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
Hewlett-Packard Labs
Hitachi, Ltd.
IBM Corporation
Intel Corporation
Microsoft Corporation
Network Appliance
Oracle Corporation
Panasas, Inc.
Seagate Technology
Symantec Corporation

PDL Packet

NEWSLETTER ON PDL ACTIVITIES AND EVENTS • FALL 2006
http://www.pdl.cmu.edu/

Putting Home Storage in Perspective

Digital content is now common in the home. An increasing number of home and personal electronic devices create, use, and display digitized forms of music, images, videos, as well as more conventional files. People are increasingly shifting important content from other forms of media, such as photographs and personal records, to online digital storage. The transition to digital homes with diverse collections of devices for accessing content is exciting, but does bring challenges.

On the device management front, most home users have little or no computer experience and limited patience with set-up and maintenance tasks. In order to be useful to home users, devices must be self-configuring and require minimal input from the user to work effectively. As well, the home environment device set is extremely heterogeneous. It includes resource-rich devices like desktop machines, power-limited devices like portable music players, and resource-scarce devices like cell phones. Many of these characteristics are also dynamic. Many devices, such as car accessories or portable music players, will frequently leave the home. Many power-constrained devices will alternate between power-scarce disconnected periods and power-rich connected periods. Any infrastructure will need to utilize the resources provided by these devices at any given time, and anticipate their availability in the future.

The biggest challenge in the home environment, however, is data management. Most consumer devices (e.g., digital video recorders and digital cameras) are specialized, each with simplified interfaces to help users interact with it and the data it stores. But, realizing the promise of the digital home requires moving past per-device interactions to easily coordinated data management across collections of storage and access devices.

Automating storage management for the home will require a home storage infrastructure with manageability as a central focus. Manageability is not something that can be added onto a system after the fact, but must be part of the design of a

continued on page 10

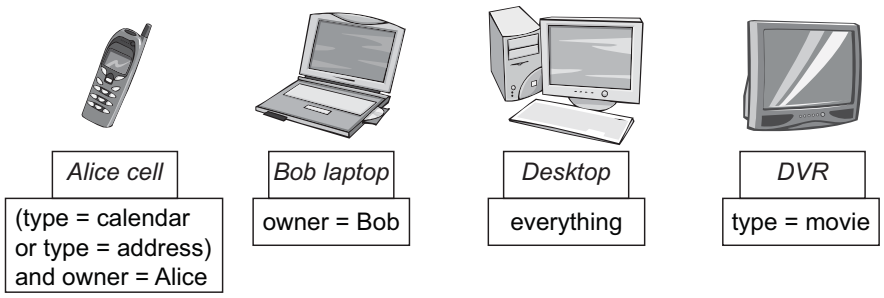


Figure 1: An example set of devices. In this diagram, we have four devices, each with a customized view. Alice's cell phone is only interested in "calendar" and "address" objects belonging to Alice, Bob's laptop is interested in all objects owned by Bob, the house desktop is interested in all data, and the DVR is interested in all "movie" objects.

FROM THE DIRECTOR'S CHAIR

Greg Ganger

Hello from fabulous Pittsburgh!

2006 has been a big year for the Parallel Data Lab, with the opening of the Data Center Observatory (DCO), two Best Paper awards at the File and Storage Technologies (FAST) conference, and significant new funding. Along the way, a new postdoc and several new students have joined PDL, several students have graduated and taken jobs with PDL Consortium companies, and many papers have been published.

For me, the most exciting thing has been the opening of the Data Center Observatory (DCO) in Spring 2006. The DCO will be a shared computing and storage utility for Carnegie Mellon researchers. At the same time, it is being deeply instrumented, from the power/cooling infrastructure to the administrator activities, to allow study of data center operations. In addition, it acts as a showcase for advanced data center infrastructure and a testbed for our ideas on automation and enhancement. As a showcase and a big investment for a University, the DCO opening led to a good number of magazine/web articles and significant press coverage. On so many fronts, the DCO is important for the PDL and Carnegie Mellon. The support of PDL Consortium companies, including APC Corporation's help with engineering plans and novel power/cooling approaches, have helped us in taking this big step on our path of research into operational costs of large-scale infrastructures.

Another very visible occurrence was the creation of the Petascale Data Storage Institute (PDSI), led by our own Garth Gibson. PDSI is a SciDAC Institute supported by the Department of Energy at $11 million over 5 years. Eight institutions, including two other Universities and five national labs, are founding members of PDSI. The PDSI brings together this expertise to help address the technology challenges faced in scaling storage systems to petascale sizes. This includes education/outreach, data dissemination (e.g., of failure and trace data), standardization, and technology transfer.

The PDL continues to pursue a broad array of storage systems research, and this past year brought much progress on many fronts and on some new projects as well. Let me highlight a few things.

Of course, first up is the our primary umbrella project, Self-* Storage, which explores the design and implementation of self-organizing, self-configuring, self-tuning, self-healing, self-managing systems of storage servers. For years, PDL Retreat attendees pushed us to attack "storage management of large installations", and this project is our response. Convinced that support for manageability must be designed into storage architecture from the beginning, rather than added after the fact, we gave ourselves a clean slate in pursuing the ideal. Our initial design (a system dubbed Ursa Major) combined a number of recent PDL projects (e.g., PASIS and self-securing storage) with heavy doses of instrumentation and agents for processing observations and enacting decisions. As a first step, we have constructed Ursa Minor, which is a versatile cluster storage system that can act as the base for Ursa Major. In December 2005, our FAST paper on Ursa Minor's design was named Best Paper of the conference. The system itself is solidifying to the point of becoming ready for initial deployment. In fact, anyone visiting my office will hear music in the background… the mp3 files are stored on an

THE PDL PACKET

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.

CONTACT US
PDL Home: http://www.pdl.cmu.edu/
PDL People: http://www.pdl.cmu.edu/PEOPLE/

FACULTY

Greg Ganger (pdl director)
412•268•1297
ganger@ece.cmu.edu
Anastassia Ailamaki
natassa@cs.cmu.edu
David Andersen
dga@cs.cmu.edu
Anthony Brockwell
abrock@stat.cmu.edu
Chuck Cranor
chuck@ece.cmu.edu
Christos Faloutsos
christos@cs.cmu.edu
Garth Gibson
garth@cs.cmu.edu
Seth Goldstein
seth@cs.cmu.edu
Mor Harchol-Balter
harchol@cs.cmu.edu
Todd Mowry
tcm@cs.cmu.edu
David O'Hallaron
droh@cs.cmu.edu
Priya Narasimhan
priya@ece.cmu.edu
Adrian Perrig
adrian@ece.cmu.edu
Mike Reiter
reiter@cmu.edu
Mahadev Satyanarayanan
satya@cs.cmu.edu
Srinivasan Seshan
srini@cmu.edu
Dawn Song
dawnsong@ece.cmu.edu
Chenxi Wang
chenxi@ece.cmu.edu
Hui Zhang
hui.zhang@cs.cmu.edu

POST DOCTORAL RESEARCHER

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
Gregg Economou
Manish Prasad
Michael Stroucken

GRADUATE STUDENTS

Zainul Abbasi
Michael Abd-El-Malek
Mukesh Agrawal
Kinman Au
Sachin Bhamare
Swaroop Choudhari
Debabrata Dash
Shobit Dayal
Ivan Dobrić
Jason Ganetsky
Kun Gao
Nikos Hardavellas
James Hendricks
Ryan Johnson
Andrew Klosterman
Jure Leskovec
Julio López
Michael Mesnier
Jim Newsome
Jia-Yu Pan
Ippokratis Pandis
Stratos Papadomanolakis
Adam Pennington
Soila Pertet
Brandon Salmon
Raja Sambasivan
Akshay Shah
Minglong Shao
Shafeeq Sinnamohideen
Joseph Slember
John Strunk
Jimeng Sun
Ajay Surie
Eno Thereska
Eric Toan
Niraj Tolia
Tiankai Tu
Matthew Wachs
Andrew Williams
Andrew Wolbach

FROM THE DIRECTOR'S CHAIR

instance of Ursa Minor, representing another step towards our goal of deploying a large-scale storage infrastructure in the DCO to test our "self-*" ideas.

With the Ursa Minor foundation, research into many components of achieving self-*-ness is being attacked, including performance insulation, performance instrumentation and predictability, metadata scalability and recoverability, multi-metric goal-based tuning, and new approaches to device modeling. Initial results on progress in each of these areas are very promising, and many can be found among the publication abstracts listed in this PDL Packet. The newest work tackles aspects of automated problem diagnosis, including failure prediction, fingerpointing within the distributed system, and debugging assistance.

