enable efficient spatial monitoring in a mobile distributed sensing scenario. Technology trends have led to the integration of positioning, communications, and sensing capabilities into mobile entities such as cars and cellular phones, enabling them to monitor and report on their surroundings. We consider scenarios where events of interest must be detected from aggregated readings of multiple devices. For example, road monitoring could infer road defects from increased vibrations or road hazards from repeated emergency braking at the same location. This raises the challenge of aggregating sensor data from multiple mobiles that have passed the same location with efficient usage of communication resources. We introduce the Geocache concept, which allows anchoring sensor information at specific spatial coordinates, rather than storing it in a designated node. The Geocache protocol in either its relayed or delayed variant will then opportunistically determine a storage vehicle near the Geocache and hand the data off as vehicles pass by. We show through simulations with synthetic and realistic automotive position traces that relayed Geocache reduce messaging overhead by 66% compared to a baseline periodic broadcasting scheme and a Boomerang Geocache can provide a further 71% reduction when only few of the passing cars are able to sense the event.

Initial Characterization of Parallel NFS Implementations

Weikuan Yu[1,2] and Jeffrey S. Vetter[2]
[1]Auburn University [2]Oak Ridge National Lab
wkyu@auburn.edu vetter@ornl.gov

Abstract

Parallel NFS (pNFS) is touted as an emergent standard protocol for parallel I/O access in various storage environments. Several pNFS prototypes have been implemented for initial validation and protocol examination. Previous efforts have focused on realizing the pNFS protocol to expose the best bandwidth potential from underlying file and storage systems. In this presentation, we provide an initial characterization of two pNFS prototype implementations, lpNFS (a Lustre-based parallel NFS implementation) and spNFS (another reference implementation from Network Appliance, Inc.). We show that both lpNFS and spNFS can faithfully achieve the primary goal of pNFS, i.e., aggregating I/O bandwidth from many storage servers. However, they both face the challenge of scalable metadata management. Particularly, the throughput of sp-NFS metadata operations degrades significantly with an increasing number of data servers. Even for the better-performing lpNFS, we discuss its architecture and propose a direct I/O request flow protocol to improve its performance.

Streaming, Low-latency Communication in On-line Trading Systems

Hari Subramoni[1,2], Fabrizio Petrini[1], Virat Agarwal[1] and Davide Pasetto[3]

[1]IBM TJ Watson, Yorktown Heights, NY 10598 USA
[2]The Ohio State University, Columbus, OH 43210 USA
[3]IBM Computational Science Center, Dublin (Ireland)
fpetrin@us.ibm.com, viratagarwal@us.ibm.com, pasetto_davide@ie.ibm.com

Abstract

This paper presents and evaluates the performance of a prototype of an on-line OPRA data feed decoder. Our work demonstrates that, by using best-in-class commodity hardware, algorithmic innovations and careful design, it is possible to obtain the performance of custom-designed hardware solutions.

Our prototype system integrates the latest Intel Nehalem processors and Myricom 10 Gigabit Ethernet technologies with an innovative algorithmic design based on the DotStar compilation tool. The resulting system can provide low latency, high bandwidth and the flexibility of commodity components in a single framework, with an end-to-end latency of less then four microseconds and an OPRA feed processing rate of almost 3 million messages per second per core, with a packet payload of only 256 bytes.

Business-Driven Capacity Planning of a Cloud-based IT Infrastructure for the Execution of Web Applications

Raquel Lopes, Francisco Brasileiro, Paulo Ditarso Maciel Jr.
Universidade Federal de Campina Grande
Departamento de Sistemas e Computação
Laboratório de Sistemas Distribuídos
Av. Aprígio Veloso, s/n, Bloco CO
58.429-900, Campina Grande - PB, Brasil
Emails: {raquel, fubica, pmaciel}@dsc.ufcg.edu.br

Abstract

With the emergence of the cloud computing paradigm and the continuous search to reduce the cost of running Information Technology (IT) infrastructures, we are currently experiencing an important change in the way these infrastructures are assembled, configured and managed. In this research we consider the problem of managing a computing infrastructure whose processing elements are acquired from *infrastructure-as-a-service* (IaaS) providers, and used to support the execution of long-lived on-line transactions processing applications, whose workloads experience huge fluctuations over time, such as Web applications. Resources can be acquired from IaaS providers in different ways, each with its own pricing scheme and associated quality of service. One acquisition model allows resources to be reserved for long periods of usage at a reduced usage price, while others allow dedicated resources to be instantiated on demand at any time, subject to the availability of resources, with a usage price that is usually larger than that paid for reserved resources. In this context, the problem that we address in this paper is how the provider of a Web application should plan the long-term reservation contracts with an IaaS provider, in such a way that its profitability is increased. We propose a model that can be used for guiding this capacity planning activity. We use the model to evaluate the gains that can be achieved with the judicious planning of the infrastructure capacity in a number of scenarios. We show that the gains can be substantial, specially when the load variation between normal operational periods and surge periods is large.

Scalability Analysis of Embarassingly Parallel Applications on Large Clusters

Fabrício Alves Barbosa da Silva
University of Lisbon
Faculty of Sciences, LASIGE
Lisbon, Portugal
fabricio@di.fc.ul.pt

Hermes Senger
Federal University of São Carlos
Department of Computer Science
São Carlos, Brazil
hermes@dc.ufscar.br

Abstract

This work presents a scalability analysis of embarrassingly parallel applications running on cluster and multi-cluster machines. Several applications can be included in this category. Examples are Bag-of-tasks (BoT) applications and some classes of online web services, such as index processing in online web search. The analysis presented here is divided in two parts: first, the impact of front end topology on scalability is assessed through a lower bound analysis. In a second step several task mapping strategies are compared from the scalability standpoint.

Autonomic Management of Distributed Systems using Online Clustering

Andres Quiroz, Manish Parashar and Ivan Rodero
NSF Center for Autonomic Computing
Department of Electrical & Computer Engineering
Rutgers, The State University of New Jersey
Email: {aquirozh,parashar,irodero}@cac.rutgers.edu

Abstract

Distributed computational infrastructures, as well as the applications and services that they support, are increasingly becoming an integral part of society and affecting every aspect of life. As a result, ensuring their efficient and robust operation is critical. However, the scale and overall complexity of these systems is growing at an alarming rate (current data centers contain tens to hundreds of thousands of computing and storage devices running complex applications), making the management of these systems extremely challenging and rapidly exceeding human capability. Furthermore, these systems require simultaneous management along multiple dimensions, including performance, quality of service, power, and reliability.

Workshop 11

Workshop on Parallel and Distributed Scientific and Engineering Computing

PDSEC 2010

Solving large sparse linear systems in a grid environment using Java

Raphaël Couturier
Laboratoire d'Informatique de l'Université de Franche-Comté
IUT Belfort-Montb?iard, Rue Engel Gros, BP 527,
90016 Belfort CEDEX, France
Email: Raphael.Couturier@univ-fcomte.fr
Fabienne Jézéquel
UPMC Univ Paris 06, UMR 7606
Laboratoire d'Informatique de Paris 6, 4 place Jussieu
75252 Paris CEDEX 05, France
Email: Fabienne.Jezequel@lip6.fr

Abstract

In this paper, we show how to solve large sparse linear systems in a grid environment using the Java language and the MPJ library for communication. We describe a parallel version of the GMRES method which takes into account the sparsity of the matrix for message exchanges among processors. Two implementations are compared: one in Java using MPJ and one in C using MPI. The performance of both codes is also compared with that of the PETSc library. Experiments have been carried out using the GRID'5000 platform, on the one hand, on a local cluster, and, on the other hand, on clusters located in distant geographical sites. It is noticeable that the performance of our solver in Java is comparable to the same solver written in C and also to the PETSc library. Our solver in Java allowed us to solve sparse systems of size up to 2 billions with two geographically distant sites.

Issues in Adaptive Mesh Refinement

William W. Dai
High Performance Computing Division
Los Alamos National Laboratory
Los Alamos, New Mexico, USA
E-mail: dai@lanl.gov

Abstract

In this paper, we present an approach for a patch-based adaptive mesh refinement (AMR) for multi-physics simulations. The approach consists of clustering, symmetry preserving, mesh continuity, flux correction, communications, management of patches, and dynamical load balance. Among the special features of this patch-based AMR are symmetry preserving, efficiency of refinement, special implementation of flux correction, and patch management in parallel computing environments. Here, higher efficiency of refinement means less unnecessarily refined cells for a given set of cells to be refined. To demonstrate the capability of the AMR framework, hydrodynamics simulations with many levels of refinement are shown in both two- and three-dimensions.

Solving the advection PDE on the Cell Broadband Engine

Georgios Rokos, Gerassimos Peteinatos, Georgia Kouveli, Georgios Goumas,
Kornilios Kourtis and Nectarios Koziris
Computing Systems Laboratory
National Technical University of Athens
Email: {grokos,gpeteinatos,gkouv,goumas,kkourt,nkoziris}@cslab.ece.ntua.gr

Abstract

In this paper we present the venture of porting two different algorithms for solving the two-dimensional advection PDE on the CBE platform, an in-place and an out-of-place one, and compare their computational performance, completion time and code productivity. Study of the advection equation reveals data dependencies which lead to limited performance and inefficient scaling to parallel architectures. We explore programming techniques and optimizations which maximize performance for these solver versions. The out-of-place version is straightforward to implement and achieves greater raw performance than the in-place one, but requires more computational steps to converge. In both cases, achieving high computational performance relies heavily on manual source code optimization, due to compiler incapability to do data vectorization and efficient instruction scheduling. The latter proves to be a key factor in pursuit of high GFLOPS measurements.

Storage Space Reduction for the Solution of Systems of Ordinary Differential Equations by Pipelining and Overlapping of Vectors

Matthias Korch and Thomas Rauber
Department of Computer Science
University of Bayreuth
Bayreuth, Germany
Email: {korch, rauber}@uni-bayreuth.de

Abstract

Systems of ordinary differential equations (ODEs) arise from the mathematical modeling of time-dependent processes. Many sequential and parallel numerical methods have been proposed that can simulate processes described by ODE systems with known initial state. One disadvantage common to the proposed methods is the large amount of storage space required if the ODE systems consist of many equations. Not only do they have to keep the solution of the ODE system corresponding to the current time step in memory, but also several intermediate solutions or results of evaluations of the right hand side function of the ODE system. In this paper, we present an approach based on pipelining and overlapping of vectors which can reduce the storage space of typical ODE solution methods such as Runge–Kutta (RK) and extrapolation methods. We analyze and compare the scalability of different implementation variants of embedded and iterated RK methods on several modern parallel computer systems. Our experiments show that, due to an increased locality of memory references, our approach leads to a good scalability behavior even on large numbers of processors.

Designing Scalable Many-core Parallel Algorithms for Min Graphs using CUDA

Quoc-Nam Tran
Lamar University, USA

Abstract

Removing redundant edges on a large graph is a fundamental problem in many practical applications such as verification of real-time systems and network routing. In this paper, we present the designs of scalable and efficient parallel algorithms for multiple many-core GPU devices using CUDA. Our algorithms expose substantial fine-grained parallelism while maintaining minimal global communication. By using the global scope of the GPU's global memory, coalescing the global memory reads and writes, and avoiding on-chip shared memory bank conflicts, we are able to achieve a large performance benefit with a speed-up of 2,500x on a desktop computer in comparison with a single core CPU program. We report our experiments on large graphs with up to 29K vertices using multiple GPU devices.

CUDA-based AES Parallelization with Fine-Tuned GPU Memory Utilization

Chonglei Mei, Hai Jiang, Jeff Jenness
Department of Computer Science
Arkansas State University
{chonglei.mei, hjiang, jeffj}@cs.astate.edu

Abstract

Current Graphics Processing Unit (GPU) presents large potentials in speeding up computationally intensive data parallel applications over traditional parallelization approaches since there are much more hardware threads inside GPUs than the computational cores available to common CPU threads. NVIDIA developed a generic GPU programming platform, CUDA, which allows programmers to utilize GPU through C programming language and parallelize applications in a similar way as in traditional multithreading approach. However, not all applications are suitable for this new platform. Only computationally intensive applications without strong dependency are good candidates. Although Advanced Encryption Standard (AES) does not belong to this group due to the light workload in its efficient implementation, this paper proposed an approach to arranging data in different GPU memory spaces properly, overcoming the extra communication delay, and still turning GPU into an effective accelerator. Experimental results have demonstrated its effectiveness by performance gains and proved that GPU can be used to accelerate more types of applications.

Performance Study of Mapping Irregular Computations on GPUs

Steven Solomon and Parimala Thulasiraman
Department of Computer Science
University of Manitoba
Winnipeg, Manitoba, Canada
{umsolom9, thulasir}@cs.umanitoba.ca

Abstract

Recently, Graphical Processing Units (GPUs) have become increasingly more capable and well-suited to general purpose applications. As a result of the GPUs high degree of parallelism and computational power, there has been a great deal of interest directed toward the platform for parallel application development. Much of the focus, however, has been on very regular applications that exhibit a high degree of data parallelism, as these applications map well to the GPU. Irregular applications, such as the Breadth First Search discussed in this paper, have not been as extensively studied and are more difficult to implement in an efficient fashion on the GPU. We will present both an implementation of the Breadth First Search algorithm as well as that of a Matrix Parenthesization algorithm. These pair of algorithms showcase similar synchronization behavior when implemented on a GPU using CUDA, enabling a more direct comparison between them. The results obtained can be used to showcase some of the synchronization issues present with irregular algorithms on the GPU.

Simulating Anomalous Diffusion on Graphics Processing Units

Karl Heinz Hoffmann[1], Michael Hofmann[2], Jens Lang[2], Gudula Rünger[2] and Steffen Seeger[1]
[1]Department of Physics, Chemnitz University of Technology, Germany
Email: {s.seeger,hoffmann}@physik.tu-chemnitz.de
[2]Department of Computer Science, Chemnitz University of Technology, Germany
Email: {mhofma,lajen,ruenger}@cs.tu-chemnitz.de

Abstract

The computational power of modern graphics processing units (GPUs) has become an interesting alternative in high performance computing. The specialized hardware of GPUs delivers a high degree of parallelism and performance. Various applications in scientific computing have been implemented such that computationally intensive parts are executed on GPUs. In this article, we present a GPU implementation of an application for the simulation of diffusion processes using random fractal structures. It is shown how the irregular computational structure that is inherent to the application can be implemented efficiently in the regular computing environment of a GPU. Performance results are shown to demonstrate the benefits of the chosen implementation approaches.

Prototype for a Large-Scale Static Timing Analyzer running on an IBM Blue Gene

Akintayo Holder
Computer Science Department
Rensselaer Polytechnic Institute
Troy, New York 12180
Email: holdea@cs.rpi.edu

Christopher D. Carothers
Computer Science Department
Rensselaer Polytechnic Institute
Troy, New York 12180
Email: chrisc@cs.rpi.edu

Kerim Kalafala
IBM Microelectronics
IBM Corporation
Fishkill, NY
Email: kalafala@us.ibm.com

Abstract

This paper focuses on parallelization of the classic static timing analysis (STA) algorithm for verifying timing characteristics of digital integrated circuits. Given ever-increasing circuit complexities, including the need to analyze circuits with billions of transistors, across potentially thousands of process corners, with accuracy tolerances down to the picosecond range, sequential execution of STA algorithms is quickly becoming a bottleneck to the overall chip design closure process. A message passing based parallel processing technique for performing STA leveraging an IBM Blue Gene/L supercomputing platform is presented. Results are collected for a small industrial 65 nm benchmarking design, where the algorithm demonstrates speedup of nearly 39 times on 64 processors and a peak of 119 times (without partitioning costs, speedup is 263 times) on 1024 processors. With an idealized synthetic circuit, the algorithm demonstrated 259 times speedup, 925 times speedup without partitioning overhead, on 1024 processors. To the best of our knowledge, this is the first result demonstrating scalable STA on the IBM Blue Gene.

