Proceedings of

IEEE 27th International Symposium on Parallel and Distributed Processing

IPDPS 2013 Advance Program Abstracts

CPS
Conference Publishing Services

Los Alamitos, California
Washington • Tokyo

IEEE
computer
society

The papers in this book comprise the proceedings of the meeting mentioned on the cover and title page. They reflect the authors' opinions and, in the interests of timely dissemination, are published as presented and without change. Their inclusion in this publication does not necessarily constitute endorsement by the editors, the IEEE Computer Society, or the Institute of Electrical and Electronics Engineers, Inc.

IEEE Computer Society Order Number E4971
BMS Part Number CFP13023-PRT
ISBN 978-0-7685-4971-2

Additional copies may be ordered from:

IEEE Computer Society
Customer Service Center
10662 Los Vaqueros Circle
P.O. Box 3014
Los Alamitos, CA 90720-1314
Tel: + 1 800 272 6657
Fax: + 1 714 821 4641
http://computer.org/cspress
csbooks@computer.org

IEEE Service Center
445 Hoes Lane
P.O. Box 1331
Piscataway, NJ 08855-1331
Tel: + 1 732 981 0060
Fax: + 1 732 981 9667
http://shop.ieee.org/store/
customer-service@ieee.org

IEEE Computer Society
Asia/Pacific Office
Watanabe Bldg., 1-4-2
Minami-Aoyama
Minato-ku, Tokyo 107-0062
JAPAN
Tel: + 81 3 3408 3118
Fax: + 81 3 3408 3553
tokyo.ofc@computer.org

Individual paper REPRINTS may be ordered at: <reprints@computer.org>

Editorial production by Lisa O'Conner
Cover art production by Mark Bartosik

IEEE
computer
society

CPS
Conference Publishing Services

IEEE Computer Society
Conference Publishing Services **(CPS)**
http://www.computer.org/cps

2013 IEEE International Symposium on Parallel & Distributed Processing

IPDPS 2013

Table of Contents

Session 1: Checkpointing

Session 2: Cloud Computing

Session 3: Hybrid Systems

Session 4: Networks

Session 5: Graph Algorithms

Session 6: Numerical Analysis

Session 7: Parallel I/O and Server Software

Session 8: Parallel I/O and File Systems

Session 9: Potpourri Algorithms 1

Session 10: GPU Scheduling

Session 11: Fault Tolerance and Contention Resolution

Session 12: Communication and Routing 1

Session 13: Data Centers

Session 14: Energy Modeling and Scheduling

Session 15: Communication and Routing 2

Session 16: Peer to Peer Systems

Session 17: Programming Frameworks

Session 18: Scheduling 1

Plenary Session: Best Papers

Session 19: Scheduling 2

Session 20: GPU Software

Session 21: Scientific Computing

Session 22: Wireless and Sensor Systems

Session 23: Potpourri Algorithms 2

Session 24: Potpourri Applications

Session 25: Potpourri Systems

Session 26: Programming Frameworks

Session 1: Checkpointing

Adaptive Incremental Checkpointing via Delta Compression for Networked Multicore Systems

Itthichok Jangjaimon and Nian-Feng Tzeng

Center for Advanced Computer Studies
University of Louisiana at Lafayette
Lafayette, LA 70504
{ixj0704, tzeng}@cacs.louisiana.edu

Abstract

Check pointing has been widely adopted in support of fault-tolerance and job migration, with checkpoint files preferably kept also at remote storage to withstand unavailability/failures of local nodes in networked systems. Lately, I/O bandwidth to remote storage becomes the bottleneck for check pointing on a large-scale system. This paper proposes an adaptive incremental check pointing (AIC), aiming to reduce the check pointing file size considerably so that its involved overhead is lowered and thus the expected job turnaround time drops. Given production multicore systems are observed to have unused cores often available, we design AIC to make use of separate cores for carrying out multi-level check pointing with delta compression at desirable points of time adaptively. We develop a new Markov model for predicting the performance of such multi-level concurrent check pointing, with AIC performance evaluated using six SPEC benchmarks under various system sizes. AIC is observed to lower the normalized expected turnaround time substantially (by up to 47%) when compared to its static counterpart and a recent multi-level check pointing scheme with fixed checkpoint intervals.

Towards Scalable Checkpoint Restart: A Collective Inline Memory Contents Deduplication Proposal

Bogdan Nicolae

Exascale Systems Group
IBM Research, Ireland
bogdan.nicolae@ie.ibm.com

Abstract

With increasing scale and complexity of supercomputing and cloud computing architectures, faults are becoming a frequent occurrence. For a large class of applications that run for a long time and are tightly coupled, Checkpoint-Restart (CR) is the only feasible method to survive failures. However, exploding checkpoint sizes that need to be dumped to storage pose a major scalability challenge, prompting the need to reduce the amount of check pointing data. This paper contributes with a novel collective memory contents deduplication scheme that attempts to identify and eliminate duplicate memory pages before they are saved to storage. Unlike previous approaches that concentrate on the checkpoints of the same process, our approach identifies duplicate memory pages shared by different processes (regardless whether on the same or different node). We show both how to achieve such a global deduplication in a scalable fashion and how to leverage it effectively to optimize the data layout in such way that it minimizes I/O bottlenecks. Large scale experiments show significant reduction of storage space consumption and performance overhead compared to several state-of-art approaches, both in synthetic benchmarks and for a real life high performance computing application.

Optimizing Checkpoints Using NVM as Virtual Memory

Sudarsun Kannan, Ada Gavrilovska, and Karsten Schwan

College of Computing
Georgia Institute of Technology, Atlanta, Georgia, USA
sudarsun@gatech.edu,{ada, schwan}@cc.gatech.edu

Dejan Milojicic
HP Labs, Palo Alto, USA
dejan.milojicic@hp.com

Abstract

Rapid check pointing will remain key functionality for next generation high end machines. This paper explores the use of node-local nonvolatile memories (NVM) such as phase-change memory, to provide frequent, low overhead checkpoints. By adapting existing multi-level checkpoint techniques, we devise new methods, termed NVM-checkpoints, that efficiently store checkpoints on both local and remote node NVM. The checkpoint frequencies are guided by failure models that capture the expected accessibility of such data after failure. To lower overheads, NVM-checkpoints reduce the NVM and interconnect bandwidth used with a novel pre-copy mechanism, which incrementally moves checkpoint data from DRAM to NVM before a local checkpoint is started. This reduces local checkpoint cost by limiting the instantaneous data volume moved at checkpoint time, thereby freeing bandwidth for use by applications. In fact, the pre-copy method can reduce peak interconnect usage up to 46%. Since our approach treats NVM as memory rather than as 'Ram disk', pre-copying can be generalized to directly move data to remote NVMs. This results in 40% faster application execution times compared to asynchronous approaches not using pre-copying.

On Closed Nesting and Checkpointing in Fault-Tolerant Distributed Transactional Memory

Aditya Dhoke

ECE Dept.
Virginia Tech.
Email: adityad@vt.edu

Binoy Ravindran

ECE Dept.
Virginia Tech.
Email: binoy@vt.edu

Bo Zhang

ECE Dept.
Virginia Tech.
Email: alexzbzb@vt.edu

Abstract

We consider the closed nesting and checkpointing model for transactions in fault-tolerant distributed transactional memory (DTM). The closed nested model allows inner-nested transactions to be aborted (in the event of a transactional conflict) without aborting the parent transaction, while check pointing allows transactions to rollback to a previous execution state, potentially improving concurrency over flat nesting. We consider a quorum-based replicated model for fault-tolerant DTM, and present algorithms to support closed nesting and checkpointing. The algorithms use incremental validation to avoid communication overhead on commit, and ensure1-copy equivalence. Our experimental studies using a Java DTM implementation of the algorithms on micro and macro benchmarks reveal the conditions when they improve transactional throughput over flat nesting, and also their relative advantages and disadvantages.

Session 2: Cloud Computing

Reliable Service Allocation in Clouds

Olivier Beaumont, Lionel Eyraud-Dubois, and Hubert Larchevêque

INRIA Bordeaux – Sud-Ouest
University of Bordeaux
Email: olivier.beaumont@labri.fr, eyraud@labri.fr, hubert.larcheveque@labri.fr

Abstract

We consider several reliability problems that arise when allocating applications to processing resources in a Cloud computing platform. More specifically, we assume on the one hand that each computing resource is associated to a capacity constraint and to a probability of failure. On the other hand, we assume that each service runs as a set of independent instances of identical Virtual Machines, and that the Service Level Agreement between the Cloud provider and the client states that a minimal number of instances of the service should run with a given probability. In this context, given the capacity and failure probabilities of the machines, and the capacity and reliability demands of the services, the question for the cloud provider is to find an allocation of the instances of the services (possibly using replication) onto machines satisfying all types of constraints during a given time period. In this paper, our goal is to assess the impact of the reliability constraint on the complexity of resource allocation problems. We consider several variants of this problem, depending on the number of services and whether their reliability demand is individual or global. We prove several fundamental complexity results (#P' and NP-completeness results) and we provide several optimal and approximation algorithms. In particular, we prove that a basic randomized allocation algorithm, that is easy to implement, provides optimal or quasi-optimal results in several contexts, and we show through simulations that it also achieves very good results in more general settings.

Scaling and Scheduling to Maximize Application Performance within Budget Constraints in Cloud Workflows

Ming Mao

Department of Computer Science
University of Virginia
Charlottesville, VA 22904 USA
ming@cs.virginia.com

Marty Humphrey

Department of Computer Science
University of Virginia
Charlottesville, VA 22904 USA
humphrey@cs.virginia.edu

Abstract

It remains a challenge to provision resources in the cloud such that performance is maximized and financial cost is minimized. A fixed budget can be used to rent a wide variety of resource configurations for varying durations. The two steps - resource acquisition and scheduling/allocation - are dependent on each other and are particularly difficult when considering complex resource usage such as workflows, where task precedence need to be preserved and the budget constraint is assigned for the whole cloud application instead of every single job. The ability to acquire resources dynamically and trivially in the cloud - while being incredibly powerful and useful - exacerbates this particular resource acquisition and scheduling problem. In this paper, we design, implement and evaluate two auto-scaling solutions to minimize job turnaround time within budget constraints for cloud workflows. The scheduling-first algorithm distributes the application-wide budget to each individual job, determines the fastest execution plan and then acquires the cloud resources, while the scaling-first algorithm determines the size and the type of the cloud resources first and then schedules the workflow jobs on the acquired instances. The scaling-first algorithm shows better performance when the budget is low while the scheduling-first algorithm performs better when the budget is high. The two algorithms can reduce the job turnaround time by 9.6% - 45.2% compared to choosing a fixed general machine type. Moreover, they show good tolerance (between-10.2% and 16.7%) to inaccurate parameters (20% estimation error).

Optimizing Resource Allocation while Handling SLA Violations in Cloud Computing Platforms

Lionel Eyraud-Dubois and Hubert Larchevêque

INRIA Bordeaux – Sud-Ouest
University of Bordeaux
{lionel.eyraud-dubois|hubert.larcheveque}@labri.fr

Abstract

In this paper, we study a resource allocation problem in the context of Cloud Computing, in which a set of Virtual Machines (VM) has to be allocated on a set of Physical Machines (PM). Each VM has a given demand (e.g. CPU demand), and each PM has a capacity. However, VMsonly use a fraction of their demand. The aim is to exploit the difference between the demand of the VM and its actual resource usage, to achieve a higher utilization on the PMs. However, the resource consumption of the VMs might change over time (while staying under its original demand), implying sometimes expensive "SLA violations" when the demand of some VMs is not satisfied because of overloaded PMs. Thus, while optimizing the global resource utilization of the PMs, it is necessary to ensure that at any moment a VM's need evolves, a few number of migrations (moving a VM from PM to PM) is sufficient to find a new configuration in which all the VMs' consumptions are satisfied. We model this problem using a fully dynamic bin packing approach and we present an algorithm ensuring a global utilization of the resources of 66%. Moreover, each time a PM is overloaded, at most one migration is sufficient to fall back in a configuration with no overloaded PM, and at most 3 different PMs are concerned by required migrations that may occur to keep the global resource utilization correct. This allows the platform to be highly resilient to a great number of changes.

V-Cache: Towards Flexible Resource Provisioning for Multi-tier Applications in IaaS Clouds

Yanfei Guo, Palden Lama, Jia Rao, and Xiaobo Zhou

Department of Computer Science
University of Colorado, Colorado Springs, USA
Email addresses: {yguo, plama, jrao, xzhou}@uccs.edu

Abstract

Although the resource elasticity offered by Infrastructure-as-a-Service (IaaS) clouds opens up opportunities for elastic application performance, it also poses challenges to application management. Cluster applications, such as multi-tier websites, further complicates the management requiring not only accurate capacity planning but also proper partitioning of the resources into a number of virtual machines. Instead of burdening cloud users with complex management, we move the task of determining the optimal resource configuration for cluster applications to cloud providers. We find that a structural reorganization of multi-tier websites, by adding a caching tier which runs on resources debited from the original resource budget, significantly boosts application performance and reduces resource usage. We propose V-Cache, a machine learning based approach to flexible provisioning of resources for multi-tier applications in clouds. V-Cache transparently places a caching proxy in front of the application. It uses a genetic algorithm to identify the incoming requests that benefit most from caching and dynamically resizes the cache space to accommodate these requests. We develop a reinforcement learning algorithm to optimally allocate the remaining capacity to other tiers. We have implemented V-Cache on a VMware-based cloud testbed. Experiment results with the RUBiS and WikiBench benchmarks show that V-Cache outperforms a representative capacity management scheme and a cloud-cache based resource provisioning approach by at least 15% in performance, and achieves at least 11% and 21% savings on CPU and memory resources, respectively.

Session 3: Hybrid Systems

High-throughput Analysis of Large Microscopy Image Datasets on CPU-GPU Cluster Platforms

George Teodoro, Tony Pan, Tahsin M. Kurc, Jun Kong, Lee A.D. Cooper, Norbert Podhorszki, Scott Klasky, and Joel H. Saltz

Center for Comprehensive Informatics, Emory University, Atlanta, GA

Scientific Data Group, Oak Ridge National Laboratory, Oak Ridge, TN
glmteodoro@gmail.com

Abstract

Analysis of large pathology image datasets offers significant opportunities for the investigation of disease morphology, but the resource requirements of analysis pipelines limit the scale of such studies. Motivated by a brain cancer study, we propose and evaluate a parallel image analysis application pipeline for high throughput computation of large datasets of high resolution pathology tissue images on distributed CPU-GPU platforms. To achieve efficient execution on these hybrid systems, we have built runtime support that allows us to express the cancer image analysis application as a hierarchical data processing pipeline. The application is implemented as a coarse-grain pipeline of stages, where each stage may be further partitioned into another pipeline of fine-grain operations. The fine-grain operations are efficiently managed and scheduled for computation on CPUs and GPUs using performance aware scheduling techniques along with several optimizations, including architecture aware process placement, data locality conscious task assignment, data prefetching, and asynchronous data copy. These optimizations are employed to maximize the utilization of the aggregate computing power of CPUs and GPUs and minimize data copy overheads. Our experimental evaluation shows that the cooperative use of CPUs and GPUs achieves significant improvements on top of GPU-only versions (up to 1.6x) and that the execution of the application as a set of fine-grain operations provides more opportunities for runtime optimizations and attains better performance than coarser-grain, monolithic implementations used in other works. An implementation of the cancer image analysis pipeline using the runtime support was able to process an image dataset consisting of 36,848 4Kx4K-pixel image tiles (about 1.8TB uncompressed) in less than 4 minutes (150 tiles/second) on 100 nodes of a state-of-the-art hybrid cluster system.

High Performance FFT Based Poisson Solver on a CPU-GPU Heterogeneous Platform

Jing Wu

Department of Electrical and Computer Engineering
and Institute for Advanced Computer Studies
University of Maryland
College Park, MD
Email: jingwu@umiacs.umd.edu

Joseph Jaja

Department of Electrical and Computer Engineering
and Institute for Advanced Computer Studies
University of Maryland
College Park, MD
Email: joseph@umiacs.umd.edu

Abstract

We develop an optimized FFT based Poisson solver on a CPU-GPU heterogeneous platform for the case when the input is too large to fit on the GPU global memory. The solver involves memory bound computations such as 3D FFT in which the large 3D data may have to be transferred over the PCIe bus several times during the computation. We develop a new strategy to decompose and allocate the computation between the GPU and the CPU such that the 3D data is transferred only once to the device memory, and the executions of the GPU kernels are almost completely overlapped with the PCI data transfer. We were able to achieve significantly better performance than what has been reported in previous related work, including over 50 GFLOPS for the three periodic boundary conditions, and over 40 GFLOPS for the two periodic, one Neumann boundary conditions. The PCIe bus bandwidth achieved is over 5GB/s, which is close to the best possible on our platform. For all the cases tested, the single 3D PCIe transfer time, which constitutes a lower bound on what is possible on our platform, takes almost 70% of the total execution time of the Poisson solver.

Design and Implementation of the Linpack Benchmark for Single and Multi-node Systems Based on Intel Xeon Phi Coprocessor

Alexander Heinecke, Karthikeyan Vaidyanathan, Mikhail Smelyanskiy, Alexander Kobotov, Roman Dubtsov, Greg Henry, Aniruddha G. Shet, George Chrysos, and Pradeep Dubey

Department of Informatics, Technische Universitat Munchen, Munich, Germany

Parallel Computing Lab, Intel Corporation, Bangalore, India

Parallel Computing Lab, Intel Corporation, Santa Clara, USA

Software and Service Group, Intel Corporation, Novosibirsk, Russia

Software and Service Group, Intel Corporation, Hillsboro, USA

Intel Architecture Group, Intel Corporation, Hillsboro, USA

heinecke@in.tum.de

Abstract

Dense linear algebra has been traditionally used to evaluate the performance and efficiency of new architectures. This trend has continued for the past half decade with the advent of multi-core processors and hardware accelerators. In this paper we describe how several flavors of the Linpack benchmark are accelerated on Intel's recently released Intel(R) Xeon Phi(TM) co-processor (code-named Knights Corner) in both native and hybrid configurations. Our native DGEMM implementation takes full advantage of Knights Corner's salient architectural features and successfully utilizes close to 90% of its peak compute capability. Our native Linpack implementation running entirely on Knights Corner employs novel dynamic scheduling and achieves close to 80% efficiency - the highest published co-processor efficiency. Similarly to native, our single-node hybrid implementation of Linpack also achieves nearly 80% efficiency. Using dynamic scheduling and an enhanced look-ahead scheme, this implementation scales well to a 100-node cluster, on which it achieves over 76% efficiency while delivering the total performance of 107 TFLOPS.

Self-Adaptive OmpSs Tasks in Heterogeneous Environments

Judit Planas

Barcelona Supercomputing Center
Universitat Politecnica de Catalunya
judit.planas@bsc.es

Rosa M. Badia

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

Eduard Ayguadé and Jesús Labarta

Barcelona Supercomputing Center
Universitat Politecnica de Catalunya
eduard.ayguade@bsc.es
jesus.labarta@bsc.es

Abstract

As new heterogeneous systems and hardware accelerators appear, high performance computers can reach a higher level of computational power. Nevertheless, this does not come for free: the more heterogeneity the system presents, the more complex becomes the programming task in terms of resource management. OmpSs is a task-based programming model and framework focused on the runtime exploitation of parallelism from annotated sequential applications. This paper presents a set of extensions to this framework: we show how the application programmer can expose different specialized versions of tasks (i.e. pieces of specific code targeted and optimized for a particular architecture) and how the system can choose between these versions at run time to obtain the best performance achievable for the given application. From the results obtained in a multi-GPU system, we prove that our proposal gives flexibility to application's source code and can potentially increase application's performance.