A new direction of PDL research focuses on storage in the home. Personal and home electronic devices increasingly create, use, and display digital content, including music, images, video, and more conventional files. But, there is little in the way of data management assistance for users concerned with sharing data across devices, ensuring data reliability and consistency, and searching among the collection of devices. Our new home storage exploration seeks to develop data management approaches for the home environment, with its non-expert users and heterogeneous devices. We are building a prototype system called Perspective to explore an architectural approach based on views, which summarize data in which a device has interest, allowing other devices to know when to notify that device of updates or to ask it about searches. We expect this to be an exciting area of research in the coming years.

Of course, many other ongoing PDL projects are also producing cool results. The Staged Database Systems team has prototyped their ideas for improving database server efficiency, and they received the Best Demo Award at ICDE 2006. The self-securing devices project continues to explore how devices enhanced with security functionality can collaborate in spotting and surviving intrusions. The Computational Database Systems (CoDS) project continues to develop tools and methodologies for shifting scientific computing from ad hoc codes to more general-purpose database systems. Relevant to PDSI, ongoing efforts have developed trace capture and replay techniques for parallel applications, analyses of failure data from large-scale computing environments, and scalability techniques for metadata services. This newsletter and the PDL website offer more details and additional research highlights.

Eno Thereska discusses his research on "Informed Data Distribution Selection in a Self-predicting Storage System" with Ankur Kemkar of Symantec Corp. at the 2006 Spring Industry Visit Day.

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.

YEAR IN REVIEW

October 2006

- 14th Annual PDL Retreat and Workshop.

September 2006

- The formation of the Petascale Data Storage Institute under the direction of Garth Gibson was officially announced by the Department of Energy
- Joseph Slember spoke on "Nondeterminism in ORBs: The Perception and the Reality" at the High Availability of Distributed Systems Workshop in Krakow, Poland.
- Greg, Garth, Mike Mesnier, and Bianca all attended HECIWG workshop, each giving a talk about PDL activities.
- Brandon Salmon began an internship at Intel Oregon this fall, working with the Storage Technology Group and an ethnographer developing digital home technology.

August 2006

- Mike Mesnier presented //TRACE at the 2006 Interagency Working Group on High End Computing (HEC-IWG) File Systems and I/O R&D Workshop in Washington, DC. //TRACE (pronounced 'parallel trace') is a new approach to tracing and replaying the I/O from parallel applications. The //TRACE team includes James Hendricks, Julio Lopez, Mike Mesnier, Raja Sambasivan, and Matthew Wachs.
- Craig Soules successfully defended his research on "Using Context to Assist in Personal File Retrieval" and has moved to Palo Alto and HP Labs.
- Shuheng Zhou successfully defended her dissertation on "Routing, Disjoint Paths, and Classification" and has begun a Postdoctoral Fellowship at CMU.
- Bianca Schroeder gave an invited talk at the HECIWG File Systems and I/O Workshop in Washington D.C. on "Failure at scale".
- John Strunk proposed his Ph.D. research, titled "Using Utility Functions to Control a Distributed Storage System".
- Christos Faloutsos and his students presented 3 papers at KDD'06 in Philadelphia this year, including "Robust Information-theoretic Clustering", "Center-Piece Subgraphs: Problem Definition and Fast Solutions", and "Beyond Streams and Graphs: Dynamic Tensor Analysis". They also presented 2 posters.
- Christos Faloutsos presented a tutorial on "Sensor Mining at Work: Principles and a Water Quality Case-Study" at KDD'06 in Philadelphia.

July 2006

- Storage/administration leaders from Cleveland Federal Reserve Bank visited the PDL.
- Shuheng presented "Edge Disjoint Paths in Moderately Connected Graphs" at the Int'l Colloquium on Automata, Languages and Programming (ICALP06) in Venice, Italy.

June 2006

- James Hendricks spent the summer at IBM Almaden working with Ted Wong and Richard Golding.
- Soila Pertet interned with the Storage Systems Dept. of HP Labs over the summer, working with John Wilkes and Jay Wylie on building a prototype of an on-demand database service.
- Eno Thereska presented "Stardust: Tracking activity in a distributed storage system" at the Joint International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS'06) in Saint-Malo, France.
- Eno Thereska presented "Informed data distribution selection in a self-predicting storage system" at the International Conference on Autonomic Computing (ICAC-06) in Dublin, Ireland.
- Bianca Schroeder spoke at DSN'06 in Philadelphia on "A Large-Scale Study of Failures in High-Performance-Computing Systems".

May 2006

- 8th annual Spring Industry Visit Day.
- The ribbon was cut, officially opening the Data Center Observatory.
- Niraj Tolia presented "An Architecture for Internet Data Transfer" at the 3rd Symposium on Networked Systems Design and Implementation (NSDI '06) in San Jose, CA.
- Joe Slember completed his M.S. for his work on "Using Program Analysis to Identify and Compensate for Nondeterminism in Distributed, Fault-Tolerant Systems."
- Christos Faloutsos was the keynote speaker at SETN (Symposium of the Hellenic Association for Artificial Intelligence), in Herakleion, Crete. He presented "Data Mining using Fractals and Power Laws".

April 2006

- Symantec Corp. joined the PDL Consortium
- Anastassia Ailamaki and her students received the Best Demo Award at ICDE 2006 in Atlanta, GA for their work titled "Simultaneous Pipelining in QPipe: Exploiting Work Sharing Opportunities Across Queries".
- Eno Thereska was invited to participate in the Shepherding Committee of the Eurosys Authoring Workshop in Leuven, Belgium. Eurosys is the European Professional Society for Systems and its goals are to encourage systems research in Europe. The Authoring Workshop's goal was to provide guidelines to PhD students on how to write good technical papers.

February 2006

- Mike Mesnier proposed his Ph.D. research, titled "Modeling the Relative Fitness of Storage".

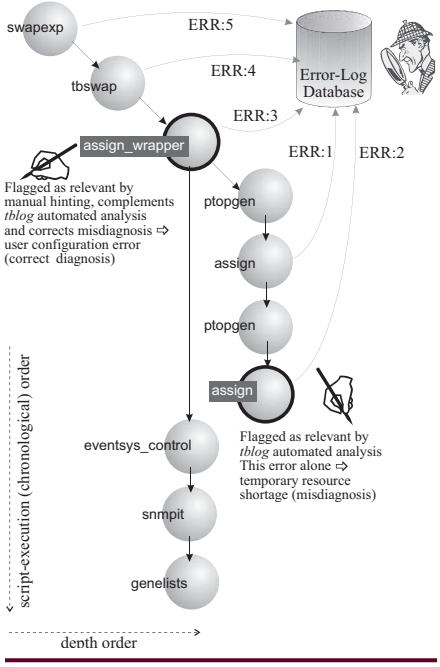
continued on page 20

Towards Fingerpointing in the Emulab Dynamic Distributed System

Kasick, Narasimhan, Atkinson & Lepreau

Proceedings of the 3rd USENIX Workshop on Real, Large Distributed Systems (WORLDS '06), Seattle, WA. Nov. 5, 2006.

In the large-scale Emulab distributed system, the many failure reports make skilled operator time a scarce and costly resource, as shown by statistics on failure frequency and root cause. We describe the lessons learned with error reporting in Emulab, along with the design, initial implementation, and results of a new local error-analysis approach that is running in production. Through structured error reporting, association of context with each error-type, and propagation of both error-type and context, our new local analysis locates the most prominent failure at the procedure, script, or session level. Evaluation of this local analysis for a targeted set of common Emulab failures suggests that this approach is generally accurate and will facilitate global fingerpointing, which will aim for reliable suggestions as to the root-cause of the failure at the system level.



Manual hinting vs. automated tblog post-processing of the script call-chain for a real failure on Emulab.

//TRACE: Parallel Trace Replay with Approximate Causal Events

Mesnier, Wachs, Sambasivan, Lopez, Hendricks & Ganger

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-06-108, September 2006.

//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-request 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 less than 5%.

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

Schroeder & Gibson

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-06-111, September 2006.

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.

Early Experiences on the Journey Towards Self-* Storage

Abd-El-Malek, Courtright, Cranor, Ganger, Hendricks, Klosterman, Mesnier, Prasad, Salmon, Sambasivan, Sinnamohideen, Strunk, Thereska, Wachs & Wylie

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, September 2006.