Performance Prediction of Weather Forecasting Software on Multicore Systems

Javier Delgado, S. Masoud Sadjadi, Marlon Bright, Malek Adjouadi
College of Engineering and Computing
Florida International University
Miami, USA
{javier.delgado, sadjadi, marlon.bright, adjouadi}@fiu.edu

Hector A. Duran-Limon
Center for Economic Administrative Sciences
University of Guadalajara
Guadalajara, Mexico
hduran@cucea.udg.mx

Abstract

Performance prediction is valuable in different areas of computing. The popularity of lease-based access to high performance computing resources particularly benefits from accurate performance prediction. Most contemporary processors are employing multiple computing cores, which complicates the task of performance prediction. In this paper, we describe the methodology employed for predicting the performance of a popular weather forecasting application on systems with between 4 and 256 processors. An average prediction error of less than 10% was achieved after testing on three different multi-node, multicore systems.

Restructuring Parallel Loops to Curb False Sharing on Multicore Architectures

Santosh Sarangkar and Apan Qasem
Department of Computer Science
Texas State University
San Marcos, TX
{ss1729,apan}@txstate.edu

Abstract

The memory hierarchy of most multicore systems contains one or more levels of cache that is shared among multiple cores. The shared-cache architecture presents many opportunities for performance gains for multi-threaded applications. However, when not handled carefully, contention for the shared-cache can lead to performance degradation. This paper addresses the issue of cache interference that occurs when concurrent threads access data that reside on a shared cache block. We propose a new compiler technique that takes advantage of hardware prefetching and thread affinity features to ameliorate performance loss due to this type of interference. Preliminary evaluation on a dual-core and a quad-core platform shows that our strategy can be effective in reducing cache interference for multi-threaded applications that exhibit inter-core spatial locality.

Parallel Task for parallelizing object-oriented desktop applications

Nasser Giacaman and Oliver Sinnen
Department of Electrical and Computer Engineering
The University of Auckland
Auckland, New Zealand
Email: ngia003@aucklanduni.ac.nz, o.sinnen@auckland.ac.nz

Abstract

As multi-cores arrive for mainstream desktop systems, developers must invest the effort to parallelize their applications. We present Parallel Task (short ParaTask), a solution to assist the parallelization of object-oriented applications, with the unique feature of including support for the parallelization of graphical user interface (GUI) applications. In the simple, but common, cases concurrency is introduced with a single keyword. Due to the wide variety of parallelization needs, ParaTask integrates different task types into the same model, provides intuitive support for dependence handling, non-blocking notification, interim progress notification and exception handling in an asynchronous environment as well as supporting a pluggable task scheduling runtime (currently work-sharing, work-stealing and a combination of the two are supported). The performance is compared to traditional Java parallelization approaches using a variety of different workloads.

Application Tuning through Bottleneck-driven Refactoring

Guogjing Cong, I-Hsin Chung, Huifang Wen, David Klepacki, Hiroki Murata,
Yasushi Negishi and Takao Moriyama
IBM research

Abstract

To fully utilize the power of current high performance computing systems, high productivity to the end user is critical. It is a challenge to map an application to the target architecture efficiently. Tuning an application for high performance remains a daunting task, and frequently involves manual changes to the program. Recently refactoring techniques are proposed to rewrite or reorganize programs for various software engineering purposes. In our research we explore combining performance analysis with refactoring techniques for automated tuning that we expect to greatly improve the productivity of application deployment. We seek to build a system that can apply appropriate refactoring according to the bottleneck discovered. We demonstrate the effectiveness of this approach through the tuning of several scientific applications and kernels.

The Pilot Approach to Cluster Programming in C

J. Carter, W. B. Gardner, G. Grewal
Department of Computing and Information Science
University of Guelph
Guelph, Ontario, Canada
jcarter@uoguelph.ca
{wgardner,gwg}@cis.uoguelph.ca

Abstract

The Pilot library offers a new method for programming parallel clusters in C. Formal elements from Communicating Sequential Processes (CSP) were used to realize a process/channel model of parallel computation that reduces opportunities for deadlock and other communication errors. This simple model, plus an application programming interface (API) fashioned on C's formatted I/O, are designed to make the library easy for novice scientific C programmers to learn. Optional runtime services including deadlock detection help the programmer to debug communication issues. Pilot forms a thin layer on top of standard Message Passing Interface (MPI), preserving the latter's portability and efficiency, with little performance impact. MPI's powerful collective operations can still be accessed within the conceptual model.

Enhancing Adaptive Middleware for Quantum Chemistry Applications with a Database Framework

Lakshminarasimhan Seshagiri,
Meng-Shiou Wu and Masha Sosonkina
Scalable Computing Laboratory
Ames Laboratory
Ames, IA 50011, USA
Email: sln,mswu,masha@scl.ameslab.gov

Zhao Zhang
Department of Electrical
and Computer Engineering,
Iowa State University,
Ames, IA 50011, USA
Email: zzhang@iastate.edu

Mark S. Gordon and Michael W. Schmidt
Department of Chemistry
and Ames Laboratory,
Iowa State University,
Ames, IA 50011, USA
Email: mark,mike@si.msg.chem.iastate.edu

Abstract

Quantum chemistry applications such as the General Atomic and Molecular Electronic Structure System (GAMESS) that can execute on a complex peta-scale parallel computing environment has a large number of input parameters that affect the overall performance. The application characteristics vary according to the input parameters. This is due to the difference in the usage of resources like network bandwidth, I/O and main memory, according to the input parameters. Effective execution of applications in a parallel computing environment that share such resources require some sort of adaptive mechanism to enable efficient usage of these resources. In our previous work, we have integrated GAMESS with an adaptive middleware NICAN (Network Information Conveyer and Application Notification) for dynamic adaptations during heavy load conditions that modify execution of GAMESS computations on a per-iteration basis. This leads to better application performance. In this research, we have expanded the structure of NICAN in order to include other input parameters based on which application performance can be controlled. The application performance has been analyzed on different architectures and a tuning strategy has been identified. A generic database framework has been incorporated in the existing NICAN mechanism so as to aid this tuning strategy.

Scheduling instructions on hierarchical machines

Florent Blachot, Guillaume Huard, Johnatan Pecero, Erik Saule and Denis Trystram
LIG, Grenoble university, 51 avenue Jean Kuntzmann, 38380 Montbonnot St Martin, France

Abstract

The aim of this work is to study the problem of scheduling fine grain task graphs on hierarchical distributed systems with communication delay. We consider as a case study how to schedule the instructions on a processor that implements incomplete bypass (ST200). We show first how this problem can be expressed as scheduling unitary tasks on a hierarchical architecture with heavy communications between clustered units. The proposed analysis is generic and can be extended to other challenging problems like scheduling in clusters of multi-cores.

Our main result is an approximation algorithm based on list scheduling whose approximation ratio is the minimum of two expressions, the first one depends on the number of clusters while the second one depends on the communication delay. Experiments run on random graphs and on structured graphs demonstrate the effectiveness of the proposed approach.

Mapping Asynchronous Iterative Applications on Heterogeneous Distributed Architectures

Raphaël Couturier, David Laiymani and Sébastien Miquée
Laboratoire d'Informatique de Franche-Comté (LIFC)
IUT de Belfort-Montbéliard,2 Rue Engel Gros, BP 27, 90016 Belfort, France
Email: {raphael.couturier,david.laiymani,sebastien.miquee}@univ-fcomte.fr

Abstract

To design parallel numerical algorithms on large scale distributed and heterogeneous platforms, the asynchronous iteration model (AIAC) may be an efficient solution. This class of algorithm is very suitable since it enables communication/computation overlapping and it suppresses all synchronizations between computation nodes. Since target architectures are composed of more than one thousand heterogeneous nodes connected through heterogeneous networks, the need for mapping algorithms is crucial. In this paper, we propose a new mapping algorithm dedicated to the AIAC model. To evaluate our mapping algorithm we implemented it in the JaceP2P programming and executing environment dedicated to AIAC applications and we conducted a set of experiments on the Grid'5000 testbed. Results are very encouraging and show that the use of our algorithm brings an important gain in term of execution time (about 40%).

Investigating the robustness of adaptive dynamic loop scheduling on heterogeneous computing systems

Srishti Srivastava and Ioana Banicescu
Department of Computer Science & Eng.
Mississippi State University
{ss878@, ioana@cse.}msstate.edu

Florina M. Ciorba
Center for Advanced Vehicular Systems
Mississippi State University
florina@cavs.msstate.edu

Abstract

Dynamic Loop Scheduling (DLS) algorithms are a powerful approach towards improving the performance of scientific applications via load balancing. The adaptive DLS (ADLS) methods have been proven to be the most appropriate for effectively balancing such applications, due to the fact that they are designed to address highly irregular, stochastic behavior caused by algorithmic and systemic variations. To guarantee certain performance levels of such DLS methods, metrics are required to measure their robustness against various unpredictable variations of factors in the computing environment. In this paper, the focus is on investigating metrics for the robustness of two Adaptive Weighted Factoring (AWF) techniques, AWFB and AWF-C, as well as of the Adaptive Factoring (AF) technique. Two robustness metrics, called flexibility and resilience, are formulated for these techniques. We also discuss their computational complexity and give notes on their usefulness.

A Framework for FPGA Functional Units in High Performance Computing

Andreas Koltes
Department of Informatics and Mathematics
University of Passau
Passau, Germany
Email: koltes@ieee.org

John T. O'Donnell
Department of Computing Science
University of Glasgow
Glasgow, United Kingdom
Email: jtod@dcs.gla.ac.uk

Abstract

FPGAs make it practical to speed up a program by defining hardware functional units that perform calculations faster than can be achieved in software. Specialised digital circuits avoid the overhead of executing sequences of instructions, and they make available the massive parallelism of the components. The FPGA operates as a coprocessor controlled by a conventional computer. An application that combines software with hardware in this way needs an interface between a communications port to the processor and the signals connected to the functional units. We present a framework that supports the design of such systems. The framework consists of a generic controller circuit defined in VHDL that can be configured by the user according to the needs of the functional units and the I/O channel. The controller contains a register file and a pipelined programmable register transfer machine, and it supports the design of both stateless and stateful functional units. Two examples are described: the implementation of a set of basic stateless arithmetic functional units, and the implementation of a stateful algorithm that exploits circuit parallelism.

FG-MPI: Fine-grain MPI for Multicore and Clusters

Humaira Kamal and Alan Wagner
Department of Computer Science
University of British Columbia
Vancouver, British Columbia
Email: {kamal,wagner}@cs.ubc.ca

Abstract

MPI (Message Passing Interface) has been successfully used in the high performance computing community for years and is the dominant programming model. Current implementations of MPI are coarse-grained, with a single MPI process per processor, however, there is nothing in the MPI specification precluding a finer-grain interpretation of the standard. We have implemented Fine-grain MPI (FG-MPI), a system that allows execution of hundreds and thousands of MPI processes on-chip or communicating between chips inside a cluster.

FG-MPI uses fibers (coroutines) to support multiple MPI processes inside an operating system process. These are full-fledged MPI processes each with their own MPI rank. We have implemented a fine-grain version of MPICH2 middleware that uses the Nemesis communication subsystem for intranode and internode communication.

We present experimental results for a real-world application that uses thousands of MPI processes and compare its performance with the following fine-grain multicore languages: Erlang, Haskell, Occam-pi and POSIX threads. Our results show that FG-MPI scales well and outperforms many of these other programming languages used for parallel programming on multicore systems while retaining MPI's intranode and internode communication abilities.

Processor Affinity and MPI Performance on SMP-CMP Clusters

Chi Zhang, Xin Yuan and Ashok Srinivasan
Department of Computer Science, Florida State University, Tallahassee, FL 32306
{czhang,xyuan,asriniva}@cs.fsu.edu

Abstract

Clusters of Symmetric MultiProcessing (SMP) nodes with multi-core Chip-Multiprocessors (CMP), also known as SMP-CMP clusters, are becoming ubiquitous today. For Message Passing interface (MPI) programs, such clusters have a multi-layer hierarchical communication structure: the performance of intra-node communication is usually higher than that of inter-node communication; and the performance of intra-node communication is not uniform with communications between cores within a chip offering higher performance than communications between cores in different chips. As a result, the mapping from Message Passing Interface (MPI) processes to cores within each compute node, that is, *processor affinity*, may significantly affect the performance of intra-node communication, which in turn may impact the overall performance of MPI applications. In this work, we study the impacts of processor affinity on MPI performance in SMP-CMP clusters through extensive benchmarking and identify the conditions when processor affinity is (or is not) a major factor that affects performance.

The Resource Locating Strategy Based on Sub-domain Hybrid P2P Network Model

Yuhua Liu[1], Yuling Li[1], Laurence T. Yang[2], Naixue Xiong[3], Longquan Zhu[1], Kaihua Xu[4]
[1]Department of Computer Science, Huazhong Normal University, Wuhan, 430079, China
[2]Department of Computer and Information Science, St.Francis Xavier University, Canada
[3]Department of Computer Science, Georgia State University, Atlanta, GA, USA
[4]Research Center of the Digital Space Technology, Huazhong Normal University, China

Abstract

P2P networks are important parts of the next generation Internet and P2P file sharing has become one of the most important internet application systems in the world. But P2P network faces a challenge in locating resource: the unstructured P2P network has complex search functions, but the efficiency of resource locating is low. However, it is efficient to locate resources in structure P2P network, but it does not support fuzzy queries and has the higher maintenance cost of network. So how to search the resources in the P2P networks efficiently is becoming the key problem.

In traditional network structure, it is difficult to simultaneously meet users' requirements in resource locating efficiency and recall ratio. So this paper proposes a subdomain hybrid P2P network model (SHPNM), which makes full use of the advantages of supporting fuzzy queries in the unstructured network and maintains efficient in resource locating in structured P2P network. Then it gives a detailed analysis about the design idea of SHPNM as well as the method of resources locating in this model. The nodes are gathered into a domain according to the approaching of nodes' physical location, and the nodes in a domain can form a structured P2P network. Each domain provides two nodes with biggest comprehensive performance value as a boundary node and a backup node, and the boundary nodes are connected together in the form of unstructured P2P networks. Simulations results show that SHPNM can both improve efficiency in locating resource and promote the recall ratio.

Workshop 12

Performance Modeling, Evaluation, and Optimisation of Ubiquitous Computing and Networked Systems

PMEO 2010

Power Assignment and Transmission Scheduling in Wireless Networks

Keqin Li
Department of Computer Science
State University of New York
New Paltz, New York 12561, USA
Email: lik@newpaltz.edu

Abstract

The problem of downlink data transmission scheduling in wireless networks is studied. It is pointed out that every downlink data transmission scheduling algorithm must have two components to solve the two subproblems of power assignment and transmission scheduling. Two types of downlink data transmission scheduling algorithms are proposed. In the first type, power assignment is performed before transmission scheduling. In the second type, power assignment is performed after transmission scheduling. The performance of two algorithms of the first type which use the equal power allocation method are analyzed. It is shown that both algorithms exhibit excellent worst-case performance and asymptotically optimal average-case performance under the condition that the total transmission power is equally allocated to the channels. In general, both algorithms exhibit excellent average-case performance. It is demonstrated that two algorithms of the second type perform better than the two algorithms of the first type due to the equal time power allocation method. Furthermore, the performance of our algorithms are very close to the optimal and the room for further performance improvement is very limited.

Performance Impact of SMP-Cluster on the On-chip Large-scale Parallel Computing Architecture

Shenggang Chen, Shuming Chen and Yaming Yin
School of Computer
National University of Defense Technology
Changsha, China
{shgchen,smchen,ymyin}@nudt.edu.cn

Abstract

To minimize the delay of the data communication, hierarchical On-chip Large-scale Parallel Computing architectures (OLPCs) with communication locality awareness are recently studied by researchers. This paper proposes a hierarchical architecture consisting of SMP clustered nodes, each of which is structured by more than one baseline cores through centrally-shared memory. The analytical speedup model of the proposed architecture is established by extending Amdahl' Law. The design space exploitation of the SMP-clustered architecture is investigated through theoretical analysis and experiential values of the parameters used in the speedup model. Finally, some useful suggestions of the future SMP-clustered OLPCs design are presented during the analysis.

