Proceedings of
26^{th} IEEE International Parallel and Distributed Processing Symposium

IPDPS 2012 Advance Program Abstracts

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

Contents

IEEE International Parallel & Distributed Processing Symposium

IPDPS 2012

Session 1
Parallel Linear Algebra Algorithms I

A Predictive Model for Solving Small Linear Algebra Problems in GPU Registers

Michael J. Anderson, David Sheffield, Kurt Keutzer
UC Berkeley: Department of Electrical Engineering and Computer Sciences
Berkeley, CA USA
{mjanders,dsheffie,keutzer}@eecs.berkeley.edu

Abstract

We examine the problem of solving many thou- sands of small dense linear algebra factorizations simultane- ously on Graphics Processing Units (GPUs). We are interested in problems ranging from several hundred of rows and columns to 4×4 matrices. Problems of this size are common, especially in signal processing. However, they have received very little attention from current numerical linear algebra libraries for GPUs, which have thus far focused only on very large problems found in traditional supercomputing applications and benchmarks. To solve small problems efficiently we tailor our implementation to the GPUs inverted memory hierarchy and multi-level parallelism hierarchy. We provide a model of the GPU memory subsystem that can accurately predict and explain the performance of our approach across different problem sizes.

As a motivating example, we look at space-time adaptive radar processing, a real-time application that requires hundreds of independent QR factorizations of small complex matrices (e.g. 240×66). For realistic matrix sizes from a standard radar processing benchmark, our implementation on an NVIDIA Quadro 6000 GPU runs 2.8× to 25× faster than Intel's Math Kernel Library (MKL) on an Intel Core i7-2600. For the QR factorizations of 5,000 56×56 single-precision matrices, our approach runs 29× faster than MKL and 140× faster than the state-of-the-art linear algebra library for GPUs. In each of these cases we are using the GPU's hardware- accelerated division and square root functions that are accurate up to 22 mantissa bits.

A Parallel Tiled Solver for Dense Symmetric Indefinite Systems on Multicore Architectures

Marc Baboulin[*†] , Dulceneia Becker[‡] and Jack Dongarra[‡§¶]
[*]Inria Saclay - Île-de-France, F-91893 Orsay, France
[†]Université Paris-Sud, F-91405 Orsay, France
[‡]University of Tennessee, Knoxville, TN 37996-3450, USA
[§]Oak Ridge National Laboratory, Oak Ridge, TN, USA
[¶]University of Manchester, Manchester, United Kingdom
marc.baboulin@inria.fr, dbecker7@eecs.utk.edu, dongarra@eecs.utk.edu

Abstract

We describe an efficient and innovative parallel tiled algorithm for solving symmetric indefinite systems on multicore architectures. This solver avoids pivoting by using a multiplicative preconditioning based on symmetric ran- domization. This randomization prevents the communication overhead due to pivoting, is computationally inexpensive and requires very little storage. Following randomization, a tiled factorization is used that reduces synchronization by using static or dynamic scheduling. We compare Gflop/s performance of our solver with other types of factorizations on a current multicore machine and we provide tests on accuracy using LAPACK test cases.

A Comprehensive Study of Task Coalescing for Selecting Parallelism Granularity in a Two-Stage Bidiagonal Reduction

Azzam Haidar*, Hatem Ltaief†, Piotr Luszczek*, and Jack Dongarra*
*Innovative Computing Laboratory, University of Tennessee, Knoxville, TN 37996, USA
Email: haidar,luszczek,dongarra@eecs.utk.edu
†KAUST Supercomputing Laboratory, Thuwal, Saudi Arabia
Email: Hatem.Ltaief@kaust.edu.sa

Abstract

We present new high performance numerical kernels combined with advanced optimization techniques that significantly increase the performance of parallel bidiagonal reduction. Our approach is based on developing efficient fine- grained computational tasks as well as reducing overheads associated with their high-level scheduling during the so-called bulge chasing procedure that is an essential phase of a scalable bidiagonalization procedure. In essence, we coalesce multiple tasks in a way that reduces the time needed to switch execution context between the scheduler and useful computational tasks. At the same time, we maintain the crucial information about the tasks and their data dependencies between the coalescing groups. This is the necessary condition to preserve numerical correctness of the computation. We show our annihilation strategy based on multiple applications of single orthogonal reflectors. Despite non-trivial characteristics in computational complexity and memory access patterns, our optimization approach smoothly applies to the annihilation scenario. The coalescing positively influences another equally important as- pect of the bulge chasing stage: the memory reuse. For the tasks within the coalescing groups, the data is retained in high levels of the cache hierarchy and, as a consequence, opera- tions that are normally memory-bound increase their ratio of computation to off-chip communication and become compute- bound which renders them amenable to efficient execution on multicore architectures. The performance for the new two-stage bidiagonal reduction is staggering. Our implementation results in up to 50-fold and 12-fold improvement (- 130 Gflop/s) compared to the equivalent routines from LAPACK V3.2 and Intel MKL V10.3, respectively, on an eight socket hexa-core AMD Opteron multicore shared-memory system with a matrix size of 24000×24000. Last but not least, we provide a comprehensive study on the impact of the coalescing group size in terms of cache utilization and power consumption in the context of this new two-stage bidiagonal reduction.

Improving the Performance of Dynamical Simulations Via Multiple Right-Hand Sides

Xing Liu Edmond Chow
School of Computational Science and Engineering
College of Computing, Georgia Institute of Technology
Atlanta, Georgia, 30332, USA
xing.liu@gatech.edu, echow@cc.gatech.edu

Karthikeyan Vaidyanathan Mikhail Smelyanskiy
Parallel Computing Lab
Intel Corporation
Santa Clara, California, 95054, USA
{karthikeyan.vaidyanathan, mikhail.smelyanskiy}@intel.com

Abstract

This paper presents an algorithmic approach for improving the performance of many types of stochastic dynamical simulations. The approach is to redesign existing algorithms that use sparse matrix-vector products (SPMV) with single vectors to instead use a more efficient kernel, the generalized SPMV (GSPMV), which computes with multiple vectors simultaneously. In this paper, we show how to redesign a dynamical simulation to exploit GSPMV in way that is not initially obvious because only one vector is available at a time. We study the performance of GSPMV as a function of the number of vectors, and demonstrate the use of GSPMV in the Stokesian dynamics method for the simulation of the motion of macromolecules in the cell. Specifically, for our application, we find that with modern multicore Intel microprocessors in clusters of up to 64 nodes, we can typically multiply by 8 to 16 vectors in only twice the time required to multiply by a single vector. After redesigning the Stokesian dynamics algorithm to exploit GSPMV, we measure a 30 percent speedup in performance in single-node, data parallel simulations.

Session 2
Bioinformatics and Performance Modeling

High-Performance Interaction-Based Simulation of Gut Immunopathologies with ENteric Immunity SImulator (ENISI)

Keith Bisset*, Md. Maksudul Alam*, Josep Bassaganya-Riera*, Adria Carbo*, Stephen Eubank*, Raquel Hontecillas*, Stefan Hoops*, Yongguo Mei*, Katherine Wendelsdorf*, Dawen Xie*, Jae-Seung Yeom*†, and Madhav Marathe*†
* Virginia Bioinformatics Institute
† Department of Computer Science
Virginia Tech, Blacksburg VA 24061
Email: {kbisset,maksud,jbassaga,acarbo,seubank,rmagarzo,shoops, ymei,dawenx,wkath83,jyeom,mmarathe}@vbi.vt.edu

Abstract

Here we present the ENteric Immunity Simulator (ENISI), a modeling system for the inflammatory and regulatory immune pathways triggered by microbe-immune cell interactions in the gut. With ENISI, immunologists and infectious disease experts can test and generate hypotheses for enteric disease pathology and propose interventions through experimental infection of an in silico gut. ENISI is an agent based simulator, in which individual cells move through the simulated tissues, and engage in context-dependent interactions with the other cells with which they are in contact. The scale of ENISI is unprecedented in this domain, with the ability to simulate 107 cells for 250 simulated days on 576 cores in one and a half hours, with the potential to scale to even larger hardware and problem sizes. In this paper we describe the ENISI simulator for modeling mucosal immune responses to gastrointestinal pathogens. We then demonstrate the utility of ENISI by recreating an experimental infection of a mouse with Helicobacter pylori 26695. The results identify specific processes by which bacterial virulence factors do and do not contribute to pathogenesis associated with H. pylori strain 26695. These modeling results inform general intervention strategies by indicating immunomodulatory mechanisms such as those used in inflammatory bowel disease may be more appropriate therapeutically than directly targeting specific microbial populations through vaccination or by using antimicrobials.

A Parallel Algorithm for Spectrum-based Short Read Error Correction

Ankit R Shah[1] , Sriram Chockalingam[1] , Srinivas Aluru[1,2]
[1]Dept. of Computer Science and Engineering
Indian Institute of Technology Bombay, Mumbai, India
[2]Dept. of Electrical and Computer Engineering
Iowa State University, Ames, IA

Abstract

Correcting sequence errors in highthroughput DNA sequencing by taking advantage of redundant sampling and low error rates is often an important first step in applications of this technology. Consequently, a number of error correction methods have been developed in the recent years. Due to an order of magnitude throughput gain per year, some of these technologies are now generating upwards of a billion reads per run. In this paper, we present an algorithm for parallelizing error correction methods that are based on frequency spectrum of kmers observed in input reads. Based on this, we present a parallelization of Reptile, a recently introduced error correction method that employs frequency spectrum of two different lengths, one for identifying correction possibilities and another for providing contextual information. Our method is well suited for distributed memory parallel computers and clusters. Experimental results indicate the method achieves near linear speedup and provides the ability to scale to larger data sets than previously demonstrated.

Enhancing the scalability of consistency-based progressive multiple sequences alignment applications

M. Orobitg, F. Cores, F. Guirado
Dept. of Computer Science
Universitat de Lleida
Lleida, Spain
Email: {orobitg, fcores, f.guirado}@diei.udl.cat

C. Kemena, C. Notredame
Centre For Genomic Regulation
Universitat Pompeu Fabra
Barcelona, Spain
Email: {carsten.kemena, cedric.notredame}@crg.cat

A. Ripoll
Computer Architecture and Operating System
Universitat Autonoma de Barcelona
Bellaterra, Spain
Email: {ana.ripoll}@uab.es

Abstract

Multiple Sequence Alignment (MSA) is an extremely powerful tool for important biological applications, such as phylogenetic analysis, identification of conserved motifs and domains and structure prediction. In this paper we propose a new approach to reduce the computational requirements of T-Coffee, a memory demanding MSA tool that uses a consistency-based scheme to produce more accurate alignments. Our goal is to minimize the memory constraints in order to increase the performance and scalability of the application. The experimental results show that our approach is able to reduce the memory consumption and increase both the time performance and the number and the length of sequences that the method can align. In summary, it is able to reduce the memory requirements by between 72% and 58% depending on the optimization level, improving the alignment scalability as a whole. Also, the library reduction yields a further improvement: the alignment execution time can be reduced by up to 92%. These results are obtained without a significant impact on the final alignment quality, that declines by less than 3%.

An Accurate GPU Performance Model for Effective Control Flow Divergence Optimization

Zheng Cui, Yun Liang, Kyle Rupnow
Advanced Digital Sciences Center, Illinois at Singapore
{cui.zheng, eric.liang, k.rupnow}@adsc.com.sg
Deming Chen
ECE Dept., University of Illinois at Urbana-Champaign
dchen@illinois.edu

Abstract

Graphics processing units (GPUs) are increasingly critical for general-purpose parallel processing performance. GPU hardware is composed of many streaming multi- processors, each of which employs the single-instruction multiple-data (SIMD) execution style. This massively parallel architecture allows GPUs to execute tens of thousands of threads in parallel. Thus, GPU architectures efficiently execute heavily data-parallel applications. However, due to this SIMD execution style, resource utilization and thus overall performance can be significantly affected if computation threads must take diverging control paths. Control flow divergence in GPUs is a well-known problem: prior approaches have attempted to reduce control flow divergence through code transformations, memory access indirection, and input data reorganization. However, as we will demonstrate, the utility of these transformations is seriously affected by the lack of a guiding metric that properly estimates how control flow divergence affects application performance. In this paper, we introduce a metric that simply and accurately estimates performance of computation-bound GPU kernels with control flow divergence, and use the metric as a value function for thread re-grouping algorithms. We measure the performance on NVIDIA GTS250 GPU. For the tested set of applications, our experiments demonstrate that the proposed metric correlates well with actual GPU application performance. Through thread re-grouping guided by our metric, control flow divergence optimization can improve application performance by up to 3.19X.

Session 3

Dynamic Pipeline and Transactional Memory Optimizations

SEL-TM: Selective Eager-Lazy Management for Improved Concurrency in Transactional Memory

Lihang Zhao, Woojin Choi, and Jeff Draper
Information Sciences Institute
University of Southern California
Marina del Rey, CA 90292
Email: {lihangzh, woojinch, draper}@isi.edu

Abstract

Hardware Transactional Memory (HTM) systems implement version management and conflict detection in hardware to guarantee that each transaction is atomic and executes in isolation. In general, HTM implementations fall into two categories, namely, eager systems and lazy systems. Lazy systems have been shown to exploit more concurrency from potentially conflicting transactions. However, lazy systems manage a transaction's entire write set lazily, which gives rise to two main disadvantages: (a) a complex cache protocol and implementation are required to maintain the speculative modifications, and; (b) the latency of committing the entire write set often leads to severe performance degradation of the whole system. It is observed in a wide range of workloads that more than 55% of the transaction aborts are due to conflicts on only three memory blocks. Thus we argue that an eager HTM system can achieve the same level of concurrency as lazy systems by managing only a small portion of a transaction's write set lazily. In this paper, we present Selective-Eager-Lazy HTM (SEL-TM), a new HTM implementation to adopt complementary version management schemes within a transaction whose write set is divided into eagerly- and lazily-managed memory addresses at runtime. An intelligent hardware scheme is designed to select the memory addresses for lazy management as well as determining whether each dynamic instance of a transaction benefits from hybrid management. Experimental results using the STAMP benchmarks show that, on average, SEL-TM improves performance by 14% over an eager system and 22% over a lazy system. The speedup demonstrates that our design is capable of harvesting the concurrency benefit of lazy version management while avoiding some of the performance penalties in lazy HTMs.

Robust SIMD: Dynamically Adapted SIMD Width and Multi-Threading Depth

Jiayuan Meng
Leadership Computing Facility Division
Argonne National Laboratory
Argonne, Illinois
jmeng@alcf.anl.gov

Jeremy W. Sheaffer
Department of Computer Science
University of Virginia
Charlottesville, Virginia
jws9c@cs.virginia.edu

Kevin Skadron
Department of Computer Science
University of Virginia
Charlottesville, Virginia
skadron@cs.virginia.edu

Abstract

Architectures that aggressively exploit SIMD often have many datapaths execute in lockstep and use multi-threading to hide latency. They can yield high throughput in terms of area- and energy-efficiency for many data- parallel applications. To balance productivity and performance, many recent SIMD organizations incorporate implicit cache hierarchies. Exaples of such architectures include Intel's MIC, AMD's Fusion, and NVIDIA's Fermi. However, unlike software-managed streaming memories used in conventional graphics processors (GPUs), hardware-managed caches are more disruptive to SIMD execution; therefore the interaction between implicit caching and aggressive SIMD execution may no longer follow the conventional wisdom gained from stream- ing memories. We show that due to more frequent memory latency divergence, lower latency in non-L1 data accesses, and relatively unpredictable L1 contention, cache hierarchies favor different SIMD widths and multi-threading depths than streaming memories. In fact, because the above effects are sub- ject to runtime dynamics, a fixed combination of SIMD width and multi-threading depth no longer works ubiquitously across diverse applications or when cache capacities are reduced due to pollution or power saving. To address the above issues and reduce design risks, this paper proposes Robust SIMD, which provides wide SIMD and then dynamically adjusts SIMD width and multi-threading depth according to performance feedback. Robust SIMD can trade wider SIMD for deeper multi-threading by splitting a wider SIMD group into multiple narrower SIMD groups. Compared to the performance generated by running every benchmark on its individually preferred SIMD organization, the same Robust SIMD organization performs similarly– sometimes even better due to phase adaptation–and outperforms the best fixed SIMD organization by 17%. When D-cache capacity is reduced due to runtime disruptiveness, Robust SIMD offers graceful performance degradation; with 25% polluted cache lines in a 32 KB D-cache, Robust SIMD performs 1.4× better compared to a conventional SIMD architecture.

Dynamic Operands Insertion for VLIW Architecture with a Reduced Bit-width Instruction Set

Jongwon Lee*, Jonghee M. Youn† , Jihoon Lee* , Minwook Ahn‡ , Yunheung Paek*
* School of EECS,
Seoul Nat'l Univ., Korea,
{jwlee,jhlee,ypaek}@sor.snu.ac.kr,
†Department of CSE,
Gangneung-Wonju Nat'l Univ., Korea
jhyoun@gwnu.ac.kr
‡Samsung Advanced Institute of Technology,
Korea
minwook.ahn@samsung.com

Abstract

Performance, code size and power consumption are all primary concern in embedded systems. To this effect, VLIW architecture has proven to be useful for embedded applications with abundant instruction level parallelism. But due to the long instruction bus width it often consumes more power and memory space than necessary. One way to lessen this problem is to adopt a reduced bit-width instruction set architecture (ISA) that has a narrower instruction word length. This facilitates a more efficient hardware implementation in terms of area and power by decreasing bus-bandwidth requirements and the power dissipation associated with instruction fetches. Also earlier studies reported that it helps to reduce the code size considerably. In practice, however, it is impossible to convert a given ISA fully into an equivalent reduced bit-width one because the narrow instruction word, due to bit-width restrictions, can encode only a small subset of normal instructions in the original ISA. Consequently, existing processors provide narrow instructions in very limited cases along with severe restrictions on register accessibility. The objective of this work is to explore the possibility of complete conversion, as a case study, of an existing 32-bit VLIW ISA into a 16-bit one that supports effectively all 32-bit instructions. To this objective, we attempt to circumvent the bit-width restrictions by dynamically extending the effective instruction word length of the converted 16-bit operations. At compile time when a 32-bit operation is converted to a 16-bit word format, we compute how many bits are additionally needed to represent the whole 32-bit operation and store the bits separately in the VLIW code. Then at run time, these bits are retrieved on demand and inserted to a proper 16-bit operation to reconstruct the original 32-bit representation. According to our experiment, the code size becomes significantly smaller after the conversion to 16-bit VLIW code. Also at a slight run time cost, the machine with the 16-bit ISA consumes much less energy than the original machine.

SUV: A Novel Single-Update Version-Management Scheme for Hardware Transactional Memory Systems

Zhichao Yan* , Hong Jiang† , Dan Feng*, Lei Tian*† and Yujuan Tan*
* School of Computer Science
Wuhan National Laboratory for Optoelectronics
Huazhong University of Science and Technology, Wuhan, China
Email: zhichao yan@smail.hust.edu.cn, {dfeng,ltian}@mail.hust.edu.cn, tanyujuan@gmail.com
† Department of Computer Science & Engineering
University of Nebraska-Lincoln, Lincoln, USA
Email: jiang@cse.unl.edu

Abstract

In order to maintain the transactional semantics, Transactional Memory (TM) must guarantee isolated read and write operations in each transaction, meaning that it must spend a non-negligible and potentially significant amount of time on keeping track of the transactional modifications in its undo or redo log and switching to the proper version at the end of each transaction. Existing TMs failed to minimize the overheads incurred by these operations that are poised to impose more significant TM overheads in current and future many-core CMPs. A direct consequence of this is that extra and different data movements are needed to manage these modifications depending on commit or abort. To address this problem, we propose a novel Single-Update Version- management (SUV) scheme to redirect each transactional store operation to another memory address, track the mapping information between the original and redirected addresses, and switch to the proper version of data upon the transaction's commit or abort. There is only one data update (movement) in our SUV regardless of commit or abort, thus significantly reducing the TM overheads while allowing it to exploit more thread parallelism. We use SUV to replace version-management schemes in some existing hardware TMs to assess SUV's performance advantages. Our extensive execution-driven experiments show that SUV-TM consistently outperforms the state-of-the- art HTM schemes LogTM-SE, FasTM and DynTM under the STAMP benchmark suite. Moreover, we use CACTI to estimate the hardware overheads of SUV and find it is feasible in hardware implementation.

Session 4
Software Scheduling

Heterogeneous Task Scheduling for Accelerated OpenMP

Thomas R. W. Scogland* Barry Rountree† Wu-chun Feng* Bronis R. de Supinski†
*Department of Computer Science, Virginia Tech, Blacksburg, VA 24060 USA
Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, Livermore, CA 94551 USA
tom.scogland@vt.edu rountree@llnl.gov feng@vt.edu bronis@llnl.gov

Abstract

Heterogeneous systems with CPUs and computa- tional accelerators such as GPUs, FPGAs or the upcoming Intel MIC are becoming mainstream. In these systems, peak performance includes the performance of not just the CPUs but also all available accelerators. In spite of this fact, the majority of programming models for heterogeneous computing focus on only one of these. With the development of Accelerated OpenMP for GPUs, both from PGI and Cray, we have a clear path to extend traditional OpenMP applications incrementally to use GPUs. The extensions are geared toward switching from CPU parallelism to GPU parallelism. However they do not preserve the former while adding the latter. Thus computational potential is wasted since either the CPU cores or the GPU cores are left idle. Our goal is to create a runtime system that can intelligently divide an accelerated OpenMP region across all available resources automatically. This paper presents our proof-of-concept runtime system for dynamic task scheduling across CPUs and GPUs. Further, we motivate the addition of this system into the proposed OpenMP for Accelerators standard. Finally, we show that this option can produce as much as a two-fold performance improvement over using either the CPU or GPU alone.