Self-* systems are self-organizing, self-configuring, self-healing, self-tuning and, in general, self-managing. Ursa Minor is a large-scale storage infrastructure being designed and deployed at Carnegie Mellon University, with the goal of taking steps towards the self-* ideal. This paper discusses our

continued on page 6

RECENT PUBLICATIONS

continued from page 5

early experiences with one specific aspect of storage management: performance tuning and projection. Ursa Minor uses self-monitoring and rudimentary system modeling to support analysis of how system changes would affect performance, exposing simple What…if query interfaces to administrators and tuning agents. We find that most performance predictions are sufficiently accurate (within 10-20%) and that the associated performance overhead is less than 6%. Such embedded support for What…if queries simplifies tuning automation and reduces the administrator expertise needed to make acquisition decisions.

Group Communication: Helping or Obscuring Failure Diagnosis?

Pertet, Gandhi & Narasimhan

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-06-107, June, 2006.

Replicated client-server systems are often based on underlying group communication protocols that provide totally ordered, reliable delivery of messages. However, in the face of a performance fault (e.g, memory leak, packet loss) at a single node, group communication protocols can cause correlated performance degradations at non-faulty nodes. We explore the impact of performance-degradation faults on token-ring and quorum-based group communication protocols in replicated systems. By empirically evaluating these protocols, in the presence of a variety of injected faults, we investigate which metrics are the most/least appropriate for failure diagnosis. We show that group communication protocols can both help and obscure root-cause analysis, and present an approach for fingerpointing the faulty node by monitoring OS-level and protocol-level metrics. Our empirical evaluation suggests that the root-cause of the failure is either the node exhibiting the most anomalies in a given window of time or the node with an "odd-man-out" behavior, e.g., if a node displays a surge in context-switch rate while the other nodes display a dip in the same metric.

Informed Data Distribution Selection in a Self-predicting Storage System

Thereska, Abd-El-Malek, Wylie, Narayanan & Ganger

Proceedings of the International Conference on Autonomic Computing (ICAC-06), Dublin, Ireland. June 12th-16th 2006.

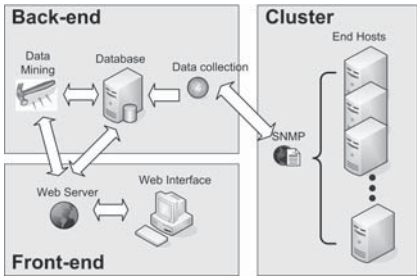
Systems should be self-predicting. They should continuously monitor themselves and provide quantitative answers to What…if questions about hypothetical workload or resource changes. Self-prediction would significantly simplify administrators' planning challenges, such as performance tuning and acquisition decisions, by reducing the detailed workload and internal system knowledge required. This paper describes and evaluates support for self-prediction in a cluster-based storage system and its application to What…if questions about data distribution selection.

InteMon: Continuous Mining of Sensor Data in Large-scale Self-* Infrastructures

Hoke, Sun, Strunk, Ganger & Faloutsos

ACM SIGOPS Operating Systems Review. Vol 40 Issue 3. July, 2006. ACM Press.

Modern data centers have a large number of components that must be monitored, including servers, switches/routers, and environmental control systems. This paper describes InteMon, a prototype monitoring and mining system for data centers. It uses the SNMP protocol to monitor a new data center at Carnegie Mellon. It stores the monitoring data in a MySQL database, allowing visualization of the time-series data using a JSP web-based frontend interface for system administrators. What sets InteMon apart from other cluster monitoring systems is its ability to automatically analyze correlations in the monitoring data in real time and alert administrators of potential anomalies. It uses efficient, state of the art stream mining methods to report broken correlations among input streams. It also uses these methods to intelligently compress historical data and avoid the need for administrators to configure threshold-based monitoring bands.



Intemon System Architecture

An Architecture for Internet Data Transfer

Tolia, Kaminsky, Andersen & Patil

Proceedings of the 3rd Symposium on Networked Systems Design and Implementation (NSDI '06), San Jose, California, May 2006.

This paper presents the design and implementation of DOT, a flexible architecture for data transfer. This architecture separates content negotiation from the data transfer itself. Applications determine what data they need to send and then use a new transfer service to send it. This transfer service acts as a common interface between applications and the lower-level network layers, facilitating innovation both above and below. The transfer service frees developers from re-inventing transfer mechanisms in each new application. New transfer mechanisms, in turn, can be easily deployed without modifying existing applications.

We discuss the benefits that arise from

continued on page 7

continued from page 6

separating data transfer into a service and the challenges this service must overcome. The paper then examines the implementation of DOT and its plugin framework for creating new data transfer mechanisms. A set of microbenchmarks shows that the DOT prototype performs well, and that the overhead it imposes is unnoticeable in the wide-area. End-to-end experiments using more complex configurations demonstrate DOT's ability to implement effective, new data delivery mechanisms underneath existing services. Finally, we evaluate a production mail server modified to use DOT using trace data gathered from a live email server. Converting the mail server required only 184 lines-of-code changes to the server, and the resulting system reduces the bandwidth needed to send email by up to 20%.

A Large-scale Study of Failures in High-performance-computing Systems

Schroeder & Gibson

Proceedings of the International Conference on Dependable Systems and Networks (DSN2006), Philadelphia, PA, June 25-28, 2006.

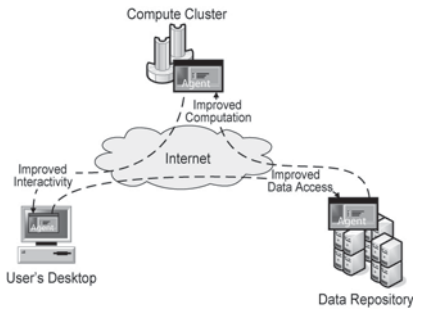
Designing highly dependable systems requires a good understanding of failure characteristics. Unfortunately little raw data on failures in large IT installations is publicly available, due to the confidential nature of this data. This paper analyzes soon-to-be public failure data covering systems at a large high-performance-computing site. The data has been collected over the past 9 years at Los Alamos National Laboratory and includes 23000 failures recorded on more than 20 different systems, mostly large clusters of SMP and NUMA nodes. We study the statistics of the data, including the root cause of failures, the mean time between failures, and the mean time to repair. We find for example that average failure rates differ wildly across systems, ranging from 20-1000 failures per year, and that time between failures is modeled well by a Weibull distribution with decreasing hazard rate. From one system to another, mean repair time varies from less than an hour to more than a day, and repair times are well modeled by a lognormal distribution.

Dimorphic Computing

Lagar-Cavilla, Tolia, Balan, de Lara, Satyanarananan & O'Hallaron

Carnegie Mellon University School of Computer Science Technical Report CMU-CS-06-123, April 2006.

Dimorphic computing is a new model of computing that switches between thick and thin client modes of execution in a completely automated and transparent manner. It accomplishes this without imposing any language or structural requirements on applications. This model greatly improves the performance of applications that alternate between phases of compute- or data-intensive processing and intense user interaction. For such applications, the thin client mode allows efficient use of remote resources such as compute servers or large datasets. The thick client mode enables crisp interactive performance by eliminating the harmful effects of Internet latency and jitter, and by exploiting local graphical hardware acceleration. We demonstrate the feasibility and value of dimorphic computing through AgentISR, a prototype that exploits virtual machine technology. Experiments with AgentISR confirm that the performance of a number of widely-used scientific and graphic arts applications can be significantly improved without requiring any modification.



Dimorphic Computing Example: An example application transitioning through data- and compute-intensive phases before returning to the user for interaction-intensive usage.

Perspective: Decentralized Data Management for the Home

Salmon, Schlosser, Mummert & Ganger

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-06-110, September 2006.

Perspective is a decentralized data management system for the growing collection of consumer electronics that store and access digital content. Perspective uses a new construct, called the view, to concisely describe which data objects each device may store and access. By knowing the views in the system, a device can know which devices may need to hear about any particular update and which devices to contact for a particular search query. By exchanging views, an ensemble of devices can coordinate and share data efficiently without relying on a centralized server; it works the same in the home or for a subset of devices outside the home. Experiments with Perspective confirm that views improve ensemble creation and search performance, by avoiding costs that are linear in the number of objects stored, without penalizing performance relative to an ideal centralized server setup.

Stardust: Tracking Activity in a Distributed Storage System

Thereska, Salmon, Strunk, Wachs, Abd-El-Malek, Lopez & Ganger

Joint International Conference on Measurement and Modeling of Computer Systems, (SIGMETRICS'06). June 26-30, 2006, Saint-Malo, France.

Performance monitoring in most distributed systems provides minimal guidance for tuning, problem diag-

continued on page 16

September 2006

Congratulations Natassa, Babak and Niki!