Parallel Isolation-Aggregation Algorithms to Solve Markov Chains Problems With Application to Page Ranking

Abderezak Touzene
College of Science
Computer Science Department
Sultan Qaboos University
P.O. Box 36, P.C. 123, Oman

Abstract

In this paper, we propose two parallel Aggregation-Isolation iterative methods for solving Markov chains. These parallel methods conserves as much as possible the benefits of aggregation, and Gauss-Seidel effects. Some experiments have been conducted testing models from queuing systems and models from Google Page Ranking. The results of the experiments show super linear speed-up for the parallel Aggregation-Isolation method.

Multicore-Aware Reuse Distance Analysis

Derek L. Schuff, Benjamin S. Parsons and Vijay S. Pai
Purdue University
West Lafayette, IN 47907
dschuff@purdue.edu, bsparson@purdue.edu, vpai@purdue.edu

Abstract

This paper presents and validates methods to extend reuse distance analysis of application locality characteristics to shared-memory multicore platforms by accounting for invalidation-based cache-coherence and inter-core cache sharing. Existing reuse distance analysis methods track the number of distinct addresses referenced between reuses of the same address by a given thread, but do not model the effects of data references by other threads. This paper shows several methods to keep reuse stacks consistent so that they account for invalidations and cache sharing, either as references arise in a simulated execution or at synchronization points. These methods are evaluated against a Simics-based coherent cache simulator running several OpenMP and transaction-based benchmarks. The results show that adding multicore-awareness substantially improves the ability of reuse distance analysis to model cache behavior, reducing the error in miss ratio prediction (relative to cache simulation for a specific cache size) by an average of 70% for per-core caches and an average of 90% for shared caches.

Clairvoyant Site Allocation of Jobs with Highly Variable Service Demands in a Computational Grid

Stylianos Zikos
Department of Informatics
Aristotle University of Thessaloniki
54124 Thessaloniki, Greece
szikos@ csd.auth.gr

Helen D. Karatza
Department of Informatics
Aristotle University of Thessaloniki
54124 Thessaloniki, Greece
karatza@ csd.auth.gr

Abstract

In this paper we evaluate performance of three different site allocation policies in a 2-level computational grid with heterogeneous sites. We consider that schedulers are aware of service demands of jobs which show high variability. A simulation model is used to evaluate performance in terms of the average response time and slowdown, under medium and high load. Simulation results show that the proposed policy outperforms the other two that are examined, especially at high load.

Resource Management of Enterprise Cloud Systems Using Layered Queuing and Historical Performance Models

David A. Bacigalupo[1], Jano van Hemert[2], Asif Usmani[3], Donna N. Dillenberger[4], Gary B. Wills[1] and Stephen A. Jarvis[5]
[1]School of Electronics and Computer Science, University of Southampton, SO17 1BJ, UK
[2]Data-Intensive Research Group, School of Informatics, University of Edinburgh, EH8 9AB, UK
[3]BRE Centre for Fire Safety Engineering, School of Engineering, University of Edinburgh, EH9 3JL, UK
[4]IBM T.J. Watson Research Centre, Yorktown Heights, New York, 10598, USA
[5]High Performance Systems Group, Department of Computer Science, University of Warwick, CV4 7AL, UK
e-mail: db1f08@ecs.soton.ac.uk

Abstract

The automatic allocation of enterprise workload to resources can be enhanced by being able to make 'what-if' response time predictions, whilst different allocations are being considered. It is important to quantitatively compare the effectiveness of different prediction techniques for use in cloud infrastructures. To help make the comparison of relevance to a wide range of possible cloud environments it is useful to consider the following. 1.) *urgent* cloud customers such as the emergency services that can demand cloud resources at short notice (e.g. for our *FireGrid* emergency response software). 2.) *dynamic* enterprise systems, that must rapidly adapt to frequent changes in workload, system configuration and/or available cloud servers. 3.) The use of the predictions in a co-ordinated manner by both the cloud infrastructure and cloud customer management systems. 4.) A broad range of criteria for evaluating each technique. However, there have been no previous comparisons meeting these requirements. This paper, meeting the above requirements, quantitatively compares the layered queuing and ("HYDRA") historical techniques – including our initial thoughts on how they could be combined. Supporting results and experiments include the following: *i.*) defining, investigating and hence providing guidelines on the use of a historical and layered queuing model; *ii.*) using these guidelines showing that both techniques can make low overhead and typically over 70% accurate predictions, for *new* server architectures for which only a small number of benchmarks have been run; and *iii.*) defining and investigating tuning a prediction-based cloud workload and resource management algorithm

Predictibility of Inter-component latency in a Software Communications Architecture Operating Environment

Gael Abgrall[1], Frédéric Le Roy[1], Jean-Philippe Diguet[2], Guy Gogniat[2] and Jean-Philippe Delahaye[3]
[1]UEB, ENSIETA / DTN, 2 rue Francois Verny, 29806 Brest Cedex 9, France
{gael.abgrall, frederic.le_roy}@ensieta.fr
[2]UEB, UBS / Lab-STICC, BP 92116, 56321 Lorient Cedex, France
{jean-philippe.diguet, guy.gogniat}@univ-ubs.fr
[3]DGA / CELAR, La Roche Marguerite, 35174 Bruz, France
jean-philippe.delahaye@dga.defense.gouv.fr

Abstract

This paper presents an in-depth analysis of the behavior of a SCA component-based waveform application in terms of ”inter-component” communication latency. The main limitation with SCA, in the context of embedded systems, is the additional cost introduced by the use of CORBA. Previous studies have already defined the major metrics of interest regarding this issue, these are CPU cost, memory requirements and ”inter-component” latency. Real-time systems can not afford high latency, in consequence, this paper focuses on this metric. The starting point of this paper is the desire of knowing if the SCA CF does not also bring an overhead. Measurements have been realized with OmniORB as CORBA distribution and OSSIE for SCA implementation. In order to perform these measurements, a SCA waveform composed of several ”empty-components” have been created. ”Empty-components” are software components compliant to SCA without any signal processing part. The study only focuses on communications between components. The same kind of ”inter-component” link has been measured between two components using CORBA without SCA. It is possible to compare the latency values between the two measurements and to show as a result that they are approximately the same. The CORBA bus is really the part which brings an overhead to the system. The final part of this paper introduces a statistical estimation of the latency distributions. It results from measurements performed with various data packet sizes and uses a fitting method based on a combination of Gaussian functions.

Analytical Performance Comparison of 2D Mesh, WK-Recursive, and Spidergon NoCs

M. Bakhouya, S. Suboh, J. Gaber and T. El-Ghazawi
UTBM, 90010 Belfort Cedex, France
{bakhouya, gaber}@utbm.fr
GWU, Washington DC. 20052, USA
{suboh, tarek}@gwu.edu

Abstract

Network-on-Chip (NoC) has been proposed as an alternative to bus-based schemes to achieve high performance and scalability in System-on-Chip (SoC) design. Performance analysis and evaluation of on-chip interconnect architectures are widely based on simulation which becomes computationally expensive, especially for large-scale NoCs. Recently, a Network Calculus-based methodology was proposed to analytically evaluate the performance of NoC-based architectures. In this paper, the 2D Mesh, Spidergong, and WK-recursive on-chip interconnects are analyzed using this methodology and main performance metrics, the end-to-end delay and buffer size requirements, are computed. Results are reported and show that WK outperforms the other on-chip interconnects in all considered performance metrics.

Adapting to NAT timeout values in P2P Overlay Networks

Richard Price and Peter Tino
School of Computer Science
University of Birmingham
Birmingham, United Kingdom
Email: {R.M.Price; P.Tino}@cs.bham.ac.uk

Abstract

Nodes within existing P2P networks typically exchange periodic keep-alive messages in order to maintain network connections between neighbours. Keep-alive messages serve a dual purpose, they're used to detect node failures and to prevent idle connections from being expired by NAT devices. However despite being widely used, the interval between messages are typically fixed below the timeout value of most NAT devices based upon crude rules of thumb.

Furthermore, although many studies have been conducted to traverse NAT devices and other studies seek to improve failure detection in P2P overlay networks; the limitations of NAT devices have received little research attention. This paper explores algorithms which allow nodes to adapt to the timeout values of individual NAT devices and investigates the resulting trade-offs.

Agent Placement in Wireless Embedded Systems: Memory Space and Energy Optimizations

Nikos Tziritas[1,2], Thanasis Loukopoulos[1,3], Spyros Lalis[1,2] and Petros Lampsas[1,3]
[1]Center for Research and Technology Thessaly (CERETETH)
Volos, Greece

[2]Computer & Communication Engineering Department
University of Thessaly
Volos, Greece
e-mail:{nitzirit, lalis}@inf.uth.gr

[3]Department of Informatics & Computer Technology
Technological & Educational Institute
(TEI) of Lamia
Lamia, Greece
e-mail: {luke, plam}@inf.teilam.gr

Abstract

Embedded applications can be structured in terms of mobile agents that are flexibly installed on available nodes. In wireless systems, such nodes typically have limited battery and memory resources; therefore it is important to place agents judiciously. In this paper we tackle the problem of placing a newcomer agent in such a system. The problem has two main components. First, enough memory space must be found or created at some node to place the agent. Second, the placement should be energy efficient. We present heuristics for tackling these two goals in a stepwise fashion, as well as a branch and bound method for achieving both goals at the same time. Our algorithms are centralized assuming a single entry point through which agents are injected into the system, with adequate knowledge of the system state and enough resources to run the proposed algorithms. The algorithms are evaluated under different simulated scenarios, and the tradeoffs across the two metrics (space, energy) are identified.

A Markov Chain Based Method for NoC End-to-End Latency Evaluation

Sahar Foroutan[1,2], Yvain Thonnart[2], Richard Hersemeule[1] and Ahmed Jerraya[2]
[1]ST-Microelectronics, [2]CEA-Leti
{sahar.foroutan, yvain.thonnart, ahmed.jerraya}@cea.fr, richard.hersemeule@st.com

Abstract

This paper presents a generic analytical method to estimate communication latency between a source and a destination of a given Network-on-Chip. This method is based on Markov chain stochastic processes. In order to solve the limiting problem of state-space explosion in complex stochastic processes, we propose to construct a reduced Markov chain model for each node of the path, and to recursively use the local mean latencies to obtain the mean latency of the complete path. Comparison between the analytical results obtained by our method and those of a corresponding SystemC CABA simulation platform shows the accuracy of our method.

An Adaptive I/O Load Distribution Scheme for Distributed Systems

Xin Chen, Jeremy Langston and Xubin He
Department of Electrical and Computer Engineering
Tennessee Technological University
Cookeville, TN 38505, USA
{xchen21, jwlangston21, hexb}@tntech.edu

Fengjiang Mao
Department of Electronic and Information Engineering
Shenzhen Polytechnic
Shenzhen, Guangdong 518055, P.R.China
mfj@oa.szpt.net

Abstract

A fundamental issue in a large-scale distributed system consisting of heterogeneous machines which vary in both I/O and computing capabilities is to distribute workloads with respect to the capabilities of each node to achieve the optimal performance. However, node capabilities are often not stable due to various factors. Simply using a static workload distribution scheme may not well match the capability of each node. To address this issue, we distribute workload adaptively to the change of system node capability.

In this paper we present an adaptive I/O load distribution scheme to dynamically capture the I/O capabilities among system nodes and to predictively determine an suitable load distribution pattern. A case study is conducted by applying our load distribution scheme into a popular distributed file system PVFS2. Experiments results show that our adaptive load distribution scheme can dramatically improve the performance: up to 70% performance gain for writes and 80% for reads, and up to 63% overall performance loss can be avoided in the presence of an unstable Object Storage Device (OSD).

Cross Layer Neighbourhood Load Routing for Wireless Mesh Networks

Liang Zhao[1], Ahmed Y. Al-Dubai[1] and Geyong Min[2]
[1]School of Computing, Edinburgh Napier University
Edinburgh, EH10 5DT, UK
Email: {l.zhao, a.al-dubai}@napier.ac.uk
[2]Department of Computing, University of Bradford
Bradford, BD7 1DP, U.K.
Email: g.min@brad.ac.uk

Abstract

Wireless Mesh Network (WMN) has been considered as a key emerging technology to construct next generation wireless communication networks. It combines the advantages of both mobile ad-hoc network (MANET) and traditional fixed network, attracting significant industrial and academic attentions. In WMN, the load balancing becomes a hot topic in enhancing the QoS provision as a load balanced WMN exhibits low delay and high quality communications. Although there are a number of proposals on using load-aware routing metrics in WMN, the neighbourhood load has not been considered within the context of load balancing and QoS aware WMNs. In this paper, we propose a Neighbourhood Load Routing scheme to further improve the performance of the existing Routing protocol such as AODV in WMN. We have conducted extensive simulation experiments. Our results confirm the superiority of our proposed scheme over its well-known counterparts, especially in grid topologies.

A New Probabilistic Linear Exponential Backoff Scheme for MANETs

Muneer Bani Yassein[1] , Saher Manaseer[2], Asmahan Abu Al-hassan[1], Zeinab Abu Taye'[1] and Ahmed Y. Al-Dubai[3]
[1]Department of Computer Science, Jordan University of Science and Technology, masadeh@just.edu.jo, {amabualhassan07|zoabutaye08}@cit.just.edu.jo
[2]Department of Computing Science, University of Glasgow, saher@dcs.gla.ac.uk
[3]School of Computing, Edinburgh Napier University, a.al-dubai@napier.ac.uk

Abstract

Broadcasting is an essential operation in Mobile ad hoc Networks (MANETs) environments. It is used in the initial phase of route discovery process in many reactive protocols. Although broadcasting is simple, it causes the well known broadcast storm problem, which is a result of packet redundancy, contention and collision. A probabilistic scheme has been proposed to overcome this problem. This work aims to study the effect of network density and network mobility on probabilistic schemes using different thresholds (fixed, 2p, 3p and 4p) with the Pessimistic Linear Exponential Backoff (PLEB) algorithm and compare the results with the standard MAC. A number of simulation experiments have been conducted to examine the performance of the proposed PLEP under different operating conditions. The simulation results show that in dense networks the normalized routing load, delay and routing packets are high and the PLEB outperforms the standard MAC in terms of delay.

A Stochastic Framework to Depict Viral Propagation in Wireless Heterogeneous Networks

Hoai-Nam Nguyen, Yasuhiro Ohara and Yoichi Shinoda
Japan Advanced Institute of Science and Technology
{namnh, yasu, shinoda}@jaist.ac.jp

Abstract

Smart phones, with the fast improvement both in hardware and software, are now able to interact and work with computer through numerous communications technologies thus expose themselves to the risk of being infected by virus. Although many researchers devote to study the spread of malware, little effort has been done to take into account the different types of devices that may concurrently occur in an wireless ad hoc network. Therefore we have developed a stochastic framework, which incorporates diversity of network nodes as well as their interactions to see how those features affect a disease's dissemination. Our models, based on a 4-compartment epidemic method, also have taken into consideration various states that a device may undergo when it gets infected by the malware. A huge result space is producible by our framework thus makes it appropriate to describe many viral proliferating scenarios. We have also conducted numerical simulations to understand changes and the equilibrium of a network under malware dissemination.

A Design Aid and Real-Time Measurement Framework for Virtual Collaborative Simulation Environment

Ming Zhang, Hengheng Xie, Azzedine Boukerche
PARADISE Research Laboratory
SITE, University of Ottawa
Ottawa, Canada
{mizhang, hxie072, boukerch}@site.uottawa.ca

Abstract

Distributed Virtual Environment (DVE) has attracted much attention in recent years due to the rapid advances in the areas of E-learning, Internet gaming, human-computer interfaces, and etc. However, the complexity of designing an efficient DVE greatly challenges today's researchers. Indeed, how to effectively measure a DVE or the design of DVE plays an important role in progressively guiding the DVE design. Meanwhile, the performance of protocols and schemes used in an existing DVE cannot be easily measured straightforwardly. Traditionally, simulation tools are involved to predict the performance of the DVE system or any used protocols; however, existing tools have limited capabilities in terms of accurately capturing the real-world performance. This is due to the statically configured simulations, hard-to-model hardware devices (such as haptic devices, mobile devices), single processor's execution of the simulations, and etc. Moreover, there exists no integrated simulating and measuring framework that can effectively support model reuse, dynamic reconfiguration of a simulation, real devices in the simulation loop (statically or run-time), and distributed simulation execution, just to name a few. In this paper, we propose and implement an integrative simulation and measuring framework; in particular, we design a generic real-time service oriented virtual simulation system which can effectively measure the real time performance of distributed virtual environments and virtual reality based applications. The main goal of our proposed framework lies in providing accurate and near-real-world measurements of a DVE and any related protocols, so that a highly cost-efficient DVE system design can be achieved. Moreover, our framework uses the Experimental Frame concept to separate simulation services from design models so that model reuses, model formalization and model validation can all be done within one layer. The idea of using experimental frame also makes possible a hardware-in-the-loop type of simulation, which is quite useful in DVE including haptic virtual environment, sensor network based virtual environment, and etc. As a case study, we investigate a QoS-aware adaptive load balance algorithm using our framework; and our real-time simulation results clearly indicate that the algorithm outperforms others in a near real-world scenario.