Session 4: Networks

RAIR: Interference Reduction in Regionalized Networks-on-Chip

Lizhong Chen, Kai Hwang, and Timothy M. Pinkston

Ming Hsieh Department of Electrical Engineering
University of Southern California
Los Angeles, CA, USA
{lizhongc, kaihwang, tpink}@usc.edu

Abstract

With the advent of many-core systems capable of hosting multiple concurrently running applications, the traffic characteristics of networks-on-chip (NoCs) may exhibit new regional behaviors. By recognizing and exploiting these traffic behaviors, the effectiveness of NoC interference reduction techniques can be greatly improved. However, few works have investigated these regional behaviors and their potential impact on interference, leaving the opportunity largely unexplored. In this paper, we identify and characterize regional behavior in NoC and propose RAIR, a region-aware interference reduction technique that not only removes any restrictions on the inter-region traffic patterns, but also captures and exploits regional behavior throughout the design, thus improving the effectiveness of interference reduction. Evaluation using a cycle-accurate simulator shows that RAIR can improve the average packet latency by up to 17% on synthetic traffic patterns and up to 26% on PARSEC benchmarks compared to state-of-the-art interference reduction techniques.

An Analytical Performance Model for Partitioning Off-Chip Memory Bandwidth

Ruisheng Wang, Lizhong Chen, and Timothy Mark Pinkston

Ming Hsieh Department of Electrical Engineering
University of Southern California
Los Angeles, California, USA
{ruishenw,lizhongc,tpink}@usc.edu

Abstract

With the emergence of multi-programmed workloads for Chip Multiprocessors (CMP), Quality of Service (QoS) of each co-scheduled application on the CMP is increasingly gaining importance. As more and more applications are consolidated into a single chip to compete for the limited off-chip memory bandwidth, off-chip memory bandwidth partitioning makes an increasing impact on system performance. Although various existing heuristic-based memory scheduling schemes have achieved significant system performance improvement by better partitioning the bandwidth, it is still not clear what are the best ways to partition off-chip bandwidth for improving different system performance objectives. The goal of this paper is to understand how off-chip memory bandwidth partitioning affects various system performance objectives. To achieve this goal, we propose an analytical model that is simple yet powerful enough to reveal the relationship between various memory bandwidth partitioning schemes and different system performance objectives. From our model, optimal memory bandwidth partitioning schemes for different system-level objectives are derived. Experimental results from a cycle-accurate full-system simulator show that, for heterogeneous workloads, performance improvements over No_partitioning/Equal_partitioning in terms of harmonic weighted speedup, minimum fairness, weighted speedup and sum of IPCs are 20.3%/2.1%, 49.8%/38.7%, 32.8%/7.6% and 64.2%/24%, on average, with our corresponding optimal partitioning schemes (i.e., Square root, Proportional, Priority_APC, Priority_API), respectively.

A Case for Handshake in Nanophotonic Interconnects

Lei Wang, Jagadish Jayabalan, Minseon Ahn, Haiyin Gu, Ki Hwan Yum, and Eun Jung Kim

Department of Computer Science and Engineering, Texas A&M University,
College Station, TX, USA
Email: {wanglei, yum, ejkim}@cse.tamu.edu

Intel
Email: jagadish.c.jayabalan@intel.com

Samsung Electronics
Email: minseon0.ahn@samsung.com

Bloomberg L.P.
Email: hgu15@bloomberg.net

Abstract

Nanophotonics has been proposed to design low latency and high bandwidth NOC for future Chip Multi-Processors (CMPs). Recent nanophotonic NOC designs adopt the token-based arbitration coupled with credit-based flow control, which leads to low bandwidth utilization. In this work, we propose two handshake schemes for nanophotonic interconnects in CMPs, Global Handshake (GHS) and Distributed Handshake (DHS), which get rid of the traditional credit based flow control, reduce the average token waiting time, and finally improve the network throughput. Furthermore, we enhance the basic handshake schemes with seta side buffer and circulation techniques to overcome the Head-Of-Line (HOL)blocking. Our evaluation shows that the proposed handshake schemes improve network throughput by up to 62% under synthetic workloads. With the extracted trace traffic from real applications, the handshake schemes can reduce the communication delay by up to 59%. The basic handshake schemes add only 0.4% hardware overhead for optical components and negligible power consumption. In addition, the performance of the handshake schemes is independent of on-chip buffer space, which makes them feasible in a large scale nanophotonic interconnect design.

P-sync: A Photonically Enabled Architecture for Efficient Non-local Data Access

David Whelihan, Jeffrey J. Hughes, Scott M. Sawyer, Eric Robinson, Michael Wolf, Sanjeev Mohindra, Julie Mullen, Anna Klein, Michelle Beard, Nadya T. Bliss, Johnnie Chan, Robert Hendry, Keren Bergman, and Luca P. Carloni

Massachusetts Institute of Technology Lincoln Laboratory, Lexington, MA

Department of Electrical Engineering, Columbia University, New York, NY

Department of Computer Science, Columbia University, New York, NY

jeffrey.hughes@ll.mit.edu

Abstract

Communication in multi- and many-core processors has long been a bottleneck to performance due to the high cost of long-distance electrical transmission. This difficulty has been partially remedied by architectural constructs such as caches and novel interconnect topologies, albeit at a steep cost in terms of complexity. Unfortunately, even these measures are rendered ineffective by certain kinds of communication, most notably scatter and gather operations that exhibit highly non-local data access patterns. Much work has gone into examining how the increased bandwidth density afforded by chip-scale silicon photonic interconnect technologies affects computing, but photonics have additional properties that can be leveraged to greatly accelerate performance and energy efficiency under such difficult loads. This paper describes a novel synchronized global photonic bus and system architecture called P-sync that uses photonics' distance independence to greatly improve performance on many important applications previously limited by electronic interconnect. The architecture is evaluated in the context of a non-local yet common application: the distributed Fast Fourier Transform. We show that it is possible to achieve high efficiency by tightly balancing computation and communication latency in P-sync and achieve upwards of a 6x performance increase on gather patterns, even when bandwidth is equalized.

Session 5: Graph Algorithms

Optimizations and Analysis of BSP Graph Processing Models on Public Clouds

Mark Redekopp, Yogesh Simmhan, and Viktor K. Prasanna

University of Southern California, Los Angeles CA 90089
{redekopp, simmhan, prasanna}@usc.edu

Abstract

Large-scale graph analytics is a central tool in many fields, and exemplifies the size and complexity of Big Data applications. Recent distributed graph processing frameworks utilize the venerable Bulk Synchronous Parallel (BSP) model and promise scalability for large graph analytics. This has been made popular by Google's Pregel, which provides an architecture design for BSP graph processing. Public clouds offer democratized access to medium-sized compute infrastructure with the promise of rapid provisioning with no capital investment. Evaluating BSP graph frameworks on cloud platforms with their unique constraints is less explored. Here, we present optimizations and analyses for computationally complex graph analysis algorithms such as betweenness-centrality and all-pairs shortest paths on a native BSP framework we have developed for the Microsoft Azure Cloud, modeled on the Pregel graph processing model. We propose novel heuristics for scheduling graph vertex processing in swaths to maximize resource utilization on cloud VMs that lead to a 3.5x performance improvement. We explore the effects of graph partitioning in the context of BSP, and show that even a well partitioned graph may not lead to performance improvements due to BSP's barrier synchronization. We end with a discussion on leveraging cloud elasticity for dynamically scaling the number of BSP workers to achieve a better performance than a static deployment, and at a significantly lower cost.

Parallel Label-Setting Multi-objective Shortest Path Search

Peter Sanders

Institute for Theoretical Informatics
Karlsruhe Institute of Technology
Karlsruhe, Germany
Email: sanders@kit.edu

Lawrence Mandow

University of Malaga
Malaga, Spain
Email: lawrence@lcc.uma.es

Abstract

We present a parallel algorithm for finding all Pareto optimal paths from a specified source in a graph. The algorithm is label-setting, i.e., it only performs work on distance labels that are optimal. The main result is that the added complexity when going from one to multiple objectives is completely parallelizable. The algorithm is based on a multi-objective generalization of a priority queue. Such a Pareto queue can be efficiently implemented for two dimensions. Surprisingly, the parallel biobjective approach yields an algorithm performing asymptotically less work than the previous sequential algorithms. We also discuss generalizations for d&gt;= 3 objective functions and for single target search.

Multi-threaded Graph Partitioning

Dominique Lasalle and George Karypis

Department of Computer Science & Engineering
University of Minnesota
Minneapolis, Minnesota 55455, USA
{lasalle,karypis}@cs.umn.edu

Abstract

In this paper we explore the design space of creating a multi-threaded graph partitioner. We present and compare multiple approaches for parallelizing each of the three phases of multilevel graph partitioning: coarsening, initial partitioning, and uncoarsening. We also explore the differences in thread lifetimes and data ownership in this context. We show that despite the options for fine-grain synchronization and task decomposition offered by current threading technologies, the best performance is achieved by preserving data ownership and minimizing synchronization. In addition to this we also presentan unprotected approach to generating a vertex matching in parallel with little overhead. We use these findings to develop an OpenMP based implementation of the Metis algorithms and compare it against MPI based partitioners on three differentmulti-core architectures. Our multi-threaded implementation not only achieves greater than a factor of two speedup over the other partitioners, but also uses significantly less memory.

High-Productivity and High-Performance Analysis of Filtered Semantic Graphs

Aydin Buluç, Erika Duriakova, Armando Fox, John R. Gilbert, Shoaib Kamil, Adam Lugowski,
Leonid Oliker, and Samuel Williams

CRD, Lawrence Berkeley National Laboratory, Berkeley, USA

School of Computer Science and Informatics, University College Dublin, Ireland

Dept. of Computer Science, University of California, Santa Barbara, USA

EECS Dept, University of California, Berkeley, USA

CSAIL, Massachusetts Institute of Technology, Cambridge, USA

Corresponding authors: abuluc@lbl.gov, skamil@mit.edu, alugowski@cs.ucsb.edu

Abstract

High performance is a crucial consideration when executing a complex analytic query on a massive semantic graph. In a semantic graph, vertices and edges carry attributes of various types. Analytic queries on semantic graphs typically depend on the values of these attributes, thus, the computation must view the graph through a filter that passes only those individual vertices and edges of interest. Knowledge Discovery Toolbox (KDT), a Python library for parallel graph computations, is customizable in two ways. First, the user can write custom graph algorithms by specifying operations between edges and vertices. These programmer-specified operations are called semiring operations due to KDT's underlying linear-algebraic abstractions. Second, the user can customize existing graph algorithms by writing filters that return true for those vertices and edges the user wants to retain during algorithm execution. For high productivity, both semiring operations and filters are written in a high-level language, resulting in relatively low performance due to the bottleneck of having to call into the Python virtual machine for each vertex and edge. In this work, we use the Selective Embedded JIT Specialization (SEJITS) approach to automatically translate semiring operations and filters defined by programmers into a lower-level efficiency language, bypassing the up call into Python. We evaluate our approach by comparing it with the high-performance Combinatorial BLAS engine, and show our approach enables users to write in high-level languages and still obtain the high performance of low-level code. We also present a new roofline model for graph traversals, and show that our high-performance implementations do not significantly deviate from the roofline. Overall, we demonstrate the first known solution to the problem of obtaining high performance from a productivity language when applying graph algorithms selectively on semantic graphs.

Session 6: Numerical Analysis

Virtual Systolic Array for QR Decomposition

Jakub Kurzak, Piotr Luszczek, Mark Gates, and Ichitaro Yamazaki

University of Tennessee
Knoxville, TN 37996, USA
{kurzak, luszczek, mgates3, iyamazak}@eecs.utk.edu

Jack Dongarra

University of Tennessee, Knoxville, TN 37996, USA
Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA
University of Manchester, Manchester, M13 9PL, UK
dongarra@eecs.utk.edu

Abstract

Systolic arrays offer a very attractive, data centric, execution model as an alternative to the von Neumann architecture. Hardware implementations of systolic arrays turned out not to be viable solutions in the past. This article shows how the systolic design principles can be applied to a software solution to deliver an algorithm with unprecedented strong scaling capabilities. Systolic array for the QR decomposition is developed and a virtualization layer is used for mapping of the algorithm to a large distributed memory system. Strong scaling properties are discovered, superior to existing solutions.

Communication-Optimal Parallel Recursive Rectangular Matrix Multiplication

James Demmel, David Eliahu, Armando Fox, Shoaib Kamil, Benjamin Lipshitz,
Oded Schwartz, and Omer Spillinger

Mathematics Department and CS Division, UC Berkeley, Berkeley, CA 94720
demmel@cs.berkeley.edu, deliahu@berkeley.edu

EECS Department, UC Berkeley, Berkeley, CA 94720
fox@cs.berkeley.edu

CSAIL, MIT, Cambridge, MA 02139
skamil@mit.edu, lipshitz@cs.berkeley.edu, odedsc@cs.berkeley.edu, omers88@berkeley.edu

Abstract

Communication-optimal algorithms are known for square matrix multiplication. Here, we obtain the first communication-optimal algorithm for all dimensions of rectangular matrices. Combining the dimension-splitting technique of Frigo, Leiserson, Prokop and Ramachandran (1999) with the recursive BFS/DFS approach of Ballard, Demmel, Holtz, Lipshitz and Schwartz (2012) allows for a communication-optimal as well as cache- and network-oblivious algorithm. Moreover, the implementation is simple: approximately 50 lines of code for the shared-memory version. Since the new algorithm minimizes communication across the network, between NUMA domains, and between levels of cache, it performs well in practice on both shared- and distributed-memory machines. We show significant speedups over existing parallel linear algebra libraries both on a 32-core shared-memory machine and on a distributed-memory supercomputer.

Improving the Performance of the Symmetric Sparse Matrix-Vector Multiplication in Multicore

Theodoros Gkountouvas, Vasileios Karakasis, Kornilios Kourtis,
Georgios Goumas, and Nectarios Koziris

School of Electrical and Computer Engineering
National Technical University of Athens, Greece
E-mail: {bkk,goumas,nkoziris}@cslab.ece.ntua.gr

Department of Computer Science
Cornell University, Ithaca, NY, USA
E-mail: tg294@cornell.edu

Department of Computer Science
ETH, Zurich, Switzerland
E-mail: kkourt@inf.ethz.ch

Abstract

Symmetric sparse matrices arise often in the solution of sparse linear systems. Exploiting the non-zero element symmetry in order to reduce the overall matrix size is very tempting for optimizing the symmetric Sparse Matrix-Vector Multiplication kernel (SpMV) for multicore architectures. Despite being very beneficial for the single-threaded execution, not storing the upper or lower triangular part of a symmetric sparse matrix complicates the multithreaded SpMV version, since it introduces an undesirable dependency on the output vector elements. The most common approach for overcoming this problem is to use local, per-thread vectors, which are reduced to the output vector at the end of the computation. However, this reduction leads to considerable memory traffic, limiting the scalability of the symmetric SpMV. In this paper, we take a two-step approach in optimizing the symmetric SpMV kernel. First, we introduce the CSX-Sym variant of the highly compressed CSX format, which exploits the non-zero element symmetry for compressing further the input matrix. Second, we minimize the memory traffic produced by the local vectors reduction phase by implementing a non-zero indexing compression scheme that minimizes the local data to be reduced. Our indexing scheme allowed the scaling of symmetric SpMV and provided a more than 2$\times$ performance improvement over the baseline CSR implementation and 83.9% over the typical symmetric SpMV kernel. The CSX-Sym variant has further increased the symmetric SpMV performance by 43.4%. Finally, we evaluate the effect of our optimizations in the context of the CG iterative method, where we achieve an 77.8% acceleration of the overall solver.

Automated Rapid Prototyping of Regular Grid-Based Numerical Applications Using Generalized Elemental Subroutines

Yingchong Situ, Ye Wang, and Zhiyuan Li

Department of Computer Science, Purdue University
West Lafayette, Indiana 47907, United States
ysitu@cs.purdue.edu

Abstract

Computational scientists and engineers commonly rely on established software libraries to achieve high performance and reliability in their numerical applications. Unfortunately, this approach does not work well if the desired functionality is absent in existing libraries or if the integration is difficult. In such scenarios, one is often forced to explore alternative algorithms and in-house implementations. Such exploration can be a challenging task for computational scientists and engineers without sufficient computer science background. To address this issue, we design and build an automated rapid prototyping tool for regular grid-based numerical applications. This new tool allows programmers to specify algorithms as composition of familiar computation patterns such as those easily found in open literature expressed as generalized elemental subroutines. The tool then automatically transforms such subroutines into code which adapts to the prescribed data structures and delivers performance expected from the underlying algorithms. We demonstrate the tool in use cases including a production-grade computational fluid dynamic application.

Session 7: Parallel I/O and Server Software

A Transparent Collective I/O Implementation

Yongen Yu, Jingjin Wu, and Zhiling Lan

Department of Computer Science
Illinois Institute of Technology, Chicago, USA
{yyu22, jwu45, lan}@iit.edu

Douglas H. Rudd

Research Computing Center
University of Chicago, Chicago, USA
drudd@uchicago.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

Abstract

I/O performance is vital for most HPC applications especially those that generate a vast amount of data with the growth of scale. Many studies have shown that scientific applications tend to issue small and noncontiguous accesses in an interleaving fashion, causing different processes to access overlapping regions. In such scenario, collective I/O is a widely used optimization technique. However, the use of collective I/O deployed in existing MPI implementations is not trivial and sometimes even impossible. Collective I/O is an optimization based on a single collective I/O access. If the data reside in different places (e.g. in different arrays), the application has to maintain a buffer to first combine these data and then perform I/O operations on the buffer rather than the original data pieces. The process is very tedious for application developers. Besides, collective I/O requires the creating of a file view to describe the noncontiguous access patterns and additional coding is needed. Moreover, for the applications with complex data access using dynamic data sizes, it is hard or even impossible to use the file view mechanism to describe the access pattern through derived data types. In this study, we develop a user-level library called transparent collective I/O (TCIO) for application developers to easily incorporate collective I/O optimization into their applications. Preliminary experiments by means of a synthetic benchmark and a real cosmology application demonstrate that the library can significantly reduce the programming efforts required for application developers. Moreover, TCIO delivers better performance at large scales as compared to the existing collective functionality provided by MPI-IO.

A Visual Network Analysis Method for Large-Scale Parallel I/O Systems

Carmen Sigovan, Chris Muelder, and Kwan-Liu Ma

University of California Davis
{cmsigovan, cwmuelder, klma}@ucdavis.edu

Jason Cope, Kamil Iskra, and Robert Ross

Argonne National Laboratory
Mathematics and Computer Science Division
{copej, iskra, rross}@mcs.anl.gov

Abstract

Parallel applications rely on I/O to load data, store end results, and protect partial results from being lost to system failure. Parallel I/O performance thus has a direct and significant impact on application performance. Because supercomputer I/O systems are large and complex, one cannot directly analyze their activity traces. While several visual or automated analysis tools for large-scale HPC log data exist, analysis research in the high-performance computing field is geared toward computation performance rather than I/O performance. Additionally, existing methods usually do not capture the network characteristics of HPC I/O systems. We present a visual analysis method for I/O trace data that takes into account the fact that HPC I/O systems can be represented as networks. We illustrate performance metrics in a way that facilitates the identification of abnormal behavior or performance problems. We demonstrate our approach on I/O traces collected from existing systems at different scales.