Natassa, Babak and big sister Niki are thrilled to welcome their new son and little brother Andreas Falsafi. He arrived on Sunday, September 24 at 10:00 a.m., weighing 9 pounds, 11 ounces, and measuring over 23 inches! He is healthy in all ways and Mom is doing fine too!

September 2006

Researchers Tackle Problem of Data Storage for Next-Generation Supercomputers: Carnegie Mellon leads DOE-sponsored Petascale Data Storage Institute

The U.S. Department of Energy (DOE) has awarded a five-year, $11 million grant to researchers at three universities and five national laboratories to find new ways of managing the torrent of data that will be produced by the coming generation of supercomputers. The Petascale Data Storage Institute (PDSI) combines the talents of computer scientists at Carnegie Mellon University, the University of California at Santa Cruz and the University of Michigan with those of researchers at the DOE's Los Alamos, Sandia, Oak Ridge, Lawrence Berkeley and Pacific Northwest national laboratories.

The innovations developed by the PDSI will enable scientists to fully exploit the power of computing systems capable of performing millions of billions of calculations each second. Increased computational power is necessary to model and simulate extremely complicated phenomena, e.g. global warming, earthquake motions, the design of fuel-efficient engines, etc., which provides scientific insights into processes that are often impossible to achieve through conventional observation or experimentation.

But, simply building computers with faster processing speeds — the new target threshold is a quadrillion (a million billion) calculations per second, or a "petaflop" — will not be sufficient to achieve those goals. Garth Gibson, who will lead the data storage institute, said new methods will be needed to handle the huge amounts of data that computer simulations both use and produce, along with all the challenges that managing a system with petaflop capabilities will involve.

The PDSI will focus its efforts in three areas: collecting field data about computer failure rates and application behaviors, disseminating knowledge through best practices and standards, and developing innovative system solutions for managing petascale data storage. The latter category could include "self-*" systems that use computers to manage computers.

For more information, see the article in this newsletter beginning on page X, and visit www.pdl.cmu.edu/PDSI/.

-- with info from the Carnegie Mellon Press Release Sept. 7, 2006

August 2005

Two Ph.D.s Awarded

Two PDL students have completed their Ph.D.s in the past year.

Craig Soules, advised by Greg Ganger, defended his work on "Using Context to Assist in Personal File Retrieval" in August and has since joined HP Labs in Palo Alto, CA.

Shuheng Zhou, jointly advised by Greg Ganger and Bruce Maggs, defended "Routing, Disjoint Paths, and Classification" in August. Shuheng will begin a postdoctoral fellowship in the School of Computer Science at Carnegie Mellon in September, working with Avrim Blum, John Lafferty and Bruce Maggs.

August 2006

Welcome Nikhil!

We are very happy to announce that Nikhil Narasimhan Gandhi was born on Saturday, August 12, 2006, to Priya Narasimhan and Rajeev Gandhi. He was born at 11.02 pm, weighed 6 lbs. 6 oz. and was 19.25 in long. The name "Nikhil" means "complete" in Sanskrit.

May 2006

Carnegie Mellon Researchers Attack Rising Costs of Data Center Operations

The Data Center Observatory (DCO) was officially opened on May 23, 2006, operating as a dual-purpose facility that is both a working data center and a research vehicle for the study of data center automation and efficiency. The dedication included

continued on page 9

continued from page 8

representatives from the university, American Power Conversion (APC), local business leaders and select members of the news media.

The DCO is a large-scale collaborative effort between Carnegie Mellon's College of Engineering and School of Computer Science. It also includes participation from a number of industry and government partners, including APC, which is providing engineering expertise and its InfraStruXure® system for powering, cooling, racking and managing equipment in the DCO.

The DCO's principle research goals are to better comprehend and mitigate human administration costs and complexities, power and cooling challenges, and failures and their consequences. It also aims to understand resource utilization patterns and opportunities to reduce costs by sharing resources among users. Among the center's major thrusts are energy efficiency and administration costs.

The 2,000 square-foot DCO has the ability to support 40 racks of computers, able to consume energy at a rate of up to 774 kW — more than the rate of consumption of 750 average-sized homes. In addition to studying dense computing environments, the DCO will support a variety of Carnegie Mellon research activities, from data mining to CAD/architecture, visualization and real networked services. The DCO joins Carnegie Mellon's long tradition of weaving infrastructure research into campus life, which keeps the university at the forefront of technology.

-- with info from the Carnegie Mellon Press Release May 23, 2006

April 2006

Brandon Salmon Awarded Intel Foundation Ph.D. Fellowship

Congratulations to Brandon on being awarded an Intel Foundation Ph.D. Fellowship. The Intel Foundation Ph.D. Fellowship Program awards two-

year fellowships to Ph.D. candidates pursuing leading-edge work in fields related to Intel's business and research interests. Fellowships are available at select U.S. universities, by invitation only, and focus on Ph.D. students who have completed at least one year of study.

April 2006

Best Demo at ICDE 2006

Congratulations to the Staged Database Systems team, who have received the Best Demo Award at ICDE 2006 in Atlanta, Georgia. Debabrata Dash, Kun Gao, Nikos Hardavellas, Stavros Harizopoulos, Ryan Johnson, Naju Mancheril, Ippokratis Pandis, Vladislav Shkapenyuk, and Anastassia Ailamaki were the collaborators on this effort titled Simultaneous Pipelining in QPipe: Exploiting Work Sharing Opportunities Across Queries. The demo paper can be found in the proceedings of the 22nd IEEE International Conference on Data Engineering (ICDE2006).

April 2006

Welcome Nina!

Congratulations to the Cranor family who welcomed Nina Veronica on April 13, 2006 at 6:24 pm. Nina was 11 days late but arrived happy and healthy, weighing in at 9 pounds, 8 oz.

April 2006

Adrian Perrig Keynote Speaker at IPSN

Adrian Perrig, assistant professor of ECE, engineering and public policy, and computer science, delivered the keynote presentation at the International Conference on Information Processing in Sensor Networks (IPSN) this April in Nashville, where he addressed his vision for security in sensor networks. IPSN is one of the two top conferences on sensor networks.

February 2006

Adrian Perrig Recipient of 2006 Sloan Award

Three CMU 2006 winners of a Sloan Research Fellowship in computer science have been announced: Carlos Guestrin, CALD and CSD, Doug James, CSD and RI, and Adrian Perrig, 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 116 fellowships are awarded annually in seven fields: chemistry, computational and evolutionary molecular biology, computer science, economics, mathematics, neuroscience, and physics. Only 14 are given in computer science each year so CMU once again shines. Congratulations to all!

February 2006

Microsoft Fellowship Helps Develop New Security Course

Faculty members Lorrie Cranor, Institute for Software Research International, Jason Hong, Human-Computer Interaction Institute, and Michael Reiter, Electrical and Computer Engineering, have received a 2005 Microsoft Research Trustworthy Computing Curriculum Award to fund the development of a new course

continued on page 21

continued from page 1

system from the outset. A home storage solution requires a distributed data management system that easily handles device addition and failure without user intervention. It must also provide data accessibility and practical search capabilities in arbitrary groups of devices, and make it easy for users to control home data storage location and protection. Users also need assistance with things like maintaining the consistency of replicated/cached data, searching for particular data within the environment, ensuring reliability of important data, and so on. These features must be well-supported and made intuitive, as most users will not tolerate the manual effort nor be capable of the expertise required today.

One approach would be to reuse distributed file systems, having each device act as a client of a conventional central file server. We find that this approach has a number of shortcomings. For example, most file systems provide little assistance with search, but this may be one of the most common actions in this environment. Also, users will often take subsets of devices elsewhere (e.g., on a family vacation) and wish to use them collectively away from the home. While a few file systems provide support for disconnected operation, very few combine that with support for ad hoc coordination and sharing among disconnected clients, and on a practical level, traditional file server administration methods are not conducive to ease-of-use by non-experts in a home environment.

Keeping your View in Perspective

We are exploring a more decentralized model based on semantic (a.k.a. attribute-based) file/object naming and metadata summaries that we call "views". Semantic naming associates descriptive metadata (e.g., keywords) with data objects and allows users to search in various ways. Much new consumer electronics data, such as music and television, fits this model immediately, since it is distributed with rich metadata about the content. Desktop file management is also moving in this direction, with tools such as Apple's Spotlight and Google's Google Desktop.