A Supplying Partner Strategy for Mobile Networks-based 3D Streaming - Proof of Concept

Haifa Raja Maamar[1,2], Richard W. Pazzi[1], Azzedine Boukerche[1] and Emil Petriu[2]
[1]PARADISE Research Laboratory
[2]SMR Research Laboratory
SITE-University of Ottawa
Email: (hmaam026, rwerner, boukerch, petriu)@site.uottawa.ca

Abstract

With the advances of wireless communication and mobile computing, there is a growing interest among researchers about augmented reality and streaming 3D graphics on mobile devices for training first responders to be better prepared in a case of disaster scenarios. However, several challenges need to be resolved before this technology become a commodity. One of the major difficulties in 3D streaming over thin mobile devices is related to the supplying partner strategy as it is not easy to discover the peer that has the correct information and that posses enough bandwidth to send the required data quickly and efficiently to the peers in need. In this paper, we propose a new supplying partner strategy for mobile networks-based 3D streaming. The primary goal of the work presented in this paper is first to address the thin mobile devices low storage capabilities; and second to avoid the flooding problem that most wireless mobile networks suffer from. Our proposed protocol is based on the quick discovery of *multiple* supplying partners, by optimizing the time required by peers to acquire data, avoiding unnecessary messages propagation and network congestion, and decreasing the latency and the network bandwidth over utilization.

Workshop 13

Dependable Parallel, Distributed and Network-Centric Systems

DPDNS 2010

Failure Prediction for Autonomic Management of Networked Computer Systems with Availability Assurance

Ziming Zhang and Song Fu
Department of Computer Science and Engineering
New Mexico Institute of Mining and Technology
{zzm, song}@nmt.edu

Abstract

Networked computer systems continue to grow in scale and in the complexity of their components and interactions. Component failures become norms instead of exceptions in these environments. Failure occurrence as well as its impact on system performance and operation costs are becoming an increasingly important concern to system designers and administrators. To achieve self-management of failures and resources in networked computer systems, we propose a framework for autonomic failure management with hierarchical failure prediction functionality for large coalition systems, such as coalition clusters and compute grids. It analyzes node, cluster and system wide failure behaviors and forecasts the prospective failure occurrences based on quantified failure dynamics. Failure correlations are inspected by the predictor. Experimental results in a computational grid on campus show the offline and online predictions by our predictors accurately forecast the failure trend and capture failure correlations in the production environment.

J2EE Instrumentation for software aging root cause application component determination with AspectJ

Javier Alonso and Jordi Torres
Barcelona Supercomputing Center
Dept. of Computer Architecture
Technical University of Catalonia
Barcelona, Spain
Email: [alonso, torres]@ac.upc.edu

Josep Ll. Berral
Dept. of Software
Dept. of Computer Architecture
Technical University of Catalonia
Barcelona, Spain
Email: berral@ac.upc.edu

Ricard Gavaldà
Dept. of Software
Technical University of Catalonia
Barcelona, Spain
Email: gavalda@lsi.upc.edu

Abstract

Unplanned system outages have a negative impact on company revenues and image. While the last decades have seen a lot of efforts from industry and academia to avoid them, they still happen and their impact is increasing. According to many studies, one of the most important causes of these outages is software aging. Software aging phenomena refers to the accumulation of errors, usually provoking resource contention, during long running application executions, like web applications, which normally cause applications/systems hang or crash. Determining the software aging root cause failure, not the resource or resources involved in, is a huge task due to the growing day by day complexity of the systems. In this paper we present a monitoring framework based on Aspect Programming to monitor the resources used by every application component in runtime. Knowing the resources used by every component of the application we can determine which components are related to the software aging. Furthermore, we present a case study where we evaluate our approach to determine in a web application scenario, which components are involved in the software aging with promising results.

Improving MapReduce Fault Tolerance in the Cloud

Qin Zheng
Institute of High Performance Computing
Advanced Computing Programme
Fusionopolis, 1 Fusionopolis Way, 16-16 Connexis, Singapore 138632
qinzheng.sg@gmail.com

Abstract

MapReduce has been used at Google, Yahoo, FaceBook etc., even for their production jobs. However, according to a recent study, a single failure on a Hadoop job could cause a 50% increase in completion time. Amazon Elastic MapReduce has been provided to help users perform data-intensive tasks for their applications. These applications may have high fault tolerance and/or tight SLA requirements. However, MapReduce fault tolerance in the cloud is more challenging as topology control and (data) rack locality currently are not possible. In this paper, we investigate how redundent copies can be provisioned for tasks to improve MapReduce fault tolerance in the cloud while reducing latency.

Tackling Consistency Issues for Runtime Updating Distributed Systems

Filippo Bannò[1], Daniele Marletta[1], Giuseppe Pappalardo[2] and Emiliano Tramontana[2]
[2]Dipartimento di Matematica e Informatica
[1]Scuola Superiore di Catania
Università di Catania, Italy
{pappalardo, tramontana}@dmi.unict.it

Abstract

Dynamic update capabilities allow a system to change some of its components without stopping execution, so as to cater for versioning and self-optimisation. In this paper, we propose FREJA to support transparent dynamic updates of classes of a distributed Java system. In doing so, FREJA purports to (i) ensure the consistency of the execution flow, and (ii) avoid that faults typical of a distributed environment interfere with the updating operations.

FREJA's operations are transparent with respect to the system to be updated, thanks to appropriate selective bytecode rewriting of classes, performed at load-time. Consistency of the execution flow, for any state of the system to be updated and even in the face of network faults, is ensured by an ad-hoc protocol which checks the validity of changes and, should a fault manifest itself, falls back to a previous configuration.

Dynamic updates, executed while safeguarding consistency, foster systems properties such as availability, maintainability and fault-tolerance, since by changing some parts of a running system bugs fixes and fault recovery can be performed. Hence, the proposed solution provides developers with valuable mechanisms which are paramount to obtain dependable systems.

Achieving Information Dependability in Grids through GDS

V. D. Cunsolo, S. Distefano, A. Puliafito and M. Scarpa
Università di Messina, Dipartimento di Matematica
Contrada Papardo, S. Sperone, 98166 Messina, Italy
vdcunsolo{sdistefano, apuliafito, mscarpa}@unime.it

Abstract

In Grid computing infrastructures, data are usually distributed among different nodes and thus shared among several users. In order to achieve satisfactory performance and adequate fault tolerance with regards to the requirements, it is necessary to replicate data, taking into account information confidentiality and integrity.

In this paper we face the problem of data dependability in Grid, by proposing a redundant lightweight cryptographic algorithm combining the strong and highly secure asymmetric cryptographic technique (RSA) with the symmetric cryptography (AES). The proposed algorithm, we named *grid dependable storage system* (GDS^2), has been implemented on top of the gLite middleware file access and replica management libraries, as a file system service with cryptographic capability, redundancy management and POSIX interface. This choice of implementing GDS^2 as a file system allows to protect also the file system structure. This paper particularly focuses on the GDS^2 algorithm and its gLite implementation.

Evaluating Database-oriented Replication Schemes in Software Transactional Memory Systems

Roberto Palmieri and Francesco Quaglia
DIS, Sapienza University, Rome, Italy

Paolo Romano and Nuno Carvalho
INESC-ID, Lisbon, Portugal

Abstract

Software Transactional Memories (STMs) are emerging as a highly attractive programming model, thanks to their ability to mask concurrency management issues to the overlying applications. In this paper we are interested in dependability of STM systems via replication. In particular we present an extensive simulation study aimed at assessing the efficiency of some recently proposed database-oriented replication schemes, when employed in the context of STM systems. Our results point out the limited efficiency and scalability of these schemes, highlighting the need for redesigning ad-hoc solutions well fitting the requirements of STM environments. Possible directions for the re-design process are also discussed and supported by some early quantitative data.

Optimizing RAID for Long Term Data Archives

Henning Klein
Fujitsu Technology Solutions GmbH
Bürgermeister-Ulrich-Strasse 100
86199 Augsburg, Germany
Henning.Klein@ts.fujitsu.com

Jörg Keller
Fernuniversität in Hagen
Dept. Of Mathematics and Computer Science
58084 Hagen, Germany
Joerg.Keller@FernUni-Hagen.de

Abstract

We present new methods to extend data reliability of disks in RAID systems for applications like long term data archival. The proposed solutions extend existing algorithms to detect and correct errors in RAID systems by preventing accumulation of undetected errors in rarely accessed disk segments. Furthermore we show how to change the parity layout of a RAID system in order to improve the performance and reliability in case of partially defect disks. All methods benefit of a hierarchical monitoring scheme that stores reliability related information. Our proposal focuses on methods that do not need significant hardware changes.

Experimental Responsiveness Evaluation of Decentralized Service Discovery

Andreas Dittrich and Felix Salfner
Institut für Informatik
Humboldt-Universität zu Berlin
Unter den Linden 6, 10099 Berlin, Germany
Email: {dittrich|salfner}@informatik.hu-berlin.de

Abstract

Service discovery is a fundamental concept in service networks. It provides networks with the capability to publish, browse and locate service instances. Service discovery is thus the precondition for a service network to operate correctly and for the services to be available. In the last decade, decentralized service discovery mechanisms have become increasingly popular. Especially in ad-hoc scenarios – such as ad-hoc wireless networks – they are an integral part of auto-configuring service networks. Albeit the fact that auto-configuring networks are increasingly used in application domains where dependability is a major issue, these environments are inherently unreliable. In this paper, we examine the dependability of decentralized service discovery. We simulate service networks that are automatically configured by Zeroconf technologies. Since discovery is a time-critical operation, we evaluate responsiveness – the probability to perform some action on time even in the presence of faults – of domain name system (DNS) based service discovery under influence of packet loss. We show that responsiveness decreases significantly already with moderate packet loss and becomes practicably unacceptable with higher packet loss.

Analysis of Network Topologies and Fault-Tolerant Routing Algorithms using Binary Decision Diagrams

Andreas C. Döring IBM Research – Zurich Säumerstrasse 4, 8803 Rüschlikon, Switzerland email: ado (at) zurich

Abstract

In the past a plethora of network topologies together with fault-tolerant routing algorithms have been proposed. Some properties have been analyzed analytically or by simulation. In most cases only some properties can be derived. There is renewed interest in the topic for application as networks-on-chip. The availability of higher computing performance and libraries for manipulating binary decision diagrams allows the complete analysis in an automated fashion. The approach is presented in this paper together with some insights on strategies to keep the computational effort reasonable when scaling the network size.

Incentive Mechanisms in Peer-to-Peer Networks

Pedro Dias Rodrigues, Carlos Ribeiro and Luís Veiga
INESC ID Lisboa / Instituto Superior Técnico C Universidade Técnica de Lisboa
Rua Alves Redol, 9
1000 Lisboa, Portugal
pedrorodrigues@ist.utl.pt, carlos.ribeiro@ist.utl.pt, luis.veiga@inesc-id.pt

Abstract

In the last few years, peer-to-peer systems became well known to the general public by allowing fast and simple exchange of resources between users. The complexity of these systems resides in the absence of a central authority and the fact that each user can act both as a client and as a server. Thus, a need for self-regulation arises in order to guarantee that every user contributes to the system. This paper describes our work developing incentive mechanisms, which enable the correct operation of peer-to-peer systems, imposing a balance between demand for resources and the existing offer. All incentive mechanisms take into account the attacks these systems are subject to, as well as the system's structure, so that they do not pose an unnecessary burden, slowing down the system excessively. We explore concepts such as reputation and currency, which are used in other systems and, most importantly, in our everyday life, enabling a coherent scheme to detect untrustworthy users and reward truthful peers with faster access to the resources. Our work is part of a larger project called GINGER, an acronym for Grid In a Non-Grid Environment, a peer-to-peer infrastructure intended to ease the sharing of computer resources between users.

Lessons Learned During the Implementation of the BVR Wireless Sensor Network Protocol on SunSPOTs

Ralph Robert Erdt and Martin Gergeleit
Hochschule RheinMain
University of Applied Sciences
Wiesbaden Rüsselsheim Geisenheim
Email: ralph@rccc.de, martin.gergeleit@hs-rm.de

Abstract

The Beacon Vector Routing (BVR) protocol [1] is a well known routing protocol for Wireless Sensor Networks (WSNs). Simulations have shown that the protocol scales well in an environment with perfect links and an ideal circular radio coverage. However, when it comes to an implementation on an embedded hardware that uses IEEE 802.15.4 2.4 GHz wireless transceivers, some problems turn out that have significant impact on the overall performance of the protocol.

The design of BVR (and most other WSN protocols) is based on certain assumptions about the underlying physics of the wireless communication links. We will show in this paper that some of these assumptions are insufficient or sometimes even unrealistic. The most important observation is that BVR is based on fairly stable, bidirectional communication links. However, our experiments with IEEE 802.15.4-based 2.4 GHz network, typical technology for WSNs, have shown, that this assumption is not always true. Here the signal strength is highly variable, even if the environment is stable. Even more severe, in many cases a link between two nodes only works reliably in one direction, i.e., it is not bidirectional.

We have implemented and evaluated the BVR protocol on SunSPOTs (experimental sensor nodes from Sun). We will give an overview and a description of all of the problems we realized during implementation (including the problems described in the BVR paper). Then we discuss the approaches and adaptations of the algorithms we used to work around these problems. Some of these ideas might be relevant for other WSN protocols as well.

Workshop 14

International Workshop on Hot Topics in Peer-to-Peer Systems

HOTP2P 2010

Estimating Operating Conditions in a Peer-to-Peer Session Initiation Protocol Overlay Network

Jouni Mäenpää and Gonzalo Camarillo
Ericsson
Finland
{jouni.maenpaa, gonzalo.camarillo}@ericsson.com

Abstract

Distributed Hash Table (DHT) based peer-to-peer overlays are decentralized, scalable, and fault tolerant. However, due to their decentralized nature, it is very hard to know the state and prevailing operating conditions of a running overlay. If the system could figure out the operating conditions, it would be easier to monitor the system and re-configure it in response to changing conditions. Many DHT-based system such as the Peer-to-Peer Session Initiation Protocol (P2PSIP) would benefit from the ability to accurately estimate the prevailing operating conditions of the overlay. In this paper, we evaluate mechanisms that can be used to do this. We focus on network size, join rate, and leave rate. We start from existing mechanisms and show that their accuracy is not sufficient. Next, we show how the mechanisms can be improved to achieve a higher level of accuracy. The improvements we study include various mechanisms improving the accuracy of leave rate estimation, use of a secondary network size estimate, sharing of estimates between peers, and statistical mechanisms to process shared estimates.

Adaptive Server Allocation for Peer-assisted Video-on-Demand

Konstantin Pussep[1], Osama Abboud[1], Florian Gerlach[1], Ralf Steinmetz[1] and Thorsten Strufe[2]
[1]Multimedia Communications Lab, Technische Universität Darmstadt
Email: {pussep,abboud,steinmetz}@kom.tu-darmstadt.de
[2]Peer-to-Peer Networks, Technische Universität Darmstadt
Email: strufe@cs.tu-darmstadt.de

Abstract

Dedicated servers are an undesirable but inevitable resource in peer-assisted streaming systems. Their provision is necessary to guarantee a satisfying quality of experience to consumers, yet they cause significant, and largely avoidable cost for the provider, which can be minimized. We propose two adaptive server allocation schemes that estimate the capacity situation and service demand of the system to adaptively optimize allocated resources. Extensive simulations support the efficiency of our approach, which, without considering any prior knowledge, allows achieving a competitive performance compared to systems that are well dimensioned using global knowledge.

