PDL Packet

NEWSLETTER ON PDL ACTIVITIES AND EVENTS • FALL 2007

http://www.pdl.cmu.edu/

AN INFORMAL PUBLICATION FROM ACADEMIA'S PREMIERE STORAGE SYSTEMS RESEARCH CENTER DEVOTED TO ADVANCING THE STATE OF THE ART IN STORAGE AND INFORMATION INFRASTRUCTURES.

CONTENTS

PDL CONSORTIUM MEMBERS

American Power Corporation
Cisco Systems, Inc.
EMC Corporation
Google
Hewlett-Packard Labs
Hitachi, Ltd.
IBM Corporation
Intel Corporation
LSI Corporation
Microsoft Corporation
Network Appliance
Oracle Corporation
Seagate Technology
Symantec Corporation

Data-Intensive Supercomputing: The Case for DISC

Randal E. Bryant & Joan Digney

Ever increasing quantities of data are available for study, as sensor, networking, and storage technologies make it possible to collect and store vast amounts of information ranging from activity logs to satellite and medical imagery. Scientific research and other computational problems increasingly rely on computing over large data sets, calling for a system design where storage and computation are co-located, and the systems are designed, programmed, and operated to enable users to interactively invoke different forms of computation over the breadth of the data. Data-Intensive Super Computing (DISC) is a new form of high-performance computing that places emphasis on data, rather than on raw computation. DISC systems are responsible for the acquisition, updating, sharing and archiving of data, and support sophisticated forms of computation, which may lead to breakthroughs in a number of scientific disciplines and other problems of societal importance. Offering us inspiration, Google and its competitors have created DISC systems that provide very high levels of search quality, response time, and availability. We believe the style of computing that has evolved to support web search can be extended to encompass a much wider set of applications, and that such systems should be designed and constructed for use by the larger research community.

A new approach to data management and computation would greatly benefit a wide variety of domains. For example, Google has demonstrated the value of performing statistical analysis on massive amounts of data to language translation in the 2005 NIST machine translation competition. They won all four categories of the competition in the first year they entered, translating Arabic to English and Chinese to English [1] using a purely statistical approach. They trained their program using, among other things, multilingual United Nations documents. No one in their machine translation group knew either Chinese or Arabic. As a second illustration, scientists are creating increasingly detailed and accurate finite-element meshes representing the geological properties of the earth's crust, enabling them to model the effect of a geological disturbance and the probabilities of earthquakes occurring in different regions of the world [2].

Data-Intensive Super Computing

This new model of computing merits a new class of machine that differs fundamentally in key principles from conventional supercomputer systems.

Intrinsic, rather than extrinsic data. Collecting and maintaining data are the duties of the system, rather than of its users. The system retrieves new and updated information over networks and computes derived forms of the data in the background. Users can use rich queries based on content and identity to access the data. Reliability mechanisms (e.g., replication, error correction) ensure data integrity and avail-

continued on page 10

FROM THE DIRECTOR'S CHAIR

Greg Ganger

Hello from fabulous Pittsburgh!

It has been another great year for the Parallel Data Lab. Some highlights include a record 12 students graduating and taking jobs with PDL Consortium companies, Best Paper awards at the File and Storage Technologies (FAST) and Sigmetrics conferences, and the start of a new broadly-collaborative research initiative (DISC). Along the way, several new students and faculty have joined PDL, great progress has been made in ongoing research, and many papers have been published. Let me highlight a few things.

The newest research initiative focuses on what we call data-intensive supercomputing, or DISC. In this emerging style of computing, applications extract deep insight from huge and dynamically-changing datasets. As Randy Bryant says: "With the massive amounts of data arising from such diverse sources as telescope imagery, medical records, online transaction records, and web pages, DISC systems have the potential to achieve major advances in science, health care, business efficiencies, and information access." Internet search companies support their data-intensive computing via a different architecture than previous data centers and supercomputers, with storage and computation integrated in each node of a large-scale cluster, allowing computation over data maintained within the cluster-based machine to occur, often without moving most of that data over the network. This approach is very similar in spirit to the Active Disk research explored in PDL in the late 90s, and new explorations into programming models and system infrastructures are warranted. Challenges include such issues as hiding complexity from programmers, coping with node failures (both crashes and corruptions), and automating tuning, diagnosis, and other cluster management issues.

Our primary ongoing umbrella project, Self-* Storage, continues to progress and experience success on many fronts. Building on our early experiences and the Ursa Minor prototype, research into algorithms for system management are now the primary focus. For example, we have developed a new approach to device modeling for cluster-based storage that we call "relative fitness" modeling — the paper on this work was named Best Paper at Sigmetrics 2007. As another example, we have developed mechanisms and algorithms for performance "insulation" among workloads sharing a storage server, bounding the efficiency loss due to interference. Excellent progress has been made (and papers written) about multi-metric utility-based tuning, metadata scalability, fingerpointing problems, and performance debugging assistance.

The Petascale Data Storage Institute (PDSI), led by Garth Gibson, has made great strides in its vision of forming a community of academic, industry, and national lab experts to address the technology challenges faced in scaling storage systems to petascale sizes. The most timely example is the second PDSI workshop, which will be held on November 11, 2007, at Supercomputing '07. This workshop brings together this community, and includes two papers from the PDL: one on the incast problem for high-performance networked storage and one on very large-scale directories. There will also be a BoF on pNFS, the extensions being standardized in NFSv4 to support direct client access to parallel storage systems. Another exciting development is the failure data repository that we are promoting based on our very positive experiences with analyzing such data (e.g., see our Best Paper from FAST 2007).

The Parallel Data Laboratory
School of Computer Science
Department of ECE
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213-3891
VOICE 412•268•6716
FAX 412•268•3010

PUBLISHER
Greg Ganger

EDITOR
Joan Digney

The PDL Packet is published once per year to update members of the PDL Consortium. A pdf version resides in the Publications section of the PDL Web pages and may be freely distributed. Contributions are welcome.

THE PDL LOGO

Skibo Castle and the lands that comprise its estate are located in the Kyle of Sutherland in the northeastern part of Scotland. Both 'Skibo' and 'Sutherland' are names whose roots are from Old Norse, the language spoken by the Vikings who began washing ashore regularly in the late ninth century. The word 'Skibo' fascinates etymologists, who are unable to agree on its original meaning. All agree that 'bo' is the Old Norse for 'land' or 'place,' but they argue whether 'ski' means 'ships' or 'peace' or 'fairy hill.'

Although the earliest version of Skibo seems to be lost in the mists of time, it was most likely some kind of fortified building erected by the Norsemen. The present-day castle was built by a bishop of the Roman Catholic Church. Andrew Carnegie, after making his fortune, bought it in 1898 to serve as his summer home. In 1980, his daughter, Margaret, donated Skibo to a trust that later sold the estate. It is presently being run as a luxury hotel.

FACULTY

Greg Ganger (pdl director)
412•268•1297
ganger@ece.cmu.edu
Anastassia Ailamaki
David Andersen
Anthony Brockwell
Chuck Cranor
Lorrie Cranor
Christos Faloutsos
Rajeev Ghandi
Garth Gibson
Seth Goldstein
Carlos Guestrin
Mor Harchol-Balter
Julio López
Todd Mowry
Priya Narasimhan
David O'Hallaron
Priya Narasimhan
Adrian Perrig
Mike Reiter
Mahadev Satyanarayanan
Srinivasan Seshan
Dawn Song
Hui Zhang

POST DOCTORAL RESEARCHERS

Bianca Schroeder bianca@cs.cmu.edu
Alice Zheng alicez@cs.cmu.edu

STAFF MEMBERS

Bill Courtright 412•268•5485
(pdl executive director) wcourtright@cmu.edu
Karen Lindenfelser, 412•268•6716
(pdl business administrator) karen@ece.cmu.edu
Mike Bigrigg
Helen Conti
Joan Digney
Adam Goode
James Moss
Manish Prasad
Michael Stevens
Michael Stroucken

GRADUATE STUDENTS

Michael Abd-El-Malek
Mukesh Agrawal
Mikhail Chainani
Jim Cipar
Debabrata Dash
Shobhit Dayal
Tudor Dumitraş
Ryan Johnson
Kun Gao
Nikos Hardavellas
James Hendricks
Mike Kasick
Andrew Klosterman
Elie Krevat
Patrick Lanigan
Jure Leskovec
Michael Mesnier
Jim Newsome
Ippokratis Pandis
Swapnil Patil
Adam Pennington
Soila Pertet
Amar Phanishayee
Milo Polte
Chandramouli Rangarajan
Rob Reeder
Brandon Salmon
Raja Sambasivan
Aditya Sethuraman
Faraz Shaikh
Tomer Shiran
Jiri Simsa
Shafeeq Sinnamohideen
Joseph Slember
John Strunk
Ajay Surie
Wittawat Tantisiriroj
Niraj Tolia
Vijay Vasudevan
Gaurav Veda
Matthew Wachs
Andrew Williams
Adam Wolbach

FROM THE DIRECTOR'S CHAIR

Our explorations into consumer storage in the home is quickly moving towards deployable prototypes. The focus is on dramatically simplifying data management and sharing among the many storage-enhanced devices (e.g., DVRs, iPODs, laptops, etc.). Our system architecture builds on semantic (attribute-based) naming and decentralized eventual consistency. The research focus is on enabling users to specify reliability and data accessibility desires and understand current system state and options when situations like device failures, travel with devices, and capacity exhaustion arise. Building on the concept of "views", which are queries against the attributes of objects, we believe that users will be much better able to manage their data. User studies and deployment in real homes will be used to explore this approach.

Of course, many other ongoing PDL projects are also producing cool results. The Data Center Observatory (DCO) continues to grow in scale, as new equipment and activities are added; together with APC, we expect to be adding a second zone soon as well as launching new research into using automation to enhance efficiency of the combined power/cooling/computation infrastructure. A new protocol for fault-tolerant storage has proven able to deliver the bandwidth of crash fault-tolerance while tolerating Byzantine faults, which is a very exciting result. We have begun exploring a better virtual machine (VM) architecture for file storage, hoping to provide much greater performance as well as extensibility. This newsletter and the PDL website offer more details and additional research highlights.

I'm always overwhelmed by the accomplishments of the PDL students and staff, and it's a pleasure to work with them. As always, their accomplishments point at great things to come.

NEW PDL FACES

Lorrie Cranor

Lorrie Faith Cranor is an Associate Research Professor in the School of Computer Science and the department of Engineering and Public Policy at Carnegie Mellon University. She is director of the CMU Usable Privacy and Security Laboratory (CUPS). She has authored over 80 research papers on online privacy, phishing and semantic attacks, spam, electronic voting, anonymous publishing, usable access control, and other topics. She has played a key role in building the usable privacy and security research community, having co-edited the seminal book Security and Usability (O'Reilly 2005) and founded the Symposium On Usable Privacy and Security (SOUPS). She also chaired the Platform for Privacy Preferences Project (P3P) Specification Working Group at the W3C and authored the book Web Privacy with P3P (O'Reilly 2002). She has served on a number of advisory boards, including the FTC Advisory Committee on Online Access and Security, and on the editorial boards of several journals. She is a member of the Board of Directors of the Electronic Frontier Foundation. In 2003 she was named one of the top 100 innovators 35 or younger by Technology Review magazine. She was previously a researcher at AT&T-Labs Research and taught in the Stern School of Business at New York University.

For the past year, Lorrie has been working with PDL students on the Perspective project. Some of the work that she and her students are currently doing aims to

continued on page 27

YEAR IN REVIEW

October 2007

- 15th Annual PDL Retreat and Workshop.
- Niraj Tolia successfully defended his Ph.D. dissertation titled "Using Content Addressable Techniques to Optimize Client-Server Systems." He is joining HP Labs in Palo Alto.

September 2007

- Natassa Ailamaki won the European Young Investigator Award.
- Brandon Salmon proposed his Ph.D. research on "Putting Home Storage Replica Management into Perspective."
- James Hendricks presented "Low-overhead Byzantine Fault-tolerant Storage" at SOSP 2007 in Stevenson, WA.
- Ryan Johnson presented "To Share Or Not To Share?" at the 33rd International Conference on Very Large Data Bases (VLDB 2007) in Vienna, Austria.
- Jimeng Sun successfully defended his Ph.D. research, titles "Incremental Pattern Discovery on Streams, Graphs and Tensors."

August 2007

- James Hendricks presented "Verifying Distributed Erasure-coded Data" at the 26th ACM SIGACT-SIGOPS Symposium on Principles of Distributed Computing (PODC 2007) in Portland, OR.
- Christos Faloutsos gave a distinguished lecture at UC Irvine, CA, on "Graph Mining: Laws, Generators and Tools."

July 2007

- Nikos Hardavellas presented "An Analysis of Database System Performance on Chip Multiprocessors" at the 6th Hellenic Data Management Symposium (HDMS2007) in Athens, Greece.