A view describes a set of objects in the semantic store in the form of a query on per-object metadata — a way of compactly expressing the data objects that a device has interest in and may store. For example, a simple view could specify every object, i.e., "*". More sophisticated views could specify all objects that are of a particular type, are newer than a particular date, or are owned by a particular user. In fact, a view can be specified by any arbitrary query on any extensible metadata in the system. Figure 1 illustrates several devices and their views.

Perspective — in seeing many views, one gains perspective — is our prototype distributed data management system designed for home/personal storage. It assumes semantic naming and uses views as a building block to efficiently support consistency, search, reliability, ad hoc creation of device ensembles, and synchronization of data between appropriate devices. By exchanging their views, which are small and change infrequently, devices can identify which other devices may need to know about particular new data (for consistency) and which other devices may have data relevant to a search. Views assist with redundancy management by making it possible to determine which devices promise to hold which replicas, which in turn makes it possible to verify appropriate reliability coverage. Device ensembles can be efficiently created by simply exchanging views among participating devices to enable appropriate consistency and search traffic. With views, these features can be supported without relying on a reliably maintained central server that stores all the data in the environment.

In the example in Figure 1, imagine that Bob (the owner of the laptop) creates a movie clip on his laptop. Perspective will send an update notification to those devices in the system with matching views, in this case the Desktop and the DVR. If Alice updates her address book on the Desktop machine, that update will only be propagated to the cell phone. When an update notification is received by a device, that device can decide whether to accept that update and store the data or ignore that update and not store the data. If a device registers a view as being complete, that device promises that it will always accept updates for objects in that view and will never delete those objects once they're stored. If a view is instead registered as being partial, the device registering that view is free to ignore updates and delete its copies of objects, if it so chooses. The distinction between complete and partial views has ramifications for both search and reliability management.

File Searching

Searching for digital files, whether the file is a document, movie, photograph, or song, is one of the most common operations in a home storage environment, so it is important to not only enable it but to make it efficient as

continued on page 11

Raja Sambasivan discusses Preserving Filesystem Namespace Locality in Object-based Storage

continued from page 10

well. Rich metadata provides a great deal of useful content for search, and views can enable some useful optimizations.

By comparing the search query to the views in the system, it is simple to exclude devices that could not store the desired data. For example, the cell phone in figure 1 will definitely not have any movies, and that can be seen in its views.

Complete views can simplify the search process even further because a device with a complete view is guaranteed to store all objects that match that view. If a complete view contains the search, the device only needs to forward the search to that device, not propagating it to other devices in the ensemble.

Efficient decentralized search is especially useful when dealing with devices from multiple administrative domains. One challenge in sharing data between domains is that they do not collaborate on name spaces, making it difficult to find data. However, by decentralizing the search process and using a semantic model, a device can see data on a device from another domain in exactly the same way it sees data from a local domain. Each device is free to manage access control as it sees fit, and "domain" could be one part of the metadata.

Reliability

Providing strong reliability is important in home storage systems, as the data that users store is often irreplaceable (e.g. family photographs) and yet the devices that they use to store it are typically inexpensive and failure-prone. Expecting the user to buy more reliable (and expensive) hardware is not a solution, since home users' purchase decisions are more likely to be driven by initial costs (i.e., "what's on sale online today?"). Instead, a home data management system should enable users to get the best reliability from the ensemble of devices they already have, using the natural redundancy of devices in the home. For example, a user with several well-provisioned devices (e.g., several PCs, a digital video recorder, etc.) should be able to automatically replicate data across those devices to provide reliability. When several devices are configured with the same views, Perspective will automatically propagate updates to replicated data on those devices. Management tools for configuring devices and establishing views could include applications that allow users to hand set views, automation tools that observe user behavior and set views appropriately, or defaults that are put onto devices at manufacture time. A management tool built on top of Perspective can be stateless, simply connecting to some device in the system to see the views in the system and propagating changes using that device. So, an automation tool could even be a Java applet downloaded over the web.

continued on page 22

Bianca Schroeder

Dr. Bianca Schroeder joined the PDL in September 2005 as a Post-Doctoral Fellow, after receiving her Ph. D. in Computer Science from CMU in August 2005. She is currently working with Garth Gibson on "empirical system reliability." This new line of research is motivated by the fact that, with the ever growing component count in large-scale IT systems, component failures are quickly becoming the norm rather than the exception. Yet, virtually no data on failures in real systems is publicly available, forcing researchers in the area to base their work on anecdotes and back of the envelope calculations rather than empirical data. The goal of Bianca's work is to collect and analyze failure data from real, large-scale production systems and to exploit the results for better system design and management. The early results of this work have been published in a DSN'06 paper and a PDL tech-report, which is currently under submission for publication.

Alice Zheng

Alice received her B.A. in Mathematics and Computer Science in 1999 and her Ph.D. in Electrical Engineering in 2005, all from U.C. Berkeley. Her background lies in machine learning algorithms and she has worked on a diverse set of projects ranging from audio signal processing to web link analysis. She is currently interested in machine learning algorithms for failure diagnosis in computer systems and software. Her Ph.D. thesis investigates automatic software debugging using statistical methods. After a brief sojourn at the Auton Lab in the Robotics Institute, she is now part of the PDL family with plans to look into problems in performance diagnosis and prediction.

Alice was born in Beijing, China, but considers Northern California her second home. She enjoys traveling, yoga, rock climbing, and is fluent in Chinese.

PETASCALE DATA STORAGE INSTITUTE LEADERSHIP AWARDED TO PDL

Garth Gibson

Launched in September of 2006, and running for at least 5 years, the Parallel Data Laboratory adds a new dimension to its research portfolio: the Petascale Data Storage Institute (PDSI). Funded by the Department of Energy in its Scientific Discovery through Advanced Computing (www.scidac.gov), PDSI brings together leading experts in high-performance file systems for high-end computing clusters for the purpose of anticipating and overcoming the challenges to be faced in the transition from terascale to petascale computing. The founding members of PDSI are CMU (G. Gibson, PI), U.C. Santa Cruz (D. Long, co- PI), U. Michigan CITI (P. Honeyman, co-PI), Los Alamos Nat. Lab. (G. Grider, co-PI), Oak Ridge Nat. Lab. (P. Roth, co-PI), Sandia Nat. Lab. (L. Ward, co-PI), Pacific Northwest Nat. Lab. (E. Felix, co-PI), and Lawrence Berkeley Nat. Lab.'s Nat. Energy Research Scientific Computing Center (W. Kramer, co-PI).

High-performance computing (HPC) today is cluster computing with better networks, not so many software licenses, and a whole lot more shared storage bandwidth than you might find in internet service provider data centers. But both share a critical characteristic: scale. Most cluster computers used as supercomputers have a 1,000 to 5,000 nodes. Commercial HPC used in applications like seismic imaging and chip simulation are as large as 10,000 to 20,000 nodes. Internet service providers like Google have well over 15,000 nodes. In the next few years large cluster installations are anticipated to reach 30,000 to 40,000 nodes. And all of these are dwarfed by the special-purpose monstrosities like BlueGene/L, at 60,000 to 130,000 nodes and growing. Considering that all nodes these days have multiple sockets and the chips have multiple cores, there is in fact parallelism of up to 100,000 and growing to 1,000,000 cores in the largest government, academic and commercial HPC clusters and in internet service provider data centers.

The web site top500.org collects self-reported cluster statistics for large scale installations. Figure 1 shows the trends in aggregate performance for the machines on its lists. The peak performance of the most massively scaled machine today is a quarter of a petaflops. The trend on this peak performance is quite predictable, and it suggests that the biggest machine's performance will cross a petaflops in 2008 or 2009. The most likely sites for this newsmaker machine include Los Alamos and Oak Ridge National Labs, both participating in the Petascale Data Storage Institute.

Horst Simon of Lawrence Berkeley National Labs, another of the PDSI's participants, points out that the 500th largest cluster is even smoother in its trend, and it should cross the petaflops boundary in 2015-2016, or 7 years after the first machine. Only a few years later, we can expect the first machine to cross the exaflops threshold, if the trend for the last 15 years continues as it has over the next 15 years. Roughly, the era of petascale looks to be 2008 through 2018.

Turning our attention to magnetic disk device trends, Seagate's Mark Kryder recently charted the areal density trends for magnetics disk drives. Over the same petascale era, he sees a continued 40% per year increase in areal density derived from perpendicular recording, then heat-assisted perpendicular recording (HAMR), then self-

continued on page 13

continued from page 12

organized or patterned media (SOMA) with HAMR. Peak data rate will also track well, though its rate of increase will be 20% per year. The principle message is that the capacity increases and data rate increases we have become used to will continue. One take on Kryder's projections is that at some time the capacity of 3.5" disks will become excessive for many applications, and the performance benefits of buying more 2.5" disk spindles per TB will become compelling. Just when this will happen in more than a few sites is an interesting topic for bar room and board room discussions.