A Source-aware Interrupt Scheduling for Modern Parallel I/O Systems

Hongbo Zou , Xian-He Sun , Siyuan Ma , and Xi Duan
Department of Computer Science
Illinois Institute of Technology
Chicago, IL, USA 60616
Email: {zouhongbao@gmail.com, sun@iit.edu, sma9@iit.edu, xduan@iit.edu}

Abstract

Recent technological advances are putting increased pressure on CPU scheduling. On one hand, processors have more cores. On the other hand, I/O systems have become more complex. Intensive research has been conducted on multi/many-core scheduling; however, most of the studies follow the conventional approach and focus on the utilization and load balance of the cores. In this study, we focus on increasing data locality by bringing source information from I/O into the core interrupt scheduling process. The premise is to group interrupts associated for the same I/O request together on the same core, and prove that data locality is more important than core utilization for many applications. Based on this idea, a source-aware affinity interrupt-scheduling scheme is introduced and a prototype system, SAIs, is implemented. Experiment results show that SAIs is feasible and promising; bandwidth shows a 23.57% improvement in a 3-Gigabit NIC environment and in the optimal case without the NIC bottleneck, the bandwidth improvement increases to 53.23%.

ExPERT: Pareto-Efficient Task Replication on Grids and a Cloud

Orna Agmon Ben-Yehuda*, Assaf Schuster* , Artyom Sharov* , Mark Silberstein* , and Alexandru Iosup‡
* Technion-Israel Institute of Technology, Haifa, 32000, Israel,
Email: {ladypine,assaf,sharov,marks}@cs.technion.ac.il
‡ Parallel and Distributed Systems Group, TU Delft, the Netherlands, Email: A.Iosup@tudelft.n

Abstract

Many scientists perform extensive computations by executing large bags of similar tasks (BoTs) in mixtures of computational environments, such as grids and clouds. Although the reliability and cost may vary considerably across these environments, no tool exists to assist scientists in the selection of environments that can both fulfill deadlines and fit budgets. To address this situation, we introduce the ExPERT BoT scheduling framework. Our framework systematically selects from a large search space the Pareto-efficient scheduling strategies, that is, the strategies that deliver the best results for both makespan and cost. ExPERT chooses from them the best strategy according to a general, user-specified utility function. Through simulations and experiments in real production environments, we demonstrate that ExPERT can substantially reduce both makespan and cost in comparison to common scheduling strategies. For bioinformatics BoTs executed in a real mixed *grid + cloud* environment, we show how the scheduling strategy selected by ExPERT reduces both makespan and cost by 30%-70%, in comparison to commonly- used scheduling strategies.

Scheduling Closed-Nested Transactions in Distributed Transactional Memory

Junwhan Kim
ECE Dept., Virginia Tech
Blacksburg, VA, 24061
Email: junwhan@vt.edu

Binoy Ravindran
ECE Dept., Virginia Tech
Blacksburg, VA, 24061
Email: binoy@vt.edu

Abstract

Distributed software transactional memory (D-STM) is an emerging, alternative concurrency control model for distributed systems that promises to alleviate the difficulties of lock-based distributed synchronization–e.g., distributed deadlocks, livelocks, and lock convoying. We consider Herlihy and Sun's dataflow D-STM model, where objects are migrated to invoking transactions, and the closed nesting model of managing inner (distributed) transactions. We present a transactional scheduler called, reactive transactional scheduler (or RTS) to boost the throughput of closed-nested transactions. RTS determines whether a conflicting parent transaction must be aborted or enqueued according to the level of contention. If a transaction is enqueued, its nested inner transactions do not have to retrieve objects again, resulting in reduced com- munication delays. Our implementation of RTS in the HyFlow D-STM framework and experimental evaluations reveal that RTS improves throughput over D-STM without RTS, by as much as 88%.

Session 5
Multicore Algorithms

Power-aware Manhattan routing on chip multiprocessors

Anne Benoit[1], Rami Melhem[2] , Paul Renaud-Goud[1] and Yves Robert[1,3]
1. École Normale Suprieure de Lyon, France, {Anne.Benoit | Paul.Renaud-Goud | Yves.Robert}@ens-lyon.fr
2. University of Pittsburgh, PA, USA, melhem@cs.pitt.edu
3. University of Tennessee Knoxville, TN, USA

Abstract

We investigate the routing of communications in chip multiprocessors (CMPs). The goal is to find a valid routing in the sense that the amount of data routed between two neighboring cores does not exceed the maximum link bandwidth while the power dissipated by communications is minimized. Our position is at the system level: we assume that several applications, described as task graphs, are executed on a CMP, and each task is already mapped to a core. Therefore, we consider a set of communications that have to be routed between the cores of the CMP. We consider a classical model, where the power consumed by a communication link is the sum of a static part and a dynamic part, with the dynamic part depending on the frequency of the link. This frequency is scalable and it is proportional to the throughput of the link. The most natural and widely used algorithm to handle all these communications is XY routing: for each communication, data is first forwarded horizontally, and then vertically, from source to destination. However, if it is allowed to use all Manhattan paths between the source and the destination, the consumed power can be reduced dramatically. Moreover, some solutions may be found while none existed with the XY routing. In this paper, we compare XY routing and Manhattan routing, both from a theoretical and from a practical point of view. We consider two variants of Manhattan routing: in single-path routing, only one path can be used for each communication, while multi-paths routing allows to split a communication between different routes. We establish the NP-completeness of the problem of finding a Manhattan routing that minimizes the dissipated power, we exhibit the minimum upper bound of the ratio power consumed by an XY routing over power consumed by a Manhattan routing, and finally we perform simulations to assess the performance of Manhattan routing heuristics that we designed.

Efficient Resource Oblivious Algorithms for Multicores with False Sharing

Richard Cole
Computer Science Dept.
Courant Institute of Mathematical Sciences, NYU
New York, NY 10012, USA
Email: cole@cs.nyu.edu

Vijaya Ramachandran
Dept. of Computer Science
University of Texas at Austin
Austin, TX 78712, USA
Email: vlr@cs.utexas.edu

Abstract

We consider algorithms for a multicore environ- ment in which each core has its own private cache and false sharing can occur. False sharing happens when two or more processors access the same block (i.e., cache-line) in parallel, and at least one processor writes into a location in the block. False sharing causes different processors to have inconsistent views of the data in the block, and many of the methods currently used to resolve these inconsistencies can cause large delays. We analyze the cost of false sharing both for variables stored on the execution stacks of the parallel tasks and for output variables. Our main technical contribution is to establish a low cost for this overhead for the class of multithreaded block- resilient HBP (Hierarchical Balanced Parallel) computations. Using this and other techniques, we develop block-resilient HBP algorithms with low false sharing costs for several fundamental problems including scans, matrix multiplication, FFT, sorting, and hybrid block-resilient HBP algorithms for list ranking and graph connected components. Most of these algorithms are derived from known multicore algorithms, but are further refined to achieve a low false sharing overhead. Our algorithms make no mention of machine parameters, and our analysis of the false sharing overhead is mostly in terms of the the number of tasks generated in parallel during the computation, and thus applies to a variety of schedulers.

Competitive Cache Replacement Strategies for Shared Cache Environments

Anil Kumar Katti
Department of Computer Science
University of Texas at Austin
Austin, TX 78712, USA
Email: akatti@cs.utexas.edu

Vijaya Ramachandran
Department of Computer Science
University of Texas at Austin
Austin, TX 78712, USA
Email: vlr@cs.utexas.edu

Abstract

We investigate cache replacement algorithms (CRAs) at a cache shared by several processes under different multicore environments. For a single shared cache, our main result is the first CRA, GLOBAL - MAXIMA, for fixed interleaving under shared full knowledge [1], where any data can be accessed by any process, and each process has full knowledge about its future request sequence. We establish that GLOBAL - MAXIMA has competitive ratio within a constant factor of optimal. This answers the major open question in [1]. We also present RR - PROC - MARK, a CRA for the disjoint full knowledge case, which is very simple and efficient, and achieves a better competitive ratio than the algorithms in [2], [1]; it is in fact optimal except when the number of processes sharing the cache is small. We then consider a cache hierarchy, both for a single process and when shared by several processes. We present CRAs for three types of caching models commonly used at a higher- level cache: inclusive, exclusive, and partially-inclusive, and we establish that several of our CRAs have optimal competitive ratio. Our results for a cache hierarchy are new even in the traditional no knowledge case and even for a single process.

A novel sorting algorithm for many-core architectures based on adaptive bitonic sort

Hagen Peters, Ole Schulz-Hildebrandt, Norbert Luttenberger
Department of Computer Science
CAU Kiel
Kiel, Germany
{hap,osh,nl}@informatik.uni-kiel.de

Abstract

Adaptive bitonic sort is a well known merge-based parallel sorting algorithm. It achieves optimal complexity using a complex tree-like data structure called a bitonic tree. Due to this, using adaptive bitonic sort together with other algorithms usually implies converting bitonic trees to arrays and vice versa. This makes adaptive bitonic sort inappropriate in the context of hybrid sorting algorithms where frequent switches between algorithms are performed. In this article we present a novel optimal sorting algorithm that is based on an approach similar to adaptive bitonic sort. Our approach does not use bitonic trees but uses the input array together with some additional information. Using this approach it is trivial to switch between adaptive bitonic sort and other algorithms. We present an implementation of a hybrid algorithm for GPUs based on bitonic sort and our novel algorithm. This implementation turns out to be the fastest comparison-based sorting algorithm for GPUs found in literature.

Session 6

Scheduling and Load Balancing Algorithms I

Optimizing Busy Time on Parallel Machines

George B. Mertzios*, Mordechai Shalom† , Ariella Voloshin‡ , Prudence W.H. Wong§ and Shmuel Zaks‡
School of Engineering and Computing Sciences, Durham University, Durham, UK
Email: george.mertzios@durham.ac.uk
† TelHai College, Upper Galilee, 12210, ISRAEL
Email: cmshalom@telhai.ac.il
‡ Department of Computer Science, Technion, Haifa, ISRAEL
Email: [variella,zaks]@cs.technion.ac.il
§ Department of Computer Science, University of Liverpool, Liverpool, UK
Email: pwong@liverpool.ac.uk

Abstract

We consider the following fundamental scheduling problem in which the input consists of n jobs to be scheduled on a set of identical machines of bounded capacity g (which is the maximal number of jobs that can be processed simultaneously by a single machine). Each job is associated with a start time and a completion time; it is supposed to be processed from the start time to the completion time (and in one of our extensions it has to be scheduled also in a continuous number of days; this corresponds to a two-dimensional version of the problem). We consider two versions of the problem. In the scheduling minimization version the goal is to minimize the total busy time of machines used to schedule all jobs. In the resource allocation maximization version the goal is to maximize the number of jobs that are scheduled for processing under a budget constraint given in terms of busy time. This is the first study of the maximization version of the problem. The minimization problem is known to be NP-Hard, thus the maximization problem is also NP-Hard. We consider various special cases, identify cases where an optimal solution can be computed in polynomial time, and mainly provide constant factor approximation algorithms for both minimization and maximization problems. Some of our results improve upon the best known results for this job scheduling problem. Our study has applications in power consumption, cloud computing and optimizing switching cost of optical networks.

WATS: Workload-Aware Task Scheduling in Asymmetric Multi-core Architectures

Quan Chen* , Yawen Chen† , Zhiyi Huang† , Minyi Guo*
* Shanghai Key Laboratory of Scalable Computing and Systems,
Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
chen-quan@sjtu.edu.cn, guo-my@cs.sjtu.edu.cn
† Department of Computer Science, University of Otago, New Zealand
yawen@cs.otago.ac.nz, hzy@cs.otago.ac.nz

Abstract

Asymmetric Multi-Core (AMC) architectures have shown high performance as well as power efficiency. However, current parallel programming environments do not perform well on AMC due to their assumption that all cores are symmetric and provide equal performance. Their random task scheduling policies, such as task-stealing, can result in unbalanced workloads in AMC and severely degrade the performance of parallel applications. To balance the workloads of parallel applications in AMC, this paper proposes a Workload- Aware Task Scheduling (WATS) scheme that adopts history- based task allocation and preference-based task stealing. The history-based task allocation is based on a near-optimal, static task allocation using the historical statistics collected during the execution of a parallel application. The preference-based task stealing, which steals tasks based on a preference list, can dynamically adjust the workloads in AMC if the task allocation is less optimal due to approximation in the history- based task allocation. Experimental results show that WATS can improve the performance of CPU-bound applications up to 82.7% compared with the random task scheduling policies.

Parametric Utilization Bounds for Fixed-Priority Multiprocessor Scheduling

Nan Guan[1,2], Martin Stigge[1], Wang Yi[1,2] and Ge Yu[2]
[1] Uppsala University, Sweden
[2] Northeastern University, China

Abstract

Future embedded real-time systems will be deployed on multi-core processors to meet the dramatically increasing high-performance and low-power requirements. This trend appeals to generalize established results on uniprocessor scheduling, particularly the various utilization bounds for schedulability test used in system design, to the multiprocessor setting. Recently, this has been achieved for the famous Liu and Layland utilization bound by applying novel task splitting techniques. However, parametric utilization bounds that can guarantee higher utilizations (up to 100%) for common classes of systems are not yet known to be generalizable to multiprocessors as well. In this paper, we solve this problem for most parametric utilization bounds by proposing new task partitioning algorithms based on exact response time analysis. In addition to the worst-case guarantees, as the exact response time analysis is used for task partitioning, our algorithms significantly improve average-case utilization over previous work.

Minimizing Weighted Mean Completion Time for Malleable Tasks Scheduling

Olivier Beaumont, Nicolas Bonichon, Lionel Eyraud-Dubois
INRIA Bordeaux Sud-Ouest F-78000 and LaBRI - Univ of Bordeaux F-33400 and LaBRI - CNRS F-33400
Loris Marchal
CNRS and ENS Lyon

Abstract

Malleable tasks are jobs that can be scheduled with preemptions on a varying number of resources. We focus on the special case of work-preserving malleable tasks, for which the area of the allocated resources does not depend on the allocation and is equal to the sequential processing time. Moreover, we assume that the number of resources allocated to each task at each time instant is limited. We consider both the clairvoyant and non-clairvoyant cases, and we focus on minimizing the weighted sum of completion times. In the weighted non-clairvoyant case, we propose an approximation algorithm whose ratio (2) is the same as in the unweighted non-clairvoyant case. In the clairvoyant case, we provide a normal form for the schedule of such malleable tasks, and prove that any valid schedule can be turned into this normal form, based only on the completion times of the tasks. We show that in these normal form schedules, the number of preemptions per task is bounded by 3 on average. At last, we analyze the performance of list schedules, and prove that optimal schedules are list schedules for a special case of homogeneous instances. We conjecture that there exists an optimal list schedule for all instances, which would greatly simplify the study of this problem. Finally, we explore the complexity of the problem restricted to homogeneous instances, which is still open despite its very simple expression.

Session 7
Scientific Applications

Load Balancing of Dynamical Nucleation Theory Monte Carlo Simulations Through Resource Sharing Barriers

Humayun Arafat, P. Sadayappan
Dept. Comp. Sci. and Eng.
The Ohio State University
arafatm,saday@cse.ohio-state.edu

James Dinan
Math. and Comp. Sci. Div.
Argonne National Lab.
dinan@mcs.anl.gov

Sriram Krishnamoorthy
Comp. Sci. and Math. Div.
Pacific Northwest National Lab.
sriram@pnl.gov

Theresa L. Windus
Dept. of Chemistry
Iowa State University
twindus@iastate.edu

Abstract

The dynamical nucleation theory Monte Carlo (DNTMC) application from the NWChem computational chemistry suite utilizes a Markov chain Monte Carlo, two-level parallel structure, with periodic synchronization points that assemble the results of independent finer-grained calculations. Like many such applications, the existing code employs a static partitioning of processes into groups and assigns each group a piece of the finer-grained parallel calculation. A significant cause of performance degradation is load imbalance among groups since the time requirements of the inner-parallel calculation varies widely with the input problem and as a result of the Monte Carlo simulation. We present a novel approach to load balancing such calculations with minimal changes to the application. We introduce the concept of a resource sharing barrier (RSB) – a barrier that allows process groups waiting on other processes' work to actively contribute to their completion. The RSB load balancing technique is applied to the production DNTMC application code, resulting in a small code change of 200 lines and a reduction in execution time of up to 37%.

Highly Efficient Performance Portable Tracking of Evolving Surfaces

Wei Yu
Citadel Investment Group
Chicago, USA
Email: yuwei.emily@gmail.com

Franz Franchetti, James C. Hoe
ECE, Carnegie Mellon University
Pittsburgh, USA
Email: {franzf, jhoe}@ece.cmu.edu

Tsuhan Chen
ECE, Cornell University
Ithaca, USA
Email: tsuhan@ece.cornell.edu

Abstract

In this paper we present a framework to obtain highly efficient implementations for the narrow band level set method on commercial off-the-shelf (COTS) multicore CPU systems with a cache-based memory hierarchy such as Intel Xeon and Atom processors. The narrow-band level set algorithm tracks wave-fronts in discretized volumes (for instance, explosion shock waves), and is computationally very demanding. At the core of our optimization framework is a novel projection-based approach to enhance data locality and enable reuse for sparse surfaces in dense discretized volumes. The method reduces stencil operations on sparse and changing sets of pixels belonging to an evolving surface into dense stencil operations on meta-pixels in a lower-dimensional projection of the pixel space. These meta-pixels are then amenable to standard techniques like time tiling. However, the complexity introduced by ever-changing meta-pixels requires us to revisit and adapt all other necessary optimizations. We apply adapted versions of SIMDization, multi-threading, DAG scheduling for basic tiles, and specialization through code generation to extract maximum performance. The system is implemented as highly parameterized code skeleton that is auto-tuned and uses program generation. We evaluated our framework on a dual-socket 2.8 GHz Xeon 5560 and a 1.6 GHz Atom N270. Our single-core performance reaches 26%–35% of the machine peak on the Xeon, and 12%– 20% on the Atom across a range of image sizes. We see up to 6.5x speedup on 8 cores of the dual-socket Xeon. For cache- resident sizes our code outperforms the best available third- party code (C pre-compiled into a DLL) by about 10x and for the largest out-of-cache sizes the speedup approaches around 200x. Experiments fully explain the high speedup numbers.

Advancing Large Scale Many-Body QMC Simulations on GPU Accelerated Multicore Systems

Andres Tomas*, Chia-Chen Chang† , Richard Scalettar† and Zhaojun Bai*
* Department of Computer Science, University of California, Davis, CA 95616, USA
{andres,bai}@cs.ucdavis.edu
† Department of Physics, University of California, Davis, CA 95616, USA
cxc639@gmail.com, scalettar@physics.ucdavis.edu

Abstract

The Determinant Quantum Monte Carlo (DQMC) method is one of the most powerful approaches for understanding properties of an important class of materials with strongly interacting electrons, including magnets and superconductors. It treats these interactions exactly, but the solution of a system of N electrons must be extrapolated to bulk values. Currently N ≈ 500 is state-of-the-art. Increasing N is required before DQMC can be used to model newly synthesized materials like functional multilayers. DQMC requires millions of linear algebra computations of order N matrices and scales as N^3. DQMC cannot exploit parallel distributed memory computers efficiently due to limited scalability with the small matrix sizes and stringent procedures for numerical stability. Today, the combination of multisocket multicore processors and GPUs provides widely available platforms with new opportunities for DQMC parallelization. The kernel of DQMC, the calculation of the Green's function, involves long products of matrices. For numerical stability, these products must be computed using graded decompositions generated by the QR decomposition with column pivoting. The high communication overhead of pivoting limits parallel efficiency. In this paper, we propose a novel approach that exploits the progressive graded structure to reduce the communication costs of pivoting. We show that this method preserves the same numerical stability and achieves 70% performance of highly optimized DGEMM on a two-socket six-core Intel processor. We have integrated this new method and other parallelization techniques into QUEST, a modern DQMC simulation package. Using 36 hours on this Intel processor, we are able to compute accurately the magnetic properties and Fermi surface of a system of N = 1024 electrons. This simulation is almost an order of magnitude more difficult than N ≈ 500, owing to the N^3 scaling. This increase in system size will allow, for the first time, the computation of the magnetic and transport properties of layered materials with DQMC. In addition, we show preliminary results which further accelerate DQMC simulations by using GPU processors.

Reducing Data Movement Costs Scalable Seismic Imaging on Blue Gene

Michael Perrone, Lurng-Kuo Liu, Ligang Lu,
Karen Magerlein, Changhoan Kim
Computational Sciences Center
IBM Research
Yorktown Heights, NY, USA
{mpp,lkliu,lul,kmager,kimchang}@us.ibm.com

Irina Fedulova, Artyom Semenikhin
Russia Systems and Technology Lab
IBM Systems & Technology Group
Moscow, Russia
{i.fedulova,artyom}@ru.ibm.com

Abstract

We present an optimized Blue Gene/P implementation of Reverse Time Migration, a seismic imaging algorithm widely used in the petroleum industry today. Our implementation is novel in that it uses large communication bandwidth and low latency to convert an embarrassingly parallel problem into one that can be efficiently solved using massive domain partitioning. The success of this seemingly counterintuitive approach is the result of several key aspects of the imaging problem, including very regular and local communication patterns, balanced compute and communication requirements, scratch data handling, multiple-pass approaches, and most importantly, the fact that partitioning the problem allows each sub-problem to fit in cache, dramatically increasing locality and bandwidth and reducing latency. This approach can be easily extended to next-generation imaging algorithms currently being developed. In this paper we present details of our implementation, including application-scaling results on Blue Gene/P.