FlexIO: I/O Middleware for Location-Flexible Scientific Data Analytics

Fang Zheng, Hongbo Zou, Greg Eisenhauer, Karsten Schwan, Matthew Wolf, Jai Dayal, Tuan-Anh Nguyen, Jianting Cao, Hasan Abbasi, Scott Klasky, Norbert Podhorszki, and Hongfeng Yu

College of Computing, Georgia Institute of Technology, Atlanta, GA, USA

Oak Ridge National Laboratory, Oak Ridge, TN, USA

Department of Computer Science and Engineering, University of Nebraska-Lincoln, Lincoln, NE, USA

fzheng@cc.gatech.edu

Abstract

Increasingly severe I/O bottlenecks on High-End Computing machines are prompting scientists to process simulation output data online while simulations are running and before storing data on disk. There are several options to place data analytics along the I/O path: on compute nodes, on separate nodes dedicated to analytics, or after data is stored on persistent storage. Since different placements have different impact on performance and cost, there is a consequent need for flexibility in the location of data analytics. The FlexIO middleware described in this paper makes it easy for scientists to obtain such flexibility, by offering simple abstractions and diverse data movement methods to couple simulation with analytics. Various placement policies can be built on top of FlexIO to exploit the trade-offs in performing analytics at different levels of the I/O hierarchy. Experimental results demonstrate that FlexIO can support a variety of simulation and analytics workloads at large scale through flexible placement options, efficient data movement, and dynamic deployment of data manipulation functionalities.

Burstiness-aware Server Consolidation via Queuing Theory Approach in a Computing Cloud

Zhaoyi Luo

State Key Laboratory for Novel Software Technology
Software Institute
Nanjing University, P.R.China
Email: luozy09@software.nju.edu.cn

Zhuzhong Qian

State Key Laboratory for Novel Software Technology
Department of Computer Science and Technology
Nanjing University, P.R.China
Email: qzz@nju.edu.cn

Abstract

Burstiness is a common pattern of virtual machines (VMs)'s workload in production data centers, where spikes usually occur a periodically with low frequency and last shortly. Since virtualization technology enables elastic resource provisioning in a computing cloud, the bursty workloads could be handled effectively through dynamically scaling up/down. However, to cut back energy consumption, VMs are usually highly consolidated with the minimum number of physical machines (PMs) used. In this case, to meet the runtime expanding demands of the resources (spikes), some VMs have to be migrated to other idle PMs, which is costly and causes performance degradation potentially. In this paper, we investigate the elastic resource provisioning problem and propose a novel VM consolidation mechanism with resource reservation which takes burstiness into consideration as well as energy consumption. We model the resource requirement pattern as the popular ON-OFF Markov chain to represent burstiness, based on which a reservation strategy via queuing theory approach is given for each PM. Next we present a complete VM consolidation scheme with resource reservation within reasonable time complexity. The experiment result show that our algorithms improve the consolidation ratio by up to 45% with large spike size and around 30% with normal spike size compared to those provisioning for peak workload, and a better balance of performance and energy consumption is achieved in comparison with other commonly used consolidation algorithms.

Session 8: Parallel I/O and File Systems

Pattern-Direct and Layout-Aware Replication Scheme for Parallel I/O Systems

Yanlong Yin, Jibing Li, Jun He, Xian-He Sun, and Rajeev Thakur

Computer Science Department
Illinois Institute of Technology, Chicago, Illinois 60616
Email: {yyin2, jli33, jhe24, sun}@iit.edu

Mathematics and Computer Science Division
Argonne National Laboratory, Argonne, Illinois 60439
Email: thakur@mcs.anl.gov

Abstract

The performance gap between computing power and the I/O system is ever increasing, and in the meantime more and more High Performance Computing (HPC) applications are becoming data intensive. This study describes an I/O data replication scheme, named Pattern-Direct and Layout-Aware (PDLA) data replication scheme, to alleviate this performance gap. The basic idea of PDLA is replicating identified data access pattern, and saving these reorganized replications with optimized data layouts based on access cost analysis. A runtime system is designed and developed to integrate the PDLA replication scheme and existing parallel I/O system, a prototype of PDLA is implemented under the MPICH2 and PVFS2 environments. Experimental results show that PDLA is effective in improving data access performance of parallel I/O systems.

Disk-Cache and Parallelism Aware I/O Scheduling to Improve Storage System Performance

Ramya Prabhakar, Mahmut Kandemir, and Myoungsoo Jung

NetApp, Inc. 2Department of CSE, Pennsylvania State University
{ramyap}@netapp.com, {kandemir, mj}@cse.psu.edu

Abstract

Modern large computing systems employ sophisticated disk I/O systems that are configured to deliver high throughput, low-latency disk I/O to multiple clients accessing them. However, due to potential interferences among concurrent I/O accesses issued by multiple clients, a disk-cache and disk-level parallelism unaware I/O scheduling algorithm employed by the operating system/storage controller may have a significant impact on both system throughput and I/O latency. In this paper, we propose two fundamentally new disk I/O scheduling techniques. The first technique, called DCAP, performs I/O scheduling in a disk cache aware and parallelism aware manner. The key idea in DCAP is to process simultaneous requests to different disks from the same application/priority class together and reorder them so that they have the highest number of hits in the disk cache. We then propose an enhanced version of DCAP called DCAP-G, that aggregates requests into service groups to alleviate the problem of request starvation that may occur in DCAP in certain cases. We evaluate both DCAP and DCAP-G using a set of I/O workloads from production-based enterprise systems as well as high-performance computing domain. In addition, we also compare the performance of our algorithms to previously proposed I/O scheduling algorithms. Our evaluation shows that, averaged across all our workloads, DCAP improves the average I/O response time, taking maximum advantage of disk access locality and exploiting parallelism among concurrent accesses to multiple disks, by 14.9% over an I/O scheduler that schedules requests on a first-come-first-served (FCFS) basis and also improves by 6.5% over a previously proposed locality-optimal I/O scheduler (SPCTF). In addition to these improvements, DCAP-G improves the average I/O response time by 6.6% over DCAP, leading to an overall 20.7% and 12.0% improvement over FCFS, and SPCTF, respectively.

Efficient and Scalable Retrieval Techniques for Global File Properties

*Dong H. Ahn, Michael J. Brim, Bronis R. de Supinski, Todd Gamblin, Gregory L. Lee,
Matthew P. Legendre, Barton P. Miller, Adam Moody, and Martin Schulz*

Lawrence Livermore National Laboratory, Computation Directorate,
Livermore, CA 94550, {ahn1, bronis, gamblin2, lee218, legendre1, moody20,
schulzm}@llnl.gov

University of Wisconsin, Computer Sciences Department,
Madison, WI 53706, {mjbrim, bart}@cs.wisc.edu

Abstract

Large-scale systems typically mount many different file systems with distinct performance characteristics and capacity. Applications must efficiently use this storage in order to realize their full performance potential. Users must take into account potential file replication throughout the storage hierarchy as well as contention in lower levels of the I/O system, and must consider communicating the results of file I/O between application processes to reduce file system accesses. Addressing these issues and optimizing file accesses requires detailed run-time knowledge of file system performance characteristics and the location(s) of files on them. In this paper, we propose Fast Global File Status (FGFS), a scalable mechanism to retrieve file information, such as its degree of distribution or replication and consistency. We use a novel node-local technique that turns expensive, non-scalable file system calls into simple string comparison operations. FGFS raises the namespace of a locally-defined file path to a global namespace with little or no file system calls to obtain global file properties efficiently. Our evaluation on a large multi-physics application shows that most FGFS file status queries on its executable and 848 shared library files complete in 272 milliseconds or faster at 32,768 MPI processes. Even the most expensive operation, which checks global file consistency, completes in under 7 seconds at this scale, an improvement of several orders of magnitude over the traditional checksum technique.

iBridge: Improving Unaligned Parallel File Access with Solid-State Drives

Xuechen Zhang, Ke Liu, Kei Davis, and Song Jiang

ECE Department
Wayne State University
Detroit, MI, 48202, US

School of Computer Science
Georgia Institute of Technology
Atlanta, GA, 30332, US

CCS Division
Los Alamos National Laboratory
Los Alamos, NM 87545, US

Abstract

When files are striped in a parallel I/O system, requests to the files are decomposed into a number of sub-requests that are distributed over multiple servers. If a request is not aligned with the striping pattern such decomposition can make the first and last sub-requests much smaller than the striping unit. Because hard-disk-based servers can be much less efficient in serving small requests than large ones, the system exhibits heterogeneity in serving sub-requests of different sizes, and the net throughput of the entire system can be severely degraded by the inefficiency of serving the smaller requests, or fragments. Because a request is not considered complete until its slowest sub-request is, the penalty is yet greater for synchronous requests. To make the situation even worse, the larger the request, or the more data servers the requested data is striped over, the larger the detrimental performance effect of serving fragments can be. This effect can become the Achilles' heel of a parallel I/O system performance seeking scalability with large sequential accesses. In this paper we propose iBridge, a scheme that uses solid-state drives to serve request fragments and thereby bridge the performance gap between serving fragments and serving large sub-requests. We have implemented iBridge in the PVFS file system. Our experimental results with representative MPI-IO benchmarks show that iBridge can significantly improve the I/O throughput of storage systems, especially for large requests with fragments.

Session 9: Potpourri Algorithms 1

Locally Self-Adjusting Tree Networks

Chen Avin, Bernhard Haeupler, Zvi Lotker, Christian Scheideler, and Stefan Schmid

Ben Gurion University, Israel; *{*avin,zvilo*}*@cse.bgu.ac.il

Massachusetts Institute of Technology (MIT), USA; hauepler@mit.edu

University of Paderborn, Germany; scheideler@upd.de

TU Berlin & Telekom Innovation Laboratories, Germany; stefan@net.t-labs.tu-berlin.de

Abstract

This paper initiates the study of self-adjusting networks (or distributed data structures) whose topologies dynamically adapt to a communication pattern σ. % (i.e., an ever changing “traffic matrix’’). We present a fully decentralized self-adjusting solution called \Splay Net. A \Splay Net\ is a distributed generalization of the classic splay tree concept. It ensures short paths (which can be found using local-greedy routing) between communication partners while minimizing topological rearrangements. We derive an upper bound for the amortized communication cost of a \Splay Net\based on empirical entropies of σ, and show that \Splay Nets\ have several interesting convergence properties. For instance, \Splay Nets\ features a provable online optimality under special requests scenarios. % and multicast tree scenarios We also investigate the optimal static network and prove different lower bounds for the average communication cost based on graph cuts and on the empirical entropy of the communication pattern σ. % which may be of independent interest. From these lower bounds it follows, e.g., that \Splay Nets\ are optimal in scenarios where the requests follow a product distribution as well. Finally, this paper shows that in contrast to the Minimum Linear Arrangement problem which is generally NP-hard, the optimal static tree network can be computed in polynomial time for any guest graph, despite the exponentially large graph family. We complement our formal analysis with a small simulation study on a Facebook graph.

A Network Configuration Algorithm Based on Optimization of Kirchhoff Index

Adam Hackett, Deepak Ajwani, Shoukat Ali, Steve Kirkland, and John P. Morrison

Hamilton Institute, National University of Ireland Maynooth, Ireland
{adam.hackett, stephen.kirkland}@nuim.ie

The Centre for Unified Computing, University College Cork, Ireland
{d.ajwani, j.morrison}@cs.ucc.ie

Exascale Systems Group, IBM Dublin Research and Development Lab, Ireland
shoukat.ali@ie.ibm.com

Abstract

Traditionally, a parallel application is partitioned, mapped and then routed on a network of compute nodes where the topology of the interconnection network is fixed and known beforehand. Such a topology often comes with redundant links to accommodate the communication patterns of a wide range of applications. With recent advances in technology for optical circuit switches, it is now possible to construct a network with much fewer links, and to make the link endpoints configurable to suit the communication pattern of a given application. While this is economical (saving both links and the power to run them), it raises the difficult problem of how to configure the network and how to reconfigure it quickly when the application's communication pattern changes. In this paper, we propose the Kirchhoff index (KI) of a certain weighted graph related to the interconnection network as a proxy for its communication throughput. Our usage of this metric is based on a theoretical analogy between resistances in an electrical network and communication loads in the interconnection network. We show how mathematical techniques for reducing KI can be used to configure a network in a dramatically shorter time as compared to the current state-of-the-art scheme.

Malleable Sorting

Patrick Flick, Peter Sanders, and Jochen Speck

Department of Informatics
Karlsruhe Institute of Technology
Karlsruhe, Germany
{sanders,speck}@kit.edu

Abstract

Malleable jobs can adapt to varying degrees of available parallelism. This is an interesting approach to more flexible usage of parallel resources. For example, malleable jobs can be scheduled optimally and efficiently where more restricted forms of parallel jobs are NP-hard to handle. However, little work has been done on how to make fundamental computations malleable. We study how this can be done for sorting. Our algorithm is an adaptive version of Multiway Merge Sort and outperforms a state-of-the art implementation in the multi core STL when the number of available cores fluctuates.

Adapting Particle Filter Algorithms to Many-Core Architectures

Mehdi Chitchian, Alexander S. van Amesfoort, Andrea Simonetto,
Tamás Keviczky, and Henk J. Sips

E-mail: mehdi.chitchian@gmail.com, {a.s.vanamesfoort, a.simonetto, t.keviczky,
h.j.sips}@tudelft.nl

Parallel and Distributed Systems Group, Delft University of Technology, Delft, The Netherlands

Delft Center for Systems and Control, Delft University of Technology, Delft, The Netherlands

Abstract

The particle filter is a Bayesian estimation technique based on Monte Carlo simulation. It is ideal for non-linear, non-Gaussian dynamical systems with applications in many areas, such as computer vision, robotics, and econometrics. Practical use has so far been limited, because of steep computational requirements. In this study, we investigate how to design a particle filter framework for complex estimation problems using many-core architectures. We develop a robotic arm application as a highly flexible estimation problem to push estimation rates and accuracy to new levels. By varying filtering and model parameters, we evaluate our particle filter extensively and derive rules of thumb for good configurations. Using our robotic arm application, we achieve a few hundred state estimations per second with one million particles. With our framework, we make a significant step towards a wider adoption of particle filters and enable studies into filtering setups for even larger estimation problems.

Session 10: GPU Scheduling

Guided Region-Based GPU Scheduling: Utilizing Multi-thread Parallelism to Hide Memory Latency

Jianmin Chen, Xi Tao, Zhen Yang, Jih-Kwon Peir, Xiaoyuan Li, and Shih-Lien Lu

Department of CISE
University of Florida
Gainesville, Florida, USA
{jichen, xtao, peir}@cise.ufl.edu

GPU Architecture
Nvidia Corp.
Santa Clara, California, USA
zhyang@nvidia.com

Department of ECE
University of Florida
Gainesville, Florida, USA
lixiaoyuan@ufl.edu

Intel Labs
Intel Corp.
Hillsboro, Oregon, USA
shih-lien.l.lu@intel.com

Abstract

Modern General-Purpose computation on Graphics Processing Units (GPGPUs) explore parallelism in applications by building massively parallel architecture and apply multithreading technology to hide the instruction and memory latencies. Such architectures become increasingly popular for parallel applications using CUDA/OpenCL programming languages. In this paper, we investigate thread scheduling algorithms on such highly-threaded GPGPUs. The traditional round-robin scheduling schemes are inefficient in handling instruction execution and memory accesses with disparate latencies. We introduce a new GPGPU thread (warp) scheduling algorithm which enables flexible round-robin distance for efficiently utilizing multithread parallelism and use program-guided priority shift among concurrent threads (warps) to allow more overlaps between short-latency compute instructions and long-latency memory accesses. Performance evaluations demonstrate that the new scheduling algorithm improves a set of kernel execution times by an average of 12% with 52% reduction on scheduler stall cycles over the fine-granularity round-robin scheme. In this paper, we also accomplish a thorough evaluation of various thread scheduling algorithms based on the amount of hardware threads, the scheduling overhead, and the global memory latency.

Optimizing and Auto-Tuning Iterative Stencil Loops for GPUs with the In-Plane Method

Wai Teng Tang, Wen Jun Tan, Ratna Krishnamoorthy, Yi Wen Wong, Shyh-Hao Kuo, Rick Siow Mong Goh, Stephen John Turner, and Weng-Fai Wong

School of Computer Engineering, Nanyang Technological University, Singapore

Department of Computer Science, School of Computing, National University of Singapore

Institute of High Performance Computing, Agency for Science, Technology and Research, Singapore

Email: wttang@ntu.edu.sg

Abstract

Stencils represent an important class of computations that are used in many scientific disciplines. Increasingly, many of the stencil computations in scientific applications are being offloaded to GPUs to improve running times. Since a large part of the simulation time is spent inside the stencil kernels, optimizing the kernel is therefore important in the context of achieving greater computation efficiencies and reducing simulation time. In this work, we proposed a novel in-plane method for stencil computations on GPUs and compared its performance with the conventional method implemented in the Nvidia SDK. We also implemented an auto-tuning framework for our method to select the optimal parameters for different GPU architectures. A performance model was developed for our proposed method, and is used to speed up the auto-tuning process. Our results show that a speedup of nearly 2X can be achieved compared to Nvidia's implementation.

Data-Driven Versus Topology-driven Irregular Computations on GPUs

Rupesh Nasre, Martin Burtscher, and Keshav Pingali

The University of Texas
Austin, Texas, USA
Email: nasre@ices.utexas.edu

Texas State University
San Marcos, Texas, USA
Email: burtscher@txstate.edu

The University of Texas
Austin, Texas, USA
Email: pingali@cs.utexas.edu

Abstract

Irregular algorithms are algorithms with complex main data structures such as directed and undirected graphs, trees, etc. A useful abstraction for many irregular algorithms is its operator formulation in which the algorithm is viewed as the iterated application of an operator to certain nodes, called active nodes, in the graph. Each operator application, called an activity, usually touches only a small part of the overall graph, so non-overlapping activities can be performed in parallel. In topology-driven implementations, all nodes are assumed to be active so the operator is applied everywhere in the graph even if there is no work to do at some nodes. In contrast, in data-driven implementations the operator is applied only to nodes at which there might be work to do. Multicore implementations of irregular algorithms are usually data-driven because current multicores only support small numbers of threads and work-efficiency is important. Conversely, many irregular GPU implementations use a topology-driven approach because work inefficiency can be counterbalanced by the large number of GPU threads. In this paper, we study data-driven and topology-driven implementations of six important graph algorithms on GPUs. Our goal is to understand the tradeoffs between these implementations and how to optimize them. We find that data-driven versions are generally faster and scale better despite the cost of maintaining a work list. However, topology-driven versions can be superior when certain algorithmic properties are exploited to optimize the implementation. These results led us to devise hybrid approaches that combine the two techniques and outperform both of them.

HQL: A Scalable Synchronization Mechanism for GPUs

Ayse Yilmazer and David Kaeli

Electrical and Computer Engineering Dept.
Northeastern University
Boston, MA
yilmazer@ece.neu.edu

Electrical and Computer Engineering Dept.
Northeastern University
Boston, MA
kaeli@ece.neu.edu

Abstract

Modern GPUs rely on atomic operations to perform global communication. These atomic operations can be used to construct finer-grained locks to provide support for mutual exclusion. However, equipped with only these basic synchronization primitives to support mutual exclusion results in inefficient use of resources. In this paper, we propose a new hardware-based blocking synchronization mechanism which uses hierarchical queuing for scalability and efficiency. We evaluate our design using a set of GPU applications for stressing synchronization mechanisms. We perform detailed simulation utilizing the Multi2Sim heterogeneous simulation infrastructure. Our results indicate that we can reduce the number of instructions executed by a GPU application by as much as 84%, while improving execution performance by as much as 73%.

Session 11: Fault Tolerance and Contention Resolution

Pluggable Watchdog: Transparent Failure Detection for MPI Programs

Keun Soo Yim, Zbigniew Kalbarczyk, and Ravishankar K. Iyer

Center for Reliable and High-performance Computing
Coordinated Science Laboratory
University of Illinois at Urbana-Champaign
Urbana, IL, 61801, USA
{yim6, kalbarcz, rkiyer}@illinois.edu

Abstract

This paper presents a framework and its techniques that can detect various types of runtime errors and failures in MPI programs. The presented framework offloads its detection techniques to an external device (e.g., extension card). By developing intelligence on the normal behavioral and semantic execution patterns of monitored parallel threads, the presented external error detectors can accurately and quickly detect errors and failures. This architecture allows us to use powerful detectors without directly using the computing power of the monitored system. The separation of hardware of the monitored and monitoring systems offers an extra advantage in terms of system reliability. We have prototyped our system on a parallel computer system by using an FPGA-based PCI extension card as a monitoring device. We have conducted a fault injection experiment to evaluate the presented techniques using eight MPI-based parallel programs. The techniques cover ~98.5% of faults, on average. The average performance overhead is 1.8% for techniques that detect crash and hang failures and 6.6% for techniques that detect SDC failures.