June 2007

- PDL graduate students Mike Mesnier, Matthew Wachs, and Raja Sambasivan, CS postdoctoral research fellow Alice Zheng, and Greg Ganger received the Best Paper award at the SIGMETRICS conference in San Diego, CA. for their work on "Modeling the Relative Fitness of Storage." The paper was presented by Mike Mesnier.
- Eno Thereska completed his Ph.D. with his defense of his dissertation "Enabling What-if Explorations in Systems." He has moved to the UK to join Microsoft Research there.
- Niraj Tolia presented "Improving Mobile Database Access Over Wide-Area Networks Without Degrading Consistency" at MobiSys'07 in San Juan, Puerto Rico.
- Shafeeq Sinnamohideen proposed his Ph.D. research on "Reusing Dynamic Redistribution to Eliminate Cross-server Operations and Maintain Semantics while Scaling Storage Systems."
- Tiankai Tu successfully defended his Ph.D. dissertation titled "Computational Databases."
- Raja presented "Categorizing and Differencing System Behaviours" at the Second Workshop on Hot Topics in Autonomic Computing in Jacksonville, FL.
- Eno presented "Observer: Keeping System Models from Becoming Obsolete" at the 2nd Workshop on Hot Topics in Autonomic Computing in Jacksonville, FL.
- Christos Faloutsos was the keynote speaker at APWeb/WAIM 2007 in Huang-Shan, China, presenting "Data Mining using Fractals and Power Laws."

May 2007

- 9th Annual PDL Spring Industry Visit Day.
- Nikos Hardavellas proposed his Ph.D. research, titled "Chip Multi-Processor Designs for Commercial Workloads."
- Niraj Tolia presented "Consistency-preserving Caching of Dynamic Database Content" at the International World Wide Web Conference in Banff, Alberta, Canada.
- Ryan Johnson interned with IBM Almaden, working with Vijayshankar Raman on the "Blink" project, which aims to give answers to virtually any business intelligence (BI) query in less than 1s.
- Matthew Wachs and Raja Sambasivan interned with Hewlett-Packard Labs in Palo Alto this summer.
- Julio López successfully defended his Ph.D. research, titled "Methods for Querying Compressed Wavefields." We are very pleased he has elected to remain with the PDL as Research Faculty.
- Efstratios Papadomanolakis successfully defended his dissertation titled" Automated Database Design for Large-Scale Scientific Applications" and has gone on to work with Oracle.
- Jure Leskovec proposed his Ph.D. research on "Dynamics of Real-world Networks."
- Christos Faloutsos and Jimeng Sun offered tutorials on "Mining Large Time-evolving Data Using Matrix and Tensor Tools" at sev-

continued on page 21

"Piled Higher and Deeper" by Jorge Cham www.phdcomics.com

In April, Raja got personal mention in Jorge Cham's Ph.D. Comics. In Raja's words, it was "a major accomplishment" in his graduate student career.

Consistency-preserving Caching of Dynamic Database Content

Tolia & Satyanarayanan

International World Wide Web Conference (WWW 2007), May 8-12, 2007, Banff, Alberta, Canada.

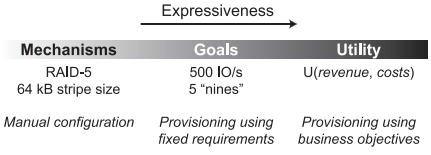
With the growing use of dynamic web content generated from relational databases, traditional caching solutions for throughput and latency improvements are ineffective. We describe a middleware layer called Ganesh that reduces the volume of data transmitted without semantic interpretation of queries or results. It achieves this reduction through the use of cryptographic hashing to detect similarities with previous results. These benefits do not require any compromise of the strict consistency semantics provided by the back-end database. Further, Ganesh does not require modifications to applications, web servers, or database servers, and works with closed-source applications and databases. Using two benchmarks representative of dynamic web sites, measurements of our prototype show that it can increase end-to-end throughput by as much as twofold for non-data intensive applications and by as much as tenfold for data intensive ones.

Using Utility to Provision Storage Systems

Strunk, Thereska, Faloutsos & Ganger

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-07-106, September 2007.

Provisioning a storage system requires balancing the costs of the solution with the benefits that the solution will provide. Previous provisioning approaches have started with a fixed set of requirements and the goal of automatically finding minimum cost solutions to meet them. Those approaches neglect the cost-benefit analysis of the purchasing decision. Purchasing a storage system involves an extensive set of trade-offs between metrics such as purchase cost, performance, reliability, availability, power, etc. Increases in one metric have consequences for others, and failing to account for these trade-offs can lead to a poor return on the storage investment. Using a collection of storage acquisition and provisioning scenarios, we show that utility functions enable this cost-benefit structure to be conveyed to an automated provisioning tool, enabling the tool to make appropriate trade-offs between different system metrics including performance, data protection, and purchase cost.



Utility provides value beyond mechanism-based and goal-based specification. Moving from mechanism-based specification to goal-based specification allowed the creation of tools for provisioning storage systems to meet fixed requirements. Moving from goal-based to utility-based specification allows tools to design storage systems that balance their capabilities against the costs of providing the service.

Lessons Learned From the Deployment of a Smartphone-Based Access-Control System

Bauer, L. Cranor, Reiter & Vaniea

2007 Symposium On Usable Privacy and Security, 18-20 July 2007, Pittsburgh, PA

Grey is a smartphone-based system by which a user can exercise her authority to gain access to rooms in our university building, and by which she can delegate that authority to other users.We present findings from a trial of Grey, with emphasis on how common usability principles manifest themselves in a smartphone-based security application. In particular, we demonstrate aspects of the system that gave rise to failures, misunderstandings, misperceptions, and unintended uses; network effects and new flexibility enabled by Grey; and the implications of these for user behavior. We argue that the manner in which usability principles emerged in the context of Grey can inform the design of other such applications.

Database Servers on Chip Multiprocessors: Limitations and Opportunities

Hardavellas, Pandis, Johnson, Mancheril, Ailamaki & Falsafi

Biennial Conference on Innovative Data Systems Research (CIDR), Asilomar, CA, January 2007.

Prior research shows that database system performance is dominated by off-chip data stalls, resulting in a concerted effort to bring data into on-chip caches. At the same time, high levels of integration have enabled the advent of chip multiprocessors and increasingly large (and slow) on-chip caches. These two trends pose the imminent technical and research challenge of adapting high-performance data management software to a shifting hardware landscape. In this paper we characterize the performance of a commercial database server running on emerging chip multiprocessor technologies. We find that the major bottleneck of current software is data cache stalls, with L2 hit stalls rising from oblivion to become the dominant execution time component in some cases. We analyze the source of this shift and derive a list of features for future database designs to attain maximum performance.

Efficient Use of the Query Optimizer for Automated Database Design

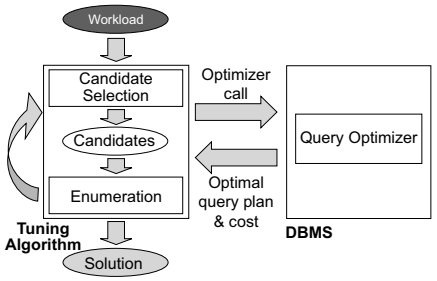
Papadomanolakis, Dash & Ailamaki

VLDB 2007: 1093-1104, September 23-27 2007, Vienna, Austria.

State-of-the-art database design tools rely on the query optimizer for comparing between physical design alternatives. Although it provides an

continued on page 6

continued from page 5



Database design tool architecture.

appropriate cost model for physical design, query optimization is a computationally expensive process. The significant time consumed by optimizer invocations poses serious performance limitations for physical design tools, causing long running times, especially for large problem instances. So far it has been impossible to remove query optimization overhead without sacrificing cost estimation precision. Inaccuracies in query cost estimation are detrimental to the quality of physical design algorithms, as they increase the chances of "missing" good designs and consequently selecting sub-optimal ones. Precision loss and the resulting reduction in solution quality is particularly undesirable and it is the reason the query optimizer is used in the first place.

In this paper we eliminate the tradeoff between query cost estimation accuracy and performance. We introduce the INdex Usage Model (INUM), a cost estimation technique that returns the same values that would have been returned by the optimizer, while being three orders of magnitude faster. Integrating INUM with existing index selection algorithms dramatically improves their running times without precision compromises.

//TRACE: Parallel Trace Replay with Approximate Causal Events

Mesnier, Wachs, Sambasivan, Lopez, Hendricks & Ganger

5th USENIX Conference on File and Storage Technologies (FAST '07), February 13–16, 2007, San Jose, CA.

//TRACE (pronounced parallel trace) is a new approach for extracting and replaying traces of parallel applications to recreate their I/O behavior. Its tracing engine automatically discovers inter-node data dependencies and inter-I/O compute times for each node (process) in an application. This information is reflected in per-node annotated I/O traces. Such annotation allows a parallel replayer to closely mimic the behavior of a traced application across a variety of storage systems. When compared to other replay mechanisms, //TRACE offers significant gains in replay accuracy. Overall, the average replay error for the parallel applications evaluated in this paper is below 6%.

Disk Failures in the Real World: What Does an MTTF of 1,000,000 Hours Mean to You?

Schroeder & Gibson

5th USENIX Conference on File and Storage Technologies (FAST '07), February 13–16, 2007, San Jose, CA.

Component failure in large-scale IT installations such as cluster supercomputers or internet service providers is becoming an ever larger problem as the number of processors, memory chips and disks in a single cluster approaches a million. In this paper, we present and analyze field-gathered disk replacement data from five systems in production use at three organizations, two supercomputing sites and one internet service provider. About 70,000 disks are covered by this data, some for an entire lifetime of 5 years. All disks were high-performance enterprise disks (SCSI or FC), whose datasheet MTTF of 1,200,000 hours suggest a nominal annual failure rate of at most 0.75%.

We find that in the field, annual disk replacement rates exceed 1%, with 2-4% common and up to 12% observed on some systems. This suggests that field replacement is a fairly different process than one might predict based on datasheet MTTF, and that it can be quite variable installation to installation.

We also find evidence that failure rate is not constant with age, and that rather than a significant infant mortality effect, we see a significant early onset of wear-out degradation. That is, replacement rates in our data grew constantly with age, an effect often assumed not to set in until after 5 years of use.

In our statistical analysis of the data, we find that time between failure is not well modeled by an exponential distribution, since the empirical distribution exhibits higher levels of variability and decreasing hazard rates. We also find significant levels of correlation between failures, including autocorrelation and long-range dependence.

Exploiting Similarity for Multi-Source Downloads using File Handprints

Pucha, Andersen, & Kaminsky

4th USENIX NSDI, Cambridge, MA, April 2007.

Many contemporary approaches for speeding up large file transfers attempt to download chunks of a data object from multiple sources. Systems such as BitTorrent quickly locate sources that have an exact copy of the desired object, but they are unable to use sources that serve similar but non-identical objects. Other systems automatically exploit cross-file similarity by identifying sources for each chunk of the object. These systems, however, require a number of lookups proportional to the number of chunks in the object and a mapping for each unique chunk in every identical and similar object to its corresponding sources. Thus, the lookups and mappings in such a system can be quite large, limiting its scalability.

This paper presents a hybrid system that provides the best of both approaches, locating identical and

continued on page 7

continued from page 6

similar sources for data objects using a constant number of lookups and inserting a constant number of mappings per object. We first demonstrate through extensive data analysis that similarity does exist among objects of popular file types, and that making use of it can sometimes substantially improve download times. Next, we describe handprinting, a technique that allows clients to locate similar sources using a constant number of lookups and mappings. Finally, we describe the design, implementation and evaluation of Similarity-Enhanced Transfer (SET), a system that uses this technique to download objects. Our experimental evaluation shows that by using sources of similar objects, SET is able to significantly out-perform an equivalently configured BitTorrent.

Fingerpointing Correlated Failures in Replicated Systems

Pertet, Gandhi & Narasimhan

USENIX Workshop on Tackling Computer Systems Problems with Machine Learning Techniques (SysML), Cambridge, MA (April 2007).

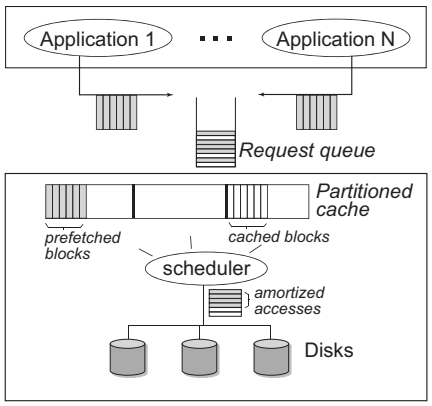
Replicated systems are often hosted over underlying group communication protocols that provide totally ordered, reliable delivery of messages. In the face of a performance problem at a single node, these protocols can cause correlated performance degradations at even non-faulty nodes, leading to potential red herrings in failure diagnosis. We propose a fingerpointing approach that combines node-level (local) anomaly detection, followed by system-wide (global) fingerpointing. The local anomaly detection relies on threshold-based analyses of system metrics, while global fingerpointing is based on the hypothesis that the root-cause of the failure is the node with an "odd-man-out" view of the anomalies. We compare the results of applying three classifiers – a heuristic algorithm, an unsupervised learner (k-means clustering), and a supervised learner (k-nearest-neighbor) – to fingerpoint the faulty node.

Argon: Performance Insulation for Shared Storage Servers

Wachs, Abd-El-Malek, Thereska, & Ganger.

5th USENIX Conference on File and Storage Technologies (FAST '07), February 13–16, 2007, San Jose, CA.

Services that share a storage system should realize the same efficiency, within their share of time, as when they have the system to themselves. The Argon storage server explicitly manages its resources to bound the inefficiency arising from inter-service disk and cache interference in traditional systems. The goal is to provide each service with at least a configured fraction (e.g., 0.9) of the throughput it achieves when it has the storage server to itself, within its share of the server—a service allocated 1 nth of a server should get nearly 1 nth (or more) of the throughput it would get alone. Argon uses automatically configured prefetch/write-back sizes to insulate streaming efficiency from disk seeks introduced by competing workloads. It uses explicit disk time quanta to do the same for non-streaming workloads with internal locality. It partitions the cache among services, based on their observed access patterns, to insulate the hit rate each achieves from the access patterns of others. Experiments show that, combined, these mechanisms and Argon's automatic configuration of each achieve the insulation goal.