Session 8

MPI Debugging and Performance Optimization

Opportunistic Data-driven Execution of Parallel Programs for Efficient I/O Services

Xuechen Zhang
ECE Department
Wayne State University
Detroit, MI, 48202, USA
xczhang@wayne.edu

Kei Davis
CCS Division
Los Alamos National Laboratory
Los Alamos, NM 87545, USA
kei.davis@lanl.gov

Song Jiang
ECE Department
Wayne State University
Detroit, MI, 48202, USA
sjiang@eng.wayne.edu

Abstract

A parallel system relies on both process scheduling and I/O scheduling for efficient use of resources, and a program's performance hinges on the resource on which it is bottlenecked. Existing process schedulers and I/O schedulers are independent. However, when the bottleneck is I/O, there is an opportunity to alleviate it via cooperation between the I/O and process schedulers: the service efficiency of I/O requests can be highly dependent on their issuance order, which in turn is heavily influenced by process scheduling. We propose a data-driven program execution mode in which process scheduling and request issuance are coordinated to facilitate effective I/O scheduling for high disk efficiency. Our implementation, DualPar, uses process suspension and resumption, as well as pre-execution and prefetching techniques, to provide a pool of pre-sorted requests to the I/O scheduler. This data-driven execution mode is enabled when I/O is detected to be the bottleneck, otherwise the program runs in the normal computation-driven mode. DualPar is implemented in the MPICH2 MPI-IO library for MPI programs to coordinate I/O service and process execution. Our experiments on a 120- node cluster using the PVFS2 file system show that DualPar can increase system I/O throughput by 31% on average, compared to existing MPI-IO with or without using collective I/O.

SyncChecker: Detecting Synchronization Errors between MPI Applications and Libraries

Zhezhe Chen[†] Xinyu Li[†] Jau-Yuan Chen[†] Hua Zhong[*] Feng Qin[†]
[†] Dept. of Computer Science and Engineering
The Ohio State University
{chenzhe, lixiny, chenja, qin}@cse.ohio-state.edu
[*] Technology Center of Software Engineering
Institute of Software, Chinese Academy of Sciences
zhongh@otcaix.iscas.ac.cn

Abstract

While improving the performance, nonblocking communication is prone to synchronization errors between MPI applications and the underlying MPI libraries. Such synchronization error occurs in the following way. After initiating nonblocking communication and performing overlapped computation, the MPI application reuses the message buffer before the MPI library completes the use of the same buffer, which may lead to sending out corrupted message data or reading undefined message data. This paper presents a new method called SyncChecker to detect synchronization errors in MPI nonblocking communication. To examine whether the use of message buffers is well synchronized between the MPI application and the MPI library, SyncChecker first tracks relevant memory accesses in the MPI application and corresponding message send/receive operations in the MPI library. Then it checks whether the correct execution order between the MPI application and the MPI library is enforced by the MPI completion check routines. If not, SyncChecker reports the error with diagnostic information. To reduce runtime overhead, we propose three dynamic optimizations. We have implemented a prototype of SyncChecker on Linux and evaluated it with seven bug cases, i.e., five introduced by the original developers and two injected, in four different MPI applications. Our experiments show that SyncChecker detects all the evaluated synchronization errors and provides helpful diagnostic information. Moreover, our experiments with seven NAS Parallel Benchmarks demonstrate that SyncChecker incurs moderate runtime overhead, 1.3-9.5 times with an average of 5.2 times, making it suitable for software testing.

Holistic Debugging of MPI Derived Datatypes

Joachim Protze* , Tobias Hilbrich* , Andreas Knüpfer* , Bronis R. de Supinski† and Matthias S. Müller*
* Center for Information Services and High Performance Computing
TU Dresden, Dresden, Germany
{joachim.protze|tobias.hilbrich|andreas.knuepfer|matthias.mueller}@tu-dresden.de
† Center for Applied Scientific Computing
LLNL, Livermore, USA
bronis@llnl.gov

Abstract

The Message Passing Interface (MPI) specifies an API that allows programmers to create efficient and scalable parallel applications. The standard defines multiple constraints for each function parameter. For performance reasons, no MPI implementation checks all of these constraints at runtime. Derived datatypes are an important concept of MPI and allow users to describe an application's data structures for efficient and convenient communication. Using existing infrastructure we present scalable algorithms to detect usage errors of basic and derived MPI datatypes. We detect errors that include constraints for construction and usage of derived datatypes, matching their type signatures in communication, and detecting erroneous overlaps of communication buffers. We implement these checks in the MUST runtime error detection framework. We provide a novel representation of error locations to highlight usage errors. Further, approaches to buffer overlap checking can cause unacceptable overheads for non-contiguous datatypes. We present an algorithm that uses patterns in derived MPI datatypes to avoid these overheads without losing precision. Application results for the benchmark suites SPEC MPI2007 and NAS Parallel Benchmarks for up to 2048 cores show that our approach applies to a broad range of applications and that our extended overlap check improves performance by two orders of magnitude. Finally, we augment our runtime error detection component with a debugger extension to support in-depth analysis of the errors that we find as well as semantic errors. This extension to gdb provides information about MPI datatype handles and enables gdb – and other debuggers based on gdb – to display the content of a buffer as used in MPI communications.

Hierarchical Local Storage:Exploiting Flexible User-Data Sharing Between MPI Tasks

Marc Tchiboukdjian[*‡] , Patrick Carribault[†*] and Marc Pérache[†*]
marc.tchiboukdjian@exascale-computing.eu, patrick.carribault@cea.fr, marc.perache@cea.fr
[*] Exascale Computing Research, Versailles, France
[†] CEA, DAM, DIF, F-91297, Arpajon, France
[‡] Université de Versailles-Saint-Quentin-en-Yvelines, France

Abstract

With the advent of the multicore era, the number of cores per computational node is increasing faster than the amount of memory. This diminishing memory to core ratio sometimes even prevents pure MPI applications to exploit all cores available on each node. A possible solution is to add a shared memory programming model like OpenMP inside the application to share variables between OpenMP threads that would otherwise be duplicated for each MPI task. Going to hybrid can thus improve the overall memory consumption, but may be a tedious task on large applications. To allow this data sharing without the overhead of mixing multiple programming models, we propose an MPI extension called Hierarchical Local Storage (HLS) that allows application developers to share common variables between MPI tasks on the same node. HLS is designed as a set of directives that preserve the original parallel semantics of the code and are compatible with C, C++ and Fortran languages and the OpenMP programming model. This new mechanism is implemented inside a state-of-the-art MPI 1.3 compliant runtime called MPC. Experiments show that the HLS mechanism can effectively reduce memory consumption of HPC applications. Moreover, by reducing data duplication in the shared cache of modern multicores, the HLS mechanism can also improve performances of memory intensive applications.

Session 9
Parallel Graph Algorithms I

Fast and Efficient Graph Traversal Algorithm for CPUs : Maximizing Single-Node Efficiency

Jatin Chhugani, Nadathur Satish, Changkyu Kim, Jason Sewall, and Pradeep Dubey
Parallel Computing Lab, Intel Corporation

Abstract

Graph-based structures are being increasingly used to model data and relations among data in a number of fields. Graph-based databases are becoming more popular as a means to better represent such data. Graph traversal is a key component in graph algorithms such as reachability and graph matching. Since the scale of data stored and queried in these databases is increasing, it is important to obtain high performing implementations of graph traversal that can efficiently utilize the processing power of modern processors. In this work, we present a scalable Breadth-First Search Traversal algorithm for modern multi-socket, multi-core CPUs. Our algorithm uses lock- and atomic-free operations on a cache- resident structure for arbitrary sized graphs to filter out expensive main memory accesses, and completely and efficiently utilizes all available bandwidth resources. We propose a work distribution approach for multi-socket platforms that ensures load-balancing while keeping cross-socket communication low. We provide a detailed analytical model that accurately projects the performance of our single- and multi-socket traversal algorithms to within 5- 10% of obtained performance. Our analytical model serves as a useful tool to analyze performance bottlenecks on modern CPUs. When measured on various synthetic and real-world graphs with a wide range of graph sizes, vertex degrees and graph diameters, our implementation on a dual-socket Intel R Xeon R X5570 (Intel microarchitecture code name Nehalem) system achieves 1.5X–13.2X performance speedup over the best reported numbers. We achieve around 1 Billion traversed edges per second on a scalefree R-MAT graph with 64M vertices and 2 Billion edges on a dual-socket Nehalem system. Our optimized algorithm is useful as a building block for efficient multi-node implementations and future exascale systems, thereby allowing them to ride the trend of increasing per-node compute and bandwidth resources.

SAHAD: Subgraph Analysis in Massive Networks Using Hadoop

Zhao Zhao*, Guanying Wang†, Ali R. Butt†, Maleq Khan*, V. S. Anil Kumar* and Madhav V. Marathe*
* Network Dynamics and Simulation Science Laboratory, Virginia Tech, Blacksburg, VA, 24060, U.S.
Email: {zhaozhao,maleq,akumar,mmarathe}@vbi.vt.edu
† Department of Computer Science, Virginia Tech, Blacksburg, VA, 24060, U.S.
Email: {wanggy,butta}@cs.vt.edu

Abstract

Relational subgraph analysis, e.g. finding labeled subgraphs in a network, which are isomorphic to a template, is a key problem in many graph related applications. It is computationally challenging for large networks and complex templates. In this paper, we develop SAH AD, an algorithm for relational subgraph analysis using Hadoop, in which the subgraph is in the form of a tree. SAH AD is able to solve a variety of problems closely related with subgraph isomorphism, including counting labeled/unlabeled subgraphs, finding supervised motifs, and computing graphlet frequency distribution. We prove that the worst case work complexity for SAH AD is asymptotically very close to that of the best sequential algorithm. On a mid-size cluster with about 40 compute nodes, SAH AD scales to networks with up to 9 million nodes and a quarter billion edges, and templates with up to 12 nodes. To the best of our knowledge, SAH AD is the first such Hadoop based subgraph/subtree analysis algorithm, and performs significantly better than prior approaches for very large graphs and templates. Another unique aspect is that SAH AD is also amenable to running quite easily on Amazon EC2, without needs for any system level optimization.

Accelerating Nearest Neighbor Search on Manycore Systems

Lawrence Cayton
Max Planck Institute
Tübingen, Germany
and Microsoft Corp. Email: work@lcayton.com

Abstract

We develop methods for accelerating metric similarity search that are effective on modern hardware. Our algorithms factor into easily parallelizable components, making them simple to deploy and efficient on multicore CPUs and GPUs. Despite the simple structure of our algorithms, their search performance is provably sublinear in the size of the database, with a factor dependent only on its intrinsic dimensionality. We demonstrate that our methods provide substantial speedups on a range of datasets and hardware platforms. In particular, we present results on a 48-core server machine, on graphics hardware, and on a multicore desktop.

Optimizing large-scale graph analysis on multithreaded, multicore platforms

Guojing Cong, Konstantin Makarychev
IBM TJ Watson Research Center
1101 Kitchawan Road, Yorktown Heights, NY, 10598
{gcong,konstantin}@us.ibm.com

Abstract

The erratic memory access pattern of graph algorithms makes it hard to optimize on cache-based architectures. While multithreading hides memory latency, it is unclear how hardware threads combined with caches impact the performance of typical graph workload. As modern architectures strike different balances between caching and multithreading, it remains an open question whether the benefit of optimizing locality behavior outweighs the cost.

We study parallel graph algorithms on two different multithreaded, multi-core platforms, that is, IBM Power7 and Sun Niagara2. Our experiments first demonstrate their performance advantage over prior architectures. We find nonetheless the number of hardware threads in either platform is not sufficient to fully mask memory latency. Our cache-friendly scheduling of memory accesses improves performance by up to 2.6 times on Power7 and prior cache-based architectures, yet the same technique significantly degrades performance on Niagara2. Software prefetching and manipulating the storage of the input to improve spatial locality improve performance by up to 2.1 times and 1.3 times on both platforms. Our study reveals interesting interplay between architecture and algorithm.

Session 10
High Performance Computing Algorithms

Low-Cost Parallel Algorithms for 2:1 Octree Balance

Tobin Isaac[*] , Carsten Burstedde[*†] , Omar Ghattas[*†§]
[*] Institute for Computational Engineering and Sciences (ICES)
The University of Texas at Austin, USA
Email: {tisaac,carsten,omar}@ices.utexas.edu
[†] Present address: Institut für Numerische Simulation (INS)
Rheinische Friedrich-Wilhelms-Universität Bonn, Germany
[‡] Jackson School of Geosciences, The University of Texas at Austin, USA
[§] Department of Mechanical Engineering, The University of Texas at Austin, USA

Abstract

The logical structure of a forest of octrees can be used to create scalable algorithms for parallel adaptive mesh refinement (AMR), which has recently been demonstrated for several petascale applications. Among various frequently used octree-based mesh operations, including refinement, coarsening, partitioning, and enumerating nodes, ensuring a 2:1 size balance between neighboring elements has historically been the most expensive in terms of CPU time and communication volume. The 2:1 balance operation is thus a primary target to optimize. One important component of a parallel balance algorithm is the ability to determine whether any two given octants have a consistent distance/size relation. Based on new logical concepts we propose fast algorithms for making this decision for all types of 2:1 balance conditions in 2D and 3D. Since we are able to achieve this without constructing any parent nodes in the tree that would otherwise need to be sorted and communicated, we can significantly reduce the required memory and communication volume. In addition, we propose a lightweight collective algorithm for reversing the asymmetric communication pattern induced by non-local octant interactions. We have implemented our improvements as part of the open- source "p4est" software. Benchmarking this code with both synthetic and simulation-driven adapted meshes we are able to demonstrate much reduced runtime and excellent weak and strong scalability. On our largest benchmark problem with 5.13 x 1011 octants the new 2:1 balance algorithm executes in less than 8 seconds on 112,128 CPU cores of the Jaguar Cray XT5 supercomputer.

A Case Study of Designing Efficient Algorithm-based Fault Tolerant Application for Exascale Parallelism

Erlin Yao, Rui Wang, Mingyu Chen, Guangming Tan, Ninghui Sun
State Key Laboratory of Computer Architecture
Institute of Computing Technology, Chinese Academy of Sciences
Beijing, China
{yaoerlin, wangrui2009, cmy, tgm, snh}@ict.ac.cn

Abstract

Fault tolerance overhead of high performance computing (HPC) applications is becoming critical to the efficient utilization of HPC systems at large scale. Today's HPC applications typically tolerate fail-stop failures by checkpointing. However, checkpointing will lose its efficiency when system becoming very large. An alternative method is algorithm-based fault recovery which has been proved to be more efficient than checkpointing. In this paper, we first point out by theoretical analysis that algorithm-based fault recovery will also lose its efficiency when systems scale up to Exaflops. Then, a more efficient algorithm- based fault tolerance scheme for HPC applications at large scale is presented. The new method has two novel skills. One is algorithm-based hot replacement, which avoids the stop-and-wait time after failure. Second is background accelerated recovery, which guarantees the system to endure multiple failures in succession. As a case study, this method is incorporated to High Performance Linpack (HPL). Theoretical analysis shows that the fault tolerance overhead can be reduced to $\frac{2}{log_2^p}$ of that of algorithm- based fault recovery method (p is the number of computation processes), so that the new method will still be efficient in Exascale. Experimental results for up to 1800 processes show that the overhead of the new method is about 25% of that of algorithm-based fault recovery method, which is close to the theoretical prediction.

High Performance Non-uniform FFT on Modern x86-based Multi-core Systems

Dhiraj D. Kalamkar*, Joshua D. Trzasko‡, Srinivas Sridharan*, Mikhail Smelyanskiy†,
Daehyun Kim†, Armando Manduca‡, Yunhong Shu§, Matt A. Bernstein§, Bharat Kaul* and Pradeep Dubey†
* Parallel Computing Lab, Intel Labs, Bangalore, KA, India
† Parallel Computing Lab, Intel Labs, Santa Clara, CA, USA
‡ Department of Physiology and Biomedical Engineering, Mayo Clinic, Rochester, MN, USA
§ Department of Radiology, Mayo Clinic, Rochester, MN, USA

Abstract

The Non-Uniform Fast Fourier Transform (NUFFT) is a generalization of FFT to non-equidistant samples. It has many applications which vary from medical imaging to radio astronomy to the numerical solution of partial differential equations. Despite recent advances in speeding up NUFFT on various platforms, its practical applications are still limited, due to its high computational cost, which is significantly dominated by the convolution of a signal between a non- uniform and uniform grids. The computational cost of the NUFFT is particularly detrimental in cases which require fast reconstruction times, such as iterative 3D non-Cartesian MRI reconstruction. We propose novel and highly scalable parallel algorithm for performing NUFFT on x86-based multi-core CPUs. The high performance of our algorithm relies on good SIMD utilization and high parallel efficiency. On convolution, we demonstrate on average 90% SIMD efficiency using SSE, as well up to linear scalability using a quad-socket 40-core Intel® Xeon® E7-4870 Processors based system. As a result, on dual socket Intel® Xeon® X5670 based server, our NUFFT implementation is more than 4x faster compared to the best available NUFFT3D implementation, when run on the same hardware. On Intel® Xeon® E5-2670 processor based server, our NUFFT implementation is 1.5X faster than any published NUFFT implementation today. Such speed improvement opens new usages for NUFFT. For example, iterative multichannel reconstruction of a 240x240x240 image could execute in just over 3 minutes, which is on the same order as contemporary non-iterative (and thus less-accurate) 3D NUFFT-based MRI reconstructions.

NUMA Aware Iterative Stencil Computations on Many-Core Systems

Mohammed Shaheen and Robert Strzodka
Integrative Scientific Computing Group
Max Planck Institut Informatik
Saarbrücken, Germany
Email: {mshaheen, strzodka} @mpi-inf.mpg.de

Abstract

Temporal blocking in iterative stencil compu- tations allows to surpass the performance of peak system bandwidth that holds for a single stencil computation. However, the effectiveness of temporal blocking depends strongly on the tiling scheme, which must account for the contradicting goals of spatio-temporal data locality, regular memory access pat- terns, parallelization into many independent tasks, and data- to-core affinity for NUMA-aware data distribution. Despite the prevalence of cache coherent non-uniform memory access (ccNUMA) in todays many-core systems, this latter aspect has been largely ignored in the development of temporal blocking algorithms. Building upon previous cache-aware [1] and cache-oblivious [2] schemes, this paper develops their NUMA-aware variants, explaining why the incorporation of data-to-core affinity as an equally important goal necessitates also new tiling and parallelization strategies. Results are presented on an 8 socket dual-core and a 4 socket oct-core systems and compared against an optimized naive scheme, various peak performance characteristics, and related schemes from literature.

Session 11
Parallel Numerical Computation

Algebraic Block Multi-Color Ordering Method for Parallel Multi-Threaded Sparse Triangular Solver in ICCG Method

Takeshi Iwashita and Hiroshi Nakashima
Academic Center for Computing and Media Studies
Kyoto University
Kyoto, Japan
Email: {iwashita, h.nakashima}@media.kyoto-u.ac.jp

Yasuhito Takahashi
Department of Electrical Engineering
Doshisha University
Kyoto, Japan
Email: ytakahashi@mail.doshisha.ac.jp

Abstract

This paper covers the multi-threaded parallel processing of a sparse triangular solver for a linear system with a sparse coefficient matrix, focusing on its application to a parallel ICCG solver. We propose algebraic block multi- color ordering, which is an enhanced version of block multi- color ordering for general unstructured analysis. We present blocking and coloring strategies that achieve a high cache hit ratio and fast convergence. Five numerical tests on a shared memory parallel computer verify that the computation time of the proposed method is between 1.7 and 2.6 times faster than that of the conventional multi-color ordering method.

Parallel Computation of Morse-Smale Complexes

Attila Gyulassy, Valerio Pascucci
Scientific Computing and Imaging Institute
Dept. of Computer Science, University of Utah
Salt Lake City, United States of America
jediati@sci.utah.edu, pascucci@sci.utah.edu

Tom Peterka, Robert Ross
Mathematics and Computer Science Division
Argonne National Laboratory
Argonne, United States of America
tpeterka@mcs.anl.gov, rross@mcs.anl.gov

Abstract

Topology-based techniques are useful for multiscale exploration of the feature space of scalar-valued functions, such as those derived from the output of large-scale simulations. The Morse-Smale (MS) complex, in particular, allows robust identification of gradient-based features, and therefore is suitable for analysis tasks in a wide range of application domains. In this paper, we develop a two-stage algorithm to construct the 1-skeleton of the Morse-Smale complex in parallel, the first stage independently computing local features per block and the second stage merging to resolve global features. Our implementation is based on MPI and a distributed-memory architecture. Through a set of scalability studies on the IBM Blue Gene/P supercomputer, we characterize the performance of the algorithm as block sizes, process counts, merging strategy, and levels of topological simplification are varied, for datasets that vary in feature composition and size. We conclude with a strong scaling study using scientific datasets computed by combustion and hydrodynamics simulations.

Hybrid static/dynamic scheduling for already optimized dense matrix factorization

Simplice Donfack*, Laura Grigori*, William D. Gropp† and Vivek Kale†
Saclay-Ile de France, Universite Paris-Sud 11, 91405 Orsay,France
E-mail: {simplice.donfack, laura.grigori}@lri.fr
† Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
E-mail: {wgropp, vivek}@illinois.edu

Abstract

We present the use of a hybrid static/dynamic scheduling strategy of the task dependency graph for direct methods used in dense numerical linear algebra. This strategy provides a balance of data locality, load balance, and low dequeue overhead. We show that the usage of this scheduling in communication avoiding dense factorization leads to significant performance gains. On a 48 core AMD Opteron NUMA machine, our experiments show that we can achieve up to 64% improvement over a version of CALU that uses fully dynamic scheduling, and up to 30% improvement over the version of CALU that uses fully static scheduling. On a 16-core Intel Xeon machine, our hybrid static/dynamic scheduling approach is up to 8% faster than the version of CALU that uses a fully static scheduling or fully dynamic scheduling. Our algorithm leads to speedups over the corresponding routines for computing LU factorization in well known libraries. On the 48 core AMD NUMA machine, our best implementation is up to 110% faster than MKL, while on the 16 core Intel Xeon machine, it is up to 82% faster than MKL. Our approach also shows significant speedups compared with PLASMA on both of these systems.