Heterogeneity in Data-Driven Live Streaming: Blessing or Curse?

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

Abstract

Distributed live streaming has brought a lot of interest in the past few years. In the homogeneous case (all nodes having the same capacity), many algorithms have been proposed, which have been proven almost optimal or optimal. On the other hand, the performance of heterogeneous systems is not completely understood yet.

In this paper, we investigate the impact of heterogeneity on the achievable delay of chunk-based live streaming systems. We propose several models for taking the atomicity of a chunk into account. For all these models, when considering the transmission of a single chunk, heterogeneity is indeed a "blessing", in the sense that the achievable delay is always shorter than an equivalent homogeneous system. But for a stream of chunks, we show that it can be a "curse": there are systems where the achievable delay can be arbitrary greater compared to equivalent homogeneous systems. However, if the system is slightly bandwidth-overprovisioned, optimal single chunk diffusion schemes can be adapted to a stream of chunks, leading to near-optimal, faster than homogeneous systems, heterogeneous live streaming systems.

Techniques for Low-latency Proxy Selection in Wide-Area P2P networks

Arijit Ganguly, P. Oscar Boykin and Renato Figueiredo
Advanced Computing and Information Systems Laboratory
University of Florida
Gainesville, Florida - 32611
Email: {aganguly, boykin, renato}@acis.ufl.edu

Abstract

Connectivity constraints due to Internet route outages and symmetric NATs create situations when direct communication is not possible between nodes in P2P deployments, often leading to high latency between high-traffic nodes. In such cases, it is possible to reduce end-to-end latency by routing their communication through another node in the P2P system that is selected based on its Internet latencies to end-nodes. In this paper, we present and compare two decentralized algorithms for discovering low-latency proxy nodes between high-traffic nodes: (1) selecting a node with least Internet latency from the set of randomly-chosen neighbors to which one of the end-nodes is directly connected, and (2) discovery using latency estimates based on network coordinates. We evaluate these techniques in context of IP-over-P2P virtual networks through experiments on realistic wide-area testbeds. Our results indicate that in a network with over 400 nodes on PlanetLab, both the techniques select proxies such that the median penalties over the global optimal are within 16% and 21%, respectively, when all nodes are able to serve as proxies. We also investigate scenarios where subsets of nodes have connectivity constraints preventing them from serving as proxies.

Mobile-Friendly Peer-to-Peer Client Routing Using Out-of-Band Signaling

Wei Wu, Jim Womack
Advanced Technology
Research In Motion, Limited
Irving, TX 75039, USA
e-mail: {wwu, jwomack}@rim.com

Xinhua Ling
BlackBerry System Architecture Research
Research In Motion, Limited
Waterloo, ON N2L 3W8, Canada
e-mail: xling@rim.com

Abstract

It is expected that Peer-to-Peer (P2P) services will co-exist with the client-server based services such as IMS. Mobile users may subscribe to the traditional wireless cellular services while participating in P2P overlay networks. In this paper, a method is proposed to reduce the signaling overhead in a mobile P2P system. With the help of the underlying infrastructure, a mobile device in the P2P overlay can be located using out-of-band non-P2P signaling. This reduces its P2P signaling for location update while a mobile device is changing the point of attachment in the P2P overlay. As the signaling cost depends on both the client's mobility and traffic models, an analytical model has been developed to determine the optimal threshold for the registration update. Analytical results have shown that the proposed method could save up to 70% signaling cost when the Call-to-Mobility Ratio (CMR) is low. On the other hand, it would be better to fall back to the base client routing method when the CMR is high, i.e., perform the registration update whenever the client changes the point of attachment in the P2P overlay.

Deetoo: Scalable Unstructured Search Built on a Structured Overlay

Tae Woong Choi
Advanced Computing Information Systems Lab
University of Florida
Email: twchoi@ufl.edu

P. Oscar Boykin
Advanced Computing Information Systems Lab
University of Florida
Email: boykin@acis.ufl.edu

Abstract

We present Deetoo, an algorithm to perform completely general queries, for instance high-dimensional proximity queries or regular expression matching, on a P2P network. Deetoo is an efficient unstructured query system on top of existing structured P2P ring topologies. Deetoo provides a reusable search tool to work alongside a DHT, thus, it provides new capabilities while reusing existing P2P models and software. Since our algorithm is for unstructured search, there is no structural relationship between the queries and the network topology and hence no need to provide a mapping of queries onto a fixed DHT structure. Deetoo is optimal in terms of the trade-off in querying and caching cost. For networks of size N, $O(\sqrt{N})$ cost for both caching and querying is required to achieve a constant (in N) search success probability. Queries execute a time of $O(\log^2 N)$.

Using query transformation to improve Gnutella search performance

Surendar Chandra
surendar@acm.org

William Acosta
william.acosta@utoledo.edu

Abstract

Gnutella peers independently choose the way in which objects are named as well as queried. Using a long term analysis of the files shared and queries issued, we show that this flexibility leads to a mismatch between the way that objects were named and the way that users were issuing search queries. Thirty percent of the failed queries contained keywords that were not present in any file name while the remaining queries failed because no file name contained all the keywords in a particular query. Our earlier analysis of files shared in the popular iTunes music file sharing system showed that standardizing the file names to make them easier to search is not a viable alternative. Instead, we transform the queries to better match the objects available in the system. We investigated spell correction (using file name information from the neighborhood) as well as remove query keywords. We consider the results from the transformed query to be relevant to the intent of the original query if the transformed query used many of the original keywords and the number of matching files closely matched the number of matches for typical successful queries. Our approach is practical and uses information available within the immediate neighborhood of an ultra-peer. An overlay agnostic analysis shows that our transformation improves success rates from 45% to between 72.5% and 91.2%. Using our *Hybrid* mechanism as a Gnutella middleware, our transformation produced relevant results for about 61% of the failed queries.

Tagging with DHARMA, a *DH*T-based *A*pproach for *R*esource *M*apping through *A*pproximation

Luca Maria Aiello, Marco Milanesio, Giancarlo Ruffo and Rossano Schifanella
Computer Science Department - Università degli Studi di Torino
{aiello, milane, ruffo, schifane}@di.unito.it

Abstract

We introduce collaborative tagging and faceted search on structured P2P systems. Since a trivial and brute force mapping of an entire folksonomy over a DHT-based system may reduce scalability, we propose an approximated graph maintenance approach. Evaluations on real data coming from Last.fm prove that such strategies reduce vocabulary noise (i.e., representation's overfitting phenomena) and hotspots issues.

Modeling and Analyzing the Effects of Firewalls and NATs in P2P Swarming Systems

L. D'Acunto, M. Meulpolder, R. Rahman, J.A. Pouwelse and H.J. Sips
Department of Computer Science
Delft University of Technology, The Netherlands
Email: l.dacunto@tudelft.nl

Abstract

Many P2P systems have been designed without taking into account an important factor: a large fraction of Internet users nowadays are located behind a network address translator (NAT) or a firewall, making them unable to accept incoming connections (i.e. unconnectable). Peers suffering from this limitation cannot fully enjoy the advantages offered by the P2P architecture and thus they are likely to get a poor performance.

In this work, we present a mathematical model to study the performance of a P2P swarming system in the presence of unconnectable peers. We quantify the average download speeds of peers and find that unconnectable peers achieve a *lower* average download speed compared to connectable peers, and this difference increases hyperbolically as the percentage of unconnectable peers grows. More interestingly, we notice that connectable peers actually *benefit* from the existence of peers behind NATs/firewalls, since they alone can enjoy the bandwidth that those peers offer to the system. Inspired by these observations, we propose a new policy for the allocation of the system's bandwidth that can mitigate the performance issues of unconnectable peers. In doing so, we also find an intrinsic limitation in the speed improvement that they can possibly achieve.

Efficient DHT attack mitigation through peers' ID distribution

Thibault Cholez, Isabelle Chrisment and Olivier Festor
MADYNES - INRIA Nancy-Grand Est, France
{thibault.cholez, isabelle.chrisment, olivier.festor}@loria.fr

Abstract

We present a new solution to protect the widely deployed KAD DHT against localized attacks which can take control over DHT entries. We show through measurements that the IDs distribution of the best peers found after a lookup process follows a geometric distribution. We then use this result to detect DHT attacks by comparing real peers' ID distributions to the theoretical one thanks to the Kullback-Leibler divergence. When an attack is detected, we propose countermeasures that progressively remove suspicious peers from the list of possible contacts to provide a safe DHT access. Evaluations show that our method detects the most efficient attacks with a very small false-negative rate, while countermeasures successfully filter almost all malicious peers involved in an attack. Moreover, our solution completely fits the current design of the KAD network and introduces no network overhead.

Degree Hunter: on the Impact of Balancing Node Degrees in de Bruijn-Based Overlay Networks

Pierre Fraigniaud
CNRS and University Paris Diderot
Email: pierre.fraigniaud@liafa.jussieu.fr
http://www.liafa.jussieu.fr/~pierref/

Hoang-Anh Phan
CNRS and University Paris Diderot
Email: hoang-anh.phan@liafa.jussieu.fr

Abstract

This paper presents several mechanisms for balancing the node degrees in de Bruijn-based overlay networks for peer-to-peer systems. One of these mechanisms is shown to perform almost as well as an ideal centralized mechanism, but it is based on the size of the key-spaces assigned to the nodes, and thus it may interfere with protocols aiming at balancing the load of the nodes. We therefore present two other mechanisms that are solely based on the structure of the connections between the nodes. The performances of these two mechanisms depend on the environment. One of them achieves the best performances in file-sharing systems, while the other achieves the best performances in media streaming systems.

BitTorrent and Fountain Codes: Friends or Foes?

Salvatore Spoto, Rossano Gaeta, Marco Grangetto and Matteo Sereno
Department of Computer Science
University of Turin
Turin, Italy
{spoto, rossano, grangetto, matteo}@di.unito.it

Abstract

BitTorrent is the most popular file sharing protocol on the Internet. It is proved that its performance are near-optimal for generic file distribution when the overlay is in a steady state. The two main BitTorrent strategies, tit-for-tat and local rarest fist, work at best: the former assures reciprocity in downloading and uploading rates between peers and the latter distributes the different file pieces equally among the overlay. Both assure good performance in terms of resource utilization and the practical consequence is that the peers achieve good downloading times. The best network condition for the protocol is a network characterized by roughly fixed arrival rates and no flash crowds. Nevertheless, many research works argue that the performance of the protocol quickly degrades when the peers join and leave at high rates, the network is affected by flash crowds phenomenon and the number of peer that shares the complete file is only a little fraction of the total population. This is the case of many real-world peer-to-peer applications like video-on-demand or live-streaming. In this scenario the introduction of some kind of network coding can mitigate the adverse network behavior and improve the overall performance of the protocol. In this paper we develop a modification of the BitTorrent protocol with the introduction of Luby-Transform codes, that belong to the class of the erasure rateless codes. Using a modified version of GPS (General Purpouse Simulator), we set up simulations that prove how these changes make the original protocol more robust to adverse network conditions and speed up its performance in such situations.

Both authors are supported by the ANR project “ALADDIN”, and by the INRIA project “GANG”.

High Performance Peer-to-Peer Distributed Computing with Application to Obstacle Problem

The Tung Nguyen[1,2,6], Didier El Baz[1,2,6], Pierre Spitéri[3,7], Guillaume Jourjon[4,8] and Ming Chau[5,9]
[1]CNRS ; LAAS ; 7 avenue du colonel Roche, F-31077 Toulouse, France.
[2]Universit de Toulouse ; UPS, INSA, INP, ISAE ; LAAS ; F-31077 Toulouse France.
[3]ENSEEIHT-IRIT, 2 rue Charles Camichel, 31071 Toulouse, France
[4]NICTA, Australian Technology Park, Eveleigh, NSW, Australia
[5]Advanced Solutions Accelerator, 199 rue de l'Oppidum, 34170 Castelnau le Lez, France
[6]{elbaz, ttnguyen}@laas.fr, [7]Pierre.Spiteri@enseeiht.fr,
[8]guillaume.jourjon@nicta.com.au, [9]mchau@advancedsolutionsaccelerator.com

Abstract

This paper deals with high performance Peer-to-Peer computing applications. We concentrate on the solution of large scale numerical simulation problems via distributed iterative methods. We present the current version of an environment that allows direct communication between peers. This environment is based on a self-adaptive communication protocol. The protocol configures itself automatically and dynamically in function of application requirements like scheme of computation and elements of context like topology by choosing the most appropriate communication mode between peers. A first series of computational experiments is presented and analyzed for the obstacle problem.

Analysis of Random Time-Based Switching for File Sharing in Peer-to-Peer Networks

Keqin Li
Department of Computer Science
State University of New York
New Paltz, New York 12561, USA
Email: lik@newpaltz.edu

Abstract

The expected file download time of the randomized time-based switching algorithm for peer selection and file downloading in a peer-to-peer (P2P) network is still unknown. The main contribution of this paper is to analyze the expected file download time of the time-based switching algorithm for file sharing in P2P networks when the service capacity of a source peer is totally correlated over time, namely, the service capacities of a source peer in different time slots are a fixed value. A recurrence relation is developed to characterize the expected file download time of the time-based switching algorithm. Is is proved that for two or more heterogeneous source peers and sufficiently large file size, the expected file download time of the time-based switching algorithm is less than and can be arbitrarily less than the expected download time of the chunk-based switching algorithm and the expected download time of the permanent connection algorithm. It is shown that the expected file download time of the time-based switching algorithm is in the range of the file size divided by the harmonic mean of service capacities and the file size divided by the arithmetic mean of service capacities. Numerical examples and data are presented to demonstrate our analytical results.

Workshop 15

Workshop on Multi-Threaded Architectures and Applications

MTAAP 2010

Modeling Bounds on Migration Overhead for a Traveling Thread Architecture

Patrick A. La Fratta and Peter M. Kogge
Dept. of Computer Science and Engineering
University of Notre Dame
Notre Dame, IN USA
plafratt@nd.edu, kogge@cse.nd.edu

Abstract

Heterogeneous multicore architectures have gained widespread use in the general purpose and scientific computing communities, and architects continue to investigate techniques for easing the burden of parallelization from the programmer. This paper presents a new class of heterogeneous multicores that leverages past work in architectures supporting the execution of traveling threads. These traveling threads execute on simple cores distributed across the chip and can move up the hierarchy and between cores based on data locality. This new design offers the benefits of improved performance at lower energy and power density than centralized counterparts through intelligent data placement and cooperative caching policies. We employ a methodology consisting of mathematical modeling and simulation to estimate the upper bounds on migration overhead for various architectural organizations. Results illustrate that the new architecture can match the performance of a conventional processor with reasonable thread sizes. We have observed that between 0.04 and 7.09 instructions per migration (IPM) (1.88 IPM on average) are sufficient to match the performance of the conventional processor. These results confirm that this distributed architecture and corresponding execution model offer promising potential in overcoming the design challenges of centralized counterparts.

TiNy Threads on BlueGene/P: Exploring Many-Core Parallelisms Beyond The Traditional OS

Handong Ye, Robert Pavel, Aaron Landwehr and Guang R. Gao
University of Delaware
Department of Electrical and Computer Engineering
Newark, Delaware
{handong, pavel, alandweh, ggao}@capsl.udel.edu

Abstract

Operating Systems have been considered as a cornerstone of the modern computer system, and the conventional operating system model targets computers designed around the sequential execution model. However, with the rapid progress of the multi-core/manycore technologies, we argue that OSes must be adapted to the underlying hardware platform to fully exploit parallelism. To illustrate this, our paper reports a study on how to perform such an adaptation for the IBM BlueGene/P multi-core system.

This paper's major contributions are threefold. First, we have proposed a strategy to isolate the traditional OS functions to a single core of the BG/P multi-core chip, leaving the management of the remaining cores to a runtime software that is optimized to realize the parallel semantics of the user application according to a parallel program execution model. Second, we have ported the TNT (TiNy Thread) execution model to allow for further utilization of the BG/P compute cores. Finally, we have expanded the design framework described above to a multi-chip system designed for scalability to a large number of chips.