The Petascale Data Storage Institute is chartered as a community of researchers addressing the data storage issues facing HPC systems in the petascale era. Its efforts break down into three types: dissemination of best practices, collection of raw data needed to design solutions for ever larger scale systems, and innovation of designs, mechanisms and management for the large scale systems.

Dissemination projects include about two workshops per year for researchers with new ideas, applications programmers with problems and vendors with evolving solutions. The first such workshop will take place at SC06 (www.sc06.org) in Tampa, FL, on Friday Nov. 17, 2006. We anticipate SCxy to be an annual venue for PDSI workshops. The other workshop each year may vary its venue to broaden our reach. For 2007, we are exploring ACM's Federated Computing Research Conference (www.acm.org/fcrc) to be held in San Diego June 8-16, 2007. Other candidates, depending on the timing in each year, include USENIX FAST and IEEE MSST. Beyond the workshops, PDSI dissemination projects are interested in the problem of helping programmers write better IO code, by better understanding how HPC file systems and storage systems work. We anticipate developing tutorial and classroom materials.

The second dimension of the dissemination projects is standardization of best practices. In this activity, PDSI members work with standards bodies such as IETF, T10, POSIX, SNW, MPI Forum, Global/ Open Grid Forum, etc., to codify best practices and facilitate deployment. One example is the leadership role our partners at the U. Michigan, CITI, are playing in the Linux reference implementation of NFSv4.1 (www.ietf.org/internet-drafts/draft-ietf-nfsv4-minorversion1-06.txt), especially the Parallel NFS (pNFS) protocol. Many of us see pNFS as a strategy to get the best of the proprietary parallel file systems technology into mainstream solutions. A second example of the leadership role our partners at the Los Alamos Nat. Lab. have played in getting POSIX to commission a High-End Computing Extensions working group (www.opengroup.org/platform/hecewg) whose web page is shown in Figure 2. This is a vehicle to allow applications to enable file systems to provide better performance, security, and interoperability.

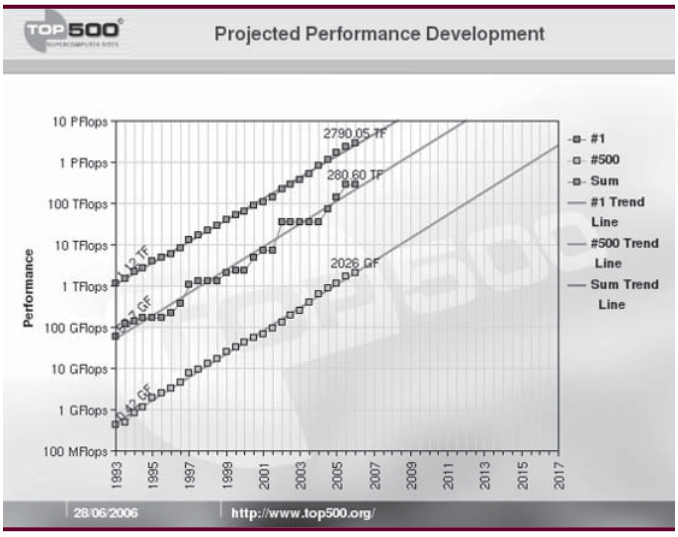


Figure 1: Self-reported cluster statistics for large scale installations.

In the collection projects of PDSI, the focus is on gathering raw data from field use of large scale computing installations to equip researchers with new insights, speeds'n'feeds and differentiations between processes/modes. PDSI partners, with leadership from Sandia Nat. Lab., will instrument file system interfaces for interesting applications on large scale systems to capture workload traces. With trace replay, trace scaling and visualization tools being developed with leadership from Institute partner Oak Ridge Nat. Labs, this will enable much higher resolution for performance tuning research than could be obtained with traditional giant-copy benchmarks.

The other major collection activity is outage, failure, error and exception history gathering. With increasing scale as a basic issue for PDSI, we expect increasing failure rates, increasing frequencies of multiple concurrent failures, and an increasing need for prediction and proactive management. Our partners at Los Alamos Nat. Lab. are way out in front on this. They have been recording a root-cause analysis for every outage on all of their HPC clusters for 9 years, and they have made this data available to the public. PDSI is expanding the scope of our failure data collection with data from other HPC sites and from internet service providers, the other community of very large scale systems.

The third type of PDSI projects are the mechanism innovation projects. While much of this will develop as our community responds to the collection projects and collaboration with DOE SciDAC applications, members of the Institute have started on promising directions. Foremost among these is the application of automated man-

continued on page 24

DISSERTATION ABSTRACT:
Using Context to Assist in Personal File Retrieval

Craig A. N. Soules, CS

Carnegie Mellon University School of Computer Science Ph.D. Dissertation CMU-CS-06-147, August 25, 2006.

Personal data is growing at ever increasing rates, fueled by a growing market for personal computing solutions and dramatic growth of available storage space on these platforms. Users, no longer limited in what they can store, are now faced with the problem of organizing their data such that they can find it again later. Unfortunately, as data sets grow the complexity of organizing these sets also grows. This problem has driven a sudden growth in search tools aimed at the personal computing space, designed to assist users in locating data within their disorganized file space.

Despite the sudden growth in this area, local file search tools are often inaccurate. These inaccuracies have been a long-standing problem for file data, as evidenced by the downfall of attribute-based naming systems that often relied on content analysis to provide meaningful attributes to files for automated organization.

While file search tools have lagged behind, search tools designed for the world wide web have found widespread acclaim. Interestingly, despite significant increases in non-textual data on the web (e.g., images, movies), web search tools continue to be effective. This is because the web contains key information that is currently unavailable within file systems: context. By capturing context information, e.g., the links describing how data on the web is inter-related, web search tools can significantly improve the quality of search over content analysis techniques alone.

This work describes Connections, a context-enhanced search tool that utilizes temporal locality among file accesses to provide inter-file relationships to the local file system. Once identified, these inter-file relationships provide context information, similar to that available in the world wide web. Connections leverages this context to improve the quality of file search results. Specifically, user studies with Connections see improvements in both precision and recall (i.e., fewer false-positives and false-negatives) over content-only search, and a live deployment found that users experienced reduced search time with Connections when compared to content-only search.

DISSERTATION ABSTRACT:
Routing, Disjoint Paths, and Classification

Shuheng Zhou, ECE

Carnegie Mellon University Parallel Data Lab Ph.D. Dissertation CMU-PDL-06-109, August 2006.

In this thesis, we study two classes of problems: routing and classification. Routing problems include those that concern the tradeoff between routing table size and short-path forwarding (Part I), and the classic Edge Disjoint Paths problem (Part II). Both have applications in communication networks, especially in overlay network, and in large and high-speed networks, such as optical networks. The third part of this thesis concerns a type of classification problem that is motivated by a computational biology problem, where it is desirable that a small amount of genotype data from each individual is sufficient to classify individuals according to their populations of origin.

In hierarchical routing, we obtain "near-optimal" routing table size and path stretch through a randomized hierarchical decomposition scheme in the metric space induced by a graph. We say that a metric (X,d) has *doubling dimension* $\dim(X)$ at most α if every set of diameter D can be covered by 2^{α} sets of diameter $D/2$. (A *doubling metric* is one whose doubling dimension $\dim(X)$ is a constant.) For a connected graph G, whose shortest path distances d_G induce the doubling metric (X, d_G), we show how to perform $(1+\tau)$-stretch routing on G for any $0 < \tau \leq 1$ with routing tables of size at most $(\alpha/\tau)^{O(\alpha)} \log\Delta\log\delta$ bits with only $(\alpha/\tau)^{O(\alpha)} \log\Delta$ entries, where Δ is the diameter of G and δ is the maximum degree of G. Hence, the number of routing table entries is just $\tau^{-O(1)}\log\Delta$ for doubling metrics.

The Edge Disjoint Paths (EDP) problem in undirected graphs refers to the following: Given a graph G with n nodes and a set $\mathcal{T}$ of pairs of terminals, connect as many terminal pairs as possible using paths that are mutually edge disjoint.