Session 12
Architecture Modeling and Scheduling

Understanding Cache Hierarchy Contention in CMPs to Improve Job Scheduling

Josué Feliu, Julio Sahuquillo, Salvador Petit, and José Duato
Department of Computer Engineering (DISCA)
Universitat Politècnica de València
València, Spain
jofepre@fiv.upv.es, {jsahuqui,spetit,jduato}@disca.upv.es

Abstract

In order to improve CMP performance, recent research has focused on scheduling to mitigate contention produced by the limited memory bandwidth. Nowadays, commercial CMPs implement multi-level cache hierarchies where last level caches are shared by at least two cache structures located at the immediately lower cache level. In turn, these caches can be shared by several multithreaded cores. In this microprocessor design, contention points may appear along the whole memory hierarchy. Moreover, this problem is expected to aggravate in future technologies, since the number of cores and hardware threads, and consequently the size of the shared caches increases with each microprocessor generation. In this paper we characterize the impact on performance of the different contention points that appear along the memory subsystem. Then, we propose a generic scheduling strategy for CMPs that takes into account the available bandwidth at each level of the cache hierarchy. The proposed strategy selects the processes to be co-scheduled and allocates them to cores in order to minimize contention effects. The proposal has been implemented and evaluated in a commercial single-threaded quad-core processor with a relatively small two-level cache hierarchy. Despite these potential contention limitations are less than in recent processor designs, compared to the Linux scheduler, the proposal reaches performance improvements up to 9% while these benefits (across the studied benchmark mixes) are always lower than 6% for a memory-aware scheduler that does not take into account the cache hierarchy. Moreover, in some cases the proposal doubles the speedup achieved by the memory-aware scheduler.

Optimization of Parallel Discrete Event Simulator for Multi-core Systems

Deepak Jagtap, Nael Abu-Ghazaleh and Dmitry Ponomarev
Computer Science Department
State University of New York at Binghamton
{djagtap1,nael,dima@cs.binghamton.edu}

Abstract

Parallel Discrete Event Simulation (PDES) can substantially improve performance and capacity of simulation, allowing the study of larger, more detailed models, in shorter times. PDES is a fine-grained parallel application whose performance and scalability are limited by communication latencies. Traditionally, PDES simulation kernels use processes that communicate using message passing; shared memory is used to optimize message passing for processes running on the same machine. We report on our experiences in implementing a thread-based version of the ROSS simulator. The multithreaded implementation eliminates multiple message copying and significantly minimizes synchronization delays. We study the performance of the simulator on two hardware platforms: a Core i7 machine and a 48-core AMD Opteron Magny-Cours system. We identify performance bottlenecks and propose and evaluate mechanisms to overcome them. Results show that multithreaded implementation improves performance over the MPI version by up to a factor of 3 for the Core i7 machine and 1.2 on Magny-cours for 48-way simulation.

Using the Translation Lookaside Buffer to Map Threads in Parallel Applications Based on Shared Memory

Eduardo H. M. Cruz, Matthias Diener, Philippe O. A. Navaux
Informatics Institute
Federal University of Rio Grande do Sul
Porto Alegre, Brazil
{ehmcruz, mdiener, navaux}@inf.ufrgs.br

Abstract

The communication latency between the cores in multiprocessor architectures differs depending on the memory hierarchy and the interconnections. With the increase of the number of cores per chip and the number of threads per core, this difference between the communication latencies is increasing. Therefore, it is important to map the threads of parallel applications taking into account the communication between them. In parallel applications based on the shared memory paradigm, the communication is implicit and occurs through accesses to shared variables. For this reason, it is difficult to detect the communication pattern between the threads. Traditional approaches use simulation to monitor the memory accesses performed by the application, requiring modifications to the source code and drastically increasing the overhead.

In this paper, we introduce a new light-weight mechanism to detect the communication pattern of threads using the Translation Lookaside Buffer (TLB). Our mechanism relies entirely on hardware features, which makes the thread mapping transparent to the programmer and allows it to be performed dynamically by the operating system. Moreover, no time consuming task, such as simulation, is required.

We evaluated our mechanism with the NAS Parallel Benchmarks (NPB) and achieved an accurate representation of the communication patterns. Using the detected communication patterns, we generated thread mappings using a heuristic method based on the Edmonds graph matching algorithm. Running the applications with these mappings resulted in performance improvements of up to 15.3%, reducing the number of cache misses by up to 31.1%.

Session 13
GPU-Based Computing

Automatic Resource Scheduling with Latency Hiding for Parallel Stencil Applications on GPGPU Clusters

Kumiko Maeda*, Masana Murase*, Munehiro Doi†, Hideaki Komatsu*, Shigeho Noda‡ and Ryutaro Himeno‡
* IBM Research - Tokyo, IBM Japan, Ltd.
Email: {kumaeda, mmasana, komatsu}@jp.ibm.com
† Systems and Technology Group, IBM Japan, Ltd.
Email: munepi@jp.ibm.com
‡ Advanced Center for Computing and Communication, RIKEN
Email: {shigeho, himeno}@riken.jp

Abstract

Overlapping computations and communication is a key to accelerating stencil applications on parallel computers, especially for GPU clusters. However, such programming is a time-consuming part of the stencil application development. To address this problem, we developed an automatic code generation tool to produce a parallel stencil application with latency hiding automatically from its dataflow model. With this tool, users visually construct the workflows of stencil applications in a dataflow programming model. Our dataflow compiler determines a data decomposition policy for each application, and generates source code that overlaps the stencil computations and communication (MPI and PCIe). We demonstrate two types of overlapping models, a CPU-GPU hybrid execution model and a GPU-only model. We use a CFD benchmark computing 19-point 3D stencils to evaluate our scheduling performance, which results in 1.45 TFLOPS in single precision on a cluster with 64 Tesla C1060 GPUs.

Productive Programming of GPU Clusters with OmpSs

Javier Bueno, Judit Planas, Alejandro Duran
Barcelona Supercomputing Center
{javier.bueno,judit.planas,alex.duran}@bsc.es

Rosa M. Badia
Barcelona Supercomputing Center
Artificial Intelligence Research Institute (IIIA)
Spanish National Research Council (CSIC)
rosa.m.badia@bsc.es

Xavier Martorell, Eduard Ayguadé, Jesús Labarta
Barcelona Supercomputing Center
Universitat Politècnica de Catalunya
{xavier.martorell,eduard.ayguade,jesus.labarta}@bsc.es

Abstract

Clusters of GPUs are emerging as a new computational scenario. Programming them requires the use of hybrid models that increase the complexity of the applications, reducing the productivity of programmers. We present the implementation of OmpSs for clusters of GPUs, which supports asynchrony and heterogeneity for task parallelism. It is based on annotating a serial application with directives that are translated by the compiler. With it, the same program that runs sequentially in a node with a single GPU can run in parallel in multiple GPUs either local (single node) or remote (cluster of GPUs). Besides performing a task-based parallelization, the runtime system moves the data as needed between the different nodes and GPUs minimizing the impact of communication by using affinity scheduling, caching, and by overlapping communication with the computational task. We show several applicactions programmed with OmpSs and their performance with multiple GPUs in a local node and in remote nodes. The results show good tradeoff between performance and effort from the programmer.

Generating Device-specific GPU code for Local Operators in Medical Imaging

Richard Membarth, Frank Hannig, and Jürgen Teich
Department of Computer Science,
University of Erlangen-Nuremberg, Germany.
{richard.membarth,hannig,teich}@cs.fau.de
Mario Körner and Wieland Eckert
Siemens Healthcare Sector, H IM AX,
Forchheim, Germany.
{mario.koerner,wieland.eckert}@siemens.com

Abstract

To cope with the complexity of programming GPU accelerators for medical imaging computations, we developed a framework to describe image processing kernels in a domain- specific language, which is embedded into C++. The description uses decoupled access/execute metadata, which allow the programmer to specify both execution constraints and memory access patterns of kernels. A source-to-source compiler translates this high-level description into low-level CUDA and OpenCL code with automatic support for boundary handling and filter masks. Taking the annotated metadata and the characteristics of the parallel GPU execution model into account, two-layered parallel implementations–utilizing SPMD and MPMD parallelism– are generated. An abstract hardware model of graphics card architectures allows to model GPUs of multiple vendors like AMD and NVIDIA, and to generate device-specific code for multiple targets. It is shown that the generated code is faster than manual implementations and those relying on hardware support for boundary handling. Implementations from RapidMind, a commercial framework for GPU programming, are outperformed and similar results achieved compared to the GPU backend of the widely used image processing library OpenCV.

Performance Portability with the Chapel Language

Albert Sidelnik*, Saeed Maleki*, Bradford L. Chamberlain†, María J. Garzarán*, David Padua*
* Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA
† Cray Inc., Seattle, Washington, USA
Email: {asideln2,maleki1,garzaran,padua}@illinois.edu*, bradc@cray.com†

Abstract

It has been widely shown that high-throughput computing architectures such as GPUs offer large performance gains compared with their traditional low-latency counterparts for many applications. The downside to these architectures is that the current programming models present numerous challenges to the programmer: lower-level languages, loss of portability across different architectures, explicit data movement, and challenges in performance optimization. This paper presents novel methods and compiler transformations that increase programmer productivity by enabling users of the language Chapel to provide a single code implementation that the compiler can then use to target not only conventional multiprocessors, but also high-throughput and hybrid machines. Rather than resorting to different parallel libraries or annotations for a given parallel platform, this work leverages a language that has been designed from first principles to address the challenge of programming for parallelism and locality. This also has the advantage of providing portability across different parallel architectures. Finally, this work presents experimental results from the Parboil benchmark suite which demonstrate that codes written in Chapel achieve performance comparable to the original versions implemented in CUDA on both GPUs and multicore platforms.

Session 14
Parallel Matrix Factorizations

Mapping Dense LU Factorization on Multicore Supercomputer Nodes

Jonathan Lifflander, Phil Miller,
Ramprasad Venkataraman, Anshu Arya, Laxmikant Kale
University of Illinois Urbana-Champaign
Email: {jliffl2, mille121, ramv, arya3, kale}@illinois.edu

Terry Jones
Oak Ridge National Laboratory
Email: trj@ornl.gov

Abstract

Dense LU factorization is a prominent benchmark used to rank the performance of supercomputers. Many implementations use block-cyclic distributions of matrix blocks onto a two-dimensional process grid. The process grid dimensions drive a trade-off between communication and computation and are architecture- and implementation-sensitive. The critical panel factorization steps can be made less communication-bound by overlapping asynchronous collectives for pivoting with the computation of rank-k updates. By shifting the computation-communication trade-off, a modified block-cyclic distribution can beneficially exploit more available parallelism on the critical path, and reduce panel factorization's memory hierarchy contention on now-ubiquitous multicore architectures. During active panel factorization, rank-1 updates stream through memory with minimal reuse. In a column-major process grid, the performance of this access pattern degrades as too many streaming processors contend for access to memory. A block- cyclic mapping in the row-major order does not encounter this problem, but consequently sacrifices node and network locality in the critical pivoting steps. We introduce striding to vary between the two extremes of row- and column-major process grids. The maximum available parallelism in the critical path work (active panel factorization, triangular solves, and subsequent broadcasts) is bounded by the length or width of the process grid. Increasing one dimension of the process grid decreases the number of distinct processes and nodes in the other dimension. To increase the harnessed parallelism in both dimensions, we start with a tall process grid. We then apply periodic rotation to this grid to restore exploited parallelism along the row to previous levels. As a test-bed for further mapping experiments, we describe a dense LU implementation that allows a block distribution to be defined as a general function of block to processor. Other mappings can be tested with only small, local changes to the code.

Hierarchical QR factorization algorithms for multi-core cluster systems

Jack Dongarra[1,2,3], Mathieu Faverge[1], Thomas Herault[1], Julien Langou[4] and Yves Robert[1,5]
1. University of Tennessee Knoxville, USA
2. Oak Ridge National Laboratory, USA
3. Manchester University, UK
4. University of Colorado Denver, USA;
supported by the National Science Foundation grant # NSF CCF 1054864.
5. Ecole Normale Supérieure de Lyon, France
{dongarra – mfaverge – therault}@eecs.utk.edu, Julien.Langou@ucdenver.edu, Yves.Robert@ens-lyon.fr

Abstract

This paper describes a new QR factorization algorithm which is especially designed for massively parallel platforms combining parallel distributed multi-core nodes. These platforms make the present and the foreseeable future of high-performance computing. Our new QR factorization algorithm falls in the category of the tile algorithms which naturally enables good data locality for the sequential kernels executed by the cores (high sequential performance), low number of messages in a parallel distributed setting (small latency term), and fine granularity (high parallelism). Each tile algorithm is uniquely characterized by its sequence of reduction trees. In the context of a cluster of multicores, in order to minimize the number of inter-processor communications (aka, "communication-avoiding" algorithm), it is natural to consider two-level hierarchical trees composed of an "inter- node" tree which acts on top of "intra-node" trees. At the intra-node level, we propose a hierarchical tree made of three levels: (0) "TS level" for cache-friendliness, (1) "low level" for decoupled highly parallel inter-node reductions, (2) "coupling level" to efficiently resolve interactions between local reductions and global reductions. Our hierarchical algorithm and its implementation are flexible and modular, and can accommodate several kernel types, different distribution layouts, and a variety of reduction trees at all levels, both inter-cluster and intra-cluster. Numerical experiments on a cluster of multicore nodes (1) confirm that each of the four levels of our hierarchical tree contributes to build up performance and (2) build insights on how these levels influence performance and interact within each other. Our implementation of the new algorithm with the DAGuE scheduling tool significantly outperforms currently available QR factorization softwares for all matrix shapes, thereby bringing a new advance in numerical linear algebra for petascale and exascale platforms.

New Scheduling Strategies and Hybrid Programming for a Parallel Right-looking Sparse LU Factorization Algorithm on Multicore Cluster Systems

Ichitaro Yamazaki
Computational Research Division
Lawrence Berkeley National Laboratory
Berkeley, CA 94720, U.S.A.
Email: ic.yamazaki@gmail.com

Xiaoye S. Li
Computational Research Division
Lawrence Berkeley National Laboratory
Berkeley, CA 94720, U.S.A.
Email: xsli@lbl.gov

Abstract

Parallel sparse LU factorization is a key computational kernel in the solution of a large-scale linear system of equations. In this paper, we propose two strategies to address some scalability issues of a factorization algorithm on modern HPC systems. The first strategy is at the algorithmic- level; we schedule independent tasks as soon as possible to reduce the idle time and the critical path of the algorithm. We demonstrate using thousands of cores that our new scheduling strategy reduces the runtime by nearly three-fold from that of a state-of-the-art pipelined factorization algorithm. The second strategy is at both programming- and architecture-levels; we incorporate light-weight OpenMP threads in each MPI process to reduce both memory and time overheads of a pure MPI implementation on manycore NUMA architectures. Using this hybrid programming paradigm, we obtain a significant reduction in memory usage while achieving a parallel efficiency competitive with that of a pure MPI paradigm. As a result, in comparison to a pure MPI paradigm which failed due to the per-core memory constraint, the hybrid paradigm could utilize more cores on each node and reduce the factorization time on the same number of nodes. We show extensive performance analysis of the new strategies using thousands of cores of the two leading HPC systems, a Cray-XE6 and an IBM iDataPlex.

ShyLU: A Hybrid-Hybrid Solver for Multicore Platforms

Sivasankaran Rajamanickam[1], Erik G. Boman[1], and Michael A. Heroux[1],
E-mail:{srajama@sandia.gov, egboman@sandia.gov and maherou@sandia.gov}
[1] Sandia National Laboratories.

Abstract

With the ubiquity of multicore processors, it is crucial that solvers adapt to the hierarchical structure of modern architectures. We present ShyLU, a "hybrid-hybrid" solver for general sparse linear systems that is hybrid in two ways: First, it combines direct and iterative methods. The iterative part is based on approximate Schur complements where we compute the approximate Schur complement using a value-based dropping strategy or structure-based probing strategy. Second, the solver uses two levels of parallelism via hybrid programming (MPI+threads). ShyLU is useful both in shared- memory environments and on large parallel computers with distributed memory. In the latter case, it should be used as a subdomain solver. We argue that with the increasing complexity of compute nodes, it is important to exploit multiple levels of parallelism even within a single compute node. We show the robustness of ShyLU against other algebraic preconditioners. ShyLU scales well up to 384 cores for a given problem size. We also study the MPI-only performance of ShyLU against a hybrid implementation and conclude that on present multicore nodes MPI-only implementation is better. However, for future multicore machines (96 or more cores) hybrid/ hierarchical algorithms and implementations are important for sustained performance.

Session 15

Distributed Computing and Programming Models

MATE-CG: A MapReduce-Like Framework for Accelerating Data-Intensive Computations on Heterogeneous Clusters

Wei Jiang Gagan Agrawal
Department of Computer Science and Engineering
The Ohio State University Columbus OH 43210
{jiangwei,agrawal}@cse.ohio-state.edu

Abstract

Clusters of GPUs have rapidly emerged as the means for achieving extreme-scale, cost-effective, and power- efficient high performance computing. At the same time, high- level APIs like map-reduce are being used for developing several types of high-end and/or data-intensive applications. Map-reduce, originally developed for data processing applications, has been successfully used for many classes of applications that involve a significant amount of computations, such as machine learning, image processing, and data mining applications. Because such applications can be accelerated using GPUs (and other accelerators), there has been interest in supporting map-reduce-like APIs on GPUs. However, while the use of map-reduce for a single GPU has been studied, developing map-reduce-like models for programming a heterogeneous CPU-GPU cluster remains an open challenge. This paper presents the MATE-CG system, which is a map- reduce-like framework based on the generalized reduction API. We develop support for enabling scalable and efficient implementation of data-intensive applications in a heterogeneous cluster of multi-core CPUs and many-core GPUs. Our contributions are three folds: 1) we port the generalized reduction model on clusters of modern GPUs with a map-reduce-like API, dealing with very large datasets; 2) we further propose three schemes to better utilize the computing power of CPUs and/or GPUs and develop an auto-tuning strategy to achieve the best-possible heterogeneous configuration for iterative applications; 3) we show how analytical models can be used to optimize important parameters in our system. We evaluate our system using three representative data- intensive applications and report results on a heterogeneous cluster of 128 CPU cores and 16 GPUs (7168 GPU cores). We show an average speedup of 87 on this cluster over execution with 2 CPU-cores. Our applications also achieve an average improvement of 25% by using CPU cores and GPUs simultaneously, over the best performance achieved from using only one of the types of resources in the cluster.

Automated and Agile Server Parameter Tuning with Learning and Control

Yanfei Guo, Palden Lama and Xiaobo Zhou
Department of Computer Science
University of Colorado, Colorado Springs, USA
Email addresses: {yguo, plama, xzhou}@uccs.edu

Abstract

Server parameter tuning in virtualized data centers is crucial to performance and availability of hosted Internet applications. It is challenging due to high dynamics and burstiness of workloads, multi-tier service architecture, and virtualized server infrastructure. In this paper, we investigate automated and agile server parameter tuning for maximizing effective throughput of multi-tier Internet applications. A recent study proposed a reinforcement learning based server parameter tuning approach for minimizing average response time of multi-tier applications. Reinforcement learning is a decision making process determining the parameter tuning direction based on trial-and-error, instead of quantitative values for agile parameter tuning. It relies on a predefined adjustment value for each tuning action. However it is nontrivial or even infeasible to find an optimal value under highly dynamic and bursty workloads. We design a neural fuzzy control based approach that combines the strengths of fast online learning and self-adaptiveness of neural networks and fuzzy control. Due to the model independence, it is robust to highly dynamic and bursty workloads. It is agile in server parameter tuning due to its quantitative control outputs. We implement the new approach on a testbed of virtualized HP ProLiant blade servers hosting RUBiS benchmark applications. Experimental results demonstrate that the new approach significantly outperforms the reinforcement learning based approach for both improving effective system throughput and minimizing average response time.

A Self-tuning Failure Detection Scheme for Cloud Computing Service

Naixue Xiong
Dept. of Computer Science
Georgia State Univ., USA
E-mail: {nxiong, wsong, pan}@gsu.edu

Athanasios V. Vasilakos
Dept. of Comp. and Tele. Engi.
Univ. of Western Macedonia, Greece
E-mail: vasilako@ath.forthnet.gr

Jie Wu
Dept. of Comp. and Info. Scie.
Temple Univ., USA.
E-mail: jiewu@temple.edu

Y. Richard Yang
Dept. of Computer Science
Yale Univ.
New Haven, USA
E-mail: yry@cs.yale.edu

Andy Rindos
IBM Corp., Dept. W4DA/Bldg 503
Research Triangle Park,
Durham, NC, USA
E-mail: rindos@us.ibm.com

Yuezhi Zhou[1], Wen-Zhan Song[2], Yi Pan[2]
[1] Dept. of Comp. Scie. & Tech.
Tsinghua Univ., China
E-mail: zhouyz@mail.tsinghua.edu.cn
[2] Dept. of Computer Science Georgia State Univ., USA
E-mail: {wsong, pan}@cs.gsu.edu

Abstract

Cloud computing is an increasingly important solution for providing services deployed in dynamically scalable cloud networks. Services in the cloud computing networks may be virtualized with specific servers which host abstracted details. Some of the servers are active and available, while others are busy or heavy loaded, and the remaining are offline for various reasons. Users would expect the right and available servers to complete their application requirements. Therefore, in order to provide an effective control scheme with parameter guidance for cloud resource services, failure detection is essential to meet users' service expectations. It can resolve possible performance bottlenecks in providing the virtual service for the cloud computing networks. Most existing Failure Detector (FD) schemes do not automatically adjust their detection service parameters for the dynamic network conditions, thus they couldn't be used for actual application. This paper explores FD properties with relation to the actual and automatic fault-tolerant cloud computing networks, and find a general non-manual analysis method to self-tune the corresponding parameters to satisfy user requirements. Based on this general automatic method, we propose a specific and dynamic Self-tuning Failure Detector, called SFD, as a major breakthrough in the existing schemes. We carry out actual and extensive experiments to compare the quality of service performance between the SFD and several other existing FDs. Our experimental results demonstrate that our scheme can automatically adjust SFD control parameters to obtain corresponding services and satisfy user requirements, while maintaining good performance. Such an SFD can be extensively applied to industrial and commercial usage, and it can also significantly benefit the cloud computing networks.