An implementation of our method has been completed on the Surveyor BG/P machine operated by Argonne National Laboratory. Our experimental results provide insight into the strengths of this approach: (1) The performance of the TNT thread system shows comparable speedup to that of Pthreads running on the same hardware; (2) The distributed shared memory operates at 95% of the experimental peak performance of the machine, with distance between nodes not being a sensitive factor; (3) The cost of thread creation shows a linear relationship as threads increase; (4) The cost of waiting at a barrier is constant and independent of the number of threads involved.

Scheduling complex streaming applications on the Cell processor

Matthieu Gallet, Mathias Jacquelin and Loris Marchal
LIP laboratory, UMR 5668, ENS Lyon – CNRS – INRIA – UCBL – Université de Lyon
Lyon, France
{Matthieu.Gallet|Mathias.Jacquelin|Loris.Marchal}@ens-lyon.fr

Abstract

In this paper, we consider the problem of scheduling streaming applications described by complex task graphs on a heterogeneous multicore processor, the STI Cell BE processor. We first present a theoretical model of the Cell processor. Then, we use this model to express the problem of maximizing the throughput of a streaming application on this processor. Although the problem is proven NP-complete, we present an optimal solution based on mixed linear programming. This allows us to compute the optimal mapping for a number of applications, ranging from a real audio encoder to complex random task graphs. These mappings are then tested on two platforms embedding Cell processors, and compared to simple heuristic solutions. We show that we are able to achieve a good speed-up, whereas the heuristic solutions generally fail to deal with the strong memory and communication constraints.

User Level DB: a Debugging API for User-Level Thread Libraries

Kevin Pouget[1], Marc Pérache[2] , Patrick Carribault[2] and Hervé Jourdren[2]
[1]kevin.pouget@gmail.com; [2]{marc.perache, patrick.carribault, herve.jourdren}@cea.fr
CEA, DAM, DIF
F-91297 Arpajon France

Abstract

With the advent of the multicore era, parallel programming is becoming ubiquitous. Multithreading is a common approach to benefit from these architectures. Hybrid *M*:*N* libraries like MultiProcessor Communication (MPC) or MARCEL reach high performance expressing fine-grain parallelism by mapping *M* user-level threads onto *N* kernel-level threads. However, such implementations skew the debuggers' ability to distinguish one thread from another, because only kernel threads can be handled. SUN MICROSYSTEMS' THREAD_DB API is an interface between the debugger and the thread library allowing the debugger to inquire for thread semantics details. In this paper we introduce the USER LEVEL DB (ULDB) library, an implementation of the THREAD_DB interface abstracting the common features of user-level thread libraries. ULDB gathers the generic algorithms required to debug threads and provide the thread library with a small and focused interface. We describe the usage of our library with widely-used debuggers (GDB, DBX) and the integration into a user-level thread library (GNUPTH) and two high-performance hybrid libraries (MPC, MARCEL).

A Multi-Threaded Approach for Data-Flow Analysis

Marcus Edvinsson and Welf Löwe
Department of Computer Science
Linnaeus University
Växjö, Sweden
marcus.edvinsson@lnu.se, welf.lowe@lnu.se

Abstract

Program analysis supporting software development is often part of edit-compile-cycles, and precise program analysis is time consuming. With the availability of parallel processing power on desktop computers, parallelization is a way to speed up program analysis. This requires a parallel data-flow analysis with sufficient work for each processing unit. The present paper suggests such an approach for object-oriented programs analyzing the target methods of polymorphic calls in parallel. With carefully selected thresholds guaranteeing sufficient work for the parallel threads and only little redundancy between them, this approach achieves a maximum speed-up of 5 (average 1.78) on 8 cores for the benchmark programs.

Experimental Comparison of Emulated Lock-free vs. Fine-grain Locked Data Structures on the Cray XMT

Rob Farber
Pacific Northwest National Laboratory
Richland, WA
rmfarber@pnl.gov

David Mizell
Cray, Inc.
Seattle, WA
dmizell@cray.com

Abstract

Three implementations of a concurrently-updateable linked list were compared, one that emulates a lock-free approach based on a compare-and-swap instruction, one that makes direct use of the Cray XMT's full-empty synchronization bits on every word of memory, and a third that uses the XMT's atomic int_fetch_add instruction. The relative performance of the three implementations was experimentally compared on a 512-processor XMT. The direct implementation approach performed up to twice as fast as the other two approaches under conditions of low contention, but the three implementations performed about the same when the amount of contention was high.

Large Scale Complex Network Analysis using the Hybrid Combination of a MapReduce cluster and a Highly Multithreaded System

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

Abstract

Complex networks capture interactions among entities in various application areas in a graph representation. Analyzing large scale complex networks often answers important questions—e.g. estimate the spread of epidemic diseases— but also imposes computing challenges mainly due to large volumes of data and the irregular structure of the graphs. In this paper, we aim to solve such a challenge: finding relationships in a subgraph extracted from the data. We solve this problem using three different platforms: a MapReduce cluster, a highly multithreaded system, and a hybrid system of the two. The MapReduce cluster and the highly multithreaded system reveal limitations in efficiently solving this problem, whereas the hybrid system exploits the strengths of the two in a synergistic way and solves the problem at hand. In particular, once the subgraph is extracted and loaded into memory, the hybrid system analyzes the subgraph five orders of magnitude faster than the MapReduce cluster.

On the Parallelisation of MCMC by Speculative Chain Execution

Jonathan M. R. Byrd
Department of Computer Science
University of Warwick
Coventry, CV4 7AL, UK
Email: J.M.R.Byrd@dcs.warwick.ac.uk

Stephen A. Jarvis
Department of Computer Science
University of Warwick
Coventry, CV4 7AL, UK
Email: Stephen.Jarvis@dcs.warwick.ac.uk

Abhir H. Bhalerao
Department of Computer Science
University of Warwick
Coventry, CV4 7AL, UK
Email: Abhir.Bhalerao@dcs.warwick.ac.uk

Abstract

The increasing availability of multi-core and multiprocessor architectures provides new opportunities for improving the performance of many computer simulations. Markov Chain Monte Carlo (MCMC) simulations are widely used for approximate counting problems, Bayesian inference and as a means for estimating very high-dimensional integrals. As such MCMC has had a wide variety of applications in fields including computational biology and physics, financial econometrics, machine learning and image processing.

One method for improving the performance of Markov Chain Monte Carlo simulations is to use SMP machines to perform 'speculative moves', reducing the runtime whilst producing statistically identical results to conventional sequential implementations. In this paper we examine the circumstances under which the original speculative moves method performs poorly, and consider how some of the situations can be addressed by refining the implementation. We extend the technique to perform Markov Chains speculatively, expanding the range of algorithms that maybe be accelerated by speculative execution to those with non-uniform move processing times. By simulating program runs we can predict the theoretical reduction in runtime that may be achieved by this technique. We compare how efficiently different architectures perform in using this method, and present experiments that demonstrate a runtime reduction of up to 35-42% where using conventional speculative moves would result in execution as slow, if not slower, than sequential processing.

Out-of-Core Distribution Sort in the FG Programming Environment

Priya Natarajan and Thomas H. Cormen
Dartmouth College
Department of Computer Science
{priya, thc}@cs.dartmouth.edu

Elena Riccio Strange
A9.com
laneyd@gmail.com

Abstract

We describe the implementation of an out-of-core, distribution-based sorting program on a cluster using FG, a multithreaded programming framework. FG mitigates latency from disk-I/O and interprocessor communication by overlapping such high-latency operations with other operations. It does so by constructing and executing a coarse-grained software pipeline on each node of the cluster, where each stage of the pipeline runs in its own thread. The sorting program distributes data among the nodes to create sorted runs, and then it merges sorted runs on each node. When distributing data, the rates at which a node sends and receives data will differ. When merging sorted runs, each node will consume data from each of its sorted runs at varying rates. Under these conditions, a single pipeline running on each node is unwieldy to program and not necessarily efficient. We describe how we have extended FG to support multiple pipelines on each node in two forms. When a node might send and receive data at different rates during interprocessor communication, we use disjoint pipelines on each node: one pipeline to send and one pipeline to receive. When a node consumes and produces data from different streams on the node, we use multiple pipelines that intersect at a particular stage. Experimental results show that by using multiple pipelines, an out-of-core, distribution-based sorting program outperforms an out-of-core sorting program based on columnsort—taking approximately 75%–85% of the time—despite the advantages that the columnsort-based program holds.

Massive Streaming Data Analytics: A Case Study with Clustering Coefficients

David Ediger, Karl Jiang, Jason Riedy and David A. Bader
Georgia Institute of Technology
Atlanta, GA, USA

Abstract

We present a new approach for parallel massive graph analysis of streaming, temporal data with a dynamic and extensible representation. Handling the constant stream of new data from health care, security, business, and social network applications requires new algorithms and data structures. We examine data structure and algorithm trade-offs that extract the parallelism necessary for high-performance updating analysis of massive graphs. Static analysis kernels often rely on storing input data in a specific structure. Maintaining these structures for each possible kernel with high data rates incurs a significant performance cost. A case study computing clustering coefficients on a general-purpose data structure demonstrates incremental updates can be more efficient than global recomputation. Within this kernel, we compare three methods for dynamically updating local clustering coefficients: a brute-force local recalculation, a sorting algorithm, and our new approximation method using a Bloom filter. On 32 processors of a Cray XMT with a synthetic scale-free graph of $2^{24} \approx 16$ million vertices and $2^{29} \approx 537$ million edges, the brute-force method processes a mean of over 50,000 updates per second and our Bloom filter approaches 200,000 updates per second.

Hashing Strategies for the Cray XMT

Eric L. Goodman[1], David J. Haglin[2], Chad Scherrer[2],
Daniel Chavarría-Miranda[2], Jace Mogill[2] and John Feo[2]
[1]Sandia National Laboratories
Albuquerque, NM, 87123, USA
elgoodm@sandia.gov
[2]Pacific Northwest National Laboratory
Richland, WA, 99354, USA
{david.haglin, chad.scherrer, daniel.chavarria, jace.mogill, john.feo}@pnl.gov

Abstract

Two of the most commonly used hashing strategies—linear probing and hashing with chaining—are adapted for efficient execution on a Cray XMT. These strategies are designed to minimize memory contention. Datasets that follow a power law distribution cause significant performance challenges to shared memory parallel hashing implementations. Experimental results show good scalability up to 128 processors on two power law datasets with different data types: integer and string. These implementations can be used in a wide range of applications.

Workshop 16

Workshop on Parallel and Distributed Computing in Finance

PDCoF 2010

Parallelizing a Black-Scholes Solver based on Finite Elements and Sparse Grids

Hans-Joachim Bungartz, Alexander Heinecke, Dirk Pflüer and Stefanie Schraufstetter
Institut für Informatik
Technische Universität München
Boltzmannstr. 3, 85748 Garching, Germany
Email: {bungartz,heinecke,pflueged,schraufs}@in.tum.de

abstract We present the parallelization of a sparse grid finite element discretization of the Black-Scholes equation, which is commonly used for option pricing. Sparse grids allow to handle higher dimensional options than classical approaches on full grids, and can be extended to a fully adaptive discretization method. We introduce the algorithmical structure of efficient algorithms operating on sparse grids, and demonstrate how they can be used to derive an efficient parallelization with OpenMP of the Black-Scholes solver. We show results on different commodity hardware systems based on multi-core architectures with up to 8 cores, and discuss the parallel performance using Intel and AMD CPUs.

Pricing of Cross-Currency Interest Rate Derivatives on Graphics Processing Units

Duy Minh Dang
Department of Computer Science
University of Toronto
Toronto, ON, Canada
dmdang@cs.toronto.edu

Abstract

We present a Graphics Processing Unit (GPU) parallelization of the computation of the price of cross-currency interest rate derivatives via a Partial Differential Equation (PDE) approach. In particular, we focus on the GPUbased parallel computation of the price of long-dated foreign exchange interest rate hybrids, namely Power Reverse Dual Currency (PRDC) swaps with Bermudan cancelable features. We consider a three-factor pricing model with foreign exchange skew which results in a time-dependent parabolic PDE in three spatial dimensions. Finite difference methods on uniform grids are used for the spatial discretization of the PDE, and the Alternating Direction Implicit (ADI) technique is employed for the time discretization. We then exploit the parallel architectural features of GPUs together with the Compute Unified Device Architecture (CUDA) framework to design and implement an efficient parallel algorithm for pricing PRDC swaps. Over each period of the tenor structure, we divide the pricing of a Bermudan cancelable PRDC swap into two independent pricing subproblems, each of which can efficiently be solved on a GPU via a parallelization of the ADI scheme at each timestep. Using this approach on two NVIDIA Tesla C870 GPUs of an NVIDIA 4-GPU Tesla S870 to price a Bermudan cancelable PRDC swap having a 30 year maturity and annual exchange of fund flows, we have achieved an asymptotic speedup by a factor of 44 relative to a single thread on a 2.0GHz Xeon processor.

A Parallel Particle Swarm Optimization Algorithm for Option Pricing

Hari Prasain, Girish Kumar Jha, Parimala Thulasiraman and Ruppa Thulasiram
Department of Computer Science
University of Manitoba
Winnipeg, Manitoba, Canada
{hprasain, girish, thulasir, tulsi}@cs.umanitoba.ca

Abstract

Option pricing is one of the challenging problems of computational finance. Nature-inspired algorithms have gained prominence in real world optimization problems such as in mobile ad hoc networks. The option pricing problem fits very well into this category of problems due to the ad hoc nature of the market. Particle swarm optimization (PSO) is one of the novel global search algorithms based on swarm intelligence.

We first show that PSO could be effectively used for the option pricing problem. The results are compared with standard classical Black-Scholes-Merton model for simple European options. In this study, we developed two algorithms for option pricing using Particle Swarm Optimization (PSO). The first algorithm we developed is synchronous option pricing algorithm using PSO (SPSO), and the second is parallel synchronous option pricing algorithm. The pricing results of these algorithms are close when compared with classical Black-Scholes-Merton model for simple European options. We test our Parallel Synchronous PSO algorithm in three architectures: shared memory machine using OpenMP, distributed memory machine using MPI and on a homogeneous multicore architecture running MPI and OpenMP (hybrid model). The results show that the hybrid model handles the load well as we increase the number of particles in simulation while maintaining equivalent accuracy.

Workshop 17
Workshop on Large-Scale Parallel Processing
LSPP 2010

Efficient Lists Intersection by CPU-GPU Cooperative Computing

Di Wu, Fan Zhang, Naiyong Ao, Gang Wang, Xiaoguang Liu, Jing Liu
Nankai-Baidu Joint Lab, College of Information Technical Science, Nankai University
Weijin Road 94, Tianjin, 300071, China
Email: {wakensky, zhangfan555}@gmail.com, {aonaiyong, wgzwp}@163.com,
liuxg74@yahoo.com.cn, jingliu@mail.nankai.edu.cn

Abstract

Lists intersection is an important operation in modern web search engines. Many prior studies have focused on the single-core or multi-core CPU platform or many-core GPU. In this paper, we propose a CPU-GPU cooperative model that can integrate the computing power of CPU and GPU to perform lists intersection more efficiently. In the so-called synchronous mode, queries are grouped into batches and processed by GPU for high throughput. We design a query-parallel GPU algorithm based on an element-thread mapping strategy for load balancing. In the traditional asynchronous model, queries are processed one-by-one by CPU or GPU to gain perfect response time. We design an online scheduling algorithm to determine whether CPU or GPU processes the query faster. Regression analysis on a huge number of experimental results concludes a regression formula as the scheduling metric. We perform exhaustive experiments on our new approaches. Experimental results on the TREC Gov and Baidu datasets show that our approaches can improve the performance of the lists intersection significantly.