This leads to a variety of classic NP-complete problems, for which approximability is not well understood. We show a polylogarithmic approximation algorithm for the undirected EDP problem in general graphs with a moderate restriction on graph connectivity: we require the global minimum cut of G to be $\Omega(\log^5 n)$. Previously, constant

continued on page 15

Matthew Wachs explains Performance Insulation and Predictability at the 2005 PDL Workshop and Retreat.

continued from page 14

or polylogarithmic approximation algorithms were known for trees with parallel edges, expanders, grids and grid-like graphs, and, most recently, even-degree planar graphs. These graphs either have special structure (e.g., they exclude minors) or there are large numbers of short disjoint paths. Our algorithm extends previous techniques in that it applies to graphs with high diameters and asymptotically large minors.

In the classification problem, we are given a set of $2N$ diploid individuals from population P_1 and P_2 (with no admixture), and a small amount of multilocus genotype data from the same set of K loci for all $2N$ individuals, and we aim to partition P_1 and P_2 perfectly. Each population P_a, where $a \in \{1,2\}$, is characterized by a set of allele frequencies at each locus. In our model, given the population of origin of each individual, the genotypes are assumed to be generated by drawing alleles independently at random across the K loci, each from its own distribution. For example, each SNP (or Single Nucleotide Polymorphism) has two alleles, which we denote with bit 1 and bit 0 respectively. In addition, each locus contains two bits (one from each parent) that are assumed to be two random draws from the same Bernoulli distribution.

We use p_1^k and p_2^k, for all $k = 1, \ldots, K$ to denote frequency of an allele mapping to bit 1 at locus k in P_1 and P_2, respectively. We use $\gamma = \sum_1^K {}_{=1} (p_1^k - p_2^k)^2 / K$ as the dissimilarity measure between P_1 and P_2. We compute the number of loci K that we need to perform different tasks, versus N and γ, and prove several theorems. Ultimately, we show that with probability $1-1/\text{poly}(N)$, given that $K = \Omega(\log N \log\log N / N\gamma^2)$ and $K = \Omega(\log N / \gamma)$, we can *recognize* the perfect partition (P_1, P_2) from among all other balanced partitions of the $2N$ individuals. We proved this theorem for two cases: either we are given two random draws for each attribute along each dimension, or only one.

Todd Mowry (CMU and Intel Research Pittsburgh) and Limor Fix (Intel Research Pittsburgh enjoy a round of mini-gold at Nemacolin.

THESIS ABSTRACT:
Design and Implementation of Self-Securing Network Interface Applications

Stanley M. Bielski, ECE

M.S. Thesis. Electrical and Computer Engineering, Carnegie Mellon University. December 2005.

This thesis presents a novel security platform that narrows the architectural gaps between traditional network security perimeters in a highly scalable and fault-isolated manner while providing administrators with a simple and powerful interface for configuration and coordination of security policies across multiple network components. The heart of this platform is the concept of self-securing network interfaces (SS-NIs), components that sit between a host system and the rest of the intranet, moving packets between the system's components and the network. Additionally SS-NIs examine the packets being moved and enforce network security policies.

This thesis makes four main contributions: First, it makes a case for NI-embedded intrusion detection and containment functionality. Second, it describes the design of NI system software for supporting such functionality. Third, it discusses our implementation of NI system software and the Castellan administrative console. Fourth, it describes several promising applications for detecting and containing network threats enabled by the placement of self-securing NIs at the host's LAN access point.

THESIS ABSTRACT:
Using Program Analysis to Identify and Compensate for Nondeterminism in Distributed, Fault-Tolerant Systems

Joseph Slember, ECE

M.S. Thesis. Electrical and Computer Engineering, Carnegie Mellon University. April 2006.

Fault-tolerant replicated applications are typically assumed to be deterministic, in order to ensure reproducible, consistent behavior and state across a distributed system. Real applications often contain nondeterministic features that cannot be eliminated. Through the novel application of program analysis to distributed CORBA applications, we decompose an application into its constituent structures, and discover the kinds of nondeterminism present within the application. We target the instances of nondeterminism that can be compensated for automatically, and highlight to the application programmer those instances of nondeterminism that need to be manually rectified. We demonstrate our approach by compensating for specific forms of nondeterminism and by quantifying the associated performance overheads. The resulting code growth is typically limited to one extra line for every instance of nondeterminism, and the runtime overhead is minimal, compared to a fault-tolerant application with no compensation for nondeterminism.

continued on page 18

continued from page 7

nosis, and decision making. Stardust is a monitoring infrastructure that replaces traditional performance counters with end-to-end traces of requests and allows for efficient querying of performance metrics. Such traces better inform key administrative performance challenges by enabling, for example, extraction of per-workload, per-resource demand information and per-workload latency graphs. This paper reports on our experience building and using end-to-end tracing as an online monitoring tool in a distributed storage system. Using diverse system workloads and scenarios, we show that such fine-grained tracing can be made efficient (less than 6% overhead) and is useful for on- and off-line analysis of system behavior. These experiences make a case for having other systems incorporate such an instrumentation framework.

Scheduling Speculative Tasks in a Compute Farm

Petrou, Gibson & Ganger

Proceedings of the ACM/IEEE Supercomputing 2005 Conference, Seattle, Washington, November, 2005.

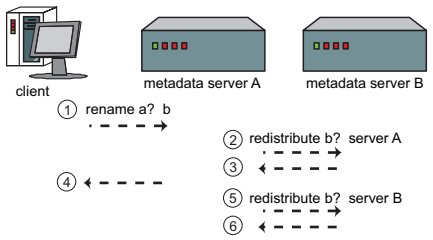
Users often behave speculatively, submitting work that initially they do not know is needed. Farm computing often consists of single node speculative tasks issued by, e.g., bioinformaticists comparing DNA sequences and computer graphics artists rendering scenes who wish to reduce their time waiting for needed tasks and the amount they will be charged for unneeded speculation. Existing schedulers are not effective for such behavior. Our 'batchactive' scheduling exploits speculation: users submit explicitly labeled batches of speculative tasks, interactively request outputs when ready to process them, and cancel tasks found not to be needed. Users are encouraged to participate by a new pricing mechanism charging for only requested tasks no matter what ran. Over a range of simulated user and task characteristics, we show that: batchactive scheduling improves visible response time—a new metric for speculative domains—by at least 2X for 20% of the simulations; batchactive scheduling supports higher billable load at lower visible response time, encouraging adoption by resource providers; and a batchactive policy favoring users who use more of their speculative tasks provides additional performance and resists a denial-of-service.

Eliminating Cross-server Operations in Scalable File Systems

Hendricks, Sinnamohideen, Sambasivan & Ganger

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-06-105, May 2006.

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 paper suggests an alternate approach 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 files onto one server and have it service the request. Although such redistribution is more expensive than a dedicated cross-server protocol, the rareness of such operations makes the overall performance impact minimal. Analysis of NFS traces indicates that cross-server operations make up fewer than 0.001% of client requests, and experiments with a prototype implementation show that the performance impact is negligible when such operations make up as much as 0.01% of operations. Thus, when dynamic redistribution functionality exists in the system, cross-server operations can be handled with almost no additional implementation complexity.



Design for eliminating cross-server operations. The sequence of operations required to handle RENAME a to b is shown. Returning to the original state is similar.

Improving Small File Performance in Object-based Storage

Hendricks, Sambasivan, Sinnamohideen & Ganger

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-06-104, May 2006.

We propose architectural refinements, server-driven metadata prefetching and namespace flattening, for improving the efficiency of small file workloads in object-based storage systems. Server-driven metadata prefetching has the metadata server provide information and capabilities for multiple objects, rather than just one, in response to each lookup. Doing so allows clients to access the contents of many small files for each metadata server interaction, reducing access latency and metadata server load. Namespace flattening encodes the directory hierarchy into object IDs such that namespace locality translates to object ID similarity. Doing so exposes namespace relationships among objects (e.g., as hints to storage devices), improves locality in metadata indices, and enables use of ranges for exploiting them. Trace-driven simulations and experiments with a prototype implementation show significant performance benefits for small file workloads.

continued on page 17

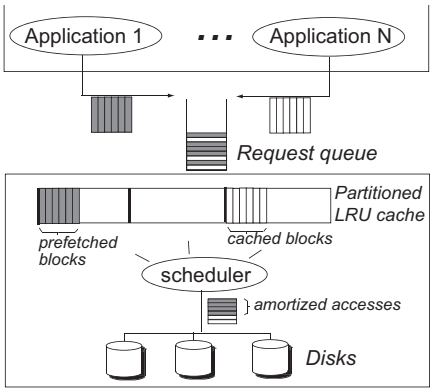
continued from page 16

Argon: Performance Insulation for Shared Storage Servers

Wachs, Abd-El-Malek, Thereska & Ganger

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-06-106, May 2006.