Improving the Computing Efficiency of HPC Systems Using a Combination of Proactive and Preventive Checkpointing

Mohamed Slim Bouguerra, Ana Gainaru, Leonardo Bautista Gomez, Franck Cappello, Satoshi Matsuoka, and Naoya Maruyam

INRIA/UIUC Joint Laboratory for Petascale Computing
Tokyo Institute of Technology

RIKEN AICS
slim.bouguerra@imag.fr

Abstract

As the failure frequency is increasing with the components count in modern and future supercomputers, resilience is becoming critical for extreme scale systems. The association of failure prediction with proactive check pointing seeks to reduce the effect of failures in the execution time of parallel applications. Unfortunately, proactive check pointing does not systematically avoid restarting from scratch. To mitigate this issue, failure prediction and proactive check pointing can be coupled with periodic check pointing. However, blind use of these techniques does not always improves system efficiency, because everyone of them comes with a mix of overheads and benefits. In order to study and understand the combination of these techniques and their improvement in the system's efficiency, we developed: (i) a prototype combining state of the art failure prediction, fast proactive check pointing and preventive check pointing, (ii) a mathematical model that reflects the expected computing efficiency of the combination and computes the optimal check pointing interval in this context, (iii) a discrete event simulator to evaluate the computing efficiency of the combination for system parameters corresponding to the current and projected large scale HPC systems. We evaluate our proposed technique on a large supercomputer (i.e. TSUBAME2) with production-level HPC applications and we show that failure prediction, proactive and preventive check pointing can be coupled successfully, imposing only about 2\% to 6\% of overhead in comparison with preventive check pointing only. Moreover, our model-based simulations show that the optimal solution improves the computing efficiency up to 30\% in comparison with classic periodic check pointing. We show that the prediction recall has a much higher impact on execution efficiency than the prediction precision. This result suggests that researchers on failure prediction algorithms should focus on improving the recall. We also show that the combination of these techniques can significantly improve (by a factor 2, for a particular configuration) the mean time between failures (MTBF) perceived by the application.

CASTED: Core-Adaptive Software Transient Error Detection for Tightly Coupled Cores

Konstantina Mitropoulou, Vasileios Porpodas, and Marcelo Cintra

School of Informatics
University of Edinburgh
{K.Mitropoulou@sms., v.porpodas@, mc@staffmail.}ed.ac.uk

Abstract

Aggressive silicon process scaling over the last years has made transistors faster and less power consuming. Meanwhile, transistors have become more susceptible to errors. The need to maintain high reliability has led to the development of various software-based error detection methodologies which target either single-core or multi-core processors. In this work, we present CASTED, a Core-Adaptive Software Transient Error Detection methodology that focuses on improving the impact of error detection overhead on single-chip scalable architectures that are composed of tightly coupled cores. The proposed compiler methodology adaptively distributes the error detection overhead to the available resources across multiple cores, fully exploiting the abundant ILP of these architectures. CASTED adapts to a wide range of architecture configurations (issue-width, inter-core delay). We evaluate our technique on a range of architecture configurations using the Mediabench II video and SPEC CINT2000 benchmark suites. Our approach successfully adapts to (and regularly outperforms by up to 21.2%) the best fixed state-of-the-art approach while maintaining the same fault coverage.

Contention Resolution in a Non-synchronized Multiple Access Channel

Gianluca De Marco

Universita di Salerno
84084 Fisciano (SA), Italy

Dariusz R. Kowalski

University of Liverpool
L69 3BX Liverpool, UK

Abstract

Multiple access channel is a well-known communication model that deploys properties of many network systems, such as Aloha multi-access systems, local area Ethernet networks, satellite communication systems, packet radio networks. The fundamental aspect of this model is to provide efficient communication and computation in the presence of restricted access to the communication resource: at most one station can successfully transmit at a time, and a wasted round occurs when more than one station attempts to transmit at the same time. In this work we consider the problem of contention resolution in a multiple access channel in a realistic scenario when up to k stations out of n join the channel at different times. The goal is to let at least one station to transmit alone, which results in successful delivery of the message through the channel. We present three deterministic algorithms: two of them working under some constrained scenarios, and achieving asymptotically optimal time complexity $\Theta(k\log(n/k))$, while the third general algorithm accomplishes the goal in time $O(k\log n \log\log n)$.

Session 12: Communication and Routing 1

Generalized Hierarchical All-to-All Exchange Patterns

Bogdan Prisacari, German Rodriguez, and Cyriel Minkenberg

IBM Research Ruschlikon, Switzerland
bpr@zurich.ibm.com

Abstract

The personalized all-to-all collective exchange is one of the most challenging communication patterns in HPC applications in terms of performance and scalability. We present a framework for the design of optimized collective patterns for generic hierarchical topologies. Our proposal can be applied, among others, to two types of topologies of great importance today: (i) the family of extended generalized fat tree networks (including k-ary n-trees and their variations) which are extensively used today in both HPC and commercial data centers, and (ii) direct low-diameter scalable hierarchical architectures such as the recently proposed dragonfly networks. We argue that exchange patterns that are congruent with the underlying structure of the network have inherent advantages compared to patterns that are oblivious to this structure. However, the current commonly used hierarchical pattern, the XOR exchange, has limited applicability, because it requires that the number of communicating nodes equals an integral power of two, making it suitable only for few tree designs and unsuitable for any dragonfly network. We propose several new, generic, universally applicable approaches to perform such exchanges in a hierarchical fashion that outperform current state of the art approaches. We support our claims by means of both mathematical proofs and simulation results that show that we can achieve an improvement of almost two-fold in dragonflies, and a two-to three-fold improvement in fat tree networks in cases where the XOR exchange cannot be applied.

Minimizing Communication in All-Pairs Shortest Paths

Edgar Solomonik

Univ. of California, Berkeley
Department of EECS
solomon@eecs.berkeley.edu

Aydin Buluç

Lawrence Berkeley Nat. Lab.
Computational Research Division
abuluc@lbl.gov

James Demmel

Univ. of California, Berkeley
Department of EECS
demmel@eecs.berkeley.edu

Abstract

We consider distributed memory algorithms for the all-pairs shortest paths (APSP) problem. Scaling the APSP problem to high concurrencies requires both minimizing inter-processor communication as well as maximizing temporal data locality. The 2.5D APSP algorithm, which is based on the divide-and-conquer paradigm, satisfies both of these requirements: it can utilize any extra available memory to perform asymptotically less communication, and it is rich in semiring matrix multiplications, which have high temporal locality. We start by introducing a block-cyclic 2D (minimal memory) APSP algorithm. With a careful choice of block-size, this algorithm achieves known communication lower-bounds for latency and bandwidth. We extend this 2D block-cyclic algorithm to a 2.5D algorithm, which can use c extra copies of data to reduce the bandwidth cost by a factor of sqrt(c), compared to its 2D counterpart. However, the 2.5Dalgorithm increases the latency cost by sqrt(c). We provide a tighter lower bound on latency, which dictates that the latency overhead is necessary to reduce bandwidth along the critical path of execution. Our implementation achieves impressive performance and scaling to 24,576 cores of a Cray XE6 supercomputer by utilizing well-tuned intra-node kernels within the distributed memory algorithm.

Programmable and Scalable Reductions on Clusters

Jan Ciesko, Javier Bueno, Nikola Puzovic, Alex Ramirez, Rosa M. Badia, and Jesús Labarta

Barcelona Supercomputing Center, Barcelona, Spain
{jan.ciesko, nikola.puzovic, rosa.m.badia, alex.ramirez, jesus.labarta}@bsc.es

Universitat Politecnica de Catalunya, Spain

Artificial Intelligence Research Institute (IIIA) - Spanish National Research Council (CSIC), Spain

Abstract

Reductions matter and they are here to stay. Wide adoption of parallel processing hardware in a broad range of computer applications has encouraged recent research efforts on their efficient parallelization. Furthermore, trends towards high productivity languages in mainstream computing increases the demand for efficient programming support. In this paper we present a new approach on parallel reductions for distributed memory systems that provides both scalability and programmability. Using OmpSs, a task-based parallel programming model, the developer has the ability to express scalable reductions through a single pragma annotation. This pragma annotation is applicable for tasks as well as for work-sharing constructs (with implicit tasking) and instructs the compiler to generate the required runtime calls. The supporting runtime handles data and task distribution, parallel execution and data reduction. Scalability is achieved through a software cache that maximizes local and temporal data reuse and allows overlapped computation and communication. Results confirm scalability for up to 32 12-core cluster nodes.

JVM-Bypass for Efficient Hadoop Shuffling

Yandong Wang, Cong Xu, Xiaobing Li, and Weikuan Yu

Department of Computer Science, Auburn University, AL 36849, USA
{wangyd,congxu,xbli,wkyu}@auburn.edu

Abstract

Hadoop employs Java-based network transport stack on top of the Java Virtual Machine (JVM) for its data shuffling and merging purposes. Our examination reveals that JVMintroduces a significant amount of overhead to data processing capability of the native interface. Furthermore, JVM constrains the use of high-performance networking mechanisms such as RDMA (Remote Direct Memory Access) which has established itself as an effective data movement technology in many networking environments because of its low-latency, high bandwidth, low CPU utilization, and energy efficiency. In this paper, we introduce a plug-in library called JVM-Bypass Shuffling (JBS) for Hadoopdata shuffling. JBS helps Hadoop data shuffling by avoiding Java-based transport protocols, removing the overhead and limitations of the JVM. In addition, we design JBS as a portable library that can leverage both TCP/IP and RDMA on different network systems such as InfiniBand and 1/10 Gigabit Ethernet. We have designed and implemented JBS as part of Hadoop acceleration. It has been transferred to Mellanox as the software product UDA(Unstructured Data Accelerator) and used to enable our studies on a variety of merging algorithms. Our performance evaluation demonstrates that JBS can effectively reduce the execution time of Hadoop jobs by up to 66.3% and lower the CPU utilization by 48.1%.

Session 13: Data Centers

Oversubscription Bounded Multicast Scheduling in Fat-Tree Data Center Networks

Zhiyang Guo, Jun Duan, and Yuanyuan Yang

Department of Electrical and Computer Engineering, Stony Brook University, Stony Brook, NY 11794, USA
yuanyuan.yang@stonybrook.edu

Abstract

Multicast benefits numerous data center applications that require group communication by eliminating sending unnecessary duplicated packets in the network, thus significantly reduces network traffic and improves application throughput. Meanwhile, many data center networks (DCNs) adopt a multi-rooted tree structure called fat-tree, which utilizes rich path multiplicity to deliver high bisection bandwidth. However, currently there is no efficient flow scheduling algorithm for the fat-tree that can route multicast flows appropriately to achieve traffic load balance, thus cannot fully take advantage of this high degree of link parallelism. Besides low bandwidth utilization, unbalanced traffic load distribution also leads to unpredictable network performance and degraded data center agility. In this paper, we study multicast traffic load balance problem in fat-tree DCNs. First, we derive a minimum link oversubscription upper bound in multicast fat-tree DCNs based on a network model that accurately describes the DCN communication environment. Then, we present Oversubscription Bounded Multicast Scheduling (OBMS), a low-complexity multicast flow scheduling algorithm that guarantees bounded link oversubscription and efficient network utilization even under the most congested traffic patterns. Finally, we evaluate the performance of OBMS in an event-driven DCN simulator under various types of traffic patterns, and show that OBMS significantly outperforms other load-balance methods in terms of network throughput and evenness of traffic load distribution.

Replicate and Bundle (RnB) - A Mechanism for Relieving Bottlenecks in Data Centers

Shachar Raindel and Yitzhak Birk

Electrical Engineering Dept.
Technion
Haifa 32000, Israel
raindel@tx.technion.ac.il, birk@ee.technion.ac.il

Abstract

This work addresses the scalability and efficiency of RAM-based storage systems wherein multiple objects must be retrieved per user request. Here, much of the CPU work is per server transaction, not per requested item. Adding servers and spreading the data across them also spreads any given set of requested items across more servers, thereby increasing the total number of server transactions per user request. The resulting poor scalability, dubbed the Multi-get Hole, has been reported in Web 2.0 systems using memcached - a popular memory-based key-value storage system. We present Replicate and Bundle (RnB), a somewhat unintuitive approach: rather than add CPUs, we add memory. Object replicas are mapped “randomly” to servers, and requested objects are bundled, selecting replicas so as to minimize the number of servers accessed per user request and thus the total CPU work per request. We studied RnB via simulation in the context of DRAM-based storage, utilizing micro benchmarks and implemented RnB modules for calibration. Our results show that RnB substantially reduces the number of transactions per request, making operation more efficient. Also, unlike most alternatives, RnB permits flexible growth and relatively easy deployment. Finally, in systems wherein data is replicated for other reasons, RnB is nearly free.

Profit Aware Load Balancing for Distributed Cloud Data Centers

Shuo Liu, Shaolei Ren, Gang Quan, Ming Zhao, and Shangping Ren

Department of Electrical and Computer Engineering, Florida International University,
Miami, FL, 33174

School of Computing and Information Sciences, Florida International University,
Miami, FL, 33199

Department of Computer Science, Illinois Institute of Technology, Chicago, IL, 60616
Emails: {sliu005, gang.quan}@fiu.edu, {sren, ming}@cs.fiu.edu, ren@iit.edu

Abstract

The advent of cloud systems has spurred the emergence of an impressive assortment of Internet services. Recent pressures on enhancing the profitability by curtailing surging dollar costs on energy have posed challenges to, as well as placed a new emphasis on, designing energy-efficient request dispatching and resource management algorithms. What further adds to the design challenge is the highly diverse nature of Internet service requests in terms of Quality-of-Service (QoS) constraints and business values. Nonetheless, most of the existing job scheduling and resource management solutions are for a single type of request and are profit oblivious. They are unable to reap the benefit of multi-service profit-aware algorithm designs. In this paper, we consider a cloud service provider operating geographically distributed data centers in a multi-electricity-market environment, and propose an energy-efficient, profit- and cost-aware request dispatching and resource allocation algorithm to maximize a service provider's net profit. We formulate the net profit maximization issue as a constrained optimization problem, using a unified task model capturing multiple cloud layers (e.g., SaaS, PaaS, IaaS.) The proposed approach maximizes a service provider's net profit by judiciously distributing service requests to data centers, powering on/off an appropriate number of servers, and allocating server resources to dispatched requests. We conduct extensive experiments to validate our proposed algorithm. Results show that our proposed approach can improve a service provider's net profit significantly.

Joint Host-Network Optimization for Energy-Efficient Data Center Networking

*Hao Jin, Tosmate Cheocherngngarn, Dmita Levy, Alex Smith, Deng Pan,
Jason Liu, and Niki Pissinou*

Florida International University, Miami, FL

Terra Environmental Research Institute, Miami, FL
pand@cis.fiu.edu

Abstract

Data centers consume significant amounts of energy. As severs become more energy efficient with various energy saving techniques, the data center network (DCN) has been accounting for 20% or more of the energy consumed by the entire data center. While DCNs are typically provisioned with full bisection bandwidth, DCN traffic demonstrates fluctuating patterns. The objective of this work is to improve the energy efficiency of DCNs during off-peak traffic time by powering off idle devices. Although there exist a number of energy optimization solutions for DCNs, they consider only either the hosts or network, but not both. In this paper, we propose a joint optimization scheme that simultaneously optimizes virtual machine (VM) placement and network flow routing to maximize energy savings, and we also build an Open Flow based prototype to experimentally demonstrate the effectiveness of our design. First, we formulate the joint optimization problem as an integer linear program, but it is not a practical solution due to high complexity. To practically and effectively combine host and network based optimization, we present a unified representation method that converts the VM placement problem to a routing problem. In addition, to accelerate processing the large number of servers and an even larger number of VMs, we describe a parallelization approach that divides the DCN into clusters for parallel processing. Further, to quickly find efficient paths for flows, we propose a fast topology oriented multipath routing algorithm that uses depth-first search to quickly traverse between hierarchical switch layers and uses the best-fit criterion to maximize flow consolidation. Finally, we have conducted extensive simulations and experiments to compare our design with existing ones. The simulation and experiment results fully demonstrate that our design outperforms existing host-or network-only optimization solutions, and well approximates the ideal linear program.

Session 14: Energy Modeling and Scheduling

Energy-Efficient Scheduling for Best-Effort Interactive Services to Achieve High Response Quality

Zhihui Du, Hongyang Sun, Yuxiong He, Yu He, David A. Bader, and Huazhe Zhang

Tsinghua National Laboratory for Information Science and Technology

Department of Computer Science and Technology, Tsinghua University, Beijing, China

School of Computer Engineering, Nanyang Technological University, Singapore

Microsoft Research, Redmonds, WA, USA

College of Computing, Georgia Institute of Technology, Atlanta, GA, USA

School of Information and Communication Engineering, Beijing University of Post and
Telecommunication, Beijing, China

sunh0007@ntu.edu.sg

Abstract

High response quality is critical for many best-effort interactive services, and at the same time, reducing energy consumption can directly reduce the operational cost of service providers. In this paper, we study the quality-energy tradeoff for such services by using a composite performance metric that captures their relative importance in practice: Service providers usually grant top priority to quality guarantee and explore energy saving secondly. We consider scheduling on multicore systems with core-level DVFS support and a power budget. Our solution consists of two steps. First, we employ an equal sharing principle for both job and power distribution. Specifically, we present a "Cumulative Round-Robin" policy to distribute the jobs onto the cores, and a "Water-Filling" policy to distribute the power dynamically among the cores. Second, we exploit the concave quality function of many best-effort applications, and develop Online-QE, a myopic optimal online algorithm for scheduling jobs on a single-core system. Combining the two steps together, we present a heuristic online algorithm, called DES (Dynamic Equal Sharing), for scheduling best-effort interactive services on multicore systems. The simulation results based on a web search engine application show that DES takes advantage of the core-level DVFS architecture and exploits the concave quality function of best-effort applications to achieve high service quality with low energy consumption.

Perfect Strong Scaling Using No Additional Energy

James Demmel, Andrew Gearhart, Benjamin Lipshitz, and Oded Schwartz

Electrical Engineering and Computer Sciences
University of California, Berkeley
Berkeley, USA
{demmel,agearh,lipshitz,odedsc}@cs.berkeley.edu

Abstract

Energy efficiency of computing devices has become a dominant area of research interest in recent years. Most previous work has focused on architectural techniques to improve power and energy efficiency, only a few consider saving energy at the algorithmic level. We prove that a region of perfect strong scaling in energy exists for matrix multiplication (classical and Strassen) and the direct n-body problem via the use of algorithms that use all available memory to replicate data. This means that we can increase the number of processors by some factor and decrease the runtime (both computation and communication) by the same factor, without changing the total energy use.

A Roofline Model of Energy

Jee Whan Choi

Georgia Institute of Technology
Atlanta, Georgia, USA
jee@gatech.edu

Daniel Bedard and Robert Fowler

Renaissance Computing Institute
Chapel Hill, North Carolina, USA
{danb,rjf}@renci.org

Richard Vuduc

Georgia Institute of Technology
Atlanta, Georgia, USA
richie@gatech.edu

Abstract

We describe an energy-based analogue of the time based roofline model. We create this model from the perspective of algorithm designers and performance tuners, with the intent not of making exact predictions, but rather, developing high level analytic insights into the possible relationships among the time, energy, and power costs of an algorithm. The model expresses algorithms in terms of operations, concurrency, and memory traffic, and characterizes the machine based on a small number of simple cost parameters, namely, the time and energy costs per operation or per word of communication. We confirm the basic form of the model experimentally. From this model, we suggest under what conditions we ought to expect an algorithmic time-energy trade-off, and show how algorithm properties may help inform power management.