High Precision Integer Multiplication with a Graphics Processing Unit

Niall Emmart and Charles Weems
Computer Science Department
University of Massachusetts
Amherst, MA 01003-4610, USA.
Email: nemmart@yrrid.com, weems@cs.umass.edu

Abstract

In this paper we evaluate the potential for using an NVIDIA graphics processing unit (GPU) to accelerate high precision integer multiplication. The reported peak vector performance for a typical GPU appears to offer considerable potential for accelerating such a regular computation. Because of limitations in the on-chip memory, the high cost of kernel launches, and the particular nature of the architecture's support for parallelism, we found it necessary to use a hybrid algorithmic approach to obtain good performance. On the GPU itself we use an adaptation of the Strassen FFT algorithm to multiply 32KB chunks, while on the CPU we adapt the Karatsuba divide-and-conquer approach to optimize the application of the GPU's partial multiplies, which are viewed as "digits" by our implementation of Karatsuba. Even with this approach, the result is at best a modest increase in performance, compared with executing the same multiplication using the GMP package on a CPU at a comparable technology node. We identify the sources of this lackluster performance and discuss the likely impact of planned advances in GPU architecture.

Large Neighborhood Local Search Optimization on Graphics Processing Units

Thé Van Luong, Nouredine Melab and El-Ghazali Talbi
Dolphin / Opac Team
INRIA Lille Nord Europe / CNRS LIFL
Parc Scientifique de la Haute Borne
40, avenue Halley, Bat. A, Park Plaza
59650 Villeneuve d'Ascq, France
The-Van.Luong@inria.fr, Nouredine.Melab@lifl.fr, El-Ghazali.Talbi@lifl.fr

Abstract

Local search (LS) algorithms are among the most powerful techniques for solving computationally hard problems in combinatorial optimization. These algorithms could be viewed as "walks through neighborhoods" where the walks are performed by iterative procedures that allow to move from a solution to another one in the solution space. In these heuristics, designing operators to explore large promising regions of the search space may improve the quality of the obtained solutions at the expense of a highly computationally process. Therefore, the use of graphics processing units (GPUs) provides an efficient complementary way to speed up the search. However, designing applications on GPU is still complex and many issues have to be faced. We provide a methodology to design and implement large neighborhood LS algorithms on GPU. The work has been experimented for binary problems by deploying multiple neighborhood structures. The obtained results are convincing both in terms of efficiency, quality and robustness of the provided solutions at run time.

A Fast GPU Algorithm for Graph Connectivity

Jyothish Soman, Kothapalli Kishore and P J Narayanan
IIIT-Hyderabad
Gachibowli, Hyderabad, Andhra Pradesh
India-500032
Email: {jyothish@students., kkishore@, pjn@}iiit.ac.in

Abstract

Graphics processing units provide a large computational power at a very low price which position them as an ubiquitous accelerator. General purpose programming on the graphics processing units (GPGPU) is best suited for regular data parallel algorithms. They are not directly amenable for algorithms which have irregular data access patterns such as list ranking, and finding the connected components of a graph, and the like. In this work, we present a GPU-optimized implementation for finding the connected components of a given graph. Our implementation tries to minimize the impact of irregularity, both at the data level and functional level.

Our implementation achieves a speed up of 9 to 12 times over the best sequential CPU implementation. For instance, our implementation finds connected components of a graph of 10 million nodes and 60 million edges in about 500 milliseconds on a GPU, given a random edge list. We also draw interesting observations on why PRAM algorithms, such as the Shiloach-Vishkin algorithm may not be a good fit for the GPU and how they should be modified.

An Efficient Associative Processor Solution to an Air Traffic Control Problem

Mike Yuan and Johnnie Baker
Department of Computer Science
Kent State University
Kent, Ohio 44242
myuan, jbaker@cs.kent.edu

Frank Drews and Lev Neiman
School of Electrical Engineering
and Computer Science
Ohio University
Athens, OH
drews@ohio.edu, lev.neiman@gmail.com

Will Meilander(retired)
Department of Computer Science
Kent State University
Kent, OH
willcm@charter.net

Abstract

This paper proposes a SIMD solution to air traffic control (ATC) using SIMD machine model called an Associative Processor (AP). This differs from previous ATC systems that are designed for MIMD computers and have a great deal of difficulty meeting the predictability requirements for ATC, which are critical in the process for meeting the strict certification standards required for safety critical software components. Our SIMD solution will support accurate and meaningful predictions of worst case execution times, will allow all deadline requirements to be met, and will guarantee all deadlines are met. Also, the software will be much simpler and smaller in size than the current corresponding ATC software. An important consequence of these features is that the V&V (Validation and Verification) process will be considerably simpler than for current ATC software. Additionally, the associative processor is enhanced SIMD hardware and is considerably cheaper and simpler than the MIMD hardware currently used to support ATC. The ClearSpeed CSX600 accelerator is used to emulate the AP model. A preliminary implementation of the proposed method has been developed and a number of experimental results comparing MIMD and CSX600 approaches are presented. The performance of CSX600 has better scalability, efficiency, and predictability than that of MIMD.

Analyzing the Trade-off between Multiple Memory Controllers and Memory Channels on Multi-core Processor Performance

José Carlos Sancho, Michael Lang and Darren J. Kerbyson
Performance and Architecture Laboratory (PAL)
Computer Science for HPC (CCS-1)
Los Alamos National Laboratory, NM 87545, USA
{jcsancho,mlang,djk}@lanl.gov

Abstract

The increasing core-count on current and future processors is posing critical challenges to the memory subsystem to efficiently handle concurrent memory requests. The current trend is to increase the number of memory channels available to the processor's memory controller. In this paper we investigate the effectiveness of this approach on the performance of parallel scientific applications. Specifically, we explore the trade-off between employing multiple memory channels per memory controller and the use of multiple memory controllers. Experiments conducted on two current state-of-the-art multicore processors, a 6-core AMD Istanbul and a 4-core Intel Nehalem-EP, for a wide range of production applications shows that there is a diminishing return when increasing the number of memory channels per memory controller. In addition, we show that this performance degradation can be efficiently addressed by increasing the ratio of memory controllers to channels while keeping the number of memory channels constant. Significant performance improvements can be achieved in this scheme, up to 28%, in the case of using two memory controllers each with one channel compared with one controller with two memory channels.

Multicore-aware parallel temporal blocking of stencil codes for shared and distributed memory

Markus Wittmann, Georg Hager and Gerhard Wellein
Erlangen Regional Computing Center
University of Erlangen-Nuremberg
Erlangen, Germany
{markus.wittmann,georg.hager,gerhard.wellein}@rrze.uni-erlangen.de

Abstract

New algorithms and optimization techniques are needed to balance the accelerating trend towards bandwidth-starved multicore chips. It is well known that the performance of stencil codes can be improved by temporal blocking, lessening the pressure on the memory interface. We introduce a new pipelined approach that makes explicit use of shared caches in multicore environments and minimizes synchronization and boundary overhead. For clusters of shared-memory nodes we demonstrate how temporal blocking can be employed successfully in a hybrid shared/distributed-memory environment.

Scalable Parallel I/O Alternatives for Massively Parallel Partitioned Solver Systems

Jing Fu[1], Ning Liu[1], Onkar Sahni[2], Kenneth E. Jansen[2], Mark S. Shephard[2], Christopher D. Carothers[1]
[1]Department of Computer Science, Rensselaer Polytechnic Institute, Troy, New York 12180
[2]Scientific Computation Research Center, Rensselaer Polytechnic Institute, Troy, New York 12180
E-mails:{fuj,liun2,chrisc}@cs.rpi.edu, {osahni,kjansen,shephard}@scorec.rpi.edu

Abstract

With the development of high-performance computing, I/O issues have become the bottleneck for many massively parallel applications. This paper investigates scalable parallel I/O alternatives for massively parallel partitioned solver systems. Typically such systems have synchronized "loops" and will write data in a well defined block I/O format consisting of a header and data portion. Our target use for such an parallel I/O subsystem is *checkpoint-restart* where writing is by far the most common operation and reading typically only happens during either initialization or during a restart operation because of a system failure. We compare four parallel I/O strategies: 1 POSIX File Per Processor (1PFPP), a synchronized parallel IO library (syncIO), "Poor-Man's" Parallel I/O (PMPIO) and a new "reduced blocking" strategy (rbIO). Performance tests using real CFD solver data from PHASTA (an unstructured grid finite element Navier-Stokes solver) show that the syncIO strategy can achieve a read bandwidth of 6.6GB/Sec on Blue Gene/L using 16K processors which is significantly faster than 1PFPP or PMPIO approaches. The serial "token-passing" approach of PMPIO yields a 900MB/sec write bandwidth on 16K processors using 1024 files and 1PFPP achieves 600 MB/sec on 8K processors while the "reduced-blocked" rbIO strategy achieves an actual writing performance of 2.3GB/sec and *perceived/latency hiding* writing performance of more than 21,000 GB/sec (i.e., 21TB/sec) on a 32,768 processor Blue Gene/L.

Performance analysis of Sweep3D on Blue Gene/P with the Scalasca toolset

Brian J. N. Wylie[1], David Böhme[1,2], Bernd Mohr[1], Zoltán Szebenyi[1,2], and Felix Wolf[1,2,3]
[1]Jülich Supercomputing Centre, Forschungszentrum Jülich, 52425 Jülich, Germany
[2] RWTH Aachen University, 52056 Aachen, Germany
[3]German Research School for Simulation Sciences, 52062 Aachen, Germany
{b.wylie, d.boehme, b.mohr, z.szebenyi, f.wolf}@fz-juelich.de

Abstract

In studying the scalability of the Scalasca performance analysis toolset to several hundred thousand MPI processes on IBM Blue Gene/P, we investigated a progressive execution performance deterioration of the well-known ASCI Sweep3D compact application. Scalasca runtime summarization analysis quantified MPI communication time that correlated wth computational imbalance, and automated trace analysis confirmed growing amounts of MPI waiting times. Further instrumentation, measurement and analyses pinpointed a conditional section of highly imbalanced computation which amplified waiting times inherent in the associated wavefront communication that seriously degraded overall execution efficiency at very large scales. By employing effective data collation, management and graphical presentation, Scalasca was thereby able to demonstrate performance measurements and analyses with 294,912 processes for the first time.

To Upgrade or not to Upgrade? Catamount vs. Cray Linux Environment

S.D. Hammond, G.R. Mudalige, J.A. Smith, J.A. Davis and S.A. Jarvis
High Performance Systems Group, University of Warwick, Coventry, CV4 7AL, UK

J. Holt
Tessella PLC, Abingdon Science Park, Berkshire, OX14 3YS, UK,

I. Miller, A Herdman and A Vadgama
Atomic Weapons Establishment, Aldermaston, Reading, RG7 4PR, UK

Abstract

Modern supercomputers are growing in diversity and complexity – the arrival of technologies such as multi-core processors, general purpose-GPUs and specialised compute accelerators has increased the potential scientific delivery possible from such machines. This is not however without some cost, including significant increases in the sophistication and complexity of supporting operating systems and software libraries. This paper documents the development and application of methods to assess the *potential performance* of selecting one hardware, operating system (OS) and software stack combination against another. This is of particular interest to supercomputing centres, which routinely examine prospective software/architecture combinations and possible machine upgrades. A case study is presented that assesses the potential performance of a particle transport code on AWE's Cray XT3 8,000-core supercomputer running images of the Catamount and the Cray Linux Environment (CLE) operating systems. This work demonstrates that by running a number of small benchmarks on a test machine and network, and observing factors such as operating system noise, it is possible to speculate as to the performance impact of upgrading from one operating system to another on the system as a whole. This use of performance modelling represents an inexpensive method of examining the likely behaviour of a large supercomputer before and after an operating system upgrade; this method is also attractive if it is desirable to minimise system downtime while exploring software-system upgrades. The results show that benchmark tests run on less that 256 cores would suggest that the impact (overhead) of upgrading the operating system to CLE was less than 10%; model projections suggest that this is not the case at scale.

IPDPS 2010 PhD Forum

Memory Affinity Management for Numerical Scientific Applications over Multi-core Multiprocessors with Hierarchical Memory

Christiane Pousa Ribeiro and Jean-François Méhaut
University of Grenoble
INRIA - MESCAL Research Team
Grenoble, France
{pousa, Jean-Francois.Mehaut}@imag.fr

Alexandre Carissimi
Universidade Federal do Rio Grande do Sul
Porto Alegre, Brazil
asc@inf.ufrgs.br

Abstract

Nowadays, on Multi-core Multiprocessors with Hierarchical Memory (Non-Uniform Memory Access (NUMA) characteristics), the number of cores accessing memory banks is considerably high. Such accesses produce stress on the memory banks, generating load-balancing issues, memory contention and remote accesses. In this context, how to manage memory accesses in an efficient fashion remains an important concern. To reduce memory access costs, developers have to manage data placement on their application assuring memory affinity. The problem is: How to guarantee memory affinity for different applications/NUMA platforms and assure efficiency, portability, minimal or none source code changes (transparency) and fine control of memory access patterns?

In this Thesis, our research have led to the proposal of Minas: an efficient and portable *memory affinity management framework* for NUMA platforms. Minas provides both explicit memory affinity management and automatic one with good performance, architecture abstraction, minimal or none application source code modifications and fine control. We have evaluated its efficiency and portability by performing some experiments with numerical scientific HPC applications on NUMA platforms. The results have been compared with other solutions to manage memory affinity.

Performance Improvements of Real-Time Crowd Simulations

Guillermo Vigueras
PhD Student. 4 years in PhD program
University of Valencia. Spain
guillermo.vigueras@uv.es

Juan M. Orduña and Miguel Lozano
PhD advisors. Departamento de Informática
University of Valencia. Spain
{juan.orduna,miguel.lozano}@uv.es

Abstract

The current challenge for crowd simulations is the design and development of a scalable system that is capable of simulating the individual behavior of millions of complex agents populating large scale virtual worlds with a good frame rate. In order to overcome this challenge, this thesis proposes different improvements for crowd simulations. Concretely, we propose a distributed software architecture that can take advantage of the existing distributed and multicore architectures. In turn, the use of these distributed architectures requires partitioning strategies and workload balancing techniques for distributed crowd simulations. Also, these architectures allow the use of GPUs not only for rendering images but also for computing purposes. Finally, the design and implementation of distributed visual clients is another research topic that can help to overcome this challenge.

Parallel Applications Employing Pairwise Computations on Emerging Architectures

Abhinav Sarje and *Advisor:* Srinivas Aluru
Department of Electrical and Computer Engineering, Iowa State University, Ames, IA, USA
Years in PhD program: 5

Abstract

Today's emerging architectures have higher levels of parallelism incorporated within a processor. They require efficient strategies to extract the performance they have to offer. In our work, we develop architecture-aware parallel strategies to perform various kinds of pairwise computations – pairwise genomic alignments, and scheduling large number of general pairwise computations with applications to computational systems biology and materials science. We present our schemes in the context of the IBM Cell BE, an example of a heterogeneous multicore, but are nevertheless applicable to any similar architecture, as well as general multicores with our strategies being cache-aware.

Fault Tolerant Linear Algebra: Recovering from Fail-Stop Failures without Checkpointing

Teresa Davies and Zizhong Chen
Colorado School of Mines
{zchen,tdavies}@mines.edu

Abstract

Today's long running high performance computing applications typically tolerate fail-stop failures by checkpointing. While checkpointing is a very general technique and can be applied in a wide range of applications, it often introduces a considerable overhead especially when applications modify a large amount of memory between checkpoints. In this research, we will design highly scalable low overhead fault tolerant schemes according to the specific characteristics of an application. We will focus on linear algebra operations and re-design selected algorithms to tolerate fail-stop failures without checkpointing. We will also incorporate the developed techniques into the widely used numerical linear algebra library package ScaLAPACK.

Highly Scalable Checkpointing for Exascale Computing

Christer Karlsson
Colorado School of Mines
Golden, CO, USA
Email: ckarlsso@mines.edu

Zizhong Chen
Colorado School of Mines
Golden, CO, USA
Email: zchen@mines.edu

Abstract

A consequence of the fact that the number of processors in High Performance Computers (HPC) continues to increase is demonstrated by the correlation between *Mean-Time-To-Failure*(T_{MTTF}) and application execution time. The T_{MTTF} is becoming shorter than the expected execution time for many next generation HPC applications. There is an ability to handle failure without a system-wide breakdown in most architecture, but many of the applications do not have a built-in ability to survive node failures. The purpose of this paper is to present an approach to develop a highly scalable technique to allow the next generation applications to survive node and/or link failure without aborting the computation. We will develop several strategies to improve the scalability of diskless checkpointing. The technique is scalable in the sense that when the number of processes increases, the overhead to handle k failures on p processes should remain as constant as possible. We will present the proposed technique, initial results together with remaining objectives and challenges.