Argon's high-level architecture. Argon makes use of cache partitioning, request amortization, and quanta-based disk time scheduling.

Observer: Keeping System Models from Becoming Obsolete

Thereska, Narayanan, Ailamaki, & Ganger.

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-07-101, January 2007.

To be effective for automation, in practice, system models used for performance prediction and behavior checking must be robust. They must be able to cope with component upgrades, misconfigurations, and workload-system interactions that were not anticipated. This paper promotes making models self-evolving, such that they continuously evaluate their accuracy and adjust their predictions accordingly. Such self-evaluation also enables confidence values to be provided with predictions, including identification of situations where no trustworthy prediction can be produced. With a combination of expectation-based and observation-based techniques, we believe that such self-evolving models can be achieved and used as a robust foundation for tuning, problem diagnosis, capacity planning, and administration tasks.

MultiMap: Preserving Disk Locality for Multidimensional Datasets

Shao, Papadomanolakis, Schlosser, Schindler, Ailamaki & Ganger

IEEE 23rd International Conference on Data Engineering (ICDE 2007) Istanbul, Turkey, April 2007.

MultiMap is an algorithm for mapping multidimensional datasets so as to preserve the data's spatial locality on disks. Without revealing disk-specific details to applications, MultiMap

continued on page 16

September 2007

Anastasia Ailamaki wins European Science Prize

Anastasia Ailamaki, associate professor of computer science at Carnegie Mellon University, is one of 20 scientists chosen for this year's highly selective European Young Investigator (EURYI) Awards.

The EURYI program is designed to attract outstanding young scientists from around the world to create their own research teams at European research centers and includes five-year grants of 1 million to 1.25 million euros.

"It's very exciting," said Ailamaki, who joined Carnegie Mellon's Computer Science Department in 2001. "Europe doesn't have many prizes of this magnitude, so winning it is a huge distinction."

Ailamaki will receive 1 million euros – almost $1.4 million – to establish a research team at Ecole Polytechnique Fédérale de Lausanne (EPFL) in Switzerland, where she has been a visiting professor since February.

-- excerpt from CMU Press Release by Byron Spice

July 2007

Anastasia Ailamaki Awarded Finmeccanica Chair in Computer Science

Anastasia Ailamaki, associate professor in the School of Computer Science , has been awarded the Finmeccanica chair in the Computer Science Department for 2007-2009.

Endowed in 1989, the Italian Finmeccanica Fellowship "acknowledges promising young faculty members in the field of computer science."

The Finmeccanica Group ranks among the largest international firms in its operating sectors of aerospace, defense, energy, transportation and information technology. They are active in the design and manufacture of aircraft, helicopters, satellites, radar, power generation components, trains, information and technology services - to name but a few - and the realization of these systems via engineering and managerial skills, electronics, information technology and innovative materials.

July 2007

O'Hallaron Named New Director of Intel Research Pittsburgh

David O'Hallaron, associate professor of computer science and electrical and computer engineering at Carnegie Mellon University, is the new director of Intel Research Pittsburgh. O'Hallaron, whose research focuses on scientific supercomputing, computational database systems and virtualization, assumed leadership of the Pittsburgh lab July 1. He succeeds Todd Mowry, who has returned to the university as an associate professor of computer science and electrical and computer engineering.

-- ECE News Online

July 2007

Eleven CMU Professors Participate in Microsoft Faculty Summit

Eleven Carnegie Mellon professors, including eight from the School of Computer Science (SCS) and three from the Electrical and Computer Engineering (ECE) Department, attended Microsoft Research's 8th annual Faculty Summit July 16-17 at the Microsoft campus in Redmond, Wash. The event, which draws about 350 academics from 175 institutions worldwide, is an opportunity for professors to meet with Microsoft researchers and product group engineers for in-depth discussions of computing problems and trends.

Carnegie Mellon attendees included Jamie Callan, Alexei Efros, Seth Copen Goldstein, Robert Kraut, Peter Lee, Roni Rosenfeld and Jeannette Wing from SCS, and Bruce Krogh, Jose Moura and Dawn Song from ECE. Luis von Ahn, a Microsoft Research New Faculty Fellow, also participated.

-- CMU 8 1/2 x 11 News

July 2007

Best Wishes Soila and Mulwa!

Soila Milanoi Pertet and Geoffrey Mulwa Kavulya's were married on 28th July, 2007 at St. Christopher's Secondary School, in Nairobi, Kenya.

June 2007

Priya Narasimhan Participates in Frontiers of Engineering Symposium

Assistant Professor of Electrical and Computer Engineering and Computer Science Priya Narasimhan and Assistant Professor of Mechanical Engineering Burak Ozdoganlar were selected to participate in the National Academy of Engineering's 13th annual U.S. Frontiers of Engineering Symposium, Sept. 24 - 26 at Microsoft Research in Redmond, Wash.

From http://www.nae.edu/nae/NAEFOE.nsf: "The Frontiers of Engineering program brings together through

continued on page 9

continued from page 8

three-day meetings a select group of emerging engineering leaders from industry, academe, and government labs to discuss pioneering technical work and leading edge research in various engineering fields and industry sectors. The goal of the meetings is to introduce these outstanding engineers (ages 30-45) to each other, and through this interaction facilitate collaboration in engineering, the transfer of new techniques and approaches across fields, and establishment of contacts among the next generation of engineering leaders."

--CMU 8 1/2 x 11 News

May 2007

Best Paper Awards for Jimeng Sun, Hui Zhang and Christos Faloutsos

Teams from Carnegie Mellon won the best paper award in both the research and application tracks at the Society for Industrial and Applied Mathematics Conference on Data Mining this past April. The research track winner was "Less Is More: Compact Matrix Decomposition for Large Sparse Graphs" by Jimeng Sun, Yinglian Xie, Hui Zhang and Christos Faloutsos. The application track winner was "Harmonium Models for Semantic Video Representation and Classification" by Jun Yang, Yan Liu, Eric Xing and Alexander Hauptmann.

-- CMU 8 1/2 x 11 News

April 2007

Elie Krevat Awarded NDSEG Fellowship

Congratulations to Elie Krevat, who has been selected to receive a 2007 National Defense Science and Engineering Graduate (NDSEG) Fellowship. The NDSEG Fellowship is sponsored and funded by the Department of Defense (DoD).

NDSEG selections were made from a pool of more than 3,400 applications by the Air Force Research Laboratory/Air Force Office of Scientific Research (AFRL/AFOSR), the Office of Naval Research (ONR), the Army Research Office (ARO), and the DoD High Performance Computing Modernization Program Office (HPCMP). The NDSEG Fellowship covers tuition and required fees for three years at any accredited U.S. college or university that offers advanced degrees in science and engineering. In addition, the NDSEG Fellowship will provide a yearly stipend.

April 2007

Photo Exhibit by PDL Student

Eno Thereska, who during graduate school led a parallel life as a photographer, has a photo exhibition running from April–June. The exhibition, titled "Species," showed at the Pittsburgh Filmmakers gallery.

March 2007

PDL Researchers Awarded Best Paper at SIGMETRICS 2007

The program chairs of SIGMETRICS 2007, held from June 12-16 in San Diego, CA, have announced that the Best Paper Award will be given to a team of researchers from the Parallel Data Lab (PDL) for their work, "Modeling the Relative Fitness of Storage." The authors are graduate students Michael Mesnier (ECE), Matthew Wachs (CS), Raja Sambasivan (ECE), CS postdoctoral research fellow Alice Zheng, and their faculty advisor, Greg Ganger, PDL director and Professor of ECE and CS.

The paper was chosen from among the 29 accepted (and many others submitted) for publication at the conference, which focuses on the measurement and modeling of computer systems.

March 2007

A New Johnson!

Ryan, his wife Leah, and two daughters Ariana and Summer welcomed Holly Johnson on March 20, 2007, weighing 7lb 9oz and measuring 20.5". Congratulations!

March 2007

Mike Kasick Awarded NSF Graduate Fellowship

ECE student Mike Kasick has been awarded an NSF Graduate Research Fellowship. CMU had 8 awardees in all, with only one from ECE.

The fellowship provides funding for a maximum of three years that can be used over a five-year period, including a stipend of $30,000 per twelve-month fellowship period. Mike is advised in his research on problem diagnosis by Priya Narasimhan.

February 2007

Two PDL Researchers Awarded Sloan Fellowships

Two Sloan Fellowships in computer science have been awarded to PDL faculty members: Priya Narasimhan,

continued on page 21

DATA-INTENSIVE SUPER COMPUTING

continued from page 1

ability as part of the system function. By contrast, current supercomputer centers provide short-term, large-scale storage to their users, high-bandwidth communication to get data to and from the system, and plenty of computing power, but no support for data management. Users must collect and maintain data on their own systems, ship the data to the supercomputer for evaluation, and then return the results back for further analysis and updates.

High-level programming models for expressing computations over the data. Current supercomputers must be programmed at a very low level to make maximal use of the resources. Wresting maximum performance from the machine requires hours of tedious optimization. DISC application developers are provided with powerful, high-level programming primitives that express natural forms of parallelism and that do not specify a particular machine configuration. It is then the job of the compiler and runtime system to map these computations onto the machine efficiently.

Interactive access. DISC system users are able to execute programs interactively and with widely varying computation and storage requirements. The system responds to user queries and simple computations on the stored data in less than one second, while more involved computations take longer but do not degrade performance for the queries and simple computations of others. By contrast, existing supercomputers are operated in batch mode to maximize processor utilization.

Scalable mechanisms to ensure high reliability and availability. Current supercomputers provide reliability by periodically checkpointing the state of a program, and then rolling back to the most recent checkpoint when an error occurs. More serious failures require bringing the machine down, running diagnostic tests, replacing failed components, and only then restarting the machine. Instead, a DISC system should employ nonstop reliability mechanisms, where all original and intermediate data are stored in redundant forms, and selective re-computation can be performed in event of component or data failures. Furthermore, the machine should automatically diagnose and disable failed components, which would only be replaced when enough had accumulated to impair system performance significantly.

DISC Predecessors

We envision that different research communities will emerge to use DISC systems, each organized around a particular shared data repository. These communities will devise different policies for the collection and maintenance of their particular data types, for what computations can be performed and how they will be expressed, and for how access to the data is granted.

We draw much of our inspiration for DISC from the infrastructure that companies have created to support web search. Many credit Inktomi (later acquired by Yahoo!) for initiating the trend of constructing specialized, large-scale systems to support web search [3]. Their 300-processor system in 1998 pointed the way to the much larger systems used today of which Google has become the most visible exemplar.

Google does not disclose the size of their server infrastructure, but reports range from 450,000 [4] to several million [5] processors, spread around at least 25 data centers worldwide. Machines are grouped into clusters of "a few thousand processors," with disk storage associated with each processor. The system makes use of low-cost, commodity parts to minimize per unit costs, including using processors that favor low power over maximum speed. Standard Ethernet communication links are used to connect the processors. This style of design stands in sharp contrast to the exotic technology found in existing supercomputers, which use the fastest possible processors and specialized, high-performance interconnection networks, consuming large amounts of power and requiring costly cooling systems.

Constructing a General-Purpose DISC System

Suppose we wanted to construct a general purpose DISC system that could be made available to a research community for solving data-intensive problems. Below we list some of the issues to be addressed.

Hardware design. There are a wide range of choices here, from assembling a system out of low-cost commodity parts, to using off-the-shelf systems designed for data centers, to using supercomputer-class hardware, with more processing power, memory, and disk storage per processing node and a much higher bandwidth interconnection network. These choices

continued on page 11

Brandon Salmon presents "Putting Home Data Management into Perspective" at the 2006 PDL Workshop & Retreat.

continued from page 10

will greatly affect the system cost, with prices ranging between around $2,000 to $10,000 per node. The fundamental problem is understanding the tradeoffs between the different hardware configurations and how well different applications would run on different systems.

Programming model. There should be a small number of program abstractions that enable users to specify their desired computations at a high level, and then the runtime system should provide an efficient and reliable implementation, handling such issues as scheduling, load balancing, and error recovery. One important software concept for scaling parallel computing beyond 100 or so processors is to incorporate error detection and recovery into the runtime system and to isolate programmers from both transient and permanent failures as much as possible. We must assume that every computation or information retrieval step can fail to complete or can return incorrect answers, so we must devise strategies to recover from errors, allowing the system to operate continuously. We also believe it is important to avoid the tightly synchronized parallel programming notations used for current supercomputers. Supporting these forces the system to use resource management and error recovery mechanisms that would be hard to integrate with the interactive scheduling and flexible error handling schemes we envision. Instead, we want programming models that dynamically adapt to the available resources and that perform well in a more asynchronous execution environment.

Resource management. We want DISC systems to be available in an interactive mode and yet able to handle very large-scale computing tasks. Different approaches to scheduling processor and storage resources can be considered, with the optimal decisions depending on the programming models and reliability mechanisms to be supported.

Supporting program development. Developing parallel programs is notoriously difficult, both in terms of correctness and getting good performance. We must provide software development tools that allow correct programs to be written easily, while also enabling more detailed monitoring, analysis, and optimization of program performance. Most likely, DISC programs should be written to be "self-optimizing," adapting strategies and parameters to the available processing, storage, and communications resources, and also on the rates and nature of failing components.

Minglong Shao presents "Exploring Multidimensional Access in DBMS" at the 2006 PDL Workshop & Retreat.