A Simplified and Accurate Model of Power-Performance Efficiency on Emergent GPU Architectures

Shuaiwen Song

Virginia Tech, Blacksburg, VA
Email: s562673@vt.edu

Chunyi Su

Virginia Tech, Blacksburg, VA
Email: sonicat@vt.edu

Barry Rountree

Lawrence Livermore National Lab, Livermore, CA
Email:rountree@llnl.gov

Kirk W. Cameron

Virginia Tech, Blacksburg, VA
Email: cameron@cs.vt.edu

Abstract

Emergent heterogeneous systems must be optimized for both power and performance at exascale. Massive parallelism combined with complex memory hierarchies form a barrier to efficient application and architecture design. These challenges are exacerbated with GPUs as parallelism increases orders of magnitude and power consumption can easily double. Models have been proposed to isolate power and performance bottlenecks and identify their root causes. However, no current models combine simplicity, accuracy, and support for emergent GPU architectures (e.g. NVIDIA Fermi). We combine hardware performance counter data with machine learning and advanced analytics to model power-performance efficiency for modern GPU-based systems. Our performance counter based approach is simpler than previous approaches and does not require detailed understanding of the underlying architecture. The resulting model is accurate for predicting power (within 2.1%) and performance (within 6.7%) for application kernels on modern GPUs. Our model can identify power-performance bottlenecks and their root causes for various complex computation and memory access patterns (e.g. global, shared, texture). We measure the accuracy of our power and performance models on a NVIDIA Fermi C2075 GPU for more than a dozen CUDA applications. We show our power model is more accurate and robust than the best available GPU power models - multiple linear regression models MLR and MLR+. We demonstrate how to use our models to identify power-performance bottlenecks and suggest optimization strategies for high-performance codes such as GEM, a biomolecular electrostatic analysis application. We verify our power-performance model is accurate on clusters of NVIDIA Fermi M2090s and useful for suggesting optimal runtime configurations on the Keene land supercomputer at Georgia Tech.

Session 15: Communication and Routing 2

Acceleration of an Asynchronous Message Driven Programming Paradigm on IBM Blue Gene/Q

Sameer Kumar

IBM T.J. Watson Research Center
Yorktown Heights, NY 10598, USA
{sameerk}@us.ibm.com

Yanhua Sun and Laximant V. Kalé

University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA
{sun51, kale}@illinois.edu

Abstract

IBM Blue Gene/Q is the next generation Blue Gene machine that can scale to tens of Peta Flops with 16 cores and 64 hardware threads per node. However, significant efforts are required to fully exploit its capacity on various applications, spanning multiple programming models. In this paper, we focus on the asynchronous message driven parallel programming model -Charm++. Since its behavior (asynchronous) is substantially different from MPI, that presents a challenge in porting it efficiently to BG/Q. On the other hand, the significant synergy between BG/Q software and Charm++ creates opportunities for effective utilization of BG/Q resources. We describe various novel fine-grained threading techniques in Charm++ to exploit the hardware features of the BG/Qcompute chip. These include the use of L2 atomics to implement lockless producer-consumer queues to accelerate communication between threads, fast memory allocates, hardware communication threads that are awakened via low overhead interrupts from the BG/Q wakeup unit. Burst of short messages is processed by using the Many to Many interface to reduce runtime overhead. We also present techniques to optimize NAMD computation via Quad Processing Unit (QPX) vector instructions and the acceleration of message rate via communication threads to optimize the Particle Mesh Ewald (PME) computation. We demonstrate the benefits of our techniques via two benchmarks, 3D Fast Fourier Transform, and the molecular dynamics application NAMD. For the 92,000-atom ApoA1 molecule, we achieved 683us/stepwith PME every 4 steps and 782us/step with PME every step.

Communication-Based Mapping Using Shared Pages

Matthias Diener, Eduardo H.M. Cruz, and Philippe O.A. Navaux

Informatics Institute
Federal University of Rio Grande do Sul
Porto Alegre, Brazil
{mdiener, ehmcruz, navaux}@inf.ufrgs.br

Abstract

In current shared memory architectures, the complexity of the cache and memory hierarchies is increasing. Therefore, it is becoming more important to analyze the communication behavior of parallel applications when mapping threads to cores, to improve performance and energy efficiency. However, communication is implicit in most programming models for shared memory, which makes it difficult to detect the communication pattern between the threads in an accurate and low-overhead way. We propose a new mechanism to detect the communication pattern of shared memory applications by monitoring page table accesses. Combining this mechanism with a dynamic migration algorithm allows mapping to be performed dynamically by the operating system. We implemented our mechanism in the Linux kernel and performed experiments with applications from the NAS~Parallel~Benchmarks. Results show a reduction of up to 16.7% of the execution time and 63% of the cache misses, compared to the original scheduler of the operating system. Furthermore, we decrease total processor and DRAM energy consumption by up to 14.7% and 28.5%, respectively.

Integrating Asynchronous Task Parallelism with MPI

Sanjay Chatterjee, Sagnak Tasirlar, Zoran Budimlic, Vincent Cavé, Milind Chabbi,
Max Grossman, and Vivek Sarkar

Department of Computer Science
Rice University
Houston, USA
Email: {cs20, sagnak, zoran, vc8, mc29, jmg3, vsarkar}@rice.edu

Yonghong Yan

Department of Computer Science
University of Houston
Houston, USA
Email: yanyh@cs.uh.edu

Abstract

Effective combination of inter-node and intra-node parallelism is recognized to be a major challenge for future extreme-scale systems. Many researchers have demonstrated the potential benefits of combining both levels of parallelism, including increased communication-computation overlap, improved memory utilization, and effective use of accelerators. However, current "hybrid programming'' approaches often require significant rewrites of application code and assume a high level of programmer expertise. Dynamic task parallelism has been widely regarded as a programming model that combines the best of performance and programmability for shared-memory programs. For distributed-memory programs, most users rely on efficient implementations of MPI. In this paper, we propose HCMPI (Habanero-C MPI), an integration of the Habanero-C dynamic task-parallel programming model with the widely used MPI message-passing interface. All MPI calls are treated as asynchronous tasks in this model, thereby enabling unified handling of messages and tasking constructs. For programmers unfamiliar with MPI, we introduce distributed data-driven futures (DDDFs), a new data-flow programming model that seamlessly integrates intra-node and inter-node data-flow parallelism without requiring any knowledge of MPI. Our novel runtime design for HCMPI and DDDFs uses a combination of dedicated communication and computation specific worker threads. We evaluate our approach on a set of micro-benchmarks as well as larger applications and demonstrate better scalability compared to the most efficient MPI implementations, while offering a unified programming model to integrate asynchronous task parallelism with distributed-memory parallelism.

DTN-FLOW: Inter-Landmark Data Flow for High-Throughput Routing in DTNs

Kang Chen and Haiying Shen

Department of Electrical and Computer Engineering
Clemson University, Clemson, SC 29631
Email: {kangc, shenh}@clemson.edu

Abstract

In this paper, we focus on the efficient routing of data among different areas in Delay Tolerant Networks (DTNs). In current algorithms, packets are forwarded gradually through nodes with higher probability of visiting the destination node or area. However, the number of such nodes usually is limited, leading to insufficient throughput performance. To solve this problem, we propose an inter-landmark data routing algorithm, namely DTN-FLOW. It selects popular places that nodes visit frequently as landmarks and divides the entire DTN area into sub-areas represented by landmarks. Nodes transiting between landmarks relay packets among landmarks, even though they rarely visit the destinations of these packets. Specifically, the number of node transits between two landmarks is measured tore present the forwarding capacity between them, based on which routing tables are built on each landmark to guide packet routing. Each node predicts its transits based on its previous landmark visiting records using the order-k Markov predictor. In a packet routing, a landmark determines the next hop landmark based on its routing table, and forwards the packet to the node with the highest probability of transiting to the selected landmark. Thus, DTN-FLOW fully utilizes all node movements to route packets along landmark paths to their destinations. We analyzed two real DTN traces to support the design of DTN-FLOW. We also deployed a small DTN-FLOW system in our campus for performance evaluation. This deployment and trace-driven simulation demonstrate the high efficiency of DTN-FLOW in comparison with state-of-the-art DTN routing algorithms.

Session 16: Peer to Peer Systems

WHATSUP: A Decentralized Instant News Recommender

Antoine Boutet, Davide Frey, Rachid Guerraoui, Arnaud Jégou, and Anne-Marie Kermarrec

INRIA Rennes, France
Email: antoine.boutet,davide.frey,arnaud.jegou,anne-marie.kermarrec@inria.fr

EPFL, Switzerland
Email: rachid.guerraoui@epfl.fr

Abstract

We present WHATSUP, a collaborative filtering system for disseminating news items in a large-scale dynamic setting with no central authority. WHATSUP constructs an implicit social network based on user profiles that express the opinions of users about the news items they receive (like-dislike). Users with similar tastes are clustered using a similarity metric reflecting long-standing and emerging (dis)interests. News items are disseminated through a novel heterogeneous gossip protocol that (1) biases the orientation of its targets towards those with similar interests, and (2) amplifies dissemination based on the level of interest in every news item. We report on an extensive evaluation of WHATSUP through(a) simulations, (b) a Model Net emulation on a cluster, and(c) a Planet Lab deployment based on real datasets. We show that WHATSUP outperforms various alternatives in terms of accurate and complete delivery of relevant news items while preserving the fundamental advantages of standard gossip: namely, simplicity of deployment and robustness.

Crowdsourcing under Real-Time Constraints

Ioannis Boutsis

Department of Informatics
Athens University of Economics and Business
Athens, Greece
mpoutsis@aueb.gr

Vana Kalogeraki

Department of Informatics
Athens University of Economics and Business
Athens, Greece
vana@aueb.gr

Abstract

In recent years we are experiencing the rapid growth of crowd sourcing systems, in which "human workers" are enlisted to perform tasks more effectively than computers, and get compensated for the work they provide. The common belief is that the wisdom of the "human crowd" can greatly complement many computer tasks which are assigned to machines. A significant challenge facing these systems is determining the most efficient allocation of tasks to workers to achieve successful completion of the tasks under real-time constraints. This paper presents REACT, a crowd sourcing system that seeks to address this challenge and proposes algorithms that aim to stimulate user participation and handle dynamic task assignment and execution in the crowd sourcing system. The goal is to determine the most appropriate workers to assign incoming tasks, in such a way so that the real-time demands are met and high quality results are returned. We empirically evaluate our approach and show that REACT meets the requested real-time demands, achieves good accuracy, is efficient, and improves the amount of successful tasks that meet their deadlines up to 61\% compared to traditional approaches like AMT.

Replication-Based Load Balancing in Distributed Content-Based Publish/Subscribe

Weixiong Rao

School of Software Engineering
Tongji University, Shanghai, China
rweixiong@gmail.com

Chao Chen

Department of Computer Science
University of Helsinki, Finland
{chao.chen, sasu.tarkoma}@cs.helsinki.fil

Pan Hui and Sasu Tarkoma

Hong Kong University of Science and Technology
Telekom Innovation Laboratories
pan.hui@telekom.de

Abstract

In recent years, content-based publish/subscribe (pub/sub) has become a popular paradigm to decouple content producers and consumers for Internet-scale content services. Many real applications show that the content workloads frequently exhibit very skewed distribution, and incur unbalanced workloads. To balance the workloads, the literature of content-based pub/sub adopted a migration scheme (Mis) to move (a subset of) subscription filters from overloaded brokers to under loaded brokers. In this way, the publications that successfully match the moved filters are then offloaded, leading to balanced workloads. Unfortunately, the Mis scheme cannot reduce the overall matching workloads. In the worst case, suppose that all brokers suffer from heavy workloads. Mis cannot find available brokers to offload the heavy workloads of those overloaded brokers, and fails to balance the workloads. To overcome the issue, the contribution of this paper is to develop a set of novel load balancing algorithms, namely a similarity-based replication scheme (Sir). The novelty of Sir is that it not only balances the workloads of brokers but also reduces the overall workloads. Based on both simulation and emulation results, the extensive experiments verify that Sir can achieve much better performance than Mis, in terms of 43.1% higher entropy value (i.e., more balanced workloads) and 46.39 lower workloads.

ZHT: A Light-Weight Reliable Persistent Dynamic Scalable Zero-Hop Distributed Hash Table

Tonglin Li, Xiaobing Zhou, Kevin Brandstatter, Dongfang Zhao, Ke Wang, Anupam Rajendran, Zhao Zhang, and Ioan Raicu

tli13@hawk.iit.edu, xzhou40@hawk.iit.edu, kbrandst@iit.edu, dzhao8@hawk.iit.edu, kwang22@hawk.iit.edu, arajend5@hawk.iit.edu, zhaozhang@uchicago.edu, iraicu@cs.iit.edu

Department of Computer Science, Illinois Institute of Technology, Chicago IL, USA

Department of Computer Science, University of Chicago, Chicago IL, USA

Mathematics and Computer Science Division, Argonne National Laboratory, Argonne IL, USA

Abstract

This paper presents ZHT, a zero-hop distributed hash table, which has been tuned for the requirements of high-end computing systems. ZHT aims to be a building block for future distributed systems, such as parallel and distributed file systems, distributed job management systems, and parallel programming systems. The goals of ZHT are delivering high availability, good fault tolerance, high throughput, and low latencies, at extreme scales of millions of nodes. ZHT has some important properties, such as being light-weight, dynamically allowing nodes join and leave, fault tolerant through replication, persistent, scalable, and supporting unconventional operations such as append (providing lock-free concurrent key/value modifications) in addition to insert/lookup/remove. We have evaluated ZHT's performance under a variety of systems, ranging from a Linux cluster with 512-cores, to an IBM Blue Gene/P supercomputer with 160K-cores. Using micro-benchmarks, we scaled ZHT up to 32K-cores with latencies of only 1.1ms and 18M operations/sec throughput. This work provides three real systems that have integrated with ZHT, and evaluate them at modest scales. 1) ZHT was used in the Fusion FS distributed file system to deliver distributed meta-data management at over 60K operations (e.g. file create) per second at 2K-core scales. 2) ZHT was used in the IStore, an information dispersal algorithm enabled distributed object storage system, to manage chunk locations, delivering more than 500 chunks/sec at 32-nodes scales. 3) ZHT was also used as a building block to MATRIX, a distributed job scheduling system, delivering 5000 jobs/sec throughputs at 2K-core scales. We compared ZHT against other distributed hash tables and key/value stores and found it offers superior performance for the features and portability it supports.

Session 17: Programming Frameworks

A Theoretical Framework for Algorithm-Architecture Co-design

Kenneth Czechowski and Richard Vuduc

School of Computational Science and Engineering
Georgia Institute of Technology, Atlanta, Georgia
{kentcz,richie}@gatech.edu

Abstract

We consider the problem of how to enable computer architects and algorithm designers to reason directly and analytically about the relationship between high-level architectural features and algorithm characteristics. We propose a modeling framework designed to help understand the long-term and high-level impacts of algorithmic and technology trends. This model connects abstract communication complexity analysis-with respect to both the inter-core and inter-processor networks and the memory hierarchy-with current technology proposals and projections. We illustrate how one might use the framework by instantiating a particular model for a class of architectures and sample algorithms (three-dimensional fast Fourier transforms, matrix multiply, and three-dimensional stencil). Then, as a suggestive demonstration, we analyze a number of what-if scenarios within the model in light of these trends to suggest broader statements and alternative futures for power-constrained architectures and algorithms.

Wait-free Hyperobjects for Task-Parallel Programming Systems

Martin Wimmer

Institute of Information Systems, Research Group Parallel Computing
Vienna University of Technology, Faculty of Informatics
Vienna, Austria
Email: wimmer@par.tuwien.ac.at

Abstract

Hyper objects are efficient mechanisms to coordinate accesses to shared variables and data-structures in task-parallel programming models, where each thread can operate on its own coordinated local view of the shared data. Synchronization between local views is restricted to occur at well-defined points in the execution, and can be left to the hyper object implementation. This paper provides a general model for hyper objects that does not require programming language or runtime support and may therefore be used with any task-parallel programming system. We show that hyper objects can be efficiently implemented in a wait-free manner, meaning that all concurrent accesses to a hyper object are guaranteed to complete in a bounded number of steps. The novel finisher hyper object presented in this paper provides transitive termination detection for task-parallel programs. It can be used to efficiently implement task synchronization primitives like finish. However, finishers can also be used to manage reference-counted resources, e.g. shared pointers and copy-on-write pointers. Finally, we provide a wait-free variant of the associative reducer hyper object known from the Cilk++ programming language.

Cyclops Tensor Framework: Reducing Communication and Eliminating Load Imbalance in Massively Parallel Contractions

Edgar Solomonik

Univ. of California, Berkeley
Dept. EECS
solomon@eecs.berkeley.edu

Devin Matthews

Univ. of Texas, Austin
Dept. Chem. and Biochem.

Jeff Hammond

Argonne National Laboratory
Leadership Computing Facility

James Demmel

Univ. of California, Berkeley
Dept. EECS

Abstract

Cyclops (cyclic-operations) Tensor Framework(CTF) is a distributed library for tensor contractions. CTF aims to scale high-dimensional tensor contractions such as those required in the Coupled Cluster (CC) electronic structure method to massively-parallel supercomputers. The framework preserves tensor structure by subdividing tensors cyclically, producing a regular parallel decomposition. An internal virtualization layer provides completely general mapping support while maintaining ideal load balance. The mapping framework decides on the best mapping for each tensor contraction at run-time via explicit calculations of memory usage and communication volume. CTF employs a general redistribution kernel, which transposes tensors of any dimension between arbitrary distributed layouts, yet touches each piece of data only once. Sequential symmetric contractions are reduced to matrix multiplication calls via tensor index transpositions and partial unpacking. The user-level interface elegantly expresses arbitrary-dimensional generalized tensor contractions in the form of a domain specific language. We demonstrate performance of CC with single and double excitations on 8192 nodes of Blue Gene/Q and show that CTF outperforms NWChem on Cray XE6 supercomputers for benchmarked systems.

Scaling Techniques for Massive Scale-Free Graphs in Distributed (External) Memory

Roger Pearce, Maya Gokhale, and Nancy M. Amato

Parasol Laboratory; Dept. of Computer Science and Engineering
Texas A&M University; College Station, TX

Center for Applied Scientific Computing
Lawrence Livermore National Laboratory; Livermore, CA

{rpearce, maya}@llnl.gov {rpearce, amato}@cse.tamu.edu

Abstract

We present techniques to process large scale-free graphs in distributed memory. Our aim is to scale to trillions of edges, and our research is targeted at leadership class supercomputers and clusters with local non-volatile memory, e.g., NAND Flash. We apply an edge list partitioning technique, designed to accommodate high-degree vertices (hubs) that create scaling challenges when processing scale-free graphs. In addition to partitioning hubs, we use ghost vertices to represent the hubs to reduce communication hotspots. We present a scaling study with three important graph algorithms: Breadth-First Search (BFS), K-Core decomposition, and Triangle Counting. We also demonstrate scalability on BG/P Intrepid by comparing to best known Graph500 results. We show results on two clusters with local NVRAM storage that are capable of traversing trillion-edge scale-free graphs. By leveraging node-local NAND Flash, our approach can process thirty-two times larger datasets with only a 39% performance degradation in Traversed Edges Per Second (TEPS).