PGAS for Distributed Numerical Python Targeting Multi-core Clusters

Mads Ruben Burgdorff Kristensen
Niels Bohr Institute
University of Copenhagen
Denmark
madsbk@nbi.dk

Yili Zheng
Lawrence Berkeley National Lab
Berkeley, CA 94720
USA
yzheng@lbl.gov

Brian Vinter
Niels Bohr Institute
University of Copenhagen
Denmark
vinter@nbi.dk

Abstract

In this paper we propose a parallel programming model that combines two well-known execution models: Single Instruction, Multiple Data (SIMD) and Single Program, Multiple Data (SPMD). The combined model supports SIMD- style data parallelism in global address space and supports SPMD-style task parallelism in local address space. One of the most important features in the combined model is that data communication is expressed by global data assignments instead of message passing. We implement this combined programming model into Python, making parallel programming with Python both highly productive and performing on distributed memory multi-core systems. We base the SIMD data parallelism on DistNumPy, an auto-parallelizing version of the Numerical Python (NumPy) package that allows sequential NumPy programs to run on distributed memory architectures. We implement the SPMD task parallelism as an extension to DistNumPy that enables each process to have direct access to the local part of a shared array. To harvest the multi-core benefits in modern processors we exploit multi-threading in both SIMD and SPMD execution models. The multi-threading is completely transparent to the user – it is implemented in the runtime with OpenMP and by using multi-threaded libraries when available. We evaluate the implementation of the combined programming model with several scientific computing benchmarks using two representative multi-core distributed memory systems – an Intel Nehalem cluster with Infiniband interconnects and a Cray XE-6 supercomputer – up to 1536 cores. The benchmarking results demonstrate scalable good performance.

Session 16
Memory Architectures

Miss-Correlation Folding: Encoding Per-Block Miss Correlations in Compressed DRAM for Data Prefetching

Gang Liu, Jih-Kwon Peir
Department of Computer & Information Science & Eng
University of Florida
Gainesville, FL, USA
galiu, peir@cise.ufl.edu

Victor Lee
Parallel Computing Lab, Intel Labs
Intel Corporation,
Santa Clara, CA, USA
victor.w.lee@intel.com

Abstract

Cache misses frequently exhibit repeated streaming behavior, i.e. a sequence of cache misses has a high tendency of being repeated. Correlation-based prefetchers record the missing streams in a history table for accurate prefetching. Saving a large miss history in off-chip DRAM is a practical implementation, but incurs access latency and consumes memory bandwidth which leads to performance degradation. In this paper, we investigate a new data prefetching mechanism based on per-block miss correlation where a miss is correlated with an earlier miss when the two misses are closely encountered both in time and space. The miss correlations are captured dynamically and saved along with the content of the data block using a simple data compression technique. As a result of this novel combination, our scheme provides unbounded correlation history and its prefetch metadata can be fetched together with demand data without incurring additional latency nor consuming any memory bandwidth. Performance evaluations using data-parallel applications demonstrate that prefetchers based on per-block miss correlations can improve IPC by 42-139% with an average of 88% compared to the IPC without prefetching. In comparison with regular stream prefetcher, sampled temporal streaming prefetcher and spatial-temporal memory streaming prefetcher, up to 115%, 99% and 98% IPC improvement can be obtained with an average about 36%, 26% and 27% respectively.

On the role of NVRAM in data-intensive architectures: an evaluation

Brian Van Essen[†] Roger Pearce[†‡] Sasha Ames[†] Maya Gokhale[†]
[†]Center for Applied Scientific Computing
Lawrence Livermore National Laboratory, Livermore, CA 94550
{vanessen1, pearce7, ames4, gokhale2}@llnl.gov
[‡]Department of Computer Science and Engineering, Texas A&M University

Abstract

Data-intensive applications are best suited to high- performance computing architectures that contain large quan- tities of main memory. Creating these systems with DRAM- based main memory remains costly and power-intensive. Due to improvements in density and cost, non-volatile random access memories (NVRAM) have emerged as compelling storage technologies to augment traditional DRAM. This work explores the potential of future NVRAM technologies to store program state at performance comparable to DRAM. We have developed the PerMA NVRAM simulator that allows us to explore applications with working sets ranging up to hundreds of gigabytes per node. The simulator is implemented as a Linux device driver that allows application execu- tion at native speeds. Using the simulator we show the impact of future technology generations of I/O-bus-attached NVRAM on an unstructured-access, level-asynchronous, Breadth-First Search (BFS) graph traversal algorithm. Our simulations show that within a couple of technology generations, a system architecture with local high performance NVRAM will be able to effectively augment DRAM to support highly concurrent data-intensive applications with large memory footprints. However, improvements will be needed in the I/O stack to deliver this performance to applications. The simulator shows that future technology generations of NVRAM in conjunction with an improved I/O runtime will enable parallel data-intensive applications to offload in-memory data structures to NVRAM with minimal performance loss.

iTransformer: Using SSD to Improve Disk Scheduling for High-performance I/O

Xuechen Zhang
ECE Department
Wayne State University
Detroit, MI, 48202, USA
xczhang@wayne.edu

Kei Davis
CCS Division
Los Alamos National Laboratory
Los Alamos, NM 87545, USA
kei.davis@lanl.gov

Song Jiang
ECE Department
Wayne State University
Detroit, MI, 48202, USA
sjiang@eng.wayne.edu

Abstract

The parallel data accesses inherent to large-scale data-intensive scientific computing require that data servers handle very high I/O concurrency. Concurrent requests from different processes or programs to hard disk can cause disk head thrashing between different disk regions, resulting in unacceptably low I/O performance. Current storage systems either rely on the disk scheduler at each data server, or use SSD as storage, to minimize this negative performance effect. However, the ability of the scheduler to alleviate this problem by scheduling requests in memory is limited by concerns such as long disk access times, and potential loss of dirty data with system failure. Meanwhile, SSD is too expensive to be widely used as the major storage device in the HPC environment. We propose iTransformer, a scheme that employs a small SSD to schedule requests for the data on disk. Being less space- constrained than with more expensive DRAM, iTransformer can buffer larger amounts of dirty data before writing it back to the disk, or prefetch a larger volume of data in a batch into the SSD. In both cases high disk efficiency can be maintained even for concurrent requests. Furthermore, the scheme allows the scheduling of requests in the background to hide the cost of random disk access behind serving process requests. Finally, as a non-volatile memory, concerns about the quantity of dirty data are obviated. We have implemented iTransformer in the Linux kernel and tested it on a large cluster running PVFS2. Our experiments show that iTransformer can improve the I/O throughput of the cluster by 35% on average for MPI/IO benchmarks of various data access patterns.

Switching Optically-Connected Memories in a Large-Scale System

Abhirup Chakraborty*†‡ , Eugen Schenfeld†, Dilma Da Silva†
* Member, ACM
abhirupc@acm.org
† IBM T. J. Watson Research Center
Yorktown Heights, NY, USA 10598
{eugen, dilmasilva}@us.ibm.com

Abstract

Recent trends in processor and memory systems in large-scale computing systems reveal a new "memory wall" that prompts investigation on alternate main memory organization separating main memory from processors and arranging them in separate ensembles. In this paper, we study the feasibility of transferring data across processors by using the optical interconnection fabric that acts as a bridge between processor and memory ensembles. We propose a memory switching protocol that transfers data across processors without physically moving the data across electrical switches. Such a mechanism allows large-scale data communication across processors through transfer of a few tiny blocks of meta-data. We present detailed techniques for supporting two communication patterns prevalent in any large-scale scientific and data management applications. We present experimental results analyzing the feasibility of memory switching in a wide range of applications, and characterize applications based on the impact of the memory switching on their performance.

Session 17

High Performance Communication and Networking

Supporting the Global Arrays PGAS Model Using MPI One-Sided Communication

James Dinan, Pavan Balaji, Jeff R. Hammond
Argonne National Laboratory
{dinan,balaji,jhammond}@anl.gov

Sriram Krishnamoorthy
Pacific Northwest National Laboratory
sriram@pnnl.gov

Vinod Tipparaju
IEEE Member[†]
tipparajuv@ieee.org

Abstract

The industry-standard Message Passing Interface (MPI) provides one-sided communication functionality and is available on virtually every parallel computing system. However, it is believed that MPI's one-sided model is not rich enough to support higher-level global address space parallel programming models. We present the first successful application of MPI one-sided communication as a runtime system for a PGAS model, Global Arrays (GA). This work has an immediate impact on users of GA applications, such as NWChem, who often must wait several months to a year or more before GA becomes available on a new architecture. We explore challenges present in the application of MPI-2 to PGAS models and motivate new features in the upcoming MPI-3 standard. The performance of our system is evaluated on several popular high-performance computing architectures through communication benchmarking and application bench- marking using the NWChem computational chemistry suite.

A uGNI-based Asynchronous Message-driven Runtime System for Cray Supercomputers with Gemini Interconnect

Yanhua Sun, Gengbin Zheng, Laximant V. Kalé
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, Illinois
{sun51, gzheng, kale}@illinois.edu

Terry R. Jones
Computer Science and Mathematics Division
Oak Ridge National Lab
Oak Ridge, Tennessee
trjones@ornl.gov

Ryan Olson
Cray Inc
ryan@cray.com

Abstract

Gemini, the network for the new Cray XE/XK systems, features low latency, high bandwidth and strong scalability. Its hardware support for remote direct memory access enables efficient implementation of the global address space programming languages. Although the user Generic Network Interface (uGNI) provides a low-level interface for Gemini with support to the message-passing programming model (MPI), it remains challenging to port alternative programming models with scalable performance. CHARM++ is an object-oriented message-driven programming model. Its applications have been shown to scale up to the full Jaguar Cray XT machine. In this paper, we present an implementation of this programming model on uGNI for the Cray XE/XK systems. Several techniques are presented to exploit the uGNI capabilites by reducing memory copy and registration overhead, taking advantage of the persistent commu- nication, and improving intra-node communication. Our micro- benchmark results demonstrate that the uGNI-based runtime system outperforms the MPI-based implementation by up to 50% in terms of message latency. For communication intensive applications such as N-Queens, this implementation scales up to 15, 360 cores of a Cray XE6 machine and is 70% faster than the MPI-based implementation. In molecular dynamics application NAMD, the performance is also considerably improved by as much as 18%.

PAMI: A Parallel Active Message Interface for the Blue Gene/Q Supercomputer

Sameer Kumar[1], Amith R. Mamidala[1], Daniel A. Faraj[2], Brian Smith[2], Michael Blocksome[2],
Bob Cernohous[2], Douglas Miller[2], Jeff Parker, Joseph Ratterman[2],
Philip Heidelberger[1], Dong Chen[1] and Burkhard Steinmacher-Burow[3]

{sameerk,amithr,philiph,chendong}@us.ibm.com
[1] IBM T.J. Watson Research Center
Yorktown Heights, NY 10598, USA

{faraja,smithbr,blocksom,bobc,dougmill,jjparker,jratt}@us.ibm.com
[2] IBM Systems and Technology Group
Rochester, MN, 55901

steinmac@de.ibm.com
[3]IBM Research and Development
Boeblingen, Germany, 71032

Abstract

The Blue Gene/Q machine is the next generation in the line of IBM massively parallel supercomputers, designed to scale to 262144 nodes and sixteen million threads. With each BG/Q node having 68 hardware threads, hybrid programming paradigms, which use message passing among nodes and multi-threading within nodes, are ideal and will enable applications to achieve high throughput on BG/Q. With such unprecedented massive parallelism and scale, this paper is a groundbreaking effort to explore the design challenges for designing a communication library that can match and exploit such massive parallelism In particular, we present the Parallel Active Messaging Interface (PAMI) library as our BG/Q library solution to the many challenges that come with a machine at such scale. PAMI provides (1) novel techniques to partition the application communication overhead into many contexts that can be accelerated by communication threads; (2) client and context objects to support multiple and different programming paradigms; (3) lockless algorithms to speed up MPI message rate; and (4) novel techniques leveraging the new BG/Q architectural features such as the scalable atomic primitives implemented in the L2 cache, the highly parallel hardware messaging unit that supports both point-to-point and collective operations, and the collective hardware acceleration for operations such as broadcast, reduce, and allreduce. We experimented with PAMI on 2048 BG/Q nodes and the results show high messaging rates as well as low latencies and high throughputs for collective communication operations.

High-Performance Design of HBase with RDMA over InfiniBand

Jian Huang[1], Xiangyong Ouyang[1], Jithin Jose[1], Md. Wasi-ur-Rahman[1], Hao Wang[1],
Miao Luo[1], Hari Subramoni[1], Chet Murthy[2], and Dhabaleswar K. Panda[1]
[1] Department of Computer Science and Engineering,
The Ohio State University
{huangjia, ouyangx, jose, rahmanmd, wangh, luom, subramon, panda}
@cse.ohio-state.edu
[2] IBM T.J Watson Research Center
Yorktown Heights, NY
{chet} @watson.ibm.com

Abstract

HBase is an open source distributed Key/Value store based on the idea of BigTable. It is being used in many data-center applications (e.g. Facebook, Twitter, etc.) because of its portability and massive scalability. For this kind of system, low latency and high throughput is expected when supporting services for large scale concurrent accesses. However, the existing HBase implementation is built upon Java Sockets Interface that provides sub-optimal performance due to the overhead to provide cross-platform portability. The byte- stream oriented Java sockets semantics confine the possibility to leverage new generations of network technologies. This makes it hard to provide high performance services for data-intensive applications. The High Performance Computing (HPC) domain has exploited high performance and low latency networks such as InfiniBand for many years. These interconnects provide advanced network features, such as Remote Direct Memory Access (RDMA), to achieve high throughput and low latency along with low CPU utilization. RDMA follows memory-block semantics, which can be adopted efficiently to satisfy the object transmission primitives used in HBase. In this paper, we present a novel design of HBase for RDMA capable networks via Java Native Interface (JNI). Our design extends the existing open-source HBase software and makes it RDMA capable. Our performance evaluation reveals that latency of HBase Get operations of 1KB message size can be reduced to 43.7 μs with the new design on QDR platform (32 Gbps). This is about a factor of 3.5 improvement over 10 Gigabit Ethernet (10 GigE) network with TCP Offload. Throughput evaluations using four HBase region servers and 64 clients indicate that the new design boosts up throughput by 3 X times over 1 GigE and 10 GigE networks. To the best of our knowledge, this is first HBase design utilizing high performance RDMA capable interconnects.

Session 18

Scheduling and Load Balancing Algorithms II

Virtual Machine Resource Allocation for Service Hosting on Heterogeneous Distributed Platforms

Mark Stillwell[1,2,3,4], Frédéric Vivien[2,3,4], and Henri Casanova[5]
[1] Department of Engineering Computing
Cranfield University, Cranfield, UK
[2] INRIA, France
[3] Université de Lyon, France
[4] LIP, Ecole Normale Supérieure de Lyon, France
[5] Department of Information and Computer Sciences
University of Hawai´i at Mānoa, Honolulu, USA

Abstract

We propose algorithms for allocating multiple resources to competing services running in virtual machines on heterogeneous distributed platforms. We develop a theoretical problem formulation and compare these algorithms via simulation experiments based in part on workload data supplied by Google. Our main finding is that vector packing approaches proposed in the homogeneous case can be extended to provide high-quality solutions in the heterogeneous case, and combined to provide a single efficient algorithm. We also consider the case when there may be bounded errors in estimates of performance-related resource needs. We provide a heuristic for compensating for such errors that performs well in simulation, as well as a proof of the worst-case competitive ratio for the single-resource, single-node case when there is no bound on the error.

Consistency-aware Partitioning Algorithm in Multi-server Distributed Virtual Environments

Yusen Li, Wentong Cai
Parallel and Distributed Computing Center
Nanyang Technological University
Singapore 639798
S080007@e.ntu.edu.sg, ASWTCAI@ntu.edu.sg

Abstract

In DVEs, the primary task is to maintain a consistent view of the virtual world among all users. Multi-server architecture has been shown to have good scalability to support a large population of users in DVEs. One of the key issues in the design of an efficient and scalable Multi-server Distributed Virtual Environment (MSDVE) is the partitioning, which concerns with efficiently distributing the workload generated in the virtual environment among multiple servers in the system. Most of the existing work on the partitioning issue in MSDVE aims to either balance workload among servers, reduce inter-server communication, and/or improve the interactivity of DVE. In this paper, we study the partitioning issue from a new perspective and aim to reduce the time-space inconsistency of a DVE. Time-space inconsistency is a consistency metric, which has been proven to be an effective performance measure of DVEs. Using the time-space inconsistency metric, we formally formulate our partitioning problem as a mix integer programming problem and propose a solution based on Alternating Optimization (AO) technique. An iterative partitioning algorithm is also developed accordingly. The algorithm gives a partition as well as the corresponding update schedule to minimize the total time-space inconsistency. Different from most of the existing work, the resulted partition is avatar-based rather than zone/region-based. To evaluate the performance of the proposed partitioning algorithm, extensive experiments were conducted and results are reported in the paper.

Optimal Resource Rental Planning for Elastic Applications in Cloud Market

Han Zhao*, Miao Pan†, Xinxin Liu*, Xiaolin Li† and Yuguang Fang†
* Department of Computer & Information Science & Engineering
University of Florida, Gainesville, FL 32611
† Department of Electrical and Computer Engineering
University of Florida, Gainesville, FL 32611

Abstract

This paper studies the optimization problem of minimizing resource rental cost for running elastic applications in cloud while meeting application service requirements. Such a problem arises when excessive generated data incurs significant monetary cost on transfer and inventory in cloud. The goal of planning is to make resource rental decisions in response to varying application progress in the most cost-effective way. To address this problem, we first develop a Deterministic Resource Rental Planning (DRRP) model, using a mixed integer linear program, to generate optimal rental decisions given fixed cost parameters. Next, we systematically analyze the predictability of the time-varying spot instance prices in Amazon EC2 and find that the best achievable prediction is insufficient to provide a close approximation to the actual prices. This fact motivates us to propose a Stochastic Resource Rental Planning (SRRP) model that explicitly considers the price uncertainty in rental decision making. Using empirical spot price data sets and realistic cost parameters, we conduct simulations over a wide range of experimental scenarios. Results show that DRRP achieves as much as 50% cost reduction compared to the no-planning scheme. Moreover, SRRP consistently outperforms its DRRP counterpart in terms of cost saving, which demonstrates that SRRP is highly adaptive to the unpredictable nature of spot price in cloud resource market.

Improved Bounds for Discrete Diffusive Load Balancing

Clemens P. J. Adolphs
Lehrstuhl fr Informatik 1
RWTH Aachen University
Aachen, Germany
clemens.adolphs@gmail.com

Petra Berenbrink
Department of Computer Science
Simon Fraser University
Burnaby, Canada
petra@sfu.ca

Abstract

In this paper we consider load balancing in a static and discrete setting where a fixed number of indivisible tasks have to be allocated to processors. We assume uniform tasks but the processors may have different speeds. The load of a processor is the number of tasks assigned to it divided by its speed. We consider diffusion load balancing which works in rounds. In every round the processors are allowed to compare their own load with the load of their neighbors and to balance the load with the neighbors, using their local information only. The question is how many rounds does it take until the whole processor network is balanced, meaning the load discrepancy (difference between maximum load and m/n) is minimized. Our balancing algorithm is deterministic and extends the algorithm studied in [1] from the case of uniform speeds to non-uniform speeds. We use a potential function argument to show that a better load balance can be obtained when the algorithm is allowed to run longer compared to the algorithm of [1].

Session 19
Parallel Graph Algorithms II

Multi-core spanning forest algorithms using the disjoint-set data structure

Md. Mostofa Ali Patwary*
*EECS Department,
Northwestern University,
Evanston, IL 60208, USA,
m-patwary@northwestern.edu

Peder Refsnes** Fredrik Manne**
**Department of Informatics,
University of Bergen,
N-5020 Bergen, Norway,
Peder.Refsnes@gmail.com, fredrikm@ii.uib.no

Abstract

We present new multi-core algorithms for computing spanning forests and connected components of large sparse graphs. The algorithms are based on the use of the disjoint-set data structure. When compared with the previous best algorithms for these problems our algorithms are appealing for several reasons: Extensive experiments using up to 40 threads on several different types of graphs show that they scale better. Also, the new algorithms do not make use of any hardware specific routines, and thus are highly portable. Finally, the algorithms are quite simple and easy to implement.

Graph Partitioning for Reconfigurable Topology

Deepak Ajwani
The Centre for Unified Computing,
University College Cork,
Cork, Ireland

Shoukat Ali
Exascale Systems Group,
IBM Research and Development Lab,
Dublin, Ireland

John P. Morrison
The Centre for Unified Computing,
University College Cork,
Cork, Ireland

Abstract

Optical circuit switches have recently been proposed as a low-cost, low-power and high-bandwidth alternative to electronic switches for the design of high-performance compute clusters. An added advantage of these switches is that they allow for a reconfiguration of the network topology to suit the requirements of the application. To realize the full potential of a high-performance computing system with a reconfigurable interconnect, there is a need to design algorithms for computing a topology that will allow for a high-throughput load distribution, while simultaneously partitioning the computational task graph of the application for the computed topology. In this paper, we propose a new frame- work that exploits such reconfigurable interconnects to achieve these interdependent goals, i.e., to iteratively co-optimize the network topology configuration, application partitioning and network flow routing to maximize throughput for a given application. We also present a novel way of computing a high- throughput initial topology based on the structural properties of the application to seed our co-optimizing framework. We show the value of our approach on synthetic graphs that emulate the key characteristics of a class of stream computing applications that require high throughput. Our experiments show that the proposed technique is fast and computes high- quality partitions of such graphs for a broad range of hardware parameters that varies the bottleneck from computation to communication.