System Software. Besides supporting application programs, system software is required for a variety of tasks, including fault diagnosis and isolation, system resource control, and data migration and replication. Many of these issues are being explored by PDL's Self-* systems project, but the detailed solutions will depend greatly on the specifics of system organization.

Other Issues. The challenges to creating a DISC system will all require ongoing, multidisciplinary research efforts. Some additional topics to consider include processor design for use in cluster machines, effectively supporting different scientific communities, reducing the energy requirements of large-scale systems, appropriately balancing performance and cost, and coping with the realities of component failures and repair times, system workload, access control, bad data handling, and choice of system components.

Turning Ideas into Reality

We envision a prototype system of around 1000 processing nodes. Such a system would be large enough to demonstrate the performance potential of DISC and to pose some of the challenges in resource management and error handling extant in very large systems. For example, if we provision each node with at least one terabyte of storage (terabyte disks will be available within the next year or so), the system would have a storage capacity of over one petabyte.

By having two dual-core processors in each node, the resulting machine would have 4,000 total processor cores. In order to support both system and application researchers simultaneously, the system would be divisible into multiple partitions, where the different partitions could operate independently. This multi-partition strategy would resolve the dilemma of how to get systems and applications researchers working together on a project; application developers want a stable and reliable machine, but systems researchers keep changing things.

For the program development partitions, we would initially use available software, such as the open source code from the Hadoop project [6], to implement the file system and support for application programming. For the systems research partitions, we would create our own design, studying the different layers of hardware and system software required to get high performance and reliability. Initially, we propose using relatively high-end

continued on page 22

FAILURE TOLERANCE IN PETASCALE COMPUTERS

Garth Gibson, Bianca Schroeder & Joan Digney

The market for high-performance computing (HPC) systems has been experiencing an impressive growth over the past years. With total revenues of $10B in 2006, the HPC server market now makes up 20% of the total server market. 26% of all processors shipped are going into HPC clusters. And these trends are expected to continue in the future. While overall server revenues have been growing at a modest 4% per year, the HPC market has seen an annual revenue increase of more than 20% over the past 4 years. With the HPC market becoming one of the driving forces in the server industry, more and more of industry is concerned with how to solve the challenges expected in future HPC systems.

We believe that one of the most difficult problems in future HPC installation will be providing reliability at scale. With hundreds of thousands or even millions of processing, storage and networking elements, failure will be frequent, making it increasingly difficult for applications to make forward progress. In our recent work [4] we therefore ask what we can learn about reliability and availability of future HPC systems based on failure data of previous HPC systems.

Our projections are based on two key observations. First, technology trends lead us to believe that the speedup of future systems will come from an increase in the number of processor cores per chip (or socket) and an increase in the total number of sockets per system, rather than an increase in clock speed. Second, our analysis of failure data [3] shows that the failure rate of a system grows in proportion to the number of sockets in the system (maybe faster with increased number of threads from more cores per socket). Moreover, there is little indication that systems and their hardware get more reliable over time as technology changes. Therefore, as the number of sockets in future systems increases to achieve top500.org performance trends, we expect system wide failure rates will increase.

Figure 1: High-performance computer cluster at LANL.

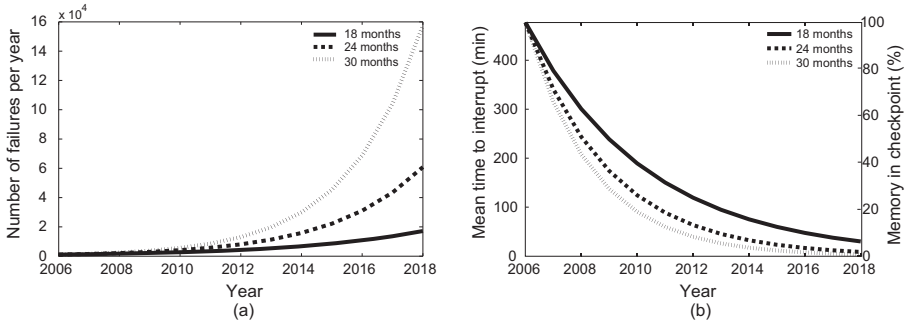


Figure 2: (a) The expected growth in failure rate and (b) decrease in MTTI, assuming that the number of cores per socket grows by a factor of two every 18, 24 and 30 months, respectively, and the number of sockets increases so that performance conforms to top500.org.

Figure 2 shows our projections for the failure rates and mean time to interrupt (MTTI) of future HPC systems. We consider three projected rates of growth, with numbers of cores doubling every 18, 24 and 30 months, and make the (optimistic) assumption that failure rates will increase only with number of chips, and not with the number of cores per chip. As Figure 3 illustrates, the failure rates across the top 500 biggest machines of the future can be expected to grow dramatically.

Observing this dramatic increase in failure rates brings up the question of how the utility of future systems will be affected. Fault tolerance in HPC systems is typically implemented with checkpoint restart programming. With failures becoming more frequent, applications will have to write back checkpoints more frequently and will restart more frequently, causing work to be recomputed more frequently.

Based on the models of Figure 2, Figure 3 shows a prediction that the effective resource utilization by an application will drastically decrease over time. For example, in the case where the number of cores per chip doubles every 30 months, the utilization drops to zero by 2013, meaning the system is spending 100% of its time writing checkpoints or recovering lost work, a situation that is clearly unacceptable. Below we consider possible ways to stave off this projected drop in resources utilization.

Better Fault Tolerance for Petascale Computers

As failure rates grow proportionally to the number of sockets, keeping the number of sockets constant should stave off an increase in the failure rate. However, doing this, while continuing top500.org performance trends, requires the performance of each processor chip to grow faster than currently projected, a trend that chip designers consider unlikely. Therefore, we think the number of sockets will continue to increase.

Socket Reliability: The increase in

continued on page 13

continued from page 12

failure rates could be prevented if individual processor chip sockets were made more reliable, i.e. if the per socket MTTI would increase proportionally to the number of sockets per system over time. Unfortunately, LANL's data does not indicate that hardware has become more reliable over time, suggesting that as long as the number of sockets is rising, the system-wide MTTI will drop.

Partitioning: The number of interrupts an application sees depends on the number of sockets it is using in parallel. One way to stave off the drop in MTTI per application would be to run it only on a constant-sized subpartition of the machine, rather than on all nodes of a machine. Unfortunately, this solution is not appealing for the most demanding "hero" applications, for which the largest new computers are often justified.

Faster Checkpointing: One way to cope with increasing failure rates is to make checkpoints faster by increasing storage bandwidth. Our projections show that the required increase in bandwidth is orders of magnitude higher than the commonly expected increase in bandwidth per disk drive. Therefore, an increase in bandwidth would have to come from a rapid growth in the total number of drives, well over 100% per year, increasing the cost of the storage system much faster than any other part of petascale computers. This might be possible, but it is not very desirable.

Another option is to decrease the amount of memory being checkpointed, either by not growing total memory as fast or by better compression of application data leading to only a smaller fraction of memory being written in each checkpoint. Compression seems to be the more appealing approach, since growing total memory at a slower rate may not be acceptable for the most demanding applications.

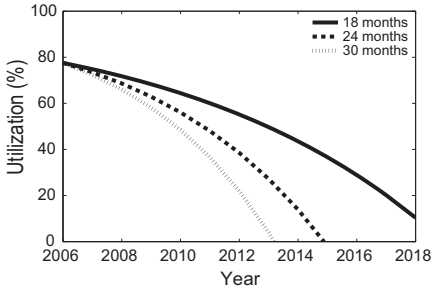


Figure 3: Effective application utilization drops because MTTI is dropping and more time will be lost to taking checkpoints and restarting from checkpoints.

Achieving higher checkpoint speedups purely by compression will require increasingly better compression ratios. As early as in 2010, an application will be limited to checkpoint at most 50% of the total memory. Once the 50% mark is crossed, other options become viable, such as diskless checkpointing where the checkpoint is written to the volatile memory of another node. We recommend that any application capable of compressing its checkpoint size should pursue this path; considering the increasing number of cycles that will go into checkpointing, the compute time needed for compression may be time well spent.

A third approach to taking checkpoints faster is to introduce special devices, which will accept a checkpoint at speeds that scale with memory, then relay the checkpoint to storage after the application has resumed computing. Although such an intermediate memory could be very expensive, it might be a good application for write-limited technologies such as flash memory, because checkpoints are written infrequently.

Non-Checkpoint-based Fault Tolerance: Process-pairs duplication is a traditional method for fault tolerance that hasn't been applied to HPC systems. Basically, every operation is done twice in different nodes so the later failure of a node does not destroy the operation's results. Process pairs would eliminate both the cost associated with writing checkpoints, as well as lost work in the case of failure. However, using process pairs is expensive in that it requires giving up 50% of the hardware to compute each operation twice and it introduces overheads to keep process pairs in sync. However, if no other method keeps utilization above 50%, this sacrifice might become appropriate.

Petascale Storage Projections

Future petascale systems will pose difficult scaling issues for storage system designers.

First, individual disk bandwidth grows at a rate of about 20% per year, which is significantly slower than the 100% per year growth in system performance that top500.org predicts. To keep systems balanced, the number of disk drives in a system will have to increase at an impressive rate. Figure 4 projects the number of drives in a system necessary to (just) maintain balance. The figure shows that, if, by 2018 an HPC system at the top of top500.org will need to have more than 800,000 disk drives. Managing this number of independent disk drives, much less delivering all of their bandwidth to an application, will be extremely challenging for storage system designers.

Second, disk drive capacity will keep growing by about 50% per year, thereby continuously increasing the amount of work and the time needed to reconstruct a failed drive. While other trends, such as decrease in physical size (diameter) of drives, will help to limit the increase in reconstruction time, these are single step decreases limited by the poorer cost effectiveness of smaller disks. Overall, we think it is realistic to expect an increase in reconstruction time of at least 10% per year. Assuming that today reconstruction times are often about 30 hours

continued on page 28

DISSERTATION ABSTRACT:
Computational Databases

Tiankai Tu

Carnegie Mellon University School of Computer Science Ph.D. Dissertation, June 25, 2007.

As we ride on a wave of technology innovation towards the age of petascale computing, the vast amount of data produced by computer simulations and scientific instruments will soon—if not already—overwhelm our ability to manipulate and interpret them. Traditional database management systems, though capable of managing huge information stores for the commercial and governmental sectors, have lagged in supporting core scientific applications. This dissertation identifies a crucial structural mismatch between the inherent unstructured nature of scientific datasets and the built-in tabular abstraction of databases.

We propose a Computational Database approach to the problem. The key is to add a computational cache on top of a standard database buffer pool and provide a mechanism to translate data between the inherent unstructured representation (stored in the computational cache) and the native database format (stored in slotted data pages). Applications then operate directly on the computational cache instead of on the native storage format.

We have implemented the methodology within a prototype system called Abacus that deals with 2D and 3D triangulation datasets. Not only is Abacus capable of storing and indexing pre-generated triangulations, but it also has the ability to support dynamic datasets—generating and indexing massive Delaunay triangulations from scratch on commodity servers. Performance evaluation shows that:

• Computing Delaunay triangulation using Abacus is more than three orders of magnitude faster than an implementation that uses standard database techniques.

• The performance of Abacus matches that of the state-of-the-art incore Delaunay triangulators when triangulating datasets that fit in memory.

• Abacus delivers scalable performance even when triangulating datasets four orders of magnitude larger than the main memory (while other software has stopped working long ago).

Furthermore, Abacus demonstrates its scalability in the context of a grand challenge application where it supports the generation of large-scale 3D Delaunay triangulated finite element meshes with multi-billion tetrahedral elements.

We conclude that in order to (1) bridge the structural mismatch between scientific datasets and traditional databases, and (2) deliver necessary performance and scalability for manipulating massive unstructured scientific datasets, we must seek a computational database solution.

DISSERTATION ABSTRACT:
Efficient Data Organization and Management on Heterogeneous Storage Hierarchies

Minglong Shao

Carnegie Mellon University School of Computer Science Ph.D. Dissertation, April 30, 2007.

Modern storage technology has advanced far beyond today's over simplified, out-of-date, or in some cases even wrong, assumptions taken by user applications about storage devices. While the simple abstraction of storage devices may work well for many applications, it falls short in data intensive applications such as database systems. It is time to revisit the overlooked rich features offered by hardware and design new data organization strategies that can exploit them and thus deliver higher performance.

Current data organization, based on the conventional linear abstract of storage devices, linearizes multidimensional data along a pre-selected

Kim Keeton, of HP Labs, gives feedback at the 2006 Workshop and Retreat.

dimension when storing them to disks. Therefore existing data organizations have inherent performance trade-offs in that they can only be optimized for workloads that access data along a single dimension while severely compromising the others. In addition, existing data management abstractions oversimplify memory hardware devices, which should be exploited to mitigate the performance problems caused by the increasing speed gap between CPUs and the memory hierarchy.