Session 18: Scheduling 1

Scheduling Tree-Shaped Task Graphs to Minimize Memory and Makespan

Loris Marchal

CNRS and University of Lyon
Lyon, France
loris.marchal@ens-lyon.fr

Oliver Sinnen

University of Auckland
Auckland, New Zealand
o.sinnen@auckland.ac.nz

Frédéric Vivien

INRIA and University of Lyon
Lyon, France
frederic.vivien@inria.fr

Abstract

This paper investigates the execution of tree-shaped task graphs using multiple processors. Each edge of such a tree represents a large IO file. A task can only be executed if all input and output files fit into memory, and a file can only be removed from memory after it has been consumed. Such trees arise, for instance, in the multifrontal method of sparse matrix factorization. The maximum amount of memory needed depends on the execution order of the tasks. With one processor the objective of the tree traversal is to minimize the required memory. This problem was well studied and optimal polynomial algorithms were proposed. Here, we extend the problem by considering multiple processors, which is of obvious interest in the application area of matrix factorization. With the multiple processors comes the additional objective to minimize the time needed to traverse the tree, i.e., to minimize the make span. Not surprisingly, this problem proves to be much harder than the sequential one. We study the computational complexity of this problem and provide an inapproximability result even for unit weight trees. Several heuristics are proposed, each with a different optimization focus, and they are analyzed in an extensive experimental evaluation using realistic trees.

On Graphs, GPUs, and Blind Dating: A Workload to Processor Matchmaking Quest

Abdullah Gharaibeh, Lauro Beltrão Costa, Elizeu Santos-Neto, and Matei Ripeanu

Department of Electrical and Computer Engineering, The University of British Columbia
{abdullah, lauroc, elizeus, matei}@ece.ubc.ca

Abtract

Graph processing has gained renewed attention. The increasing large scale and wealth of connected data, such as those accrued by social network applications, demand the design of new techniques and platforms to efficiently derive actionable information from large scale graphs. Hybrid systems that host processing units optimized for both fast sequential processing and bulk processing (e.g., GPU-accelerated systems) have the potential to cope with the heterogeneous structure of real graphs and enable high performance graph processing. Reaching this point, however, poses multiple challenges. The heterogeneity of the processing elements (e.g., GPUs implement a different parallel processing model than CPUs and have much less memory) and the inherent irregularity of graph workloads require careful graph partitioning and load assignment. In particular, the workload generated by a partitioning scheme should match the strength of the processing element the partition is allocated to. This work explores the feasibility and quantifies the performance gains of such low-cost partitioning schemes. We propose to partition the workload between the two types of processing elements based on vertex connectivity. We show that such partitioning schemes offer a simple, yet efficient way to boost the overall performance of the hybrid system. Our evaluation illustrates that processing a 4-billion edges graph on a system with one CPU socket and one GPU, while offloading as little as 25% of the edges to the GPU, achieves 2x performance improvement over state-of-the-art implementations running on a dual-socket symmetric system. Moreover, for the same graph, a hybrid system with dual-socket and dual-GPU is capable of 1.13 Billion breadth-first search traversed edge per second, a performance rate that is competitive with the latest entries in the Graph500 list, yet at a much lower price point.

Non Linear Divisible Loads: There is No Free Lunch

Olivier Beaumont and Hubert Larchevêque

INRIA Bordeaux Sud-Ouest, LaBRI, CNRS and Univ. of Bordeaux
Bordeaux, France
{olivier.beaumont|hubert.larcheveque}@labri.fr

Loris Marchal

CNRS, LIP and Univ. of Lyon
Lyon, France
loris.marchal@ens-lyon.fr

Abstract

Divisible Load Theory (DLT) has received a lot of attention in the past decade. A divisible load is a perfect parallel task, that can be split arbitrarily and executed in parallel on a set of possibly heterogeneous resources. The success of DLT is strongly related to the existence of many optimal resource allocation and scheduling algorithms, what strongly differs from general scheduling theory. Moreover, recently, close relationships have been underlined between DLT, that provides a fruitful theoretical framework for scheduling jobs on heterogeneous platforms, and MapReduce, that provides a simple and efficient programming framework to deploy applications on large scale distributed platforms. The success of both have suggested to extend their framework to non-linear complexity tasks. In this paper, we show that both DLT and MapReduce are better suited to workloads with linear complexity. In particular, we prove that divisible load theory cannot directly be applied to quadratic workloads, such as it has been proposed recently. We precisely state the limits for classical DLT studies and we review and propose solutions based on a careful preparation of the dataset and clever data partitioning algorithms. In particular, through simulations, we show the possible impact of this approach on the volume of communications generated by MapReduce, in the context of Matrix Multiplication and Outer Product algorithms.

SIPMaP: A Tool for Modeling Irregular Parallel Computations in the Super Instruction Architecture

Nakul Jindal

Comp. & Info. Sci. & Eng.
University of Florida, Gainesville, FL
nakul02@ufl.edu

Victor Lotrich

ACES QC
Gainesville, FL
lotrich@qtp.ufl.edu

Erik Deumens

Dept. of Chemistry
University of Florida, Gainesville, FL
deumens@qtp.ufl.edu

Beverly A. Sanders

Comp. & Info. Sci. & Eng.
University of Florida, Gainesville, FL
sanders@cise.ufl.edu

Abstract

Performance modeling is becoming an increasingly important part of the parallel application development process, particulary for expensive computations that will be run on very high-end systems where resources are scarce. We describe a performance modeling tool SIP Map (Super Instruction Processor Modeling and Prediction) developed for the Super-Instruction Architecture(SIA). The SIA is designed for applications where the dominant data structures are large multi-dimensional arrays and it comprises a DSL, the Super-Instruction Assembly Language(SIAL) that supports expressing algorithms in terms of blocks(tiles), and its runtime system Super Instruction Processor (SIP)that manages distribution and disk storage of the arrays. SIPMaP generates performance models from the SIAL source code. In comparison with many applications where useful performance models have been developed and reported, these programs are irregular and have other difficult to model characteristics such as extensive overlapping of communication and computation.

Plenary Session: Best Papers

Implementing a Blocked Aasen's Algorithm with a Dynamic Scheduler on Multicore Architectures

Grey Ballard, Dulceneia Becker, James Demmel, Jack Dongarra, Alex Druinsky, Inon Peled, Oded Schwartz, Sivan Toledo, and Ichitaro Yamazaki

University of Tennessee, Knoxville, USA

Tel-Aviv University, Tel-Aviv, Israel

University of California, Berkeley, Berkeley, USA

Oak Ridge National Laboratory, Oak Ridge, USA

University of Manchester, Manchester, UK

ballard@cs.berkeley.edu, dbecker7@eecs.utk.edu, demmel@cs.berkeley.edu,
dongarra@eecs.utk.edu, alexdrui@post.tau.ac.il, inon.peled@gmail.com,
odedsc@cs.berkeley.edu, stoledo@tau.ac.il, iyamazak@eecs.utk.edu

Abstract

Factorization of a dense symmetric indefinite matrix is a key computational kernel in many scientific and engineering simulations. However, there is no scalable factorization algorithm that takes advantage of the symmetry and guarantees numerical stability through pivoting at the same time. This is because such an algorithm exhibits many of the fundamental challenges in parallel programming like irregular data accesses and irregular task dependencies. In this paper, we address these challenges in a tiled implementation of a blocked Aasen's algorithm using a dynamic scheduler. To fully exploit the limited parallelism in this left-looking algorithm, we study several performance enhancing techniques, e.g., parallel reduction to update a panel, tall-skinny LU factorization algorithms to factorize the panel, and a parallel implementation of symmetric pivoting. Our performance results on up to 48 AMD Opteron processors demonstrate that our implementation obtains speedups of up to 2.8 over MKL, while losing only one or two digits in the computed residual norms.

DLOOP: A Flash Translation Layer Exploiting Plane-Level Parallelism

Abdul R. Abdurrab

Microsoft Corporation
555 110th Ave NE
Bellevue, WA 98004, USA
Email: abdula@microsoft.com

Tao Xie

San Diego State University
5500 Campanile Drive
San Diego, CA 92182, USA
Email: txie@mail.sdsu.edu

Wei Wang

San Diego State University
5500 Campanile Drive
San Diego, CA 92182, USA
Email: wang@rohan.sdsu.edu

Abstract

A flash translation layer (FTL) is a software layer running in the flash controller of a NAND flash memory solid-state disk (hereafter, flash SSD). It translates logical addresses received from a file system to physical addresses in flash SSD so that the linear flash memory appears to the system like a block storage device. Since the effectiveness of an FTL significantly impacts the performance and durability of a flash SSD, FTL design has attracted significant attention from both industry and academy in recent years. In this research, we propose a new FTL called DLOOP (Data Log On One Plane), which fully exploits plane-level parallelism supported by modern flash SSDs. The basic idea of DLOOP is to allocate logs (updates) onto the same plane where their associated original data resides so that valid page copying operations triggered by garbage collection can be carried out by intra-plane copy-back operations without occupying the external I/O bus. Further, we largely extend a validated simulation environment DiskSim3.0/FlashSim to implement DLOOP. Finally, we conduct comprehensive experiments to evaluate DLOOP using realistic enterprise-scale workloads. Experimental results show that DLOOP consistently outperforms a classical hybrid FTL named FAST and a morden page-mapping FTL called DFTL.

Exploring Traditional and Emerging Parallel Programming Models Using a Proxy Application

Ian Karlin, Abhinav Bhatele, Jeff Keasler, Bradford L. Chamberlain, Jonathan Cohen, Zachary Devito, Riyaz Haque, Dan Laney, Edward Luke, Felix Wang, David Richards, Martin Schulz, and Charles H. Still

Lawrence Livermore National Laboratory, PO Box 808, Livermore, California 94551 USA

Cray Inc., Seattle, Washington 98164 USA

Stanford University, Palo Alto, California 94305 USA

University of California, Los Angeles, California 90095 USA

Mississippi State University, Mississippi State, Mississippi 39762 USA

University of Illinois at Urbana-Champaign, Urbana, Illinois 61801 USA

E-mail: karlin1@llnl.gov, bhatele@llnl.gov, keasler@llnl.gov

Abstract

Parallel machines are becoming more complex with increasing core counts and more heterogeneous architectures. However, the commonly used parallel programming models, C/C++ with MPI and/or OpenMP, make it difficult to write source code that is easily tuned for many targets. Newer language approaches attempt to ease this burden by providing optimization features such as automatic load balancing, overlap of computation and communication, message-driven execution, and implicit data layout optimizations. In this paper, we compare several implementations of LULESH, a proxy application for shock hydrodynamics, to determine strengths and weaknesses of different programming models for parallel computation. We focus on four traditional (OpenMP, MPI, MPI+OpenMP, CUDA) and four emerging (Chapel, Charm++, Liszt, Loci) programming models. In evaluating these models, we focus on programmer productivity, performance and ease of applying optimizations.

Extending the Generality of Molecular Dynamics Simulations on a Special-Purpose Machine

Daniele P. Scarpazza, Douglas J. Ierardi, Adam K. Lerer, Kenneth M. Mackenzie, Albert C. Pan, Joseph A. Bank, Edmond Chow, Ron O. Dror, J.P. Grossman, Daniel Killebrew, Mark A. Moraes, Cristian Predescu, John K. Salmon, and David E. Shaw

D.E. Shaw Research, New York, NY 10036, USA
David.Shaw@DEShawResearch.com

Abstract

Special-purpose computing hardware can provide significantly better performance and power efficiency for certain applications than general-purpose processors. Even within a single application area, however, a special-purpose machine can be far more valuable if it is capable of efficiently supporting a number of different computational methods that, taken together, expand the machine's functionality and range of applicability. We have previously described a massively parallel special-purpose supercomputer, called Anton, and have shown that it executes traditional molecular dynamics simulations orders of magnitude faster than the previous state of the art. Here, we describe how we extended Anton's software to support a more diverse set of methods, allowing scientists to simulate a broader class of biological phenomena at extremely high speeds. Key elements of our approach, which exploits Anton's tightly integrated hardwired pipelines and programmable cores, are applicable to the hardware and software design of various other specialized or heterogeneous parallel computing platforms.

Session 19: Scheduling 2

Algorithms for the Thermal Scheduling Problem

Koyel Mukherjee, Samir Khuller, and Amol Deshpande

Department of Computer Science
University of Maryland
College Park, USA
{koyelm,samir,amol}@cs.umd.edu

Abstract

The energy costs for cooling a data center constitute a significant portion of the overall running costs. Thermal imbalance and hot spots that arise due to imbalanced workloads lead to significant wasted cooling effort - in order to ensure that no equipment is operating above a certain temperature, the data center may be cooled more than necessary. Therefore it is desirable to schedule the workload in a data center in a thermally aware manner, assigning jobs to machines not just based on local load of the machines, but based on the overall thermal profile of the data center. This is challenging because of the spatial cross-interference between machines, where a job assigned to a machine may impact not only that machine's temperature, but also nearby machines. Here, we continue formal analysis of the thermal scheduling problem that we initiated recently. In that work, the notion of effective load of a machine which is a function of the local load on the machine as well as the load on nearby machines, was introduced, and optimal scheduling policies for a simple model (where cross-effects are restricted within a rack) were presented, under the assumption that jobs can be split among different machines. Here we consider the more realistic problem of integral assignment of jobs, and allow for cross-interference among different machines in adjacent racks in the data center. The integral assignment problem with cross-interference is NP-hard, even for a simple two machine model. We consider three different heat flow models, and give constant factor approximation algorithms for maximizing the number (or total profit) of jobs assigned in each model, without violating thermal constraints. We also consider the problem of minimizing the maximum temperature on any machine when all jobs need to be assigned, and give constant factor algorithms for this problem.

Lock-Free and Wait-Free Slot Scheduling Algorithms

Pooja Aggarwal

Computer Science Department
IIT Delhi
New Delhi, India
pooja.mcs11@cse.iitd.ac.in

Smruti R. Sarangi

Computer Science Department
IIT Delhi
New Delhi, India
srsarangi@cse.iitd.ac.in

Abstract

Scalable scheduling is being increasingly regarded as an important requirement in high performance systems. There is a demand for high throughput schedulers in servers, data-centers, networking hardware, large storage systems, and in multi-cores of the future. In this paper, we consider an important subset of schedulers namely slot schedulers that discretize time into quanta called slots. Slot schedulers are commonly used for scheduling jobs in a large number of applications. Current implementations of slot schedulers are either sequential, or use locks. Sadly, lock based synchronization can lead to blocking, and deadlocks, and effectively reduces concurrency. To mitigate these problems, we propose a set of parallel lock-free and wait-free slot scheduling algorithms. Our algorithms are immune to operating system jitter, and guarantee forward progress. Additionally, all our algorithms are linearizable and expose the scheduler's interface as a shared data structure with standard semantics. We empirically demonstrate the scalability of our algorithms for a setup with thousands of requests per second on a 24 thread server. The wait free algorithms are most of the time as fast as the lock-free versions (3X-8X slower in the worst case).

Distributed Algorithms for Scheduling on Line and Tree Networks with Non-uniform Bandwidths

Venkatesan T. Chakaravarthy, Anamitra R. Choudhury, Sambuddha Roy, and Yogish Sabharwal

IBM Research, New Delhi, India
anamchou@in.ibm.com

Abstract

In this paper we study the unsplittable flow problem (UFP) on tree networks in a distributed setting. We have a set of processors (or agents) and a set of tree networks defined over some vertex set. Each processor can access a subset of the tree networks. Each edge in each of the tree networks is associated with a capacity. Each processor has a demand specified as a pair of vertices u and v, along with a profit and a height, the processor wishes to send data between u and v and requires bandwidth equal to its height. Towards that goal, the processor needs to select a tree network accessible to it. A feasible solution selects a subset of demands and schedules each selected demand on a tree network accessible to the processor owning the demand. The requirement is that for any tree network and any edge in the network, the sum of heights of demands scheduled on the network and passing through the edge must not exceed the capacity offered by the edge. The goal is to output a solution having the maximum aggregate profit. Prior work has addressed the above problem in a distributed setting for the special case where all the edge capacities are uniform, say one unit. The main contributions of this paper is to address the general case where the edge capacities can be non-uniform and arbitrary. For this case, we present distributed algorithms with poly-logarithmic approximation ratio.

Analysis of Randomized Work Stealing 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

This paper analyzes the overhead due to false sharing when parallel tasks are scheduled using randomized work stealing (RWS). We obtain high-probability bounds on the cache miss overhead, including the overhead due to false sharing, for several parallel cache-efficient algorithms when scheduled using RWS. These include algorithms for fundamental problems, such as matrix computations, FFT, sorting, basic dynamic programming, list ranking and graph connected components. Our main technical contribution, from which these results follow, is the derivation of nontrivial high-probability bounds on the number of steals incurred by these algorithms in the presence of false sharing, when using RWS.

Session 20: GPU Software

Extending OpenSHMEM for GPU Computing

S. Potluri, D. Bureddy, H. Wang, H. Subramoni, and D.K. Panda

Department of Computer Science and Engineering, The Ohio State University
{potluri,bureddy,wangh,subramon,panda}@cse.ohio-state.edu

Abstract

Graphics Processing Units (GPUs) are becoming an integral part of modern supercomputer architectures due to their high compute density and performance per watt. In order to maximize utilization, it is imperative that applications running on these clusters have low synchronization and communication overheads. Partitioned Global Address Space (PGAS) models provide an attractive approach for developing parallel scientific applications. Such models simplify programming through the abstraction of a shared memory address space while their one-sided communication primitives allow for efficient implementation of applications with minimum synchronization. OpenSHMEM is a library-based programming model that is gaining popularity. However, the current OpenSHMEM standard does not support direct communication from GPU device buffers. It requires data to be copied to the host memory before OpenSHMEM calls can be made. Similarly, data has to moved to the GPU explicitly by remote processes. This severely limits the programmability and performance of GPU applications. In this paper we provide extensions to the OpenSHMEM model which allow communication calls to be made directly on the GPU memory. The proposed extensions are interoperable with the two most popular GPU programming frameworks: CUDA and OpenCL. We present designs for an efficient OpenSHMEM runtime which transparently provide high-performance communication between GPUs in different inter-node and intra-node configurations. To the best of our knowledge this is the first work that enables GPU-GPU communication using the OpenSHMEM model for both CUDA and OpenCL computing frameworks. The proposed extensions to OpenSHMEM, coupled with the high-performance runtime, improve the latency of GPU-GPU shmem getmem operation by 90%, 40% and 17%, for intra-IOH (I/O Hub), inter-IOH and inter-node configurations. It improves the performance of OpenSHMEM atomics by up to 55% and 52%, for intra-IOH and inter-node GPU configurations respectively. The proposed enhancements improve the performance of Stencil2D kernel by 65% on a cluster of 192 GPUs and the performance of BFS kernel by 12% on a cluster of 96 GPUs.

Deploying Graph Algorithms on GPUs: An Adaptive Solution

Da Li

Dept. of Electrical and Computer Engineering
University of Missouri - Columbia
dlx7f@mail.missouri.edu

Michela Becchi

Dept. of Electrical and Computer Engineering
University of Missouri - Columbia
becchim@missouri.edu

Abstract

Thanks to their massive computational power and their SIMT computational model, Graphics Processing Units (GPUs) have been successfully used to accelerate a wide variety of regular applications (linear algebra, stencil computations, image processing and bioinformatics algorithms, among others). However, many established and emerging problems are based on irregular data structures, such as graphs. Examples can be drawn from different application domains: networking, social networking, machine learning, electrical circuit modeling, discrete event simulation, compilers, and computational sciences. It has been shown that irregular applications based on large graphs do exhibit runtime parallelism, moreover, the amount of available parallelism tends to increase with the size of the datasets. In this work, we explore an implementation space for deploying a variety of graph algorithms on GPUs. We show that the dynamic nature of the parallelism that can be extracted from graph algorithms makes it impossible to find an optimal solution. We propose a runtime system able to dynamically transition between different implementations with minimal overhead, and investigate heuristic decisions applicable across algorithms and datasets. Our evaluation is performed on two graph algorithms: breadth first search and single source shortest paths. We believe that our proposed mechanisms can be extended and applied to other graph algorithms that exhibit similar computational patterns.