Multithreaded Clustering for Multi-level Hypergraph Partitioning

Ümit V. Çatalyürek, Mehmet Deveci, Kamer Kaya
The Ohio State University
Dept. of Biomedical Informatics
{umit,mdeveci,kamer}@bmi.osu.edu

Bora Uçar
CNRS and LIP, ENS Lyon
Lyon 69364, France
bora.ucar@ens-lyon.fr

Abstract

Requirements for efficient parallelization of many complex and irregular applications can be cast as a hypergraph partitioning problem. The current-state-of-the art software libraries that provide tool support for the hypergraph partitioning problem are designed and implemented before the game-changing advancements in multi-core computing. Hence, analyzing the structure of those tools for designing multithreaded versions of the algorithms is a crucial tasks. The most successful partitioning tools are based on the multi-level approach. In this approach, a given hypergraph is coarsened to a much smaller one, a partition is obtained on the the smallest hypergraph, and that partition is projected to the original hypergraph while refining it on the intermediate hypergraphs. The coarsening operation corresponds to clustering the vertices of a hypergraph and is the most time consuming task in a multi- level partitioning tool. We present three efficient multithreaded clustering algorithms which are very suited for multi-level partitioners. We compare their performance with that of the ones currently used in today's hypergraph partitioners. We show on a large number of real life hypergraphs that our implementations, integrated into a commonly used partitioning library PaToH, achieve good speedups without reducing the clustering quality.

Multithreaded Algorithms for Maximum Matching in Bipartite Graphs

Ariful Azad[1], Mahantesh Halappanavar[2], Sivasankaran Rajamanickam[3], Erik G. Boman[3],
Arif Khan[1], and Alex Pothen[1]
E-mail: {aazad,khan58,apothen}@purdue.edu, mahantesh.halappanavar@pnnl.gov,
and {srajama,egboman}@sandia.gov
[1] Purdue University [2] Pacific Northwest National Laboratory [3] Sandia National Laboratories

Abstract

We design, implement, and evaluate algorithms for computing a matching of maximum cardinality in a bipartite graph on multicore and massively multithreaded computers. As computers with larger numbers of slower cores dominate the commodity processor market, the design of multithreaded algorithms to solve large matching problems becomes a necessity. Recent work on serial algorithms for the matching problem has shown that their performance is sensitive to the order in which the vertices are processed for matching. In a multithreaded environment, imposing a serial order in which vertices are considered for matching would lead to loss of concurrency and performance. But this raises the question: Would parallel matching algorithms on multithreaded machines improve performance over a serial algorithm? We answer this question in the affirmative. We report efficient multithreaded implementations of three classes of algorithms based on their manner of searching for augmenting paths: breadth-first-search, depth-first-search, and a combination of both. The Karp-Sipser initialization algorithm is used to make the parallel algorithms practical. We report extensive results and insights using three shared-memory platforms (a 48-core AMD Opteron, a 32-core Intel Nehalem, and a 128-processor Cray XMT) on a representative set of real-world and synthetic graphs. To the best of our knowledge, this is the first study of augmentation-based parallel algorithms for bipartite cardinality matching that demonstrates good speedups on multithreaded shared memory multiprocessors.

Session 20
Data Intensive and Peer-to-Peer Computing

Multi-level Layout Optimization for Efficient Spatio-temporal Queries on ISABELA-compressed Data

Zhenhuan Gong[1,2], Sriram Lakshminarasimhan[1,2], John Jenkins[1,2], Hemanth Kolla[3]
Stephane Ethier[4], Jackie Chen[3], Robert Ross[5], Scott Klasky[2], Nagiza F. Samatova[1,2,*]
[1] North Carolina State University, NC 27695, USA
[2] Oak Ridge National Laboratory, TN 37831, USA
[3] Sandia National Laboratory, Livermore, CA 94551, USA
[4] Princeton Plasma Physics Laboratory, Princeton, NJ 08543, USA
[5] Argonne National Laboratory, Argonne, IL 60439, USA
[*] Corresponding author: samatova@csc.ncsu.edu

Abstract

The size and scope of cutting-edge scientific simulations are growing much faster than the I/O subsystems of their runtime environments, not only making I/O the primary bottleneck, but also consuming space that pushes the storage capacities of many computing facilities. These problems are exacerbated by the need to perform data-intensive analytics applications, such as querying the dataset by variable and spatio- temporal constraints, for what current database technologies commonly build query indices of size greater than that of the raw data. To help solve these problems, we present a parallel query- processing engine that can handle both range queries and queries with spatio-temporal constraints, on B–spline compressed data with user-controlled accuracy. Our method adapts to widening gaps between computation and I/O performance by querying on compressed metadata separated into bins by variable values, utilizing Hilbert space-filling curves to optimize for spatial constraints and aggregating data access to improve locality of per-bin stored data, reducing the false positive rate and latency-bound I/O operations (such as seek) substantially. We show our method to be efficient with respect to storage, computation, and I/O compared to existing database technologies optimized for query processing on scientific data.

Evaluating Mesh-based P2P Video-on-Demand Systems

Yingwu Zhu
Department of Computer Science and Software Engineering
Seattle University
Seattle, WA, USA
Email: zhuy@seattleu.edu

Abstract

Three stakeholders come into play in peer-to- peer (P2P) video-on-demand (VoD), namely peers/viewers, content providers and ISPs. Different design choices have been proposed to improve quality of user experience, to bring down content server bandwidth cost and to reduce ISP-unfriendly traffic. However, it is unclear whether the ability of these design choices to meet interests of one stakeholder comes at the expense of their ability to satisfy the other stakeholders. Yet, another question remains open for two different types of mesh-based P2P VoD protocols: How well do the regular P2P VoD protocol and the network coding-based P2P VoD protocol perform compared to each other? In this paper, we present a simple performance versus cost framework (PCF) to evaluate impact of different design choices. Via detailed PCF- based simulations, we compare the two P2P VoD protocols and show that the two protocols each has its own territory on which it excels at efficiency. We reveal many important findings about P2P VoD design and hope they are useful to P2P VoD designers.

Query optimization and execution in a parallel analytics DBMS

Todd Eavis
Department of Computer Science
Concordia University
Montreal, Canada
Email: eavis@cs.concordia.ca

Ahmad Taleb
Department of Computer Science
Najran University
Najran, Saudi Arabia
Email: ahmadtaleb@hotmail.com

Abstract

Over the past 15 years, data warehousing and OLAP technologies have matured to the point whereby they have become a cornerstone for the decision making process in organizations of all sizes. With the underlying databases growing enormously in size, parallel DBM systems have become a popular target platform. Perhaps the most "obvious" approach to scalable warehousing is to combine a small collection of conventional relational DBMSs into a loosely connected parallel DBMS. Such systems, however, benefit little, if at all, from advances in OLAP indexing, storage, compression, modeling, or query optimization. In the current paper, we discuss a parallel analytics server that has been designed from the ground up as a high performance OLAP query engine. Moreover, its indexing and query processing model directly exploits an OLAP-specific algebra that enables performance optimizations beyond the reach of simple relational DBMS clusters. Taken together, the server provides class-leading query performance with the scalability of shared nothing databases and, perhaps most importantly, achieves this balance with a modest physical architecture.

Dynamic Message Ordering for Topic-Based Publish/Subscribe Systems

Roberto Baldoni, Silvia Bonomi, Marco Platania, Leonardo Querzoni
Dipartimento di Ingegneria Informatica Automatica e Gestionale "A. Ruberti"
University of Rome "La Sapienza", Rome, Italy
{baldoni–bonomi–platania–querzoni}@dis.uniroma1.it

Abstract

A distributed event notification service (ENS) is a middleware architecture commonly used to provide applications with scalable and robust publish/subscribe communi- cation primitives. A distributed ENS can route events toward subscribers using multiple paths with different lengths and latencies; as a consequence, subscribers can receive events out of order. In this paper, we propose a novel solution for out-of-order notification detection on top of an existing topic- based ENS. Our solution guarantees that events published on different topics will be either delivered in the same order to all the subscribers of those topics or tagged as out-of-order. The proposed algorithm is completely distributed and is able to scale with the system size while imposing a reasonable cost in terms of notification latency. Our solution improves the current state of the art solutions by dynamically handling subscriptions/unsubscriptions and by automatically adapting with respect to topic popularity changes.

Session 21
Disk and Memory Software Optimization

iHarmonizer: Improving the Disk Efficiency of I/O-intensive Multithreaded Codes

Yizhe Wang *
ECE Department
Wayne State University
Detroit, MI, 48202, USA
yizhe.wang@gmail.com

Kei Davis
CCS Division
Los Alamos National Laboratory
Los Alamos, NM 87545, USA
kei.davis@lanl.gov

Yuehai Xu, Song Jiang
ECE Department
Wayne State University
Detroit, MI, 48202, USA
{yhxu, sjiang}@wayne.edu

Abstract

Challenged by serious power and thermal constraints and limited by available instruction-level parallelism, processor designs have evolved to multi-core architectures. These architectures, many augmented with native simultaneous multithreading, are driving software developers to use multithreaded programs to exploit thread-level parallelism. While multithreading is well known to introduce concerns of data dependency and CPU load balance, less known is that the uncertainty of relative progress of thread execution can cause patterns of I/O requests, issued by different threads, to be effectively random and so significantly degrade hard-disk efficiency. This effect can severely offset the performance gains from parallel execution, especially for I/O-intensive programs. Retaining the benefits of multithreading while not losing I/O efficiency is an urgent and challenging problem. We propose a user-level scheme, iHarmonizer, to streamline the servicing of I/O requests from multiple threads in the OpenMP programs. Specifically, we use the compiler to insert code into OpenMP programs so that data usage can be transmitted at run time to a supporting run-time library that prefetches data in a disk- friendly way and coordinates threads' execution according to the availability of their requested data. Transparent to the programmer, iHarmonizer makes a multithreaded program I/O efficient while maintaining the benefits of parallelism. Our experiments show that iHarmonizer can significantly speed up the execution of a representative set of I/O-intensive scientific benchmarks.

Improving Parallel IO Performance of Cell-based AMR Cosmology Applications

Yongen Yu
Department of Computer Science
Illinois Institute of Technology
Chicago, IL
yyu22@iit.edu

Douglas H. Rudd
Yale Center for Astronomy and Astrophysics
Yale University
New Haven, CT
douglas.rudd@yale.edu

Zhiling Lan
Department of Computer Science
Illinois Institute of Technology
Chicago, IL
lan@iit.edu

Nickolay Y. Gnedin
Theoretical Astrophysics Group
Fermi National Accelerator Laboratory
Batavia, IL
gnedin@fnal.gov

Andrey Kravtsov
Department of Astronomy and Astrophysics
The University of Chicago
Chicago, IL
andrey@oddjob.uchicago.edu

Jingjin Wu
Department of Computer Science
Illinois Institute of Technology
Chicago, IL
jwu45@iit.edu

Abstract

To effectively model various regions with different resolutions, adaptive mesh refinement (AMR) is commonly used in cosmology simulations. There are two well-known numerical approaches towards the implementation of AMR- based cosmology simulations: block-based AMR and cell-based AMR. While many studies have been conducted to improve performance and scalability of block-structured AMR applications, little work has been done for cell-based simulations. In this study, we present a parallel IO design for cell-based AMR cosmology applications, in particular, the ART(Adaptive Refinement Tree) code. First, we design a new data format that incorporates a space filling curve to map between spatial and on-disk locations. This indexing not only enables concurrent IO accesses from multiple application processes, but also allows users to extract local regions without significant additional memory, CPU or disk space overheads. Second, we develop a flexible N-M mapping mechanism to harvest the benefits of N-N and N-1 mappings where N is number of application processes and M is a user-tunable parameter for number of files. It not only overcomes the limited bandwidth issue of an N-1 mapping by allowing the creation of multiple files, but also enables users to efficiently restart the application at a variety of computing scales. Third, we develop a user-level library to transparently and automatically aggregate small IO accesses per process to accelerate IO performance. We evaluate this new parallel IO design by means of real cosmology simulations on production HPC system at TACC. Our preliminary results indicate that it can not only provide the functionality required by scientists (e.g., effective extraction of local regions and flexible process-to- file mapping), but also significantly improve IO performance.

Identifying Opportunities for Byte-Addressable Non-Volatile Memory in Extreme-Scale Scientific Applications

Dong Li[†], Jeffrey S. Vetter[†], Gabriel Marin[†], Collin McCurdy[†], Cristian Cira, Zhuo Liu[*] and Weikuan Yu[*]
[†] Oak Ridge National Laboratory, Auburn University
{lid1,vetter,maring,cmcurdy}@ornl.gov, {cmc0031,zhuoliu,wkyu}@auburn.edu

Abstract

Future exascale systems face extreme power challenges. To improve power efficiency of future HPC systems, non-volatile memory (NVRAM) technologies are being investigated as potential alternatives to existing memories technologies. NVRAMs use extremely low power when in standby mode, and have other performance and scaling benefits. Although previous work has explored the integration of NVRAM into various architecture and system levels, an open question remains: do specific memory workload characteristics of scientific applications map well onto NVRAMs' features when used in a hybrid NVRAM-DRAM memory system? Furthermore, are there common classes of data structures used by scientific applications that should be frequently placed into NVRAM? In this paper, we analyze several mission-critical scientific applications in order to answer these questions. Specifically, we develop a binary instrumentation tool to statistically report memory access patterns in stack, heap, and global data. We carry out hardware simulation to study the impact of NVRAM for both memory power and system performance. Our study identifies many opportunities for using NVRAM for scientific applications. In two of our applications, 31% and 27% of the memory working sets are suitable for NVRAM. Our simulations suggest at least 27% possible power savings and reveal that the performance of some applications is insensitive to relatively long NVRAM write-access latencies.

NVMalloc: Exposing an Aggregate SSD Store as a Memory Partition in Extreme-Scale Machines

Chao Wang[1], Sudharshan S. Vazhkudai[1], Xiaosong Ma[1,2], Fei Meng[2], Youngjae Kim[1] , and Christian Engelmann[1]
[1] Oak Ridge National Laboratory, [2] North Carolina State University
{wangcn, vazhkudaiss, kimy1, engelmannc}@ornl.gov, fmeng@ncsu.edu, ma@cs.ncsu.edu

Abstract

DRAM is a precious resource in extreme-scale machines and is increasingly becoming scarce, mainly due to the growing number of cores per node. On future multi-petaflop and exaflop machines, the memory pressure is likely to be so severe that we need to rethink our memory usage models. Fortunately, the advent of non-volatile memory (NVM) offers a unique opportunity in this space. Current NVM offerings possess several desirable properties, such as low cost and power efficiency, but suffer from high latency and lifetime issues. We need rich techniques to be able to use them alongside DRAM. In this paper, we propose a novel approach for exploiting NVM as a secondary memory partition so that applications can explicitly allocate and manipulate memory regions therein. More specifically, we propose an NVMalloc library with a suite of services that enables applications to access a distributed NVM storage system. We have devised ways within NVMalloc so that the storage system, built from compute node-local NVM devices, can be accessed in a byte-addressable fashion using the memory mapped I/O interface. Our approach has the potential to re-energize out-of-core computations on large-scale machines by having applications allocate certain variables through NVMalloc, thereby increasing the overall memory capacity available. Our evaluation on a 128-core cluster shows that NVMalloc enables applications to compute problem sizes larger than the physical memory in a cost-effective manner. It can bring more performance/efficiency gain with increased computation time between NVM memory accesses or increased data access locality. In addition, our results suggest that while NVMalloc enables transparent access to NVM-resident variables, the explicit control it provides is crucial to optimize application performance.

Plenary Session
Best Papers

HierKNEM: An Adaptive Framework for Kernel-Assisted and Topology-Aware Collective Communications on Many-core Clusters

Teng Ma*, George Bosilca*, Aurelien Bouteiller*, Jack J. Dongarra†
* EECS, University of Tennessee
1122 Volunteer Blvd., Knoxville, TN 37996-3450, USA
Email: {tma, bosilca, bouteill}@eecs.utk.edu
† University of Tennessee
Oak Ridge National Laboratory, Oak Ridge, TN, USA
University of Manchester, Manchester, UK
Email: dongarra@eecs.utk.edu

Abstract

Multicore Clusters, which have become the most prominent form of High Performance Computing (HPC) systems, challenge the performance of MPI applications with non uniform memory accesses and shared cache hierarchies. Recent advances in MPI collective communications have alleviated the performance issue exposed by deep memory hierarchies by carefully considering the mapping between the collective topology and the core distance, as well as the use of single-copy kernel assisted mechanisms. However, on distributed environments, a single level approach cannot encompass the extreme variations not only in bandwidth and latency capabilities, but also in the aptitude to support duplex communications or operate multiple concurrent copies simultaneously. This calls for a collaborative approach between multiple layers of collective algorithms, dedicating to extracting the maximum degree of parallelism from the collective algorithm by consolidating the intra- and inter- node communications. In this work, we present HierKNEM a kernel-assisted topology-aware collective framework, and how this framework orchestrates the collaboration between multiple layers of collective algorithms. The resulting scheme enables perfect overlap of intra- and inter- node communications. We demonstrated experimentally, by considering three of the most used collective operations (Broadcast, Allgather and Reduction), that 1) this approach is immune to modifications of the underlying process- core binding; 2) it outperforms state-of-art MPI libraries (Open MPI, MPICH2 and MVAPICH2) demonstrating up to a 30x speedup for synthetic benchmarks, and up to a 3x acceleration for a parallel graph application (ASP); 3) it furthermore demonstrates a linear speedup with the increase of the number of cores per node, a paramount requirement for scalability on future many-core hardware.

BRISA: Combining Efficiency and Reliability in Epidemic Data Dissemination

Miguel Matos*, Valerio Schiavoni†, Pascal Felber†, Rui Oliveira*, Etienne Riviere†
* HASLab - High-Assurance Software Lab, INESC TEC & U. Minho, Portugal.
Email: {miguelmatos,rco}@di.uminho.pt
† University of Neuchâtel, Switzerland. Email: first.last@unine.ch

Abstract

There is an increasing demand for efficient and robust systems able to cope with today's global needs for intensive data dissemination, e.g., media content or news feeds. Unfortunately, traditional approaches tend to focus on one end of the efficiency/robustness design spectrum, by either leveraging rigid structures such as trees to achieve efficient distribution, or using loosely-coupled epidemic protocols to obtain robustness. In this paper we present B RISA, a hybrid approach combining the robustness of epidemic-based dissemination with the efficiency of tree-based structured approaches. This is achieved by having dissemination structures such as trees implicitly emerge from an underlying epidemic substrate by a judicious selection of links. These links are chosen with local knowledge only and in such a way that the completeness of data dissemination is not compromised, i.e., the resulting structure covers all nodes. Failures are treated as an integral part of the system as the dissemination structures can be promptly compensated and repaired thanks to the underlying epidemic substrate. Besides presenting the protocol design, we conduct an extensive evaluation in a real environment, analyzing the effectiveness of the structure creation mechanism and its robustness under faults and churn. Results confirm B RISA as an efficient and robust approach to data dissemination in the large scale.

Locality Principle Revisited: A Probability-Based Quantitative Approach

Saurabh Gupta, Ping Xiang, Yi Yang, Huiyang Zhou
Department of Electrical and Computer Engineering
North Carolina State University
Raleigh, USA
{sgupta12, pxiang, yyang14, hzhou}@ncsu.edu

Abstract

This paper revisits the fundamental concept of the locality of references and proposes to quantify it as a conditional probability: in an address stream, given the condition that an address is accessed, how likely the same address (temporal locality) or an address within its neighborhood (spatial locality) will be accessed in the near future. Based on this definition, spatial locality is a function of two parameters, the neighborhood size and the scope of near future, and can be visualized with a 3D mesh. Temporal locality becomes a special case of spatial locality with the neighborhood size being zero byte. Previous works on locality analysis use stack/reuse distances to compute distance histograms as a measure of temporal locality. For spatial locality, some ad-hoc metrics have been proposed as a quantitative measure. In contrast, our conditional probability-based locality measure has a clear mathematical meaning, offers justification for distance histograms, and provides a theoretically sound and unified way to quantify both temporal and spatial locality. The proposed locality measure clearly exhibits the inherent application characteristics, from which we can easily derive information such as the sizes of the working data sets and how locality can be exploited. We showcase that our quantified locality visualized in 3D-meshes can be used to evaluate compiler optimizations, to analyze the locality at different levels of memory hierarchy, to optimize the cache architecture to effectively leverage the locality, and to examine the effect of data prefetching mechanisms. A GPU-based parallel algorithm is also presented to accelerate the locality computation for large address traces.

Evaluating the Impact of TLB Misses on Future HPC Systems

Alessandro Morari*‡, Roberto Gioiosa*, Robert W. Wisniewski†, Bryan S. Rosenburg†,
Todd A. Inglett†, Mateo Valero*‡
* Barcelona Supercomputing Center
Barcelona, Spain
† IBM T. J. Watson Research Center
Yorktown Heights, NY, US
‡ Univesitat Politecnica de Catalunya
Barcelona, Spain
{alessandro.morari, roberto.gioiosa}@bsc.es, {bobww, rosnbrg, tinglett}@us.ibm.com, mateo@ac.upc.edu

Abstract

TLB misses have been considered an important source of system overhead and one of the causes that limit scalability on large supercomputers. This assumption lead to HPC lightweight kernel designs that usually statically map page table entries to TLB entries and do not take TLB misses. While this approach worked for petascale clusters, programming and debugging exascale applications composed of billions of threads is not a trivial task and users have started to explore novel programming models and tools, which require a richer system software support. In this study we present a quantitative analysis of the effect of TLB misses on current and future parallel applications at scale. To provide a fair evaluation, we compare a noiseless OS (CNK) with a custom version of the same OS capable of handling TLB misses on a BG/P system (up to 4096 cores). Our methodology follows a two-step approach: we first analyze the effects of TLB misses with a low-overhead, range-checking TLB miss handler, and then simulate a more complex TLB management system through TLB noise injection. We analyze the system behavior with different page sizes and increasing number of nodes and perform a sensitivity analysis. Our results show that the overhead introduced by TLB misses on complex HPC applications from the LLNL and ANL benchmarks is below 2% if the TLB pressure is contained and/or the TLB miss handler overhead is low, even with 1MB-pages and under large TLB noise injection. These results open the possibility of implementing richer OS memory management services to satisfy the requirements of future applications and users.