Toward this goal, I first propose DBmbench as a significantly reduced database microbenchmark suite which simulates OLTP and DSS workloads. DBmbench enables quick evaluation on new designs and provides forecasting for performance of real large scale benchmarks. I design and develop Clotho, which focuses on the page layout for relational tables. Clotho decouples the in-memory page layout from the storage organization by using a new query-specific layout called CSM. CSM combines the best performance of NSM and DSM, achieving good performance for both DSS and OLTP workloads. Experimentation on Clotho is based on Atropos, a new disk volume manager which exposes new efficient access paths on modern disks, and Lachesis. Then, I expand my work from two-dimensional layout design to multidimensional data mapping. MultiMap is a new mapping algorithm that stores multidimensional data onto disks without losing spacial locality. MultiMap exploits the new

continued on page 15

continued from page 14

adjacency model of disks to build a multidimensional structure on top of the linear disk space. It outperforms existing multidimensional mapping schemes on various spatial queries. After that, I propose new ways to organization intermediate results for two major query operators, hash join and external sorting, where the I/O performance of different execution phases exhibits similar trade-offs as those in 2-D table accesses. The experiments on our prototype demonstrate an up to 2X performance improvement over the existing implementation for systems with limited memory resource. And this is achieved without modifying the kernel algorithms.

DISSERTATION ABSTRACT:
Incremental Pattern Discovery on Streams, Graphs and Tensors

Jimeng Sun

Carnegie Mellon University School of Computer Science Ph.D. Dissertation, September 10, 2007.

Incremental pattern discovery targets at streaming applications where the data are arriving continuously in real-time. How to find patterns (main trends) in real-time? How to efficiently update the old patterns when new data arrive? How to utilize the pattern to solve other problem such as anomaly detection? For example, 1) a sensor network monitors a large number of distributed streams (such as temperature and humidity); 2) network forensics monitor the Internet communication patterns to identify the attacks; 3) cluster monitoring examines the system behaviors of a number of machines for potential failures; 4) social network analysis monitors a dynamic graph for communities and abnormal individuals; 5) financial fraud detection tries to find fraudulent activities from a large number of transactions in real-time. We first investigate a powerful data model tensor stream (TS) where there is one tensor per timestamp. To capture diverse data formats: we have a zero-order TS for a single time-series (stock price for Google over time), a first-order TS for multiple time-series (e.g., sensor measurement streams), a second-order TS for a matrix (e.g., graphs), and a high-order TS for a multi-array (e.g. Internet communication network, source-destination-port). Second, we develop different online algorithms on TS: 1) the centralized and distributed SPIRIT for mining a first-order TS as well as its extension on local correlation function and privacy preservation; 2) compact matrix decomposition (CMD) and GraphScope for a second-order TS; 3) the dynamic tensor analysis (DTA), streaming tensor analysis (STA) and window-based tensor analysis (WTA) for a high-order TS. All the techniques are evaluated extensively in real applications such as network forensics, cluster monitoring and financial fraud detection.

DISSERTATION ABSTRACT:
Using Content Addressable Techniques to Optimize Client-Server Systems

Niraj Tolia

Carnegie Mellon University School of Computer Science Ph.D. Dissertation, October 22, 2007.

Efficient access to bulk data over the Internet has become an critical problem in today's world. Even while bandwidth, both in the core and the edge of the network, is improving, the simultaneous growth in the use of digital media and large personal data sets is placing increasing demands on it. Further, with increasing trends towards mobility, an increasing amount of data access is over cellular and wireless networks. These trends are placing pressure on applications and the Wide-Area Network (WAN) to deliver better performance to the end user. This dissertation puts forward the claim that external resources can be used in an opportunistic manner to optimize bulk data transfer in WAN-based client-server systems. In particular, it advocates the use of content-addressable techniques, a system-independent method for naming objects, to cleanly enable these optimizations. By detecting similarity between different data sources, these techniques allow external sources to be used without weakening any attributes, including consistency, of legacy systems and with no or minor changes to the original system.

This dissertation validates this claim empirically through the use of five case studies that encompass the two traditional forms of data storage, file systems and database systems, and then generalize the claim in the form of a generic transfer service that can be shared by different applications. In each of these case studies, I focus on three questions. First, even with the addition of content-based optimizations, how does the resulting system still maintain the attributes and semantics of the original system? Second, how do these systems efficiently find similarity in data? Third, does the use of these content-addressable techniques provide a substantial performance improvement when compared to the original system? These questions are answered with a detailed description of the system design and implementation and a careful evaluation of the prototypes with both synthetic and real benchmarks.

Mike Mesnier discusses his research with Gary Grider and James Nunez, both of Los Alamos National Laboratory.

continued on page 18

continued from page 7

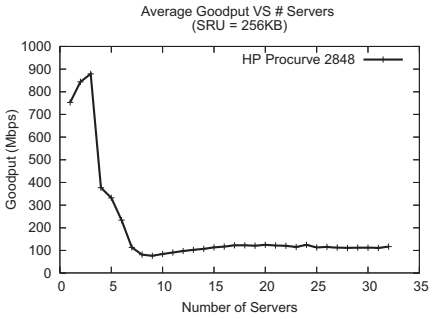
exploits modern disk characteristics to provide full streaming bandwidth for one (primary) dimension and maximally efficient non-sequential access (i.e., minimal seek and no rotational latency) for the other dimensions. This is in contrast to existing approaches, which either severely penalize non-primary dimensions or fail to provide full streaming bandwidth for any dimension. Experimental evaluation of a prototype implementation demonstrates MultiMap's superior performance for range and beam queries. On average, MultiMap reduces total I/O time by over 50% when compared to traditional linearized layouts and by over 30% when compared to space-filling curve approaches such as Z-ordering and Hilbert curves. For scans of the primary dimension, MultiMap and traditional linearized layouts provide almost two orders of magnitude higher throughput than space-filling curve approaches.

Measurement and Analysis of TCP Throughput Collapse in Cluster-based Storage Systems

Phanishayee, Krevat, Vasudevan, Andersen, Ganger, Gibson & Seshan

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-07-105, September 2007.

Cluster-based and iSCSI-based storage systems rely on standard TCP/IP-over-Ethernet for client access to data. Unfortunately, when data is striped over multiple networked storage nodes, a client can experience a TCP throughput collapse that results in much lower read bandwidth than should be provided by the available network links. Conceptually, this problem arises because the client simultaneously reads fragments of a data block from multiple sources that together send enough data to overload the switch buffers on the client's link. This paper analyzes this Incast problem, explores its sensitivity to various system parameters, and examines the effectiveness of alternative TCP- and Ethernet-level strategies in mitigating the TCP throughput collapse.



TCP throughput collapse for a synchronized reads application performed on a storage cluster.

Scheduling Threads for Constructive Cache Sharing on CMPs

Chen, Gibbons, Kozuch, Liaskovitis, Ailamaki, Blelloch, Falsafi, Fix, Hardavellas, Mowry & Wilkerson

19th ACM Symposium on Parallelism in Algorithms and Architectures (SPAA' 07), San Diego, CA, June 2007.

In chip multiprocessors (CMPs), limiting the number of off-chip cache misses is crucial for good performance. Many multithreaded programs provide opportunities for *constructive* cache sharing, in which concurrently scheduled threads share a largely overlapping working set. In this paper, we compare the performance of two state-of-the-art schedulers proposed for ne-grained multithreaded programs: Parallel Depth First (PDF), which is specifically designed for constructive cache sharing, and Work Stealing (WS), which is a more traditional design. Our experimental results indicate that PDF scheduling yields a 1.3–1.6X performance improvement relative to WS for several fine-grain parallel benchmarks on projected future CMP configurations; we also report several issues that may limit the advantage of PDF in certain applications. These results also indicate that PDF more effectively utilizes off-chip bandwidth, making it possible to trade-off on-chip cache for a larger number of cores. Moreover, we find that task granularity plays a key role in cache performance. Therefore, we present an automatic approach for selecting effective grain sizes, based on a new working set profiling algorithm that is an order of magnitude faster than previous approaches. This is the first paper demonstrating the effectiveness of PDF on real benchmarks, providing a direct comparison between PDF and WS, revealing the limiting factors for PDF in practice, and presenting an approach for overcoming these factors.

Using Provenance to Aid in Personal File Search

Shah, Soules, Ganger & Noble

Usenix Annual Technical Conference, Santa Clara, CA, June 17–22, 2007.

As the scope of personal data grows, it becomes increasingly difficult to find what we need when we need it. Desktop search tools provide a potential answer, but most existing tools are incomplete solutions: they index content, but fail to capture dynamic relationships from the user's context. One emerging solution to this is context-enhanced search, a technique that reorders and extends the results of content-only search using contextual information. Within this framework, we propose using strict *causality*, rather than temporal locality, the current state of the art, to direct contextual searches. Causality more accurately identifies data flow between files, reducing the false-positives created by context-switching and background noise. Further, unlike previous work, we conduct an online user study with a fully-functioning implementation to evaluate *user-perceived* search quality directly. Search results generated by our causality mechanism are rated a statistically-significant 17% higher, on average over all queries, than by using content-only search or

continued on page 17

continued from page 16

context-enhanced search with temporal locality.

Verifying Distributed Erasure-coded Data

Hendricks, Ganger & Reiter

Twenty-Sixth Annual ACM SIGACT-SIGOPS Symposium on Principles of Distributed Computing (PODC 2007), Portland, August 2007.

Erasure coding can reduce the space and bandwidth overheads of redundancy in fault-tolerant data storage and delivery systems. But it introduces the fundamental difficulty of ensuring that all erasure-coded fragments correspond to the same block of data. Without such assurance, a different block may be reconstructed from different subsets of fragments. This paper develops a technique for providing this assurance without the bandwidth and computational overheads associated with current approaches. The core idea is to distribute with each fragment what we call homomorphic fingerprints. These fingerprints preserve the structure of the erasure code and allow each fragment to be independently verified as corresponding to a specific block. We demonstrate homomorphic fingerprinting functions that are secure, efficient, and compact.

An Analysis of Database System Performance on Chip Multiprocessors

Hardavellas, Pandis, Johnson, Mancheril, Harizopoulos, Ailamaki & Falsafi

Sixth Hellenic Data Management Symposium (HDMS2007), Athens, Greece, July 2007.

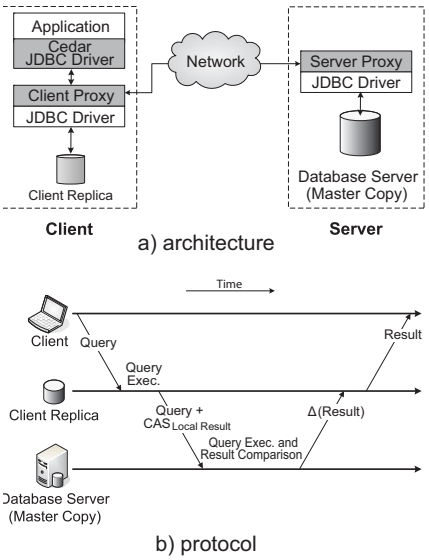
Prior research shows that database system performance is dominated by off-chip data stalls, resulting in a concerted effort to bring data into on-chip caches. At the same time, high levels of integration have enabled the advent of chip multiprocessors and increasingly large (and slow) on-chip caches. These two trends pose the imminent technical and research challenge of adapting high-performance data management software to a shifting hardware landscape. In this paper we characterize the performance of a commercial database server running on emerging chip multiprocessor technologies. We find that the major bottleneck of current software is data cache stalls, with L2 hit stalls rising from oblivion to become the dominant execution time component in some cases. We analyze the source of this shift and derive a list of features for future database designs to attain maximum performance. Towards this direction, we propose the adoption of staged database system designs to achieve high performance on chip multiprocessors. We present the basic principles of staged databases and an initial implementation of such a system, called Cordoba.

Improving Mobile Database Access Over Wide-Area Networks Without Degrading Consistency

Tolia, Satyanarayanan & Wolbach

MobiSys '07, June 11–13, 2007, San Juan, Puerto Rico, USA.

We report on the design, implementation, and evaluation of a system called Cedar that enables mobile database access with good performance over low-bandwidth networks. This is accomplished without degrading consistency. Cedar exploits the disk storage and processing power of a mobile client to compensate for weak connectivity. Its central organizing principle is that even a stale client replica can be used to reduce data transmission volume from a database server. The reduction is achieved by using content addressable storage to discover and elide commonality between client and server results. This organizing principle allows Cedar to use an optimistic approach to solving the difficult problem of database replica control. For laptop-class clients, our experiments show that Cedar



Proxy-Based Cedar Implementation. a) shows how we transparently interpose Cedar into an existing client-server system that uses Java Database Connectivity (JDBC). The gray boxes represent Cedar components. b) maps this architecture to the protocol executed for a `select` query.

improves the throughput of read-write workloads by 39% to as much as 224% while reducing response time by 28% to as much as 79%.

Low-overhead Byzantine Fault-tolerant Storage

Hendricks, Ganger & Reiter

Twenty-First ACM Symposium on Operating Systems Principles (SOSP 2007), Stevenson, WA, October 2007.

This paper presents an erasure-coded Byzantine fault-tolerant block storage protocol that is nearly as efficient as protocols that tolerate only crashes. Previous Byzantine fault-tolerant block storage protocols have either relied upon replication, which is inefficient for large blocks of data when tolerating multiple faults, or a combination of additional servers, extra computation, and versioned storage. To avoid these expensive techniques, our protocol

continued on page 20

continued from page 15

DISSERTATION ABSTRACT:
Advanced Tools for Multimedia Data Mining

Jia-Yu Pan

Carnegie Mellon University School of Computer Science Ph.D. Dissertation, April 7, 2007.

How do we automatically find patterns and do data mining in large multimedia databases, to make these databases useful and accessible? We focus on two problems: (1) mining "uni-modal patterns" that summarize the characteristics of a data modality, and (2) mining the "cross-modal correlations" among multiple modalities. Uni-modal patterns such as "news videos have static scenes and speech-like sounds", and cross-modal correlations like "the blue region at the upper part of a natural scene image is likely to be the 'sky'", could provide insights on the multimedia content and have many applications.