GPU-based Runtime Verification

Shay Berkovich

Dept. of Elec. and Comp. Eng., University of Waterloo
200 University Avenue West
Waterloo N2L 3G1, Canada
Email: sberkovi@uwaterloo.ca

Borzoo Bonakdarpour

School of Computer Science, University of Waterloo
200 University Avenue West
Waterloo N2L 3G1, Canada
Email: borzoo@cs.uwaterloo.ca

Sebastian Fischmeister

Dept. of Elec. and Comp. Eng., University of Waterloo
200 University Avenue West
Waterloo N2L 3G1, Canada
Email: sfischme@uwaterloo.ca

Abstract

Runtime verification is a monitoring technique to gain assurance about well-being of a program at run time. Most existing approaches use sequential monitors, i.e., when the state of the program with respect to an event of interest changes, the monitor interrupts the program execution, evaluates a set of logical properties, and finally resumes the program execution. In this paper, we propose a GPU-based method for design and implementation of monitors that enjoy two levels of parallelism: the monitor (1) works along with the program in parallel, and (2) evaluates a set of properties in a parallel fashion as well. Our parallel monitoring algorithms effectively exploit the many-core platform available in the GPU. In addition to parallel processing, our approach benefits from a true separation of monitoring and functional concerns, as it isolates the monitor in the GPU. Our method is fully implemented and experimental results show significant reduction in monitoring overhead, monitoring interference, and power consumption due to leveraging the GPU technology.

Kernel Specialization for Improved Adaptability and Performance on Graphics Processing Units (GPUs)

Nicholas Moore

MathWorks
3 Apple Hill Drive
Natick, MA
Nick.Moore@mathworks.com

Miriam Leeser

Dept. of Electrical and Computer Engineering
Northeastern University
Boston, MA
mel@coe.neu.edu

Laurie Smith King

Dept. of Mathematics and Computer Science
College of the Holy Cross
Worcester, MA
lking@holycross.edu

Abstract

Graphics processing units (GPUs) offer significant speedups over CPUs for certain classes of applications. However, programming for GPUs is challenging. There are many parameters that affect performance and their values may change depending on both problem instance and GPU hardware specifics. In addition, most GPU kernels are compiled once, performance optimizations are applied at application compile time. As a result, many GPU libraries and programs have limited adaptability to variations among problem instances and hardware configurations. These factors limit code reuse and the applicability of GPU computing to a wider variety of problems. This paper introduces GPGPU kernel specialization, a technique used to describe highly adaptable kernels that exhibit high performance across a wide range of programmer variables as well as different generations of GPUs. We also introduce our GPU Prototyping Framework (GPU-PF) for dynamic runtime generation of customized GPU kernels incorporating both problem and implementation-specific parameters. GPU-PF fully separates the GPU and CPU code so the GPU code can be compiled during program execution once all the parameters are known. This work explores the implementation and parameterization of two real world applications targeting two generations of NVIDIA CUDA-enabled GPUs using kernel specialization and GPU-PF: large template matching and cone-beam image reconstruction via back projection. Starting with high performance GPU kernels that compare favorably to multi-threaded reference implementations, kernel specialization is shown to increase adaptability while providing performance improvements including improved run time and reduction in resource usage. Kernel specialization offers productivity benefits, improved library code, and a means to increase the parameterizability of GPGPU implementations.

Session 21: Scientific Computing

The Bounded Data Reuse Problem in Scientific Workflows

Mohsen Zohrevandi and Rida A. Bazzi

School of Computing, Informatics, and Decision Systems Engineering
Arizona State University
699 S Mill Ave, Tempe AZ, USA
Email: fmohsen, bazzig@asu.edu

Abstract

Large datasets and time-consuming processes have become the norm in scientific computing applications. The exploration phase in the development of scientific workflows involves trial-and-error with workflow components, which can take a lot of time given the time-consuming nature of the workflow tasks. These facts suggest the possibility of reducing the development time by reusing intermediate data whenever possible. However the storage space is always limited. This introduces a problem: which intermediate datasets from one workflow should be kept to be reused in another workflow, with a limited amount of storage. For the general class of series parallel graphs, we model this problem using a non-linear integer programming formulation and show that it is NP-Hard. We provide a branch and bound optimal algorithm as well as efficient heuristics. We conducted experiments over a large set of randomly-generated workflows as well as a smaller set of synthetic workflows which are based on real-world workflows used by scientists in different disciplines. Our experiments show that the best solution produced by the heuristics only differs from the optimal value by less than 1% on average.

Performance Analysis of the Lattice Boltzmann Model Beyond Navier-Stokes

Amanda Peters Randles, Vivek Kale, Jeff Hammond, William Gropp, and Efthimios Kaxiras

School of Engineering and Applied Sciences
Harvard University, Cambridge, Massachusetts 02138
Contact Email: apeters@fas.harvard.edu

Leadership Computing Facility, Argonne National Laboratory, Argonne, IL 60439

Department of Computer Science, University of Illinois at Urbana-Champaign,
Urbana, IL, 61801

Abstract

The lattice Boltzmann method is increasingly important in facilitating large-scale fluid dynamics simulations. To date, these simulations have been built on discretized velocity models of up to 27 neighbors. Recent work has shown that higher order approximations of the continuum Boltzmann equation enable not only recovery of the Navier-Stokes hydro-dynamics, but also simulations for a wider range of Knudsen numbers, which is especially important in micro- and nano-scale flows. These higher-order models have significant impact on both the communication and computational complexity of the application. We present a performance study of the higher-order models as compared to the traditional ones, on both the IBM Blue Gene/P and Blue Gene/Q architectures. We study the tradeoffs of many optimizations methods such as the use of deep halo level ghost cells that, alongside hybrid programming models, reduce the impact of extended models and enable efficient modeling of extreme regimes of computational fluid dynamics.

A Communication-Optimal N-Body Algorithm for Direct Interactions

Michael Driscoll, Evangelos Georganas, Penporn Koanantakool,
Edgar Solomonik, and Katherine Yelick

Computer Science Division, University of California, Berkeley
Lawrence Berkeley National Laboratory, Berkeley, CA

{driscoll,egeor,penpornk,solomon,yelick}@cs.berkeley.edu

Abstract

We consider the problem of communication avoidance in computing interactions between a set of particles in scenarios with and without a cutoff radius for interaction. Our strategy, which we show to be optimal in communication, divides the work in the iteration space rather than simply dividing the particles over processors, so more than one processor may be responsible for computing updates to a single particle. Similar to a force decomposition in molecular dynamics, this approach requires up to p times more memory than a particle decomposition, but reduces communication costs by factors up to p and is often faster in practice than a particle decomposition [1]. We examine a generalized force decomposition algorithm that tolerates the memory limited case, i.e. when memory can only hold c copies of the particles for c = 1, 2, ..., p. When c = 1, the algorithm degenerates into a particle decomposition, similarly when c = p, the algorithm uses a force decomposition. We present a proof that the algorithm is communication-optimal and reduces critical path latency and bandwidth costs by factors of c2 and c, respectively. Performance results from experiments on up to 24K cores of Cray XE-6 and 32K cores of IBM Blue Gene/P machines indicate that the algorithm reduces communication in practice. In some cases, it even outperforms the original force decomposition approach because the right choice of c strikes a balance between the costs of collective and point-to-point communication. Finally, we extend the analysis to include a cutoff radius for direct evaluation of force interactions. We show that with a cutoff, communication optimality still holds. We sketch a generalized algorithm for multi-dimensional space and assess its performance for 1D and 2D simulations on the same systems.

Exploring SIMD for Molecular Dynamics, Using Intel Xeon Processors and Intel Xeon PhiTM Coprocessors

Simon J. Pennycook, Chris J. Hughes, M. Smelyanskiy, and S.A. Jarvis

Department of Computer Science, University of Warwick, Coventry, UK

Parallel Computing Lab, Intel Corporation, Santa Clara, CA

Email: sjp@dcs.warwick.ac.uk

Abstract

We analyse gather-scatter performance bottlenecks in molecular dynamics codes and the challenges that they pose for obtaining benefits from SIMD execution. This analysis informs a number of novel code-level and algorithmic improvements to Sandia's miniMD benchmark, which we demonstrate using three SIMD widths (128-, 256- and 512-bit). The applicability of these optimisations to wider SIMD is discussed, and we show that the conventional approach of exposing more parallelism through redundant computation is not necessarily best. In single precision, our optimised implementation is up to 5x faster than the original scalar code running on Intel Xeon processors with 256-bit SIMD, and adding a single Intel Xeon Phi coprocessor provides up to an additional 2x performance increase. These results demonstrate: (i) the importance of effective SIMD utilisation for molecular dynamics codes on current and future hardware, and (ii) the considerable performance increase afforded by the use of Intel Xeon Phi coprocessors for highly parallel workloads.

Session 22: Wireless and Sensor Systems

Multi-vehicle Coordination for Wireless Energy Replenishment in Sensor Networks

Cong Wang, Ji Li, Fan Ye, and Yuanyuan Yang

Dept. of Electrical and Computer Engineering, Stony Brook University,
Stony Brook, NY 11794, USA

IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, USA

Abstract

Mobile vehicles equipped with wireless energy transmission technology can recharge sensor nodes over the air. When to recharge which nodes, and in what order, critically impact the network performance. So far only a few works have studied the recharging policy for a single mobile vehicle. In this paper, we study how to coordinate the recharging activities of multiple mobile vehicles, which provide more scalability and robustness than a single vehicle. We leverage concepts and mechanisms from NDN (Named Data Networking) to design energy monitoring protocols that deliver energy status information to mobile vehicles in an efficient manner. Then we study how to minimize the total traveling cost of multiple vehicles while ensuring no node failure. We derive theoretical results on the energy neutral condition and the minimum number of mobile vehicles required for perpetual network operations. We formulate the optimization problem into a Multiple Traveling Salesman Problem with Deadlines (m-TSP with Deadlines), which is NP-hard. To accommodate the dynamic nature of node energy conditions and reduce computational overhead, we present a heuristic algorithm that selects the node with the minimum weighted sum of traveling time and residual lifetime. Our scheme not only improves network scalability but also guarantees the perpetual operation of networks. Finally, we conduct extensive simulations to demonstrate the effectiveness and efficiency of our proposed design, and validate the correctness of theoretical analysis.

On Feasibility of Fingerprinting Wireless Sensor Nodes Using Physical Properties

Xiaowei Mei and Donggang Liu

Computer Science and Engineering Department
The University of Texas at Arlington
dliu@uta.edu

Kun Sun

Center for Secure Information Systems
George Mason University

Dingbang Xu

Computer Science Department
Governors State University

Abstract

Fingerprinting wireless devices using physical properties has been recently suggested as an alternative for device identification and authentication. It has been found that the clock skew caused by the frequency discrepancy of the quartz crystals in different devices can be used as a reliable source for fingerprinting. Researchers have studied the application of the clock skew-based fingerprinting in sensor networks and claimed that it can detect fake identities, wormholes, and node replicas. However, the study in this paper draws a completely opposite conclusion, i.e., the clock skew of sensor nodes can be easily forged by adversaries to evade the detection. This paper then studies the feasibility of using the distribution of signal power in space to fingerprint sensor nodes. The result shows that a sensor node's signal power distribution in space is not only reliable for being used as a source for fingerprinting but also very hard to forge. Finally, the paper discusses the application of using signal power distribution for detecting various attacks as well as the limitations and open problems.

Distributed Algorithms for Joint Routing and Frame Aggregation in 802.11n Wireless Mesh Networks

Dawei Gong and Yuanyuan Yang

Department of Electrical and Computer Engineering, Stony Brook University,
Stony Brook, NY 11794, USA
yuanyuan.yang@stonybrook.edu

Abstract

A wireless mesh network (WMN) is a special type of wireless ad-hoc network, which consists of mesh clients, mesh routers and gateways to the Internet, organized in a mesh topology. The mesh clients are often laptops, cell phones and other wireless devices. Mesh routers forward traffic between mesh clients and gateways. Despite a number of promising features provided by WMNs, such as low deployment cost, self healing, etc., the throughput of WMNs is often limited by severe congestion and collisions, and thus cannot satisfy the increasing traffic demands of numerous applications. In this paper, we study how to maximize the throughput of IEEE 802.11n WMNs by joint routing and frame aggregation. Frame aggregation is to aggregate multiple frames into a large frame before transmission, to reduce communication overhead and alleviate collisions. We first show that previous frame aggregation strategies cannot achieve optimal network throughput. We then formulate the joint problem into a linear programming (LP) problem by considering traffic in the network as flow. As most previous algorithms for LP are centralized and difficult to deploy in large-scale WMNs, we propose a distributed algorithm to solve the formulated problem, in which each mesh router determines the amount of traffic flow for its adjacent links based on the traffic information of neighbors and interfering links. However, in realistic 802.11n WMNs, traffic is transmitted in frames instead of flow, and the traffic to different routers needs to be distinguished. Thus, we further provide an algorithm to determine the routing and frame aggregation strategy for each mesh router, using the traffic flow derived from the first algorithm. We have conducted extensive simulations to evaluate the proposed algorithms and the results demonstrate that the network throughput can be significantly improved compared with existing schemes.

Distributed Low-Latency Out-of-Order Event Processing for High Data Rate Sensor Streams

Christopher Mutschler and Michael Philippsen

Programming Systems Group, CS Dept., University of Erlangen-Nuremberg, Germany

Sensor Fusion and Event Processing Group, Locating and Comm. Systems Dept.,
Fraunhofer Institute for Integrated Circuits IIS, Erlangen, Germany

{christopher.mutschler,philippsen}@cs.fau.de

Abstract

Event-based Systems (EBS) are used to detect and analyze meaningful events in surveillance, sports, finances and many other areas. With rising data and event rates and with correlations among these events, sequential event processing becomes infeasible and needs to be distributed. Existing approaches cannot deal with the ubiquity of out-of-order event arrival that is introduced by network delays when distributing EBS. Order-less event processing may result in a system failure. We present a low-latency approach based on K-slack that achieves ordered event processing on high data rate sensor and event streams without a-priori knowledge. Slack buffers are dynamically adjusted to fit the disorder in the streams without using local or global clocks. The middleware transparently reorders the event input streams so that events can still be aggregated and processed to a granularity that satisfies the demands of the application. On a Real time Locating System (RTLS) our system performs accurate low-latency event detection under the predominance of out-of-order vent arrival and with a close to linear performance scale-up when the system is distributed over several threads and machines.

Session 23: Potpourri Algorithms 2

Agreement via Symmetry Breaking: On the Structure of Weak Subconsensus Tasks

Armando Castañeda

Department of Computer Science, Technion
Haifa, Israel
armando@cs.technion.ac.il

Sergio Rajsbaum

Instituto de Matematicas, UNAM
Mexico City, Mexico
rajsbaum@math.unam.mx

Michel Raynal

Institut Universitaire de France and IRISA-INRIA
Rennes, France
michel.raynal@irisa.fr

Abstract

This paper is on the relative power and the relations linking two important synchronization problems in n-process wait-free shared memory models, namely, set agreement and renaming, which are two of the most studied sub consensus tasks. Since the 2006 seminal paper of Gafni, Rajsbaum and Herlihy, it is known that some renaming instances are strictly weaker than set agreement. Indeed, it was later on shown that not even $(n+1)$-renaming (the strongest task in the renaming family, after perfect n-renaming) can implement$(n-1)$-set agreement (the weakest non-trivial task in the set agreement family). These and other results seem to imply that renaming and, more generally, the tasks called generalized symmetry breaking tasks (GSB) are weaker than agreement tasks. This paper shows that this is not the case, namely, it shows that there is a large family of GSB tasks that are more powerful than $(n-1)$-set agreement. Some of these tasks are equivalent to n-renaming, while others lie strictly between n-renaming and $(n+1)$-renaming. Moreover, none of these GSB tasks can solve $(n-2)$-set agreement. Hence, these sub consensus tasks have a rich structure and are interesting in their own. The proofs of these results are based on algebraic topology techniques and new ideas about different notions of non-determinism that can be associated with shared objects. Interestingly, this paper sheds a new light on the relations linking set agreement and renaming.

A Multi-Partitioning Approach to Building Fast and Accurate Counting Bloom Filters

Kun Huang, Jie Zhang, Dafang Zhang, Gaogang Xie, Kave Salamatian, Alex X. Liu, and Wei Li

Institute of Computing Technology, CAS, Beijing, China

Hunan University, Changsha, China

Universite de Savoie, Chambery, France

Michigan State University, East Lansing, USA

{huangkun09, xie}@ict.ac.cn, {jiezhang, dfzhang}@hnu.edu.cn,
kave.salamatian@univ-savoie.fr, alexliu@cse.msu.edu

Abstract

Bloom filters are space-efficient data structures for fast set membership queries. Counting Bloom Filters (CBFs) extend Bloom filters by allowing insertions and deletions to support dynamic sets. The performance of CBFs is critical for various applications and systems. This paper presents a novel approach to building a fast and accurate data structure called Multiple-Partitioned Counting Bloom Filter (MPCBF) that addresses large-scale data processing challenges. MPCBF is based on two ideas: reducing the number of memory accesses from k (for k hash functions) in the standard CBF to only one memory access in the basic MPCBF-1 case, and a hierarchical structure to improve the false positive rate. We also generalize MPCBF-1 to MPCBF-g to accommodate up to g memory accesses. Our simulation and implementation in MapReduce show that MPCBF outperforms the standard CBF in terms of speed and accuracy. Compared to CBF, at the same memory consumption, MPCBF significantly reduces the false positive rate by an order of magnitude, with a reduction of processing overhead by up to 85.9%.

Composing Relaxed Transactions

Vincent Gramoli

The University of Sydney
vincent.gramoli@sydney.edu.au

Rachid Guerraoui

EPFL
rachid.guerraoui@epfl.ch

Mihai Letia

EPFL
mihai.letia@epfl.ch

Abstract

As the classic transactional abstraction is sometimes considered too restrictive in leveraging parallelism, a lot of work has been devoted to devising relaxed transactional models with the goal of improving concurrency. Nevertheless, the quest for improving concurrency has somehow led to neglect one of the most appealing aspects of transactions: software composition, namely, the ability to develop pieces of software independently and compose them into applications that behave correctly in the face of concurrency. Indeed, a closer look at relaxed transactional models reveals that they do jeopardize composition, raising the fundamental question whether it is at all possible to devise such models while preserving composition. This paper shows that the answer is positive. We present out heritance, a necessary and sufficient condition for a (potentially relaxed) transactional memory to support composition. Basically, out heritance requires child transactions to pass their conflict information to their parent transaction, which in turn maintains this information until commit time. Concrete instantiations of this idea have been used before, classic transactions being the most prevalent example, but we believe to be the first to capture this as a general principle as well as to prove that it is, strictly speaking, equivalent to ensuring composition. We illustrate the benefits of out heritance using elastic transactions and show how they can satisfy out heritance and provide composition without hampering concurrency. We leverage this to present a new (transactional) Java package, a compos able alternative to the concurrency package of the JDK, and evaluate efficiency through an implementation that speeds up state of the art software transactional memory implementations (TL2, LSA, Swiss TM) by almost a factor of 3.