Session 22
Network Algorithms

Optimal algorithms and approximation algorithms for replica placement with distance constraints in tree networks

A. Benoit[1], H. Larchevêque[2], P. Renaud-Goud[1]
1. LIP, Ecole Normale Supérieure de Lyon, France, {Anne.Benoit|Paul.Renaud-Goud}@ens-lyon.fr
2. LABRI, University of Bordeaux I, France, hubert.larcheveque@labri.fr

Abstract

In this paper, we study the problem of replica placement in tree networks subject to server capacity and distance constraints. The client requests are known beforehand, while the number and location of the servers are to be determined. The Single policy enforces that all requests of a client are served by a single server in the tree, while in the Multiple policy, the requests of a given client can be processed by multiple servers, thus distributing the processing of requests over the platform. For the Single policy, we prove that all instances of the problem are NP-hard, and we propose approximation algorithms. The problem with the Multiple policy was known to be NP-hard with distance constraints, but we provide a polynomial time optimal algorithm to solve the problem in the particular case of binary trees when no request exceeds the server capacity.

On Nonblocking Multirate Multicast Fat-tree Data Center Networks with Server Redundancy

Zhiyang Guo and Yuanyuan Yang
Department of Electrical and Computer Engineering, Stony Brook University, Stony Brook, NY 11794, USA

Abstract

Fat-tree networks have been widely adopted as network topologies in data center networks (DCNs). However, it is costly for fat-tree DCNs to support nonblocking multicast communication, due to the large number of core switches required. Since multicast is an essential communication pattern in many cloud services and nonblocking multicast communication can ensure the high performance of such services, reducing the cost of nonblocking multicast fat-tree DCNs is very important. On the other hand, server redundancy is ubiquitous in today's data centers to provide high availability of services. In this paper, we explore server redundancy in data centers to reduce the cost of nonblocking multicast fat-tree data center networks (DCNs). First, we present a multirate network model that accurately describes the communication environment of the fat-tree DCNs. Then, we show that the sufficient number of core switches for nonblocking multicast communication under the multirate model can be significantly reduced in arbitrary 2-redundant fat-tree DCNs, i.e., each server has exactly one redundant backup in the data center. We generalize the result to practical fat-tree DCNs where servers may have different number of redundant backups depending on the availability requirements of services they provide, and show that a higher redundancy level further reduces the cost of nonblocking multicast fat-tree DCNs. Finally, we propose a multicast routing algorithm with linear time complexity to configure multicast connections in fat-tree DCNs.

Distributed Transactional Memory for General Networks

Gokarna Sharma, Costas Busch, and Srivathsan Srinivasagopalan
Department of Computer Science
Louisiana State University
Baton Rouge, LA 70803, USA
Email: {gokarna,busch}@csc.lsu.edu, ssrini1@tigers.lsu.edu

Abstract

We consider the problem of implementing transactional memory in large-scale distributed networked systems. We present and analyze Spiral, a novel distributed directory- based protocol for transactional memory. Spiral is designed for the data-flow distributed implementation of software transactional memory which supports three basic operations: publish, allowing a shared object to be inserted in the directory so that other nodes can find it; lookup, providing a read-only copy of the object to the requesting node; move, allowing the requesting node to write the object locally after the node gets it. The protocol runs on a hierarchical directory construction based on sparse covers, where clusters at each level are ordered to avoid race conditions while serving concurrent requests. Given a shared object the protocol maintains a directory path pointing to the object. The basic idea is to use "spiral" paths that grow outward to search for the directory path of the object in a bottom-up fashion. For general networks, this protocol guarantees an $O(\log^2 n \log D)$ approximation for move requests, where n is the number of nodes and D is the diameter of the network. It also guarantees poly- log approximation for lookup requests. To the best of our knowledge, this is the first consistency protocol for distributed transactional memory that achieves poly-log approximation in general networks.

On λ-Alert Problem

Marek Klonowski
Wroclaw University of Technology
Faculty of Fundamental Problems of Technology
Wroclaw, Poland
Email: Marek.Klonowski@pwr.wroc.pl

Dominik Pajak
INRIA Bordeaux Sud-Ouest
LaBRI
Talence, France
Email: Dominik.Pajak@labri.fr

Abstract

In this paper we introduce and analyse the λ-Alert problem: in a single hop radio network a subset of stations is activated. The aim of the protocol is to decide if the number of activated stations is greater or equal to λ. This problem is similar to the k-Selection problem. It can also be seen as an extension of the standard Alert problem. In our paper we consider the λ-Alert problem in various settings. We describe characteristics of oblivious and adaptive deterministic algorithms for the model with and without collision detection. We also show some results for randomized algorithms. In particular, we present a very efficient Las Vegas- type algorithm which is immune to an adversary.

Session 23
GPU Acceleration

Efficient Quality Threshold Clustering for Parallel Architectures

Anthony Danalis[*†], Collin McCurdy[†] and Jeffrey S. Vetter[†‡]
adanalis@eecs.utk.edu, {cmccurdy,vetter}@ornl.gov [*] University of Tennessee
Knoxville, Tennessee 37996, USA
[†] Oak Ridge National Laboratory
Oak Ridge, Tennessee 37831, USA
[‡] Georgia Institute of Technology
Atlanta, Georgia 30332, USA

Abstract

Quality Threshold Clustering (QTC) is an algorithm for partitioning data, in fields such as biology, where clustering of large data-sets can aid scientific discovery. Unlike other clustering algorithms, QTC does not require knowing the number of clusters a priori, however, its perceived need for high computing power often makes it an unattractive choice. This paper presents a thorough study of QTC. We analyze the worst case complexity of the algorithm and discuss methods to reduce it by trading memory for computation. We also demonstrate how the expected running time of QTC is affected by the structure of the input data. We describe how QTC can be parallelized, and discuss implementation details of our thread- parallel, GPU, and distributed memory implementations of the algorithm. We demonstrate the efficiency of our implementations through experimental data. We show how data sets with tens of thousands of elements can be clustered in a matter of minutes in a modern GPU, and seconds in a small scale cluster of multi-core CPUs, or multiple GPUs. Finally, we discuss how user selected parameters, as well as algorithmic and implementation choices, affect performance.

A Highly Parallel Reuse Distance Analysis Algorithm on GPUs

Huimin Cui
SKL Computer Architecture,
Institute of Computing Technology,
Beijing, China
cuihm@ict.ac.cn

Qing Yi
Department of Computer Science
CAS University of Texas at San Antonio
San Antonio, TX, USA
qingyi@cs.utsa.edu

Jingling Xue
School of Computer Science and Engineering
University of New South Wales
Sydney, NSW, Australia
jingling@cse.unsw.edu.au

Lei Wang
SKL Computer Architecture,
Institute of Computing Technology,
Beijing, China
wlei@ict.ac.cn

Yang Yang
SKL Computer Architecture,
CAS Institute of Computing Technology,
Beijing, China
yangyang@ict.ac.cn

Xiaobing Feng
SKL Computer Architecture,
CAS Institute of Computing Technology, CAS
Beijing, China
fxb@ict.ac.cn

Abstract

Reuse distance analysis is a runtime approach that has been widely used to accurately model the memory system behavior of applications. However, traditional reuse distance analysis algorithms use tree-based data structures and are hard to parallelize, missing the tremendous computing power of modern architectures such as the emerging GPUs. This paper presents a highly-parallel reuse distance analysis algorithm (HP-RDA) to speedup the process using the SPMD execution model of GPUs. In particular, we propose a hybrid data structure of hash table and local arrays to flatten the traditional tree representation of memory access traces. Further, we use a probabilistic model to correct any loss of precision from a straightforward parallelization of the original sequential algorithm. Our experimental results show that using an NVIDIA GPU, our algorithm achieves a factor of 20 speedup over the traditional sequential algorithm with less than 1% loss in precision.

Accelerating Large Scale Image Analyses on Parallel, CPU-GPU Equipped Systems

George Teodoro, Tahsin M. Kurc, Tony Pan, Lee A.D. Cooper, Jun Kong, Patrick Widener, and Joel H. Saltz
Center for Comprehensive Informatics
Emory University
Atlanta, GA 30322
{george.teodoro,tkurc,tony.pan,lee.cooper,jun.kong,patrick.widener,jhsaltz}@emory.edu

Abstract

The past decade has witnessed a major paradigm shift in high performance computing with the introduction of accelerators as general purpose processors. These computing devices make available very high parallel computing power at low cost and power consumption, transforming current high performance platforms into heterogeneous CPU-GPU equipped systems. Although the theoretical performance achieved by these hybrid systems is impressive, taking practical advantage of this computing power remains a very challenging problem. Most applications are still deployed to either GPU or CPU, leaving the other resource under- or un-utilized. In this paper, we propose, implement, and evaluate a performance aware scheduling technique along with optimizations to make efficient collaborative use of CPUs and GPUs on a parallel system. In the context of feature computations in large scale image analysis applications, our evaluations show that intelligently co-scheduling CPUs and GPUs can significantly improve performance over GPU-only or multi-core CPU-only approaches.

Radio Astronomy Beam Forming on Many-Core Architectures

Alessio Sclocco, Ana Lucia Varbanescu
Faculty of Sciences
Vrije Universiteit Amsterdam
Amsterdam, The Netherlands
a.sclocco@vu.nl, a.l.varbanescu@vu.nl

Jan David Mol, Rob V. van Nieuwpoort
ASTRON
Netherlands Institute for Radio Astronomy
Dwingeloo, The Netherlands
mol@astron.nl, nieuwpoort@astron.nl

Abstract

Traditional radio telescopes use large steel dishes to observe radio sources. The largest radio telescope in the world, LOFAR, uses tens of thousands of fixed, omni- directional antennas instead, a novel design that promises ground-breaking research in astronomy. Where traditional telescopes use custom-built hardware, LOFAR uses software to do signal processing in real time. This leads to an instrument that is inherently more flexible. However, the enormous data rates and processing requirements (tens to hundreds of teraflops) make this extremely challenging. The next-generation telescope, the SKA, will require exaflops. Unlike traditional instruments, LOFAR and SKA can observe in hundreds of directions simultaneously, using beam forming. This is useful, for example, to search the sky for pulsars (i.e. rapidly rotating highly magnetized neutron stars). Beam forming is an important technique in signal processing: it is also used in WIFI and 4G cellular networks, radar systems, and health-care microwave imaging instruments. We propose the use of many-core architectures, such as 48-core CPU systems and Graphics Processing Units (GPUs), to accelerate beam forming. We use two different frameworks for GPUs, CUDA and OpenCL, and present results for hardware from different vendors (i.e. AMD and NVIDIA). Additionally, we implement the LOFAR beam former on multi-core CPUs, using OpenMP with SSE vector instructions. We use auto-tuning to support different architectures and implementation frameworks, achieving both platform and performance portability. Finally, we compare our results with the production implementation, written in assembly and running on an IBM Blue Gene/P supercomputer. We compare both computational and power efficiency, since power usage is one of the fundamental challenges modern radio telescopes face. Compared to the production implementation, our auto-tuned beam former is 45–50 times faster on GPUs, and 2–8 times more power efficient. Our experimental results lead to the conclusion that GPUs are an attractive solution to accelerate beam forming.

Session 24
Interconnection Networks

Cross-layer Energy and Performance Evaluation of a Nanophotonic Manycore Processor System using Real Application Workloads

George Kurian, Chen Sun, Chia-Hsin Owen Chen, Jason E. Miller, Jurgen Michel, Lan Wei
Dimitri A. Antoniadis, Li-Shiuan Peh, Lionel Kimerling, Vladimir Stojanovic and Anant Agarwal
Massachusetts Institute of Technology, Cambridge, USA
{gkurian, sunchen, owenhsin, jasonm, jmichel, lanwei, antoniadis, peh, lckim, vlada, agarwal}@mit.edu

Abstract

Recent advances in nanophotonic device research have led to a proliferation of proposals for new architectures that employ optics for on-chip communication. However, since standard simulation tools have not yet caught up with these advances, the quality and thoroughness of the evaluations of these architectures have varied widely. This paper provides the first complete end-to-end analysis of an architecture using on- chip optical interconnect. This analysis incorporates realistic performance and energy models for both electrical and optical devices and circuits into a full-fledged functional simulator, thus enabling detailed analyses when running actual applications. Since on-chip optics is not yet mature and unlikely to see widespread use for several more years, we perform our analysis on a future 1000-core processor implemented in an 11nm technology node. We find that the proposed optical interconnect can provide between 1.8x and 4.8x better energy-delay product than conventional electrical-only interconnects. In addition, based on a detailed energy breakdown of all processor components, we conclude that athermal ring resonators and on-chip lasers that allow rapid power gating are key areas worthy of additional nanophotonic research. This will help guide future optical device research to the areas likely to provide the best payoff.

Exploring the Scope of the InfiniBand Congestion Control Mechanism

Ernst Gunnar Gran, Sven-Arne Reinemo,
Olav Lysne, Tor Skeie
Simula Research Laboratory, Fornebu, Norway
Email: {ernstgr, svenar, olavly, tskeie}@simula.no

Eitan Zahavi, Gilad Shainer
Mellanox Technologies, Israel/USA
Email: eitan@mellanox.co.il,
Shainer@Mellanox.com

Abstract

In a lossless interconnection network, network congestion needs to be detected and resolved to ensure high performance and good utilization of network resources at high network load. If no countermeasure is taken, congestion at a node in the network will stimulate the growth of a congestion tree that not only affects contributors to congestion, but also other traffic flows in the network. Left untouched, the congestion tree will block traffic flows, lead to underutilization of network resources and result in a severe drop in network performance. The InfiniBand standard specifies a congestion control (CC) mechanism to detect and resolve congestion before a congestion tree is able to grow and, by that, hamper the network performance. The InfiniBand CC mechanism includes a rich set of parameters that can be tuned in order to achieve effective CC. Even though it has been shown that the CC mechanism, properly tuned, is able to improve both throughput and fairness in an interconnection network, it has been questioned whether the mechanism is fast enough to keep up with dynamic network traffic, and if a given set of parameter values for a topology is robust when it comes to different traffic patterns, or if the parameters need to be tuned depending on the applications in use. In this paper we address both these questions. Using the three-stage fat-tree topology from the Sun Datacenter InfiniBand Switch 648 as a basis, and a simulator tuned against CC capable InfiniBand hardware, we conduct a systematic study of the efficiency of the InfiniBand CC mechanism as the network traffic becomes increasingly more dynamic. Our studies show that the InfiniBand CC, even when using a single set of parameter values, performs very well as the traffic patterns becomes increasingly more dynamic, outperforming a network without CC in all cases. Our results show throughput increases varying from a few percent, to a seventeen-fold increase.

DCAF - A Directly Connected Arbitration-Free Photonic Crossbar For Energy-Efficient High Performance Computing

Christopher Nitta, Matthew Farrens, and Venkatesh Akella
University of California, Davis
Davis, CA USA
Email: cjnitta@ucdavis.edu, farrens@cs.ucdavis.edu, akella@ucdavis.edu

Abstract

DCAF is a directly connected arbitration free photonic crossbar that is realized by taking advantage of multiple photonic layers connected with photonic vias. In order to evaluate DCAF we developed a detailed implementation model for the network and analyzed the power and performance on a variety of benchmarks, including SPLASH- 2 and synthetic traces. Our results demonstrate that the overhead required by arbitration is non-trivial, especially at high loads. Eliminating the need for arbitration, sizing the buffers carefully and retransmitting lost packets when there is contention results in a 44% reduction in average packet latency without additional power overhead. We also use an analytical model for ScaLAPACK QR decomposition and find that a 64 processor DCAF could outperform a 1024 node cluster connected with 40Gbps links on matrices up to ~500MB in size.

Designing Non-blocking Allreduce with Collective Offload on InfiniBand Clusters: A Case Study with Conjugate Gradient Solvers*

K. Kandalla[1], U. Yang[2], J. Keasler[2], T. Kolev[2], A. Moody[2], H. Subramoni[1], K. Tomko[3], J. Vienne[1], B. R. de Supinski[2] and D. K. Panda[1]

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

[2] Lawrence Livermore National Laboratory
Livermore, California
{yang11, keasler1, kolev1, moody20, bronis}
@llnl.gov

[3] Ohio Supercomputer Center
Columbus, Ohio
{ktomko}@osc.edu

Abstract

Scientists across a wide range of domains increasingly rely on computer simulation for their investigations. Such simulations often spend a majority of their run-times solving large systems of linear equations that require vast amounts of computational power and memory. It is hence critical to design solvers in a highly efficient and scalable manner. Hypre is a high performance, scalable software library that offers several optimized linear solver routines and pre-conditioners. In this paper, we study the characteristics of Hypre's Preconditioned Conjugate Gradient (PCG) solver algorithm. The PCG routine is known to spend a majority of its communication time in the MPI Allreduce operation to compute a global summation during the inner product operation. The MPI Allreduce is a blocking operation, whose latency is often a limiting factor to the overall efficiency of the PCG solver routine, and correspondingly the performance of simulations that rely on this solver. Hence, hiding the latency of the MPI Allreduce operation is critical towards scaling the PCG solver routine and improving the performance of many simulations. The upcoming revision of MPI, MPI-3, will provide support for non-blocking collective communication to enable latency-hiding. The latest InfiniBand adapter from Mellanox, ConnectX-2, enables offloading of generalized lists of commu- nication operations to the network interface. Such an interface can be leveraged to design non-blocking collective operations. In this paper, we design fully functional, scalable algorithms for the MPI_Iallreduce operation, based on the network offload technology. To the best of our knowledge, these network- offload-based algorithms are the first to be presented for the MPI_Iallreduce operation. Our designs scale beyond 512 processes and we achieve near perfect communication/computation overlap. We also re-design the PCG solver routine to leverage our proposed MPI_Iallreduce operation to hide the latency of the global reduction operations. We observe up to 21% improvements in the run-times of the PCG routine, when compared to the default PCG implementation in Hypre. We also note that about 16% of the overall benefits are due to overlapping the Allreduce operations.

Session 25
Software Reliablity

Taming of the Shrew: Modeling the Normal and Faulty Behaviour of Large-scale HPC Systems

Ana Gainaru
Computer Science Department
NCSA, UIUC, Urbana, IL, USA
againaru@illinois.edu

Franck Cappello
INRIA, France
UIUC, Urbana, IL, USA
fci@lri.fr

William Kramer
NCSA, UIUC, Urbana, IL, USA
wkramer@ncsa.illinois.edu

Abstract

HPC systems are complex machines that generate a huge volume of system state data called "events". Events are generated without following a general consistent rule and different hardware and software components of such systems have different failure rates. Distinguishing between normal system behaviour and faulty situation relies on event analysis. Being able to detect quickly deviations from normality is essential for system administration and is the foundation of fault prediction. As HPC systems continue to grow in size and complexity, mining event flows become more challenging and with the upcoming 10 Petaflop systems, there is a lot of interest in this topic. Current event mining approaches do not take into consideration the specific behaviour of each type of events and as a consequence, fail to analyze them according to their characteristics. In this paper we propose a novel way of characterizing the normal and faulty behaviour of the system by using signal analysis concepts. All analysis modules create ELSA (Event Log Signal Analyzer), a toolkit that has the purpose of modelling the normal flow of each state event during a HPC system lifetime, and how it is affected when a failure hits the system. We show that these extracted models provide an accurate view of the system output, which improves the effectiveness of proactive fault tolerance algorithms. Specifically, we implemented a filtering algorithm and short-term fault prediction methodology based on the extracted model and test it against real failure traces from a large-scale system. We show that by analyzing each event according to its specific behaviour, we get a more realistic overview of the entire system.

Meteor Shower: A Reliable Stream Processing System for Commodity Data Centers

Huayong Wang, Li-Shiuan Peh, Emmanouil Koukoumidis
Computer Science And Artificial Intelligence Lab
MIT
Email: huayongw@smart.mit.edu, {peh,koukou}@csail.mit.edu
Shao Tao, Mun Choon Chan
School of Computing
National University of Singapore
Email: {shaot,chanmc}@comp.nus.edu.sg

Abstract

Large-scale failures are commonplace in commodity data centers, the major platforms for Distributed Stream Processing Systems (DSPSs). Yet, most DSPSs can only handle single-node failures. Here, we propose Meteor Shower, a new fault-tolerant DSPS that overcomes large-scale burst failures while improving overall performance. Meteor Shower is based on checkpoints. Unlike previous schemes, Meteor Shower orchestrates operators' checkpointing activities through tokens. The tokens originate from source operators, trickle down the stream graph, triggering each operator that receives these tokens to checkpoint its own state1 . Meteor Shower is a suite of three new techniques: 1) source preservation, 2) parallel, asynchronous checkpointing, and 3) application-aware checkpointing. Source preservation allows Meteor Shower to avoid the overhead of redundant tuple saving in prior schemes; parallel, asynchronous checkpointing enables Meter Shower operators to continue processing streams during a checkpoint; while application-aware checkpointing lets Meteor Shower learn the changing pattern of operators' state size and initiate checkpoints only when the state size is minimal. All three techniques together enable Meteor Shower to improve throughput by 226% and lower latency by 57% vs prior state-of-the-art. Our results were measured on a prototype implementation running three real world applications in the Amazon EC2 Cloud.

Hybrid Transactions: Lock Allocation and Assignment for Irrevocability

Jaswanth Sreeram†
Intel Labs
jaswanth.sreeram@intel.com

Santosh Pande
College of Computing
Georgia Institute of Technology
santosh@cc.gatech.edu

Abstract

Irrevocable Software Memory Transactions provide a safe way to perform certain unrecoverable operations such as I/O or system calls inside transactions. In this paper we show that this notion of irrevocability has strong relevance and impact on the performance and throughput of a transactional program. In this paper we propose a compile-time analysis and transactional memory runtime system that allows several irrevocable transactions to execute concurrently with normal revocable transactions unlike state-of-the-art methods that allow for atmost one irrevocable transaction at a time. Our approach uses a compile-time analysis to derive a fine-grained lock-assignment scheme using a precise context-sensitive data structure analysis that is able to identify disjoint logical data structures. We describe the prototype implementation of our system in the LLVM compiler and TL2 STM system and evaluate the parallel performance of our system on high- contention transactional programs in the STAMP suite.