Performance Modeling of Heterogeneous Systems

Jan Christian Meyer (Ph.D. student, 4^{th} year)
Dr. Anne Cathrine Elster (Advisor)
Dept. of Computer and Information Science
Norwegian University of Science and Technology
Trondheim, Norway
{janchris, elster}@idi.ntnu.no

Abstract

Predicting how well applications may run on modern systems is becoming increasingly challenging. It is no longer sufficient to look at number of floating point operations and communication costs, but one also needs to model the underlying systems and how their topology, heterogeneity, system loads, etc, may impact performance. This work focuses on developing a practical model for heterogeneous computing by looking at the older BSP model, which attempts to model communication costs on homogeneous systems, and looks at how its library implementations can be extended to include a run-time system that may be useful for heterogeneous systems. Our extensions of BSPlib with MPI and GASnet mechanisms at the communication layer should provide useful tools for evaluating applications with respect to how they may run on heterogeneous systems.

Large-Scale Distributed Storage for Highly Concurrent MapReduce Applications

Diana Moise (Advisors: Luc Bougé, Gabriel Antoniu)
INRIA/IRISA, Rennes, France

Abstract

A large part of today's most popular applications are data-intensive; the data volume they process is continuously growing. Specialized abstractions like Google's MapReduce and Pig-Latin were developed to efficiently manage the workloads of data-intensive applications. These models propose high-level data processing frameworks intended to hide the details of parallelization from the user. Such frameworks rely on storing huge objects and target high performance by optimizing the parallel execution of the computation. The purpose of this PhD is to provide efficient storage for the MapReduce framework and the applications it was designed for. The research conducted so far, concerned the storage layer this type of applications require. To meet these requirements we rely on BlobSeer, a system for managing massive data in a large-scale distributed context.

Scalable Verification of MPI Programs

Anh Vo and Ganesh Gopalakrishnan
School of Computing, University of Utah, Salt Lake City, UT
{avo,ganesh}@cs.utah.edu

Abstract

Large message passing programs today are being deployed on clusters with hundreds, if not thousands of processors. Any programming bugs that happen will be very hard to debug and greatly affect productivity. Although there have been many tools aiming at helping developers debug MPI programs, many of them fail to catch bugs that are caused by non-determinism in MPI codes. In this work, we propose a distributed, scalable framework that can explore all relevant schedules of MPI programs to check for deadlocks, resource leaks, local assertion errors, and other common MPI bugs.

Ensuring Deterministic Concurrency through Compilation

Nalini Vasudevan
Department of Computer Science
Columbia University
New York, NY
naliniv@cs.columbia.edu

Stephen A. Edwards
Department of Computer Science
Columbia University
New York, NY
sedwards@cs.columbia.edu

Abstract

Multicore shared-memory architectures are becoming prevalent but bring many programming challenges. Among the biggest is non-determinism: the output of the program does not depend merely on the input, but also on scheduling choices taken by the operating system.

In this paper, we discuss and propose additional tools that provide determinism guarantees—compilers that generate deterministic code, libraries that provide deterministic constructs, and analyzers that check for determinism. Additionally, we discuss techniques to check for problems like deadlock that can result from the use of these deterministic constructs.

Use of Peer-To-Peer Technology in Internet Access Networks and its Impacts

Peter Danielis and Dirk Timmermann
University of Rostock
Institute of Applied Microelectronics and Computer Engineering
18051 Rostock, Germany
Tel.: +49 (381) 498 -7272
Email: peter.danielis@uni-rostock.de
Personal Website: http://www.imd.uni-rostock.de/ma/pd032
Internetworking Project: http://www.imd.uni-rostock.de/networking

Abstract

Objectives of the dissertation are impacts of Peer-to-Peer (P2P) traffic on Internet core networks as well as novel approaches for using P2P technology in Internet access networks. Thereby, challenges of P2P computing concerning topology awareness, scalability, and fault-tolerance are analyzed. The thesis' first part focuses on improving insufficient scalability and fault-tolerance properties of present-day Internet services like the Domain Name System (DNS) by using available resources in access networks. The second part addresses P2P mechanisms for a highly scalable, resilient, and distributed storing and computing solution in the access network. Finally, a new algorithm for allowing topology awareness in P2P networks is proposed.

Currently, the author is in his fourth PhD year and will finish in 2010. Further information, full references, and papers can be obtained from the websites given below the affiliation.

A Path Based Reliable Middleware Framework for RFID Devices

Nova Ahmed
Georgia Institute of Technology
College of Computing
Atlanta, GA 30332, USA
nova@cc.gatech.edu

Umakishore Ramachandran
Georgia Institute of Technology
College of Computing
Atlanta, GA 30332, USA
rama@cc.gatech.edu

Abstract

The rapid interest in large scale RFID deployment has introduced interesting opportunities and challenges which includes handling of massive amount of RFID generated data, taking care of the error prone nature of the data and being able to provide a scalable solution for the applications. There has been research work that focuses mainly on item tracking applications but not item location applications. We propose a general infrastructure that is able to serve the item tracking applications as well as item location applications using the path based spatial and temporal stamping of RFID generated data. The system takes advantage of the data flow information at the system level to improve system performance.

Improving Topological Mapping on NoCs

Rafael Tornero
PhD Student. 3 years in PhD Program
Departament d'Informàtica
Universitat de València, Spain
rafael.tornero@uv.es

Juan M. Orduña
PhD Advisor
Departament d'Informàtica
Universitat de València, Spain
juan.orduna@uv.es

Abstract

Networks-on-Chip (NoCs) have been proposed as an efficient solution to the complex communications on System-on-chip (SoCs). The design flow of network-on-chip (NoCs) include several key issues, and one of them is the decision of where cores have to be topologically mapped. This thesis proposes a new approach to the topological mapping strategy for NoCs. Concretely, we propose a new topological mapping technique for regular and irregular NoC platforms and its application for optimizing application specific NoC based on distributed and source routing.

Coping with Uncertainty in Scheduling Problems

Louis-Claude CANON
Nancy University, LaBRI, Bordeaux, France
Email: louis-claude.canon@labri.fr

Abstract

Large-scale distributed systems such as Grids constitute computational environments that are essential to academic and industry needs. However, they present uncertain behaviors due to their scales that increase continually. We propose to revisit traditional scheduling problematics in these environments by considering uncertainty in the models.

AuctionNet: Market Oriented Task Scheduling in Heterogeneous Distributed Environments

Han Zhao (Ph.D. Student, Fourth Year) and Xiaolin Li (Advisor)
Scalable Software Systems Laboratory, Department of Computer Science
Oklahoma State University, Stillwater, OK 74078, USA
Email: {haz, xiaolin}@cs.okstate.edu

Abstract

We propose a suite of market-oriented task scheduling algorithms to build an AuctionNet for heterogeneous distributed environments. In heterogeneous distributed environments, computing nodes are autonomous and owned by different organizations, for example peer-to-peer systems, desktop grids/clouds. To address such diverse heterogeneity and dynamism in systems, applications, and local policies, efficient and fair task scheduling becomes a challenging issue. To cope with such complexity in a distributed and noncooperative environment, we propose to use market-oriented incentive mechanisms to regulate task scheduling in a distributed manner. Further, to accommodate multiple objectives and criteria, we adopt a combined approach leveraging the advantage of both hypergraph theory and incentive mechanisms. We first formulate a general framework of market-oriented task scheduling in distributed systems. We then present two algorithms for task-bundle scheduling. Preliminary results demonstrate the satisfactory performance of our proposed algorithms. The remaining work to complete the PhD dissertation is then presented.

The proposed research carries significant intellectual merits and potential broader impacts in the following aspects. (1) We propose the notion of task-bundle for the first time in the literature. Product-bundle has been a common marketing strategy in our daily life for a long time. In the emerging commercial clouds and desktop clouds, task-bundle could be a useful concept for computing and storage markets. (2) We propose efficient distributed mechanisms that are very suitable for such distributed systems. A novel algorithm combining hypergraph and incentive mechanisms achieves multi-objective optimization. (3) We conduct rigorous analytical study and prove that our algorithms ensure efficiency and fairness and in the meantime maximize social welfare. (4) Overall, this proposal lays a solid foundation and sheds light on future research and realworld applications in the broad area of task scheduling in distributed systems.

Towards Dynamic Reconfigurable Load-balancing for Hybrid Desktop Platforms

Alécio P. D. Binotto[1,2], Carlos E. Pereira[1] and Dieter W. Fellner[2]
[1]Informatics Institute
UFRGS - Federal University of Rio Grande do Sul, Porto Alegre, Brazil
Email: abinotto@inf.ufrgs.br, cpereira@ece.ufrgs.br
[2]Fraunhofer IGD
Technische Universität Darmstadt, Darmstadt, Germany
Email: alecio.binotto@igd.fraunhofer.de, d.fellner@igd.fraunhofer.de

Abstract

High-performance platforms are required by applications that use massive calculations. Actually, desktop accelerators (like the GPUs) form a powerful heterogeneous platform in conjunction with multi-core CPUs. To improve application performance on these hybrid platforms, load-balancing plays an important role to distribute workload. However, such scheduling problem faces challenges since the cost of a task at a Processing Unit (PU) is non-deterministic and depends on parameters that cannot be known a priori, like input data, online creation of tasks, scenario changing, etc. Therefore, self-adaptive computing is a potential paradigm as it can provide flexibility to explore computational resources and improve performance on different execution scenarios.

This paper presents an ongoing PhD research focused on a dynamic and reconfigurable scheduling strategy based on timing profiling for desktop accelerators. Preliminary results analyze the performance of solvers for SLEs (Systems of Linear Equations) over a hybrid CPU and multi-GPU platform applied to a CFD (Computational Fluid Dynamics) application. The decision of choosing the best solver as well as its scheduling must be performed dynamically considering online parameters in order to achieve a better application performance.

Dynamic Fractional Resource Scheduling for Cluster Platforms

Mark Stillwell (4th year PhD student)
Department of Information & Computer Sciences
University of Hawai‘i at Mānoa, Honolulu, U.S.A.
Advisor: Henri Casanova

Abstract

We propose a novel approach, called Dynamic Fractional Resource Scheduling (DFRS), to share homogeneous cluster computing platforms among competing jobs. DFRS leverages virtual machine technology to share node resources in a precise and controlled manner. A key feature of this approach is that it defines and optimizes a user-centric metric of performance and fairness. We explain the principles behind DFRS and its advantages over the current state of the art, develop a model of resource sharing, and summarize results from two different simulation experiments: one comparing various heuristics in an off-line setting and another comparing our heuristics to current technology in an on-line setting. Finally, we summarize our conclusions and describe our plans for future research.

Energy-aware Joint Scheduling of Tasks and Messages in Wireless Sensor Networks

Benazir Fateh and G. Manimaran
Department of Electrical and Computer Engineering
Iowa State University
Ames, Iowa
{benazir, gmani}@iastate.edu

Abstract

We consider the problem of energy-aware joint scheduling of tasks and messages with real-time constraints in wireless networked embedded systems specifically wireless sensor networks.We use the mixed tree coloring approach in order to model the constraints and show that k-coloring of a mixed tree can be mapped to a non-conflicting schedule consisting of k time slots. Also, we propose to conduct testbed evaluation to quantify the performance of combined implementation of energy management techniques such as Dynamic Modulation Scaling (DMS) along with Dynamic Voltage Scaling (DVS).

BlobSeer: Efficient Data Management for Data-Intensive Applications Distributed at Large-Scale

Bogdan Nicolae
University of Rennes 1
IRISA
Rennes, France
bogdan.nicolae@irisa.fr

Advisor: Gabriel Antoniu
INRIA Rennes
IRISA
Rennes, France
gabriel.antoniu@inria.fr

Advisor: Luc Bougé
ENS Cachan, Brittany
IRISA
Rennes, France
luc.bouge@bretagne.ens-cachan.fr

Abstract

Large-scale data-intensive applications are a class of applications that acquire and maintain massive datasets, while performing distributed computations on these datasets. In this context, a a key factor is the storage service responsible for the data management, as it has to efficiently deal with massively parallel data access in order to ensure scalability and performance for the whole system itself. This PhD thesis proposes BlobSeer, a data management service specifically designed to address the needs of large-scale data-intensive applications. Three key design factors: data striping, distributed metadata management and versioning-based concurrency control enable BlobSeer not only to provide efficient support for features commonly used to exploit data-level parallelism, but also enable exploring a set of new features that can be leveraged to further improve parallel data access. Extensive experimentations, both in scale and scope, on the Grid5000 testbed demonstrate clear benefits of using BlobSeer as the underlying storage for a variety of scenarios: data-intensive grid applications, grid file systems, MapReduce datacenters, desktop grids. Further work targets providing efficient storage solutions for cloud computing as well.

Extendable Storage Framework for Reliable Clustered Storage Systems

Sumit Narayan (4th year student)
Advisor: John A. Chandy
University of Connecticut, Storrs, CT 06269-2157
{sumit.narayan,john.chandy}@uconn.edu

Abstract

The total amount of information stored on disks has increased tremendously in recent years with data storage, sharing and backup becoming more important than ever. The demand for storage has not only changed in size, but also in speed, reliability and security. These requirements create a big challenge for storage system architects who aim for a one system fits all?design. Storage policies like backup and security are typically set for an entire file system. However, this granularity is too large and can sacrifice storage efficiency and performance, particularly since different files have different storage requirements. In this work, we provide a framework for an attribute-based extendable storage system which will allow storage policy decisions to be made at file-level granularity and at all levels of the storage stack, including file system, operating system, and device managers. We propose to do this by using a file's extended attributes that will enable different tasks via plugins or functions implemented at various levels within the storage stack and provide a complete data-aware storage functionality from an application point of view. We provide examples of how our framework can be used to improve performance in a reliable clustered storage system.

The Effects on Branch Prediction when Utilizing Control Independence

Chris J. Michael and David M. Koppelman
Department of Electrical and Computer Engineering
Louisiana State University
Baton Rouge, Louisiana, USA
cmichael@cct.lsu.edu, koppel@ece.lsu.edu

Abstract

Though current general-purpose processors have several small CPU cores as opposed to a single more complex core, many algorithms and applications are inherently sequential and so hard to explicitly parallelize. Cores designed to handle these problems may exhibit deeper pipelines and wider fetch widths to exploit instruction-level parallelism via out-of-order execution. As these parameters increase, so does the amount of instructions fetched along an incorrect path when a branch is mispredicted. Some instructions are fetched regardless of the direction of a branch. In current conventional CPUs, these instructions are always squashed upon branch misprediction and are fetched again shortly thereafter. Recent research efforts explore lessening the effect of branch mispredictions by retaining these instructions when squashing or fetching them in advance when encountering a branch that is difficult to predict. Though these control independent processors are meant to lessen the damage of misprediction, an inherent side-effect of fetching out of order, branch weakening, reduces realized speedup and is in part responsible for lowering potential speedup. This study formally defines and works towards identifying the causes of branch weakening. The overall goal of the research is to determine how much weakening is avoidable and develop techniques to help reduce weakening in control independent processors.

High Performance Reconfigurable Multi-Processor-Based Computing on FPGAs

Diana Göhringer (3rd year PhD student)
Fraunhofer IOSB
Ettlingen, Germany
e-mail: dgoehringer@fom.fgan.de

Jürgen Becker (Advisor)
Karlsruher Institute of Technologie (KIT)
Karlsruhe, Germany
e-mail: becker@kit.edu

Abstract

Multi-processor architectures are a promising solution to provide the required computational performance for applications in the area of high performance computing. Multi- and many-core Systems-on-Chip offer the possibility to host an application, partitioned in a number of tasks, on the different cores on one silicon die. Unfortunately, a partitioning of the tasks near to the performance optimum is the challenge in this domain and often a show-stopper for the success story of multi- and many-core hardware. The missing feature of these architectures is runtime adaptivity of the underlying hardware, which offers to tailor the hardware to the application in order to meet the task mapping process coming from top-down development. Especially, this Meet-in-the- Middle solution offers the novel hardware and software approach of RAMPSoC, which is described in this paper.