Throughput Enhancement through Selective Time Sharing and Dynamic Grouping

Junliang Chen, Bing Bing Zhou, Chen Wang, Peng Lu, Penghao Wang, and Albert Y. Zomaya

Centre for Distributed and High Performance Computing, School of Information Technologies
The University of Sydney, NSW 2006, Australia
{jchen,bbz,pelu1144,zomaya}@it.usyd.edu.au

CSIRO ICT Center, PO Box 76, Epping, NSW 1710, Australia
chen.wang@csiro.au

Prince of Wales Clinical School, The University of New South Wales, NSW 2052, Australia
penghao.wang@unsw.edu.au

National ICT Australia Limited, 13 Garden Street, Australian Technology Park
Eveleigh, NSW 2015, Australia

Abstract

Space sharing approaches are widely used in job scheduling for HPC systems. The main drawback of these approaches is the blocking of short jobs, which results in low throughput. The research on gang scheduling has shown the potential of time sharing in improving throughput. However, traditional gang scheduling adds jobs for time sharing without selection, which may cause a higher performance degradation of existing running jobs than the performance gain of waiting jobs. Moreover, gang scheduling often adopts a contiguous buddy allocation scheme which has problems of fragmentation and low resource utilization. We design a selective time sharing technique that allows waiting jobs to be co-scheduled with existing running jobs only if the overall throughput can be improved. To alleviate the fragmentation problem, we present a dynamic grouping resource allocation mechanism that relaxes the contiguous allocation requirement imposed on gang scheduling. By integrating these techniques, our new job co-scheduling algorithm is able to simultaneously take system throughput and resource utilization into consideration. The experimental results demonstrate that our approach significantly outperforms both EASY backfilling and traditional gang scheduling in terms of both average turnaround time and bounded slowdown.

Session 24: Potpourri Applications

Novel Parallelization Schemes for Large-Scale Likelihood-based Phylogenetic Inference

Alexandros Stamatakis and Andre J. Aberer

The Exelixis Lab, Scientific Computing Group
Heidelberg Institute for Theoretical Studies
Heidelberg, Germany
Alexandros.Stamatakis@h-its.org, Andre.Aberer@h-its.org

Abstract

The molecular data avalanche generated by novel wet-lab sequencing technologies allows for reconstructing phylogenies (evolutionary trees) using hundreds of complete genomes as input data. Therefore, scalable codes are required to infer trees on these data under likelihood-based models of molecular evolution. We recently introduced a check pointable and scalable MPI-based code for this purpose called RAxML-Light and are currently using it for several real-world data analysis projects. It turned out that the scalability of RAxML-Light is nonetheless still limited because of the fork-join parallelization approach that is deployed. To this end, we introduce a novel, generally applicable, approach to computing the phylogenetic likelihood in parallel on whole-genome datasets and implement it in ExaML (Exascale Maximum Likelihood). ExaML executes up to 3.2 times faster than RAxML-Light because of the more efficient parallelization and communication scheme, while implementing exactly the same tree search algorithm. Moreover, the new parallelization approach exhibits lower code complexity and a more appropriate structure for implementing fault tolerance with respect to hardware failures.

Integrating Online Compression to Accelerate Large-Scale Data Analytics Applications

Tekin Bicer, Jian Yin, David Chiu, Gagan Agrawal, and Karen Schuchardt

Computer Science and Engineering
Ohio State University
E-mail: *{*bicer, agrawal*}*@cse.ohio-state.edu

Pacific Northwest National Laboratories
E-mail: *{*jian.yin, karen.schuchardt*}*@pnnl.gov

Washington State University
E-mail: david.chiu@wsu.edu

Abstract

Compute cycles in high performance systems are increasing at a much faster pace than both storage and wide area bandwidths. To continue improving the performance of large-scale data analytics applications, compression has therefore become promising approach. In this context, this paper makes the following contributions. First, we develop a new compression methodology, which exploits the similarities between spatial and/or temporal neighbors in a popular climate simulation dataset and enables high compression ratios and low decompression costs. Second, we develop a framework that can be used to incorporate a variety of compression and decompression algorithms. This framework also supports a simple API to allow integration with an existing application or data processing middleware. Once a compression algorithm is implemented, this framework automatically mechanizes multi-threaded retrieval, multi-threaded data decompression, and the use of informed prefetching and caching. By integrating this framework with a data-intensive middleware, we have applied our compression methodology and framework to three applications over two datasets, including the Global Cloud-Resolving Model (GCRM) climate dataset. We obtained an average compression ratio of 51.68%, and up to 53.27% improvement in execution time of data analysis applications by amortizing I/O time by moving compressed data.

Massively Parallel Model of Extended Memory Use in Evolutionary Game Dynamics

Amanda Peters Randles, David G. Rand, Christopher Lee, Greg Morrisett, Jayanta Sircar, Martin A. Nowak, and Hanspeter Pfister

School of Engineering and Applied Sciences
Harvard University, Cambridge, Massachusetts 02138
Contact Email: apeters@fas.harvard.edu

Program for Evolutionary Dynamics
Harvard University, Cambridge, Massachusetts 02138

Abstract

To study the emergence of cooperative behavior, we have developed a scalable parallel framework for evolutionary game dynamics. This is a critical computational tool enabling large-scale agent simulation research. An important aspect is the amount of history, or memory steps, that each agent can keep. When six memory steps are taken into account, the strategy space spans 24096 potential strategies, requiring large populations of agents. We introduce a multi-level decomposition method that allows us to exploit both multi-node and thread-level parallel scaling while minimizing communication overhead. We present the results of a production run modeling up to six memory steps for populations consisting of up to 10^18 agents, making this study one of the largest yet undertaken. The high rate of mutation within the population results in a non-trivial parallel implementation. The strong and weak scaling studies provide insight into parallel scalability and programmability trade-offs for large-scale simulations, while exhibiting near perfect weak and strong scaling on 16,384 tasks on Blue Gene/Q. We further show 99% weak scaling up to 294,912 processors 82% strong scaling efficiency up to 262,144 processors of Blue Gene/P. Our framework marks an important step in the study of game dynamics with potential applications in fields ranging from biology to economics and sociology.

Early Experience on the Blue Gene/Q Supercomputing System

Vitali Morozov, Kalyan Kumaran, Venkatram Vishwanath, Jiayuan Meng, and Michael E. Papka

Argonne National Laboratory, Argonne, IL, USA

Northern Illinois University, DeKalb, IL, USA

*{*morozov,kumaran,venkat,jmeng,papka*}*@anl.gov

Abstract

The Argonne Leadership Computing Facility (ALCF) is home to Mira, a10 PF Blue Gene/Q (BG/Q) system. The BG/Q system is the third generation in Blue Gene architecture from IBM and like its predecessors combines system-on-chip technology with a proprietary interconnect (5-D torus). Each compute node has 16 augmented PowerPCA2 processor cores with support for simultaneous multithreading, 4-wide double precision SIMD, and different data prefetchingmechanisms. Mira offers several new opportunities for tuning and scaling scientific applications. This paper discusses our early experience with a subset of micro-benchmarks, MPI benchmarks, and a variety of science and engineering applications running at ALCF. Both performance and power are studied and results on BG/Q is compared with its predecessor BG/P. Several lessons gleaned from tuning applications on the BG/Q architecture for better performance and scalability are shared.

Session 25: Potpourri Systems

Adaptive Cache Bypassing for Inclusive Last Level Caches

Saurabh Gupta, Hongliang Gao, and Huiyang Zhou

Department of Electrical and Computer Engineering
North Carolina State University
Raleigh, USA
E-mail: {sgupta12, hzhou}@ncsu.edu

Intel Corporation
Hillsboro, USA
E-mail: hongliang.gao@intel.com

Abstract

Cache hierarchy designs, including bypassing, replacement, and the inclusion property, have significant performance impact. Recent works on high performance caches have shown that cache bypassing is an effective technique to enhance the last level cache (LLC) performance. However, commonly used inclusive cache hierarchy cannot benefit from this technique because bypassing inherently breaks the inclusion property. This paper presents a solution to enabling cache bypassing for inclusive caches. We introduce a bypass buffer to an LLC. Bypassed cache lines skip the LLC while their tags are stored in this bypass buffer. When a tag is evicted from the bypass buffer, it invalidates the corresponding cache lines in upper level caches to ensure the inclusion property. Our key insight is that the lifetime of a bypassed line, assuming a well-designed bypassing algorithm, should be short in upper level caches and is most likely dead when its tag is evicted from the bypass buffer. Therefore, a small bypass buffer is sufficient to maintain the inclusion property and to reap most performance benefits of bypassing. Furthermore, the bypass buffer facilitates bypassing algorithms by providing the usage information of bypassed lines. We show that a top performing cache bypassing algorithm, which is originally designed for non-inclusive caches, performs comparably for inclusive caches equipped with our bypass buffer. The usage information collected from the bypass buffer also significantly reduces the cost of hardware implementation compared to the original design.

Hardware-Accelerated Regular Expression Matching with Overlap Handling on IBM PowerEN Processor

Kubilay Atasu

IBM Research - Zurich
Zurich, Switzerland
kat@zurich.ibm.com

Florian Doerfler

Supercomputing Systems AG
Zurich, Switzerland
florian.doerfler@scs.ch

Jan van Lunteren and Christoph Hagleitner

IBM Research - Zurich
Zurich, Switzerland
{jvl,hle}@zurich.ibm.com

Abstract

Programmable hardware accelerators for regular expression (regex) matching are evolving into increasingly complex stream processors, which involve multiple state machines that operate in parallel, and specialized post-processors that can process instructions dispatched by the state machines. To improve the speed and the storage-efficiency, complex regexs are decomposed into simpler sub expressions, where each sub expression can fire one or more instructions. Although the impact of regex decompositions on the storage efficiency is well-known, little has been done to address the correctness and completeness. We show that regex decompositions can resultin false positives if overlaps between sub expressions are not taken into account. We describe formal methods to recognize various types of sub expression overlaps that can arise in regex decompositions. We also describe efficient post-processing techniques to eliminate the associated false positives. To enable efficient mapping of the decomposed regexs to the post-processors, we propose integer programming based register allocation methods. Our methods pack narrow variables to reduce the register and instruction usage, and take advantage of multi-register reset instructions to reduce the number of instructions that must be executed in parallel. Experiments on regex sets obtained from open-source and proprietary network intrusion detection systems demonstrate orders of magnitude improvement in the storage efficiency over state-of-the-art.

TM-dietlibc: A TM-aware Real-world System Library

Vesna Smiljkovic, Martin Nowack, Nebojša Miletic, Timothy Harris, Osman Ünsal,
Adrián Cristal, and Mateo Valero

Barcelona Supercomputing Center, Spain
{vesna.smiljkovic, osman.unsal, adrian.cristal, mateo.valero}.bsc.es
mileticn@gmail.com

Technische Universitat Dresden, Germany
martin@se.inf.tu-dresden.de

Oracle Labs, Cambridge, UK
timothy.l.harris@oracle.com

Abstract

The simplicity of concurrent programming with Transactional Memory (TM) and its recent implementation in mainstream processors greatly motivates researchers and industry to investigate this field and propose new implementations and optimizations. However, there is still no standard C system library which a wide range of TM developers can adopt. TM application developers have been forced to avoid library calls inside of transactions or to execute them irrevocably (i.e. in serial order). In this paper, we present the first TM-aware system library, a complex software implementation integrated with TM principles and suited for software (STM), hardware (HTM) and hybrid TM (HyTM). The library we propose is derived from a modified lock-based implementation and can be used with the existing standard C API. In our work, we describe design challenges and code optimizations that would be specific to any TM-based system library or application. We argue about system call execution within transactions, highlighting the possibility of unexpected results from threads. For this reason we propose: (1) a mechanism for detecting conflicts over kernel data in user space, and (2) a new barrier to allow hybrid TM to be used effectively with system libraries. Our evaluation includes different TM implementations and the focus is on memory management and file operations since they are widely used in applications and require additional mechanisms for concurrent execution. We show the benefit we gain with our libc modifications providing parallel execution as much as possible. The library we propose shows high scalability when linked with STM and HTM. For file operations it shows on average a 1.1, 2.6 and 3.7x performance speedup for 8 cores using HyTM, STM and HTM, respectively (over a lock-based single-threaded execution). For a red-black tree it shows on average 3.14x performance speedup for 8 cores using STM (over a multi-read single-threaded execution).

Cura: A Cost-optimized Model for MapReduce in a Cloud

Balaji Palanisamy, Aameek Singh, Ling Liu, and Bryan Langston

College of Computing, Georgia Tech

IBM Research - Almaden

{balaji, lingliu}@cc.gatech.edu
{aameek.singh, bryanlan}@us.ibm.com

Abstract

We propose a new MapReduce cloud service model, Cura, for data analytics in the cloud. We argue that performing MapReduce analytics in existing cloud service models - either using a generic compute cloud or a dedicated MapReduce cloud- is inadequate and inefficient for production workloads. Existing services require users to select a number of complex cluster and job parameters while simultaneously forcing the cloud provider to use those potentially sub-optimal configurations resulting in poor resource utilization and higher cost. In contrast Cura leverages MapReduce profiling to automatically create the best cluster configuration for the jobs so as to obtain a global resource optimization from the provider perspective. Secondly, to better serve modern MapReduce workloads which constitute a large proportion of interactive real-time jobs, Cura uses a unique instant VM allocation technique that reduces response times by up to65%. Thirdly, our system introduces deadline-awareness which, by delaying execution of certain jobs, allows the cloud provider to optimize its global resource allocation and reduce costs further. Cura also benefits from a number of additional performance enhancements including cost-aware resource provisioning, VM aware scheduling and online virtual machine reconfiguration. Our experimental results using Facebook-like workload traces show that along with response time improvements, our techniques lead to more than 80% reduction in the compute infrastructure cost of the cloud data center.

Session 26: Programming Frameworks

A Scalable Heterogeneous Parallelization Framework for Iterative Local Searches

Martin Burtscher

Department of Computer Science
Texas State University-San Marcos
San Marcos, TX 78666, USA
burtscher@txstate.edu

Hassan Rabeti

Department of Mathematics
Texas State University-San Marcos
San Marcos, TX 78666, USA

Abstract

This paper describes and evaluates a highly-scalable framework for running iterative local searches on heterogeneous HPC platforms. The user only needs to provide serial CPU or single-GPU code that implements a simple interface. The framework then executes this code in parallel using MPI between compute nodes and OpenMP and multi-GPU support within nodes. It handles all parallelization aspects, seed distribution and program termination, and it regularly records the currently best solution. We evaluate our framework on three supercomputers using a heuristic iterative hill-climbing TSP solver as well as a search for good finite-state machines. The framework scales to 2048 nodes (32,768 cores) on Ranger with less than a 5% drop in efficiency, searches over 12.2 trillion TSP tours per second on Stampede using 1024 nodes, and evaluates over 21.5 trillion FSM transitions per second using 256 CPUs and 384 GPUs on Keene land.

XKaapi: A Runtime System for Data-Flow Task Programming on Heterogeneous Architectures

Thierry Gautier, João V.F. Lima, Nicolas Maillard, and Bruno Raffin

Grenoble University, France

INRIA, Grenoble, France

Federal University of Rio Grande do Sul (UFRGS), Brazil

thierry.gautier@inrialpes.fr, {joao.lima, nicolas}@inf.ufrgs.br, Bruno.Raffin@inria.fr

Abstract

Most recent HPC platforms have heterogeneous nodes composed of multi-core CPUs and accelerators, like GPUs. Programming such nodes is typically based on a combination of OpenMP and CUDA/OpenCL codes, scheduling relies on a static partitioning and cost model. We present the XKaapi runtime system for data-flow task programming on multi-CPU and multi-GPU architectures, which supports a data-flow task model and a locality-aware work stealing scheduler. XKaapi enables task multi-implementation on CPU or GPU and multi-level parallelism with different grain sizes. We show performance results on two dense linear algebra kernels, matrix product (GEMM) and Cholesky factorization (POTRF), to evaluate XKaapi on a heterogeneous architecture composed of two hexa-core CPUs and eight NVIDIA Fermi GPUs. Our conclusion is two-fold. First, fine grained parallelism and online scheduling achieve performance results as good as static strategies, and in most cases outperform them. This is due to an improved work stealing strategy that includes locality information, a very light implementation of the tasks in XKaapi, and an optimized search for ready tasks. Next, the multi-level parallelism on multiple CPUs and GPUs enabled by XKaapi led to a highly efficient Cholesky factorization. Using eight NVIDIA Fermi GPUs and four CPUs, we measure up to 2.43 TFlop/s on double precision matrix product and 1.79 TFlop/s on Cholesky factorization, and respectively 5.09 TFlop/s and 3.92 TFlop/s in single precision.

A Study of the Behavior of Synchronization Methods in Commonly Used Languages and Systems

Daniel Cederman, Bapi Chatterjee, Nhan Nguyen, Yiannis Nikolakopoulos,
Marina Papatriantafilou, and Philippas Tsigas

Computer Science and Engineering
Chalmers University of Technology, Sweden

Email: {cederman, bapic, nhann, ioaniko, ptrianta, tsigas}@chalmers.se

Abstract

Synchronization is a central issue in concurrency and plays an important role in the behavior and performance of modern programmes. Programming languages and hardware designers are trying to provide synchronization constructs and primitives that can handle concurrency and synchronization issues efficiently. Programmers have to find a way to select the most appropriate constructs and primitives in order to gain the desired behavior and performance under concurrency. Several parameters and factors affect the choice, through complex interactions among (i) the language and the language constructs that it supports, (ii) the system architecture, (iii) possible run-time environments, virtual machine options and memory management support and(iv) applications. We present a systematic study of synchronization strategies, focusing on concurrent data structures. We have chosen concurrent data structures with different number of contention spots. We consider both coarse-grain and fine-grain locking strategies, as well as lock-free methods. We have investigated synchronization-aware implementations in C++, C# (.NET and Mono) and Java. Considering the machine architectures, we have studied the behavior of the implementations on both Intel's Nehalem and AMD's Bulldozer. The properties that we study are throughput and fairness under different workloads and multiprogramming execution environments. For NUMA architectures fairness is becoming as important as the typically considered throughput property. To the best of our knowledge this is the first systematic and comprehensive study of synchronization-aware implementations. This paper takes steps towards capturing a number of guiding principles and concerns for the selection of the programming environment and synchronization methods in connection to the application and the system characteristics.

Managing Asynchronous Operations in Coarray Fortran 2.0

Chaoran Yang, Karthik Murthy, and John Mellor-Crummey

Rice University, Houston, Texas
{chaoran, ksm2, johnmc}@rice.edu

Abstract

As the gap between processor speed and network latency continues to increase, avoiding exposed communication latency is critical for high performance on modern supercomputers. One can hide communication latency by overlapping it with computation using non-blocking data transfers, or avoid exposing communication latency by moving computation to the location of data it manipulates. Co array Fortran 2.0 (CAF 2.0) - a partitioned global address space language - provides a rich set of asynchronous operations for avoiding exposed latency including asynchronous copies, function shipping, and asynchronous collectives. CAF 2.0 provides event variables to manage completion of asynchronous operations that use explicit completion. This paper describes CAF 2.0's finish and cofence synchronization constructs, which enable one to manage implicit completion of asynchronous operations. Finish ensures global completion of a set of asynchronous operations across the members of a team. Because of CAF 2.0's SPMD model, its semantics and implementation of finish differ significantly from those of finish in X10 and Habanero-C. cofence controls local data completion of implicitly-synchronized asynchronous operations. Together these constructs provide the ability to tune a program's performance by exploiting the difference between local data completion, local operation completion, and global completion of asynchronous operations, while hiding network latency. We explore subtle interactions between cofence, finish, events, asynchronous copies and collectives, and function shipping. We justify their presence in a relaxed memory model for CAF 2.0. We demonstrate the utility of these constructs in the context of two benchmarks: Unbalanced Tree Search (UTS), and HPC Challenge Random Access. We achieve 74%-77% parallel efficiency for 4K-32K cores for UTS using the T1WL spec, which demonstrates scalable performance using our synchronization constructs. Our cofence micro-benchmark shows that for a producer-consumer scenario, using local data completion rather than local operation completion yields superior performance.

Author Index

Author Index

Author Index

Author Index

Author Index

Author Index