Profiling-based Adaptive Contention Management for Software Transactional Memory

Zhengyu He, Xiao Yu and Bo Hong
School of Electrical and Computer Engineering
Georgia Institute of Technology
Email:{zhengyu.he, xyu40, bohong}@gatech.edu

Abstract

In software transactional memory (STM) systems, the contention management (CM) policy decides what action to take when a conflict occurs. CM is crucial to the performance of STM systems. However, the performance of existing CMs is sensitive to transaction workload and system platforms. A static policy is therefore unsatisfactory. In this paper, we argue that adaptive contention management is necessary and feasible. We further present a profiling-based method that can choose a suitable CM for a given workload and system platform during run-time. We also propose to use logic-time (transactional commit or abort events) to measure the profiling length and compare it with the traditional physical-time-based method. Experimental results demonstrate that our proposed adaptive contention manager (ACM) outperforms static CMs across benchmarks and platforms. In particular, the ACM that uses the number of aborts for the profiling length performs better than others.

Session 26

Communication Protocols and Benchmarking Algorithms

HydEE: Failure Containment without Event Logging for Large Scale Send-Deterministic MPI Applications

Amina Guermouche*, Thomas Ropars†, Marc Snir‡, Franck Cappello*‡
* INRIA Saclay-Île de France, F-91405 Orsay, France
† Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
‡ University of Illinois at Urbana-Champaign, Urbana, IL, USA
Email: guermou@lri.fr, thomas.ropars@epfl.ch, snir@illinois.edu, fci@lri.fr

Abstract

High performance computing will probably reach exascale in this decade. At this scale, mean time between failures is expected to be a few hours. Existing fault tolerant protocols for message passing applications will not be efficient anymore since they either require a global restart after a failure (checkpointing protocols) or result in huge memory occupation (message logging). Hybrid fault tolerant protocols overcome these limits by dividing applications processes into clusters and applying a different protocol within and between clusters. Combining coordinated checkpointing inside the clusters and message logging for the inter-cluster messages allows confining the consequences of a failure to a single cluster, while logging only a subset of the messages. However, in existing hybrid protocols, event logging is required for all application messages to ensure a correct execution after a failure. This can significantly impair failure free performance. In this paper, we propose HydEE, a hybrid rollback-recovery protocol for send-deterministic message passing applications, that provides failure containment without logging any event, and only a subset of the application messages. We prove that HydEE can handle multiple concurrent failures by relying on the send- deterministic execution model. Experimental evaluations of our implementation of HydEE in the MPICH2 library show that it introduces almost no overhead on failure free execution.

Distributed Demand and Response Algorithm for Optimizing Social-Welfare in Smart Grid

Qifen Dong
College of Information Engineering
Zhejiang University of Technology
Hangzhou, China
qdong@cs.gsu.edu

Li Yu College of Information Engineering
Zhejiang University of Technology
Hangzhou, China
lyu@zjut.edu.cn

Wen-Zhan Song
Department of Computer Science
Georgia State University
Atlanta, USA
wsong@gsu.edu

Lang Tong
School of Electrical and Computer Engineering
Cornell University
Ithaca, NY 14853
ltong@ece.cornell.edu

Shaojie Tang
Department of Computer Science
Illinois Institute of Technology
Chicago, IL 60616
stang7@hawk.iit.edu

Abstract

This paper presents a distributed Demand and Response algorithm for smart grid with the objective of optimizing social-welfare. Assuming the power demand range is known or predictable ahead of time, our proposed distributed algorithm will calculate demand and response of all participating energy demanders and suppliers, as well as energy flow routes, in a fully distributed fashion, such that the social-welfare is optimized. During the computation, each node (e.g., demander or supplier) only needs to exchange limited rounds of messages with its neighboring nodes. It provides a potential scheme for energy trade among participants in the smart girds. Our theoretical analysis proves that the algorithm converges even if there is some random noise induced in the process of our distributed Lagrange-Newton based solution. The simulation also shows that the result is close to that of centralized solution.

Scalable Distributed Consensus to Support MPI Fault Tolerance

Darius Buntinas
Argonne National Laboratory
buntinas@mcs.anl.gov

Abstract

As system sizes increase, the amount of time in which an application can run without experiencing a failure decreases. Exascale applications will need to address fault tolerance. In order to support algorithm-based fault tolerance, communication libraries will need to provide fault-tolerance features to the application. One important fault-tolerance operation is distributed consensus. This is used, for example, to collectively decide on a set of failed processes. This paper describes a scalable, distributed consensus algorithm that is used to support new MPI fault-tolerance features proposed by the MPI 3 Forum's fault-tolerance working group. The algorithm was implemented and evaluated on a 4,096-core Blue Gene/P. The implementation was able to perform a full-scale distributed consensus in 222 μs and scaled logarithmically.

ScalaBenchGen: Auto-Generation of Communication Benchmarks Traces

Xing Wu
Department of Computer Science
North Carolina State University
Raleigh, NC, USA
Email: xwu3@ncsu.edu

Vivek Deshpande
Department of Computer Science
North Carolina State University
Raleigh, NC, USA
Email: vrdeshpa@ncsu.edu

Frank Mueller
Department of Computer Science
North Carolina State University
Raleigh, NC, USA
Email: mueller@cs.ncsu.edu

Abstract

Benchmarks are essential for evaluating HPC hardware and software for petascale machines and beyond. But benchmark creation is a tedious manual process. As a result, benchmarks tend to lag behind the development of complex scientific codes. This work contributes an automated approach to the cre- ation of communication benchmarks. Given an MPI applica- tion, we utilize ScalaTrace, a lossless and scalable framework to trace communication operations and execution time while abstracting away the computations. A single trace file that reflects the behavior of all nodes is subsequently expanded to C source code by a novel code generator. This resulting benchmark code is compact, portable, human-readable, and accurately reflects the original application's communication characteristics and runtime characteristics. Experimental results demonstrate that generated source code of benchmarks preserves both the communication patterns and the wallclock- time behavior of the original application. Such automatically generated benchmarks not only shorten the transition from application development to benchmark extraction but also facilitate code obfuscation, which is essential for benchmark extraction from commercial and restricted applications.

Session 27
Parallel Algorithms

A Self-Stabilization Process for Small-World Networks

Sebastian Kniesburges
Department of Computer Science
University of Paderborn
Email: seppel@upb.de

Andreas Koutsopoulos
Department of Computer Science
University of Paderborn
Email: koutsopo@mail.upb.de

Christian Scheideler
Department of Computer Science
University of Paderborn
Email: scheideler@mail.upb.de

Abstract

Small-world networks have received significant attention because of their potential as models for the interaction networks of complex systems. Specifically, neither random networks nor regular lattices seem to be an adequate framework within which to study real-world complex systems such as chemical-reaction networks, neural networks, food webs, social networks, scientific-collaboration networks, and computer networks. Small-world networks provide some desired properties like an expected polylogarithmic distance between two processes in the network, which allows routing in polylogarithmic hops by simple greedy routing, and robustness against attacks or failures. By these properties, small-world networks are possible solutions for large overlay networks comparable to structured overlay networks like CAN, Pastry, Chord, which also provide polylogarithmic routing, but due to their uniform structure, structured overlay networks are more vulnerable to attacks or failures. In this paper we bring together a randomized process converging to a small-world network and a self-stabilization process so that a small-world network is formed out of any weakly connected initial state. To the best of our knowledge this is the first distributed self-stabilization process for building a small-world network.

Self-organizing Particle Systems

Maximilian Drees*, Martina Hüllmann†, Andreas Koutsopoulos‡ and Christian Scheideler§
* Dept. of Computer Science, University of Paderborn, maxdrees@mail.uni-paderborn.de
† Dept. of Computer Science, University of Paderborn, martinah@uni-paderborn.de
‡ Dept. of Computer Science, University of Paderborn, koutsopo@uni-paderborn.de
§ Dept. of Computer Science, University of Paderborn, scheideler@uni-paderborn.de

Abstract

Nanoparticles are getting more and more in the focus of the scientific community since the potential for the development of very small particles interacting with each other and completing medical and other tasks is getting bigger year by year. In this work we introduce a distributed local algorithm for arranging a set of nanoparticles on the discrete plane into specific geometric shapes, for instance a rectangle. The concept of a particle we use can be seen as a simple mobile robot with the following restrictions: it can only view the state of robots it is physically connected to, is anonymous, has only a constant size memory, can only move by using other particles as an anchor point on which it pulls itself alongside, and it operates in Look-Compute- Move cycles. The main result of this work is the presentation of a random distributed local algorithm which transforms any given connected set of particles into a particular geometric shape. As an example we provide a version of this algorithm for forming a rectangle with an arbitrary predefined aspect ratio. To the best of our knowledge this is the first work that considers arrangement problems for these types of robots.

PARDA: A Fast Parallel Reuse Distance Analysis Algorithm

Qingpeng Niu
The Ohio State University
niuq@cse.ohio-state.edu

James Dinan
Argonne National Laboratory
dinan@anl.gov

Qingda Lu
Intel Corporation
qingda.lu@intel.com

P. Sadayappan
The Ohio State University
saday@cse.ohio-state.edu

Abstract

Reuse distance is a well established approach to characterizing data cache locality based on the stack histogram model. This analysis so far has been restricted to offline use due to the high cost, often several orders of magnitude larger than the execution time of the analyzed code. This paper presents the first parallel algorithm to compute accurate reuse distances by analysis of memory address traces. The algorithm uses a tunable parameter that enables faster analysis when the maximum needed reuse distance is limited by a cache size upper bound. Experimental evaluation using the SPEC CPU 2006 benchmark suite shows that, using 64 processors and a cache bound of 8 MB, it is possible to perform reuse distance analysis with full accuracy within a factor of 13 to 50 times the original execution times of the benchmarks.

A Lower Bound On Proximity Preservation by Space Filling Curves

Pan Xu
Industrial and Manufacturing Systems Engg.
Iowa State University
Ames, IA, USA
Email: panxu@iastate.edu

Srikanta Tirthapura
Electrical and Computer Engg.
Iowa State University
Ames, IA, USA
Email: snt@iastate.edu

Abstract

A space filling curve (SFC) is a proximity preserving mapping from a high dimensional space to a single dimensional space. SFCs have been used extensively in dealing with multi-dimensional data in parallel computing, scientific computing, and databases. The general goal of an SFC is that points that are close to each other in high-dimensional space are also close to each other in the single dimensional space. While SFCs have been used widely, the extent to which proximity can be preserved by an SFC is not precisely understood yet. We consider natural metrics, including the "nearest- neighbor stretch" of an SFC, which measure the extent to which an SFC preserves proximity. We first show a powerful negative result, that there is an inherent lower bound on the stretch of any SFC. We then show that the stretch of the commonly used Z curve is within a factor of 1.5 from the optimal, irrespective of the number of dimensions. Further we show that a very simple SFC also achieves the same stretch as the Z curve. Our results apply to SFCs in any dimension d such that d is a constant.

Session 28

Software Performance Analysis and Optimization

Modeling and Analyzing Key Performance Factors of Shared Memory MapReduce

Devesh Tiwari and Yan Solihin
Department of Electrical and Computer Engineering
North Carolina State University
{devesh.dtiwari,solihin}@ncsu.edu

Abstract

MapReduce parallel programming model has seen wide adoption in data center applications. Recently, lightweight, fast, in-memory MapReduce runtime systems have been proposed for shared memory systems. However, what factors affect performance and what performance bottlenecks exist for a given program, are not well understood. This paper builds an analytical model to capture key performance factors of shared memory MapReduce and investigates important performance trends and behavior. Our study discovers several important findings and implications for system designers, performance tuners, and programmers. Our model quantifies relative contribution of different key performance factors for both map and reduce phases, and shows that performance of MapReduce programs are highly input-content dependent. Our model reveals that performance is heavily affected by the order in which distinct keys are encountered during the Map phase, and the frequency of these distinct keys. Our model points out cases in which reduce phase time dominates the total execution time. We also show that data-structure and algorithm design choices affect map and reduce phases differently and sometimes affecting map phase positively while affecting reduce phase negatively. Finally, we propose an application classification framework that can be used to reason about performance bottlenecks for a given application.

Predicting Potential Speedup of Serial Code via Lightweight Profiling and Emulations with Memory Performance Model

Minjang Kim Pranith Kumar Hyesoon Kim
School of Computer Science, College of Computing
Georgia Institute of Technology
Atlanta, GA, USA
{minjang, pranith, hyesoon}@gatech.edu

Bevin Brett
Software and Services Group
Intel Corporation
Nashua, NH, USA
bevin.brett@intel.com

Abstract

We present Parallel Prophet, which projects potential parallel speedup from an annotated serial program before actual parallelization. Programmers want to see how much speedup could be obtained prior to investing time and effort to write parallel code. With Parallel Prophet, programmers simply insert annotations that describe the parallel behavior of the serial program. Parallel Prophet then uses lightweight interval profiling and dynamic emulations to predict potential performance benefit. Parallel Prophet models many realistic features of parallel programs: unbalanced workload, multiple critical sections, nested and recursive parallelism, and specific thread schedulings and paradigms, which are hard to model in previous approaches. Furthermore, Parallel Prophet predicts speedup saturation resulting from memory and caches by monitoring cache hit ratio and bandwidth consumption in a serial program. We achieve very small runtime overhead: approximately a 1.2-10 times slowdown and moderate memory consumption. We demonstrate the effectiveness of Parallel Prophet in eight benchmarks in the OmpSCR and NAS Parallel benchmarks by comparing our predictions with actual parallelized code. Our simple memory model also identifies performance limitations resulting from memory system contention.

Scalable Critical-Path Based Performance Analysis

David Böhme
and Felix Wolf
German Research School
for Simulation Sciences
52062 Aachen, Germany
Email: {d.boehme,f.wolf}@grs-sim.de

Bronis R. de Supinski
and Martin Schulz
Lawrence Livermore National Laboratory
Livermore, California
Email: {bronis,schulzm}@llnl.gov

Markus Geimer
Jülich Supercomputing Centre
52425 Jülich, Germany
Email: m.geimer@fz-juelich.de

Abstract

The critical path, which describes the longest execution sequence without wait states in a parallel program, identifies the activities that determine the overall program runtime. Combining knowledge of the critical path with traditional parallel profiles, we have defined a set of compact performance indicators that help answer a variety of important performance-analysis questions, such as identifying load imbalance, quantifying the impact of imbalance on runtime, and characterizing resource consumption. By replaying event traces in parallel, we can calculate these performance indicators in a highly scalable way, making them a suitable analysis instrument for massively parallel programs with thousands of processes. Case studies with real-world parallel applications confirm that–in comparison to traditional profiles–our indicators provide enhanced insight into program behavior, especially when evaluating partitioning schemes of MPMD programs.

FractalMRC: Online Cache Miss Rate Curve Prediction on Commodity Systems

Lulu He, Zhibin Yu, Hai Jin
Service Computing Technology and System Lab
Cluster and Grid Computing Lab
Huazhong University of Science and Technology
Wuhan, 430074, China
Emails: loloseed@gmail.com, {yuzhibin, hjin}@hust.edu.cn

Abstract

Shared caches in chip multi-processors (CMPs) have important benefits such as accelerating inter-core communication, yet the inherent cache contention among multiple processes on such architectures can significantly degrade performance. To address this problem, cache partitioning has been studied based on the prediction of the cache miss rate curve (MRC) of the concurrently running programs. On-line MRC prediction, however, either requires special hardware support or incurs a high overhead when conducted purely in software. This paper presents a new MRC prediction scheme based on a fractal model and hence called FractalMRC. It uses the easily available features in performance monitoring units of modern Intel processors and predicts the MRC of a running program with low overhead and high accuracy. No changes to applications and hardware are required. The prediction is validated against the measured results for 26 applications from SPEC CPU2006 benchmark suite. The highest prediction accuracy is 99.3%, the accuracy of 12 of the applications is over 80%, and the average is 76%. The cost of prediction is 2% slowdown on average. The new, efficient and accurate MRC prediction has enabled a dynamic technique to partition cache between pairs of applications at run time to match or exceed the best performance attainable with static cache partitioning.

Session 29

Performance Optimization Frameworks and Methods

Enabling In-situ Execution of Coupled Scientific Workflow on Multi-core Platform

Fan Zhang, Ciprian Docan, Manish Parashar
Center for Autonomic Computing
Rutgers University, Piscataway NJ, USA
{zhangfan,docan,parashar}@cac.rutgers.edu

Scott Klasky, Norbert Podhorszki, Hasan Abbasi
Oak Ridge National Laboratory
P.O. Box 2008, Oak Ridge, TN, 37831, USA
{klasky,pnorbert,habbasi}@ornl.gov

Abstract

Emerging scientific application workflows are composed of heterogeneous coupled component applications that simulate different aspects of the physical phenomena being modeled, and that interact and exchange significant volumes of data at runtime. With the increasing performance gap between on-chip data sharing and off-chip data transfers in current systems based on multicore processors, moving large volumes of data using communication network fabric can significantly impact performance. As a result, minimizing the amount of inter-application data exchanges that are across compute nodes and use the network is critical to achieving overall application performance and system efficiency. In this paper, we investigate the in-situ execution of the coupled components of a scientific application workflow so as to maximize on-chip exchange of data. Specifically, we present a distributed data sharing and task execution framework that (1) employs data-centric task placement to map computations from the coupled applications onto processor cores so that a large portion of the data exchanges can be performed using the intra-node shared memory, (2) provides a shared space programming abstraction that supplements existing parallel programming models (e.g., message passing) with specialized one-sided asynchronous data access operators and can be used to express coordination and data exchanges between the coupled components. We also present the implementation of the framework and its experimental evaluation on the Jaguar Cray XT5 at Oak Ridge National Laboratory.

GTI: A Generic Tools Infrastructure for Event-Based Tools in Parallel Systems

Tobias Hilbrich*, Matthias S. Müller*, Bronis R. de Supinski†, Martin Schulz† and Wolfgang E. Nagel*
* Technische Universität Dresden, ZIH
D-01062 Dresden, Germany,
Email: {tobias.hilbrich, matthias.mueller, wolfgang.nagel}@tu-dresden.de
† Lawrence Livermore National Laboratory
Livermore, CA 94551
Email: {bronis,schulzm}@llnl.gov

Abstract

Runtime detection of semantic errors in MPI applications supports efficient and correct large-scale application development. However, current approaches scale to at most one thousand processes and design limitations prevent increased scalability. The need for global knowledge for analyses such as type matching, and deadlock detection presents a major challenge. We present a scalable tool infrastructure – the Generic Tool Infrastructure (GTI) – that we will use to implement MPI runtime error detection tools and that applies to other use cases. GTI supports simple offloading of tool processing onto extra processes or threads and provides a tree based overlay network (TBON) for creating scalable tools that analyze global knowledge. We present its abstractions and code generation facilities that ease many hurdles in tool develop- ment, including wrapper generation, tool communication, trace reductions, and filters. GTI ultimately allows tool developers to focus on implementing tool functionality instead of the surrounding infrastructure. Further, we demonstrate that GTI supports scalable tool development through a lost message detector and a phase profiler. The former provides a more scalable implementation of important base functionality for MPI correctness checking, while the latter tool demonstrates that GTI can serve as the basis of further types of tools. Experiments with up to 2048 cores show that GTI's scalability features apply to both tools.

An Efficient Framework for Multi-dimensional Tuning of High Performance Computing Applications

Guojing Cong, Huifang Wen,
I-hsin Chung, David Klepacki
IBM TJ Watson Research Center
1101 Kitchawan Road
Yorktown Heights, NY, 10598, US
{gcong,hfwen,ihchung,klepacki}@us.ibm.com

Hiroki Murata, Yasushi Negishi
IBM Tokyo Research Laboratory
1623-14 Shimotsuruma
Yamato-shi, 242-8502 Japan
{MRTHRK,NEGISHI}@jp.ibm.com

Abstract

Deploying an application onto a target platform for high performance oftentimes demands manual tuning by experts. As machine architecture gets increasingly complex, tuning becomes even more challenging and calls for systematic approaches. In our earlier work we presented a prototype that combines efficiently expert knowledge, static analysis, and runtime observation for bottleneck detection, and employs refactoring and compiler feedback for mitigation. In this study, we develop a software tool that facilitates fast searching of bottlenecks and effective mitigation of problems from major dimensions of computing (e.g., computation, communication, and I/O). The impact of our approach is demonstrated by the tuning of the LBMHD code and a Poisson solver code, representing traditional scientific codes, and a graph analysis code in UPC, representing emerging programming paradigms. In the experiments, our framework detects with a single run of the application intricate bottlenecks of memory access, I/O, and communication. Moreover, the automated solution implementation yields significant overall performance improvement on the target platforms. The improvement for LBMHD is upto 45%, and the speedup for the UPC code is upto 5. These results suggest that our approach is a concrete step towards systematic tuning of high performance computing applications.

An SMT-Selection Metric to Improve Multithreaded Applications' Performance

Justin R. Funston, Kaoutar El Maghraoui, Joefon Jann, Pratap Pattnaik, Alexandra Fedorova*
IBM T.J Watson Research Center Simon Fraser University*

Abstract

Simultaneous multithreading (SMT) increases CPU utilization and application performance in many circumstances, but it can be detrimental when performance is limited by application scalability or when there is significant contention for CPU resources. This paper describes an SMT-selection metric that predicts the change in application performance when the SMT level and number of application threads are varied. This metric is obtained online through hardware performance counters with little overhead, and allows the application or operating system to dynamically choose the best SMT level. We have validated the SMT-selection metric using a variety of benchmarks that capture various application characteristics on two different processor architectures. Our results show that the SMT-selection metric is capable of predicting the best SMT level for a given workload in 90% of the cases. The paper also shows that such a metric can be used with a scheduler or application optimizer to help guide its optimization decisions.