For uni-modal pattern discovery, we propose the method "AutoSplit". AutoSplit provides a framework for mining meaningful "independent components" in multimedia data, and can find patterns in a wide variety of data modalities (video, audio, text, time sequence). For example, in video clips, AutoSplit finds characteristic visual/auditory patterns, and can classify news and commercial clips with 81% accuracy. In time sequences like stock prices, AutoSplit finds hidden variables like "general growth trend" and "Internet bubble", and can detect outliers (e.g., lackluster stocks). Based on AutoSplit, we design a system, ViVo, for mining biomedical images. ViVo automatically constructs a visual vocabulary which is biologically meaningful and can classify 9 biological conditions with 84% accuracy. Moreover, ViVo supports data mining tasks such as highlighting biologically interesting image regions, for biomedical research.

For cross-modal correlation discovery, we propose "MAGIC", a graph-based framework for multimedia correlation mining. When applied to news video databases, MAGIC can identify relevant video shots and transcript words for event summarization. On the task of automatic image captioning, MAGIC achieves a relative improvement by 58% in captioning accuracy as compared to recent machine learning techniques.

DISSERTATION ABSTRACT:
Enabling What-if Explorations in Systems

Eno Thereska

Carnegie Mellon University Department of Electrical & Computer Engineering Ph.D. Dissertation, June 25, 2007.

With a large percentage of total system cost going to system administration tasks, ease of system management remains a difficult and important goal. As a step towards that goal, this dissertation presents a success story on building systems that are self-predicting. Self-predicting systems continuously monitor themselves and provide quantitative answers to What...if questions about hypothetical workload or resource changes. Self-prediction has the potential to simplify administrators' decision making, such as acquisition planning and performance tuning, by reducing the detailed workload and internal system knowledge required.

Self-prediction has as the primary building block mathematical models, that, once built into the system, analyze past, and predict future behavior. Because of the traditional disconnect between systems researchers and theoretical researchers, however, there are fundamental difficulties in enabling existing mathematical models to make meaningful predictions in real systems. In part, this dissertation serves as a bridge between research in theory (e.g., queuing theory and statistical theory) and research in systems (e.g., database and storage systems). It identifies ways to build systems to support use of mathematical models and addresses fundamental show-stoppers that keep models from being useful in practice. For example, we explore many opportunities to deeply understand workload-system interactions by having models be first-class system components, rather than developing and deploying them separately from the system, as is traditionally done. As another example, lack of good measurement information in a distributed system can be a show-stopper for models based on queuing analysis. This dissertation introduces a measurement framework that replaces performance counters with end-to-end activity tracing. End-to-end tracing allows contextual information to be propagated with requests so that queuing models can attribute resource demands to the correct workloads. In addition, this dissertation presents a first step towards a robust, hybrid mathematical modeling framework, based on models that reflect domain expertise and models that guide model designers to discover new, unforeseen system behavior once the system is deployed. Such robust models could con-

Greg Ganger fills in for Natassa Ailamaki, introducing her student Nikos Hardavellas, before his talk on "Database Servers on Chip Multiprocessors: Limitations and Opportunities" at the 2006 PDL Retreat.

continued on page 19

continued from page 18

tinuously evaluate their accuracy and adjust their predictions accordingly. Self-evaluation can enable confidence values to be provided with predictions, including identification of situations where no trustworthy predictions can be produced.

Through an analysis of positive and negative lessons learned, in a storage system that we designed from scratch as well as in a legacy commercial database system, this dissertation makes the case that systems can be built to accommodate mathematical models efficiently, but cautions that mathematical models are not a panacea. Models are as good as the system is; to make predictions more meaningful, systems should be built so that they are inherently more predictable to start with.

DISSERTATION ABSTRACT:
Methods for Querying Compressed Wavefields

Julio López

Carnegie Mellon University
Department of Electrical & Computer Engineering Ph.D. Dissertation.

Large wavefield datasets are becoming increasingly large due to improvements in simulation techniques and advances in computer systems. Larger storage capacities enable scientists to store more data. For example, state of the art ground-motion numerical solvers produce Terabyte-size datasets per simulation. Operating on these datasets becomes extremely challenging due to decades of declining normalized storage performance both in terms of access latency and throughput. Data access and transfer rates have not kept pace with the increase in storage capacity or CPU performance, i.e., (seek time/disk capacity) and (transfer bandwidth / disk capacity) have decreased. As dataset sizes increase, it takes much longer to access the data on disk. We present new mechanisms that allow querying and processing large wavefields in the compressed domain (i.e, directly in their compressed representation). These mechanisms combine well-known spatial-indexing techniques with novel compressed representations in order to reduce bandwidth requirements when moving data from storage to main memory. The compression technique uses frequency domain representation to take advantage of the temporal redundancy found in wave propagation data, coupled with a new representation based on boundary integral equations which takes advantage of data spatial coherence. This approach transforms a large I/O problem into a massively-parallel CPU-intensive computation. Common queries to these datasets result in difficult to handle I/O workloads with semi-random access patterns. In the proposed representation I/O access patterns exhibit larger sequential patterns. The decompression stage for this approach places heavy demands on the CPU. The good news is that the decompression can be performed in parallel, and is well-suited for the surfacing many-core processors. We evaluate our approach in the context of post-processing of dataset produced by CMU's Quake project.

DISSERTATION ABSTRACT:
Automated Database Design for Large-Scale Scientific Applications

Efstratios Papadomanolakis

Carnegie Mellon University
Department of Electrical & Computer Engineering Ph.D. Dissertation.

The need for large-scale scientific data management is today more pressing than ever, as modern sciences need to store and process terabyte-scale data volumes. Traditional systems, relying on filesystems and custom data access and processing code do not scale for multi-terabyte datasets. Therefore, supporting today's data-driven sciences requires the development of new data management capabilities.

This Ph.D dissertation develops techniques that allow modern Database Management Systems (DBMS) to efficiently handle large scientific datasets. Several recent successful DBMS deployments target applications like astronomy, that manage collections of objects or observations (e.g. galaxies, spectra) and can easily store their data in a commercial relational DBMS. Query performance for such systems critically depends on the *database physical design*, the organization of database structures such as indexes and tables. This dissertation develops algorithms and tools for *automating* the physical design process. Our tools allow databases to tune themselves, providing efficient query execution in the presence of large data volumes and complex query workloads.

For more complex applications dealing with multidimensional and time-varying data, standard relational DBMS are inadequate. Efficiently supporting such applications requires the development of novel indexing and query processing techniques. This dissertation develops an indexing technique for *unstructured tetrahedral meshes*, a multidimensional data organization used in finite element analysis applications. Our technique outperforms existing multidimensional indexing techniques and has the advantage that can easily be integrated with standard DBMS, providing existing systems with the ability to handle spatial data with minor modifications.

continued on page 24

PDL Alumnus Erik Riedel (Seagate), recent graduate Eno Thereska (Microsoft Research, UK) and Elie Krevat (PDL graduate student) at the 2007 Spring Industry visit Day.

continued from page 17

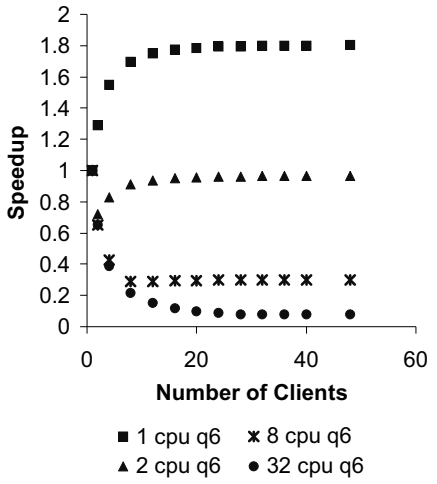
employs novel mechanisms to optimize for the common case when faults and concurrency are rare. In the common case, a write operation completes in two rounds of communication and a read completes in one round. The protocol requires a short checksum comprised of cryptographic hashes and homomorphic fingerprints. It achieves throughput within 10% of the crash-tolerant protocol for writes and reads in failure-free runs when configured to tolerate up to 6 faulty servers and any number of faulty clients.

To Share Or Not To Share?

Johnson, Hardavellas, Pandis, Mancheril, Harizopoulos, Sabirli, Ailamaki & Falsafi

Proceedings of the 33rd International Conference on Very Large Data Bases (VLDB'07), Vienna, Austria, September 2007.

Intuitively, aggressive work sharing among concurrent queries in a database system should always improve performance by eliminating redundant computation or data accesses. We show that, contrary to common intuition, this is not always the case in practice, especially in the highly parallel world of chip multiprocessors. As the number of cores in the system increases, a trade-off appears between exploiting work sharing opportunities and the available parallelism. To resolve the trade-off, we develop an analytical approach that predicts the effect of work sharing in multi-core systems. Database systems can use the model to determine, statically or at runtime, whether work sharing is beneficial and apply it only when appropriate. The contributions of this paper are as follows. First, we introduce and analyze the effects of the trade-off between work sharing and parallelism on database systems running complex decision-support queries. Second, we propose an intuitive and simple model that can evaluate the trade-off using real-world measurement approximations of the query execution processes. Furthermore, we integrate the model into a prototype database execution engine, and demonstrate that selective work sharing according to the model outperforms never-share static schemes by 20% on average and always-share ones by 2.5x.



Speedup when sharing part of a data warehousing query (TPC-H query 6) relative to never-share execution.

Categorizing and Differencing System Behaviours

Sambasivan, Zheng, Thereska & Ganger

Second Workshop on Hot Topics in Autonomic Computing. June 15, 2007. Jacksonville, FL.

Making request flow tracing an integral part of software systems creates the potential to better understand their operation. The resulting traces can be converted to per-request graphs of the work performed by a service, representing the flow and timing of each request's processing. Collectively, the graphs contain detailed and comprehensive data about the system's behavior and the workload that induced it, leaving the challenge of extracting insights. Categorizing and differencing such graphs should greatly improve our ability to understand the runtime behavior of complex distributed services and diagnose problems. Clustering the set of graphs can identify common request processing paths and expose outliers. Moreover, clustering two sets of graphs can expose differences between the two; for example, a programmer could diagnose a problem that arises by comparing current request processing with that of an earlier non-problem period and focusing on the aspects that change. Such categorizing and differencing of system behavior can be a big step in the direction of automated problem diagnosis.

Modeling the Relative Fitness of Storage

Mesnier, Wachs, Sambasivan, Zheng & Ganger

SIGMETRICS'07, June 12–16, 2007, San Diego, California, USA.

Relative fitness is a new black-box approach to modeling the performance of storage devices. In contrast with an absolute model that predicts the performance of a workload on a given storage device, a relative fitness model predicts performance *differences* between a pair of devices. There are two primary advantages to this approach. First, because a relative fitness model is constructed for a device pair, the application-device feedback of a closed workload can be captured (e.g., how the I/O arrival rate changes as the workload moves from device A to device B). Second, a relative fitness model allows performance and resource utilization to be used in place of workload characteristics. This is beneficial when workload characteristics are difficult to obtain or concisely express (e.g., rather than describe the spatio-temporal characteristics of a workload, one could use the observed cache behaviour of device A to help predict the performance of B).

This paper describes the steps necessary to build a relative fitness model, with an approach that is general enough to be used with any black-box

continued on page 22

continued from page 9

ECE and ISR, and Dawn Song, ECE and CSD.

A Sloan Fellowship is a prestigious award intended to enhance the careers of the very best young faculty members in specified fields of science. Currently a total of 118 fellowships, valued at $45,000, are awarded annually in seven fields: chemistry, computational and evolutionary molecular biology, computer science, economics, mathematics, neuroscience, and physics. Only 16 are given in computer science each year so CMU once again shines. Since the establishment of the fellowships in 1955, Thirty-two Sloan Fellows have gone on to win Nobel Prizes.

February 2007

PDL Researchers Win Best Paper at FAST 2007!

Bianca Schroeder and Garth Gibson brought home the Best Paper Award from the 5th USENIX Conference on File and Storage Technologies (FAST 2007), held in San Jose, CA this year. The award was given for their research on "Disk Failures in the Real World: What Does an MTTF of 1,000,000 Hours Mean to You?"

PDL researchers have been well received at past FAST conferences too, winning best student paper awards in 2002 ("Track-Aligned Extents: Matching Access Patterns to Disk Drive Characteristics", Schindler, et al.) and in 2004 ("A Framework for Building Unobtrusive Disk Maintenance Applications", Thereska et al.). In 2005 the PDL brought home both best paper awards for work on "Ursa Minor: Versatile Cluster-based Storage", Abd-El-Malek et al. and "On Multidimensional Data and Modern Disks", Schlosser et al.

January 2007

Dawn Song Selected for College of Engineering Award

Dawn Song is among the three ECE faculty members who won awards from the College of Engineering this year. Song, Assistant Professor of ECE and Computer Science, received a George Tallman Ladd Research Award, which is granted in recognition of outstanding research, professional accomplishments, and potential.

-- with info from ECE News Online

December 2006

Christos Faloutsos Receives 2006 Research Contributions Award

The ICDM Research Contributions Award is given to one individual or one group who has made influential contributions to the field of data mining. The 2006 IEEE ICDM Research Contributions Award goes to Prof. Christos Faloutsos of Carnegie Mellon University, to recognize his research contributions in the areas of mining for graphs and streams, and for searching and mining temporal and video data. The award was given at a ceremony during the ICDM 2006 Conference in Hong Kong.
--ACM/SIGMOD News

YEAR IN REVIEW

continued from page 4

eral conferences including SIAM Data Mining, SIGMOD 20007, ICML 2007 and KDD 2007.

- Jimeng Sun, Hui Zhang and Christos Faloutsos win best paper for "Less Is More: Compact Matrix Decomposition for Large Sparse Graphs" at SIAM'07 (Society for Indust. & Applied Mathematics).

April 2007

- Minglong Shao completed her Ph.D. with her defense of "Efficient Data Organization and Management on Heterogeneous Storage Hierarchies." She is now with Network Appliance.
- Jia-Yu Pan successfully defended his dissertation "Advanced Tools for Multimedia Data Mining."

February 2007

- Mike Mesnier presented "//TRACE: Parallel Trace Replay with Approximate Causal Events" at FAST '07 in San Jose, CA.
- Bianca Schroeder presented "Disk Failures in the Real World: What Does an MTTF of 1,000,000 Hours Mean to You?" and was awarded Best Paper at the 5th USENIX Conf. on File and Storage Technologies (FAST 2007), San Jose, CA.
- Matthew Wachs presented "Argon: Performance Insulation for Shared Storage Servers "at FAST '07 in San Jose, CA.

January 2007

- Alina Mihaela Oprea successfully defended her Ph.D. research on "Efficient Cryptographic Techniques for Securing Storage Systems." Alina is now with RSA Labs (A division of EMC).

December 2006

- Andrew Klosterman proposed his Ph.D. research, titled "Delayed Instantiation Bulk Operations in a Clustered, Object-based Storage System."
- Christos Faloutsos receives 2006 IEEE ICDM Research Contributions Award.

October 2005

- 14th Annual PDL Retreat and Workshop.

DATA INTENSIVE SUPER COMPUTING

continued from page 11

hardware—we can easily throttle back component performance to study the capabilities of lesser hardware. Over time, we would migrate the software being developed as part of the systems research to the partitions supporting applications programming.

In pursuing this evolutionary approach, we must decide what forms of compatibility we would seek to maintain. Our current thinking is that any compatibility should only be provided at a very high level, such that an application written in terms of something like MapReduce [7] or other high-level constructs can continue to operate with minimal modifications, but with no guarantees of compatibility. Otherwise, systems researchers would be overly constrained to follow the same paths set by existing projects.

Conclusion

Just as web search has become an essential tool in the lives of people ranging from schoolchildren to academic researchers to senior citizens, we believe that DISC systems could change the face of scientific research worldwide. We're also confident that any work in this area will have great impact on the many industries that rely on powerful and capable information technology. In nearly every domain, ranging from retail services to health care delivery, to the basic sciences, vast amounts of data are being collected and analyzed around the world. DISC will help realize the potential all this data provides.

References

[1] Google tops translation ranking. News@Nature, Nov. 6, 2006.

[2] V. Akcelik, J. Bielak, G. Biros, I. Epanomeritakis, A. Fernandez, O. Ghattas, E. J. Kim, J. Lopez, D. R. O'Hallaron, T. Tu, and J. Urbanic. High resolution forward and inverse earthquake modeling on terasacale computers. In Proceedings of SC2003, November 2003.

[3] E. A. Brewer. Delivering high availability for Inktomi search engines. In L. M. Haas and A. Tiwary, editors, ACM SIGMOD International Conference on Management of Data, page 538. ACM, 1998.

[4] J. Markoff and S. Hansell. Hiding in plain sight, Google seeks more power. New York Times, June 14, 2006.

[5] J.Markoff. Sun and IBM offer new class of high-end servers. New York Times, Apr. 26, 2007.

[6] The Hadoop Project. http://lucene.apache.org/hadoop/

[7] Google Map Reduce. http://labs.google.com/papers/mapreduce.html

RECENT PUBLICATIONS

continued from page 20

modeling technique. We compare relative fitness models and absolute models across a variety of workloads and storage devices. On average, relative fitness models predict bandwidth and throughput within 10–20% and can reduce prediction error by as much as a factor of two when compared to absolute models.

VMM-Independent Graphics Acceleration.

Lagar-Cavilla, Tolia , Satyanarayanan & de Lara

VEE'07, June 13–15, 2007, San Diego, California, USA.

This paper describes VMGL, a cross-platform OpenGL virtualization solution that is both VMM and GPU independent. VMGL allows applications executing within virtual machines (VMs) to leverage hardware rendering acceleration, thus solving a problem that has limited virtualization of a growing class of graphics-intensive applications. VMGL also provides applications running within VMs with suspend and resume capabilities across GPUs from different vendors. Our experimental results from a number of graphics-intensive applications show that VMGL provides excellent rendering performance, coming within 14% or better of native graphics hardware acceleration. Further, VMGL's performance is two orders of magnitude better than that of software rendering, the commonly available alternative today for graphics-intensive applications running in virtualized environments. Our results confirm VMGL's portability across VMware Workstation and Xen (on VT and non-VT hardware), and across Linux (with and without paravirtualization), FreeBSD, and Solaris. Finally, the resource demands of VMGL align well with the emerging trend of multi-core processors.

SWAP: Shared Wireless Access Protocol (using Reciprocity)

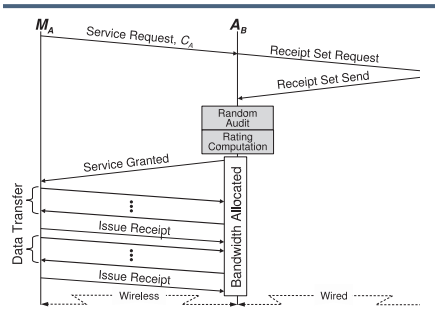
Dunlop, Perng & Andersen.

IEEE Workshop on Information Assurance, June 2007.

Wireless access points are becoming more and more prominent in the home, yet there is no incentive to encourage access point owners to share their service. We introduce SWAP, a lightweight protocol that uses reciprocity to motivate users to share service. Each node participating in SWAP stores perishable receipts that are used to calculate a user's rating (how much

continued on page 23

continued from page 23



SWAP protocol.

the user shares his or her access point). SWAP does not use a centralized authority to store or validate receipts nor does it place an excessive burden on peers. SWAP is also robust against collusion, which we show through analysis and implementation. As demonstrated by an implementation of the most computationally expensive portions of the protocol, SWAP imposes little overhead even on mobile devices.

Understanding Failure in Petascale Computers

Schroeder & Gibson

SciDAC 2007. To appear in the Journal of Physics: Conf. Ser. 78.

With petascale computers only a year or two away there is a pressing need to anticipate and compensate for a probable increase in failure and application interruption rates. Researchers, designers and integrators have available to them far too little detailed information on the failures and interruptions that even smaller terascale computers experience. The information that is available suggests that application interruptions will become far more common in the coming decade, and the largest applications may surrender large fractions of the computer's resources to taking checkpoints and restarting from a checkpoint after an interruption. This paper reviews sources of failure information for compute clusters and storage systems, projects failure rates and the corresponding decrease in application effectiveness, and discusses coping strategies such as application-level checkpoint compression and system level process-pairs fault-tolerance for supercomputing. The need for a public repository for detailed failure and interruption records is particularly concerning, as projections from one architectural family of machines to another are widely disputed. To this end, this paper introduces the Computer Failure Data Repository and issues a call for failure history data to publish in it.

Less is More: Compact Matrix Decomposition for Large Sparse Graphs

Sun, Xie, Zhang & Faloutsos

SDM'07, Minneapolis, MN, USA, April 26-28, 2007.

Given a large sparse graph, how can we find patterns and anomalies? Several important applications can be modeled as large sparse graphs, e.g., network traffic monitoring, research citation network analysis, social network analysis, and regulatory networks in genes. Low rank decompositions, such as SVD and CUR, are powerful techniques for revealing latent/hidden variables and associated patterns from high dimensional data. However, those methods often ignore the sparsity property of the graph, and hence usually incur too high memory and computational cost to be practical.

We propose a novel method, the Compact Matrix Decomposition (CMD), to compute sparse low rank approximations. CMD dramatically reduces both the computation cost and the space requirements over existing decomposition methods (SVD, CUR). Using CMD as the key building block, we further propose procedures to efficiently construct and analyze dynamic graphs from real-time application data. We provide theoretical guarantee for our methods, and present results on two real, large datasets, one on network flow data (100GB trace of 22K hosts over one month) and one on DBLP (200MB over 25 years).

We show that CMD is often an order of magnitude-more efficient than the state of the art (SVD and CUR): it is over 10X faster, but requires less than 1/10 of the space, for the same reconstruction accuracy. Finally, we demonstrate how CMD is used for detecting anomalies and monitoring time-evolving graphs, in which it successfully detects worm-like hierarchical scanning patterns in real network data.

The Computer Failure Data Repository

Schroeder & Gibson

Invited contribution to the Workshop on Reliability Analysis of System Failure Data (RAF'07) March 1-2 2007, Cambridge, UK.

System reliability is a major challenge in system design. Unreliable systems are not only major source of user frustration, they are also expensive. Avoiding downtime and the cost of actual downtime make up more than 40% of the total cost of ownership for modern IT systems. Unfortunately, with the large component count in today's large-scale systems, failures are quickly becoming the norm rather than the exception.

This paper describes an effort currently underway at CMU to create a public Computer Failure Data Repository (CFDR), sponsored by USENIX. The goal of the repository is to accelerate research on system reliability by filling the nearly empty collection of public data with detailed failure data from a variety of large production systems. Below we give a brief overview of the data sets we have collected so far, and discuss our ongoing efforts and the long-term goals of the CFDR.

continued on page 26

continued from page 19

THESIS PROPOSAL:
Reusing Dynamic Redistribution to Eliminate Cross-Server Operations and Maintain Semantics While Scaling Storage Systems

Shafeeq Sinnamohideen, SCS

Distributed file systems that scale by partitioning files and directories among a collection of servers inevitably encounter cross-server operations. A common example is a rename that moves a file from a directory managed by one server to a directory managed by another. Systems that provide the same semantics for cross-server operations as for those that do not span servers traditionally implement dedicated protocols for these rare operations.

This thesis explores an alternate approach, with simplicity as a goal, that exploits the existence of dynamic redistribution functionality (e.g., for load balancing, incorporation of new servers, and so on). When a client request would involve files on multiple servers, the system can redistribute those files onto one server and have it service the request. Although such redistribution is more expensive than a dedicated cross-server protocol, preliminary analysis of NFS traces indicates that such operations are extremely rare in file system workloads. Thus, when dynamic redistribution functionality exists in the system, cross-server operations can be handled with very little additional implementation complexity.

THESIS PROPOSAL:
Dynamics of Real-world Networks

Jure Leskovec, SCS

In our recent work we found very interesting and unintuitive patterns for time evolving networks, which change some of the basic assumptions that were made in the past. The main objective of observing the evolution patterns is to develop models that

Andrew Klosterman discusses his poster with Stephen Harpster of Network Appliance at the Spring Industry Visit Day.

explain processes which govern the network evolution. Such models can then be fitted to real networks, and used to generate realistic graphs or give formal explanations about their properties. In addition, our work has a wide range of applications: we can spot anomalous graphs and outliers, design better graph sampling algorithms, forecast future graph structure and run simulations of network evolution. Another important aspect of this research is the study of "local" patterns and structures of propagation in networks. We aim to identify building blocks of the networks and find the patterns of influence that these block have on information or virus propagation over the network. Our recent work included the study of the spread of influence in a large person-to-person product recommendation network and its effect on purchases. We also model the propagation of information on the blogosphere, and propose algorithms to efficiently find influential nodes in the network.

Further work will include three areas of research. We will continue investigating models for graph generation and evolution. Second, we will analyze large online communication networks and devise models on how user characteristics and geography relate to communication and network patterns. Third, we will extend the work on the propagation of influence in recommendation networks to blogs on the Web, studying how information spreads over the Web by finding influential blogs and analyzing their patterns of influence. We will also study how the local behavior affects the global structure of the network.

THESIS PROPOSAL:
Optimizing Chip Multi-Processors for Commercial Workloads

Nikos Hardavellas, SCS

Technological advancements in semiconductor fabrication have led to enormous levels of chip integration. Future chips will have the area budget to host hundreds of cores, tens of megabytes of on-chip cache and reach enormous clock speeds. At the same time, increases in on-chip communication delays require a departure from traditional caches with a single, uniform access time. Instead, cache designs in future chips will expose a continuum of latencies to the application, making the hit time a function of the line's physical location within the cache. Unfortunately, commercial workloads exhibit adverse memory access patterns that hinder performance and are oblivious to physical cache line placement.

The goal of this thesis is to propose chip multi-processor designs that attain maximum performance when executing commercial workloads, while conforming to area, power, and bandwidth constraints. The preliminary results provide design guidelines for chip multi-processors that optimally allocate shared on-chip hardware resources for each process technology while respecting the pertinent tradeoffs. At the same time, simple architectural mechanisms remove data stalls from the workload's critical path. The results of this thesis are applicable to conventional software server designs, while additional hardware support facilitates even higher performance for the emerging software paradigm of staged software servers.

continued on page 25

continued from page 24

THESIS PROPOSAL:
Putting Home Storage Replica Management into Perspective

Brandon Salmon, ECE

While the recent increase in the number and power of home electronic devices presents home users with exciting new functionality, it also presents new challenges in managing the data stored and used between these devices. This dissertation will present a new abstraction called views which provide a powerful framework for users to accomplish home replica management tasks. In particular, views allow users to set mobility and reliability preferences. This dissertation will evaluate views using a prototype system and a combination of lab and long-term user studies. These studies will also provide insight into the yet-unstudied data access patterns of home users.

THESIS PROPOSAL:
Delayed Instantiation Bulk Operations in a Clustered, Object-based Storage System

Andrew J. Klosterman, ECE

Many storage management tasks contain, at their heart, a step that applies the same operation to many stored data items: a bulk operation. Such bulk operations have evolved over the years in enterprise-class block- and file-based storage systems. A new breed of storage, object-based storage, will benefit from supporting bulk operations that have come to be expected in established storage systems.

This thesis proposes the investigation of ways to support the execution of specific management tasks on flexibly defined sets of objects in a clustered, object-based storage system. The semantics and performance of such tasks are expected to at least meet, and hopefully exceed, those of supporting operations on current block- and file-based storage systems. This is to be done while coping with the challenges presented by the distributed nature of storage in clustered object-based storage systems.

Through the delayed instantiation of bulk operations on objects, performance will be maintained and operation semantics upheld. The use of copy-on-write and lazy evaluation techniques, along with capability-based access control, enables the use of delayed instantiation.

By investigating the effects of delayed instantiation on different workloads and client-access scenarios, the associated costs can be measured and compared in a prototype clustered, object-based storage system. Furthermore, advantageous structuring of higher-level storage systems built atop the prototype will be demonstrated and characterized.

THESIS PROPOSAL:
Heterogeneous Intrusion Detection Fusion

Adam Pennington, ECE

Current intrusion detection systems (IDSs) are good at watching for attacks, but have a much harder time determining if an intrusion has occurred. This gives an administrator a low level of confidence that a given alert indicates an intrusion. Additionally these intrusion systems only are able to watch for a small number of activities that might be caused by an intrusion, resulting in less coverage of these indications than could be possible. Self-securing devices, along with other types of IDSs, give new vantage points from which to watch for intrusion activities and data that can better indicate if an intrusion has taken place. My thesis proposes that alert data from intrusion detection systems with different types of vantage points can be combined to give improved coverage and confidence over traditional intrusion detection systems acting alone. By using a configuration like this, called a Honeynet, we are able to gather real world attack data that is used to demonstrate this improvement.

THESIS PROPOSAL:
Maintaining Consistent Replication through Program Analysis

Joe Slember, ECE

State-machine replication is often precluded as an option for providing fault-tolerance to nondeterministic distributed applications. I seek to support the state-machine replication of both deterministic and nondeterministic applications alike. To this end, I employ the offline static analysis of the application/infrastructure source-code to derive key insights about the application. The recognition of superficial nondeterminism, and its prevalence in real applications (analysis shows that 95.6% of Apache's nondeterminism is superficial), allow my technique to be efficient in handling nondeterminism. I leverage offline and online analysis to discover the appropriate state-convergence points that can execute in a lazy, performance-sensitive manner across a service's replicas at runtime. Lazy convergence can mask some of the overheads that occur when there is an increase in the number of services in a nondeterministic distributed application.

Chuck Cranor stops to admire the artwork at Nemacolin Woodlands Resort, location of the annual PDL Retreat & Workshop.

RECENT PUBLICATIONS

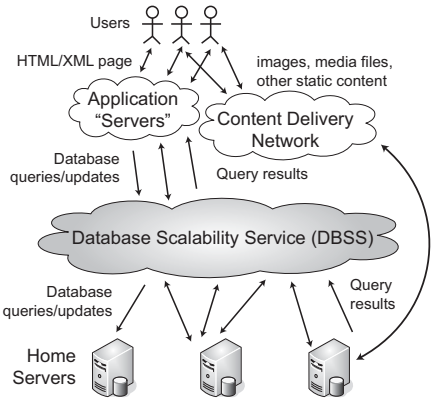
continued from page 23

Invalidation Clues for Database Scalability Services

Manjhi, Gibbons, Ailamaki, Garrod, Maggs, Mowry, Olston, Tomasic & Yu

ICDE 2007: 316-325, Istanbul, Turkey.

For their scalability needs, data-intensive Web applications can use a Database Scalability Service (DBSS), which caches applications' query results and answers queries on their behalf. To address security/privacy concerns while retaining the scalability benefits of a DBSS, applications would like to encrypt all their cached query results yet somehow enable the DBSS to invalidate these results when data updates render them obsolete. Without adequate information the DBSS is forced to invalidate large regions of its cache on an update. In this paper, we present invalidation clues, a general technique that enables applications to reveal little data to the DBSS, yet limit the number of unnecessary invalidations. Compared with previous approaches, invalidation clues provide applications significantly improved tradeoffs between security/privacy and scalability. Our experiments using three Web application benchmarks, on a prototype DBSS we have built, confirm that invalidation clues are indeed a low-overhead, effective, and general technique for applications to balance their privacy and scalability needs.



A scalable architecture for database-intensive web applications.

Stream Monitoring under the Time Warping Distance

Sakurai, Faloutsos & Yamamuro

Distance ICDE 2007, Istanbul, Turkey.

The goal of this paper is to monitor numerical streams, and to find subsequences that are similar to a given query sequence, under the DTW (Dynamic Time Warping) distance. Applications include word spotting, sensor pattern matching, and monitoring of biomedical signals (e.g., EKG, ECG), and monitoring of environmental (seismic and volcanic) signals. DTW is a very popular distance measure, permitting accelerations and decelerations, and it has been studied for finite, stored sequence sets. However, in many applications such as network analysis and sensor monitoring, massive amounts of data arrive continuously and it is infeasible to save all the historical data. We propose SPRING, a novel algorithm that can solve the problem. We provide a theoretical analysis and prove that SPRING does not sacrifice accuracy, while it requires constant space and time per time-tick. These are dramatic improvements over the naive method. Our experiments on real and realistic data illustrate that SPRING does indeed detect the qualifying subsequences correctly and that it can offer dramatic improvements in speed over the naive implementation.

Learning to Share: A Study of Data Sharing Among Home Devices

Salmon, Hady & Melican

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-07-107, October 2007.

As an increasing number of home and personal electronic devices create, use, and display digitized forms of data, it is becoming more important to be able to easily share data among these devices. This paper discusses an observation of a home data sharing system constructed from currently available technologies. We deployed this system into two households for two and half weeks and studied the impact of the system on household data usage and device management. We focus on users' perceptions of the system's advantages, the reasons home users employed multiple devices, the ways home users managed their devices, and how some distributed file system concepts were ill-suited to the home environment.

A study participant explains how she keeps her personal data private.

Information Survival Threshold in Sensor and P2P Networks

Chakrabarti, Leskovec, C. Faloutsos, Madden, Guestrin & M. Faloutsos

INFOCOM, Anchorage, Alaska, USA, May 2007.

Consider a network of, say, sensors, or P2P nodes, or bluetooth-enabled cell-phones, where nodes transmit information to each other and where links and nodes can go up or down. Consider also a 'datum', that is, a piece of information, like a report of an emergency condition in a sensor network, a national traditional song,

continued on page 27

continued from page 3

find better ways to author, visualize, and manage policies (for example, access-control policies and file permissions).

Lorrie is married to PDL faculty member Chuck Cranor. They have three children, Shane (6), Maya (4), and Nina (18 months). When she has time, Lorrie enjoys designing and creating quilts.

Carlos Guestrin

Carlos Guestrin is an assistant professor in the Machine Learning Department and in the Computer Science Department at Carnegie Mellon University. He received his Ph.D. in Computer Science from Stanford University, and received a Mechatronics Engineer (Mechanical Engineering, with emphasis in Automation and Systems) degree in 1998 from the Polytechnic School of the University of São Paulo, Brazil.

Carlos' long-term research goals are to develop efficient distributed machine learning algorithms for effective inference, learning and control in large-scale real-world distributed systems, such as sensor networks. These algorithms must perform the global inference and optimization tasks required by sensor network applications, while being robust to network losses and failures, and limiting communication and power requirements. In addition to developing theoretically-founded algorithms, he seeks to evaluate these methods on data from real sensor network deployments, and to implement some of these approaches on real deployed systems.

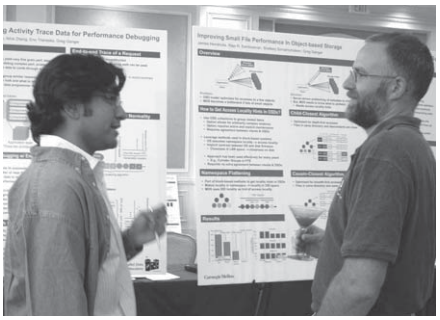
Through his collaboration with PDL, Carlos hopes to expand on the application of these algorithms to computer systems, and discover new core directions for machine learning.

Julio López

We are pleased to welcome Dr. Julio López to the PDL in his new position as a system scientist. Up until recently, Julio was a member of the PDL as a graduate student. He graduated in August with his Ph.D. from the Department of Electrical and Computer Engineering at Carnegie Mellon, following the presentation of his dissertation on "Methods for Querying Compressed Wavefields." In this work, Julio developed techniques to compress, index and query large wavefield datasets in their compressed representation, turning an I/O intensive problem into a massively parallel computational workload. He received his M.Sc. in Electrical and Computer Engineering from Carnegie Mellon University in May 2000 and his B.S. in Computer Science (Ingeniero de Sistemas) from the Universidad EAFIT, Medellín in Colombia in 1996.

Julio's current research interests are in the various aspect of systems and applications for data intensive computing at large scale, including computational databases, parallel and distributed systems, scalable I/O and indexing techniques for large multi-dimensional spatial datasets, data compression and visualization. In particular, he is interested in programming models, abstractions and supporting systems for these types of computation.

continued from page 26



Raja discusses his poster on "Clustering Activity Trace Data for Performance Debugging" with Alistair Veitch of HP Labs at the 2006 PDL Retreat.

or a mobile phone virus. How often should nodes transmit the datum to each other, so that the datum can survive (or, in the virus case, under what conditions will the virus die out)? Clearly, the link and node fault probabilities are important — what else is needed to ascertain the survivability of the datum?

We propose and solve the problem using non-linear dynamical systems and fixed point stability theorems. We provide a closed-form formula that, surprisingly, depends on only one additional parameter, the largest eigenvalue of the connectivity matrix. We illustrate the accuracy of our analysis on realistic and real settings, like mote sensor networks from Intel and MIT, as well as Gnutella and P2P networks.

continued from page 13

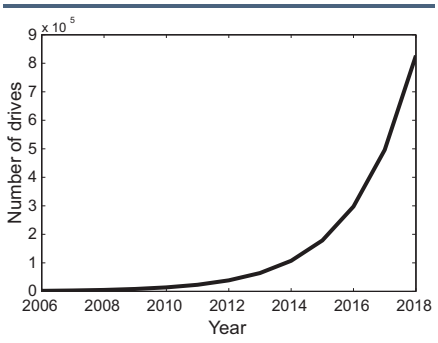


Figure 4. Number of drives in future systems.

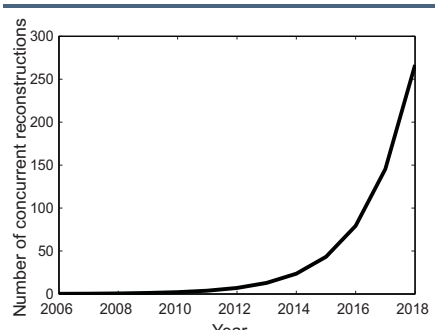


Figure 5. The number of concurrent reconstructions in the system.

and that 3% of drives in a system fail per year on average [2], we project the number of concurrent reconstructions in future HPC systems as shown in Figure 5. The figure indicates that in 2018 on average nearly 300 concurrent reconstructions will be in progress at any time!

Conclusions

The most demanding applications will see ever-increasing failure rates if the trends seen at top500.org continue. Using the standard checkpoint restart fault tolerance strategy, the efficacy of petascale machines running demanding applications will fall off. Relying on computer vendors to counter this trend is not recommended by historical data, and relying on disk storage bandwidth to counter it is likely to be expensive at best. We recommend that these applications consider spending an increasing number of cycles compressing checkpoints. We also recommend experimentation with process pairs fault tolerance for supercomputing. And if technologies such as flash memory are appropriate, we recommend experimenting with special devices devoted to checkpointing.

Finally, we would like to remark that our "crystal ball gazing" into the future of HPC systems is hard to defend as truly predictive. Our goal with this article is not to exactly predict the future, but to stimulate readers to consider the problems that need to be solved, so that our predictions will not come true.

The Computer Failure Data Repository

The work described in this article is part of our broader research agenda with the goal of analyzing and making publicly available failure data from a large variety of real production systems. We have built a public Computer Failure Data Repository (CFDR), hosted by the USENIX association [1] with the goal of accelerating research on system reliability by filling the nearly empty collection of public data with detailed failure data from a variety of large production systems. We encourage all organizations running large-scale IT systems to contribute failure data to the repository.

References

[1] The Computer Failure Data Repository (CFDR). http://cfdr.usenix.org.

[2] B. Schroeder and G. Gibson. Disk failures in the real world: What does an MTTF of 1,000,000 hours mean to you? In Proc of the 5th Usenix Conference on File and Storage Technologies (FAST 2007).

[3] B. Schroeder and G. Gibson. A large-scale study of failures in high-performance computing systems. In Proc. of the 2006 Int. Conference on Dependable Systems and Networks (DSN'06), 2006.

[4] B. Schroeder and G. Gibson. Understanding Failures in Petascale Computers, In SciDAC 2007: Journal of Physics: Conference Series 78 (2007) 012022.