Services that share a storage system should realize the same efficiency, within their share of its time, as when they have it to themselves. This paper describes mechanisms for mitigating the inefficiency arising from inter-service disk and cache interference in traditional systems and their realization in Ursa Minor's storage server, Argon, which uses multi-MB prefetching and write-back to insulate sequential stream efficiency from the disk seeks introduced by competing workloads. It explicitly partitions the cache capacity among services to insulate the hit rate each enjoys from the access patterns of others. Experiments show that, combined, these mechanisms allow Argon to provide to each client a configurable fraction (e.g., 0.9) of its standalone efficiency. With fair-share scheduling, each of n clients approaches the ideal of 1/n of its standalone throughput.



In many storage systems, requests from different applications end up mixed together in the same global queue, resulting in inefficient scheduling. Similarly, nothing prevents one application from unfairly taking most of the cache space. With enhancements for performance insulation in the Argon storage server, Ursa Minor partitions the cache to ensure each application receives space; throttles applications issuing too many requests; and uses read prefetching and write coalescing to improve disk efficiency.

Challenges and Opportunities in Internet Data Mining

Andersen & Feamster

Carnegie Mellon University Parallel Data Lab Technical Report, CMU-PDL-06-102, February 2006.

Internet measurement data provides the foundation for the operation and planning of the networks that comprise the Internet, and is a necessary component in research for analysis, simulation, and emulation. Despite its critical role, however, the management of this data—from collection and transmission to storage and its use within applications—remains primarily ad hoc, using techniques created and re-created by each corporation or researcher that uses the data. This paper examines several of the challenges faced when attempting to collect and archive large volumes of network measurement data, and outlines an architecture for an Internet data repository—the datapository—designed to create a framework for collaboratively addressing these challenges.

Towards Efficient Semantic Object Storage for the Home

Salmon, Schlosser & Ganger

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-06-103, May 2006.

The home provides a new and challenging environment for data management. Devices in the home are extremely heterogeneous in terms of computational capability, capacity, and usage. Yet, ideally, information would be shared easily across them. Current volume-based filesystems do not provide the flexibility to allow these specialized devices to keep an up-to-date view of the information they require without seeing large amounts of traffic to other, unrelated pieces of information. We propose the use of "data views" to allow devices to subscribe to particular classes of objects. Data views allow devices to see up-to-date information from available devices, and eventual consistency with unavailable devices, for objects of interest without seeing updates to other objects in the system. They also provide a basis on which to build reliability, data management and search.

Quantifying Interactive User Experience on Thin Clients

Tolia, Andersen & Satyanarayanan

IEEE Computer. March, 2006.

The adequacy of thin-client computing is highly variable and depends on both the application and the available network quality. For intensely interactive applications, a crisp user experience may be hard to guarantee. An alternative—stateless thick clients— preserves many of the benefits of thin-client computing but eliminates its acute sensitivity to network latency.

Database Servers on Chip Multiprocessors: Limitations and Opportunities

Hardavellas, Pandis, Johnson, Mancheril, Ailamaki & Falsafi

Carnegie Mellon University School of Computer Science Technical Report CMU-CS-06-153, September 2006.

With the advent of chip multiprocessors, high-performance data management software is an imminent technical and research challenge for the database community. In this paper we characterize the performance of a commercial database server on top of the emerging chip multiprocessor technologies. Using both simulations and experiments on real hardware we find that the major bottleneck of current software is data cache stalls, and briefly describe techniques to address the problem. Finally, we derive a list

continued on page 20

continued from page 15

THESIS PROPOSAL:
Using Utility Functions to Control a Distributed Storage System

John Strunk, ECE

Managing a storage system and its associated workloads is a difficult task. Each stored dataset has unique requirements and cost constraints. It is the system administrator's responsibility to configure the storage system to adequately meet the needs of each workload, subject to the constraints posed by the available system resources. This task is particularly difficult because of the complex tradeoffs involved. Each configuration decision, from the level of redundancy to the selection of device types, affects multiple workload and system metrics (e.g., reliability, capacity, and latency). This work will investigate the use of utility functions for automating these decisions to better control a distributed storage system.

Alina Oprea and Minglong Shao, both graduate students at CMU, enjoy the sunshine at the 2005 Retreat.

THESIS PROPOSAL:
Efficient Data Organization and Management on Heterogeneous Storage Hierarchies

Minglong Shao, CS

As a central part of database systems, data organization has a direct impact on functionality and performance of data management applications. Current data organization, based on the conventional linear abstract of storage devices, linearizes multidimensional data along a preselected 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.

My thesis plan is to propose new data organization and management in DBMSs that better utilize and adapt to different characteristics of storage devices across 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 have designed and developed Clotho, which focuses on the page layout for 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 model that stores multidimensional data onto disks without losing spacial locality. MultiMap exploits the new 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. In the future, I will continue working on MultiMap to support non-uniform multidimensional datasets.

Garth Gibson, leads John Wilkes of HP Labs, Gregg Economou, and others on a trek through the woodlands at the 2005 Retreat.

After that, my next step is to investigate the buffer pool management for query-specific pages.

THESIS PROPOSAL:
Modeling the Relative Fitness of Storage

Michael Mesnier, ECE

Relative fitness is a new approach to modeling the performance of storage devices. In contrast with conventional modeling techniques, a relative fitness model is trained to predict -changes- in performance between any pair of devices. Relative fitness therefore allows service observations (e.g., performance and resource utilization) from one device to be used in making predictions for another. Such observations often provide more predictability than basic workload characteristics and are beneficial when workload characteristics are difficult to accurately, yet concisely, represent.

Mike Mesnier discusses his research with Will Akin of Intel.

DATA CENTER OBSERVATORY OPENS ITS DOORS

Bill Courtright

The Data Center Observatory (DCO), originally conceived in 2003, went online in April of this year. Much more than just a machine room, the DCO is both a showcase and a testbed for data center operations research as well as a compute and storage utility for researchers across the CMU campus.

In May, the DCO was officially launched with a ribbon cutting ceremony. Speakers at this event included Dr. Jared Cohon, President of CMU; Pradeep Khosla, Dean of CIT; Ron Seftick, VP of APC's construction and facilities engineering group; and our very own Greg Ganger. IT organizations from around the area sent people to attend, and the event was well-covered in the press, most notably by articles in Information Week and Network World.

Engineering and Construction

The process of planning and building the DCO has been a lengthy one. The first task was a thorough scoping phase that lasted over a year, concluding with allocation by Carnegie Mellon of just over 3,000 square feet on the lobby level of the new Collaborative Innovation Center, having a preliminary set of design requirements in hand and anxious researchers around campus eager to participate in a large-scale shared infrastructure.

In June of 2005, working with the University's Campus Design organization and APC, our partner for power & cooling solutions, we engaged an outside engineering firm to produce a detailed design of all mechanical systems necessary to operate the DCO (power, cooling, monitoring, flooring, fire suppression, access control, etc.). This led to the preparation of a complete set of construction plans, which were put out to bid in September. Construction began in December and was substantially complete by the end of March of this year. Teams from APC arrived on site in the first week of April to install the racks, electrical and cooling equipment, which

Dedication ceremony attendees, from left to right: Bill Courtright, PDL Executive Director; Jared Cohon, Carnegie Mellon University President; Greg Ganger, ECE Professor and PDL Director; and Ron Seftick, APC Vice President.

form the first of four InfraStruXure® zones. One week later we began to install computers, storage servers and network switches.

Status

The DCO is intended to serve two purposes: as a vehicle for studying data center challenges and solutions and as a shared computing and storage utility. Although much work remains, we have made good progress in both dimensions in the first six months.

At present, the DCO houses 220 compute and storage servers. Installations to date consist of 154 1u servers (P4 and Xeon configurations), 26 3u Xeon boxes (each with 16x400GB=6.4TB of disk space) and 40 blade servers. Together with network switches, environmental monitoring gear and file servers, we have populated 8 of the 12 racks in the first zone and now draw about 47kw of power. With additional machines arriving each month, we expect the first zone to be filled by the end of the year.

Most of the computers are currently used by affiliated PDL and CyLab researchers for software development and distributed system experimentation, but there is also an external Carnegie Mellon "customer" as well. The addition of machines used by non-affiliated people is an important step in building the shared infrastructure. Professor Elias Towe, ECE faculty and Director of the Center for Nano-enabled Device and Energy Technologies (CNXT) provided the 40 blade servers in the DCO, and his group uses them for nanotechnology simulations.intends Currently, machine allocation is crude and human-determined, with little sharing, but that will change in the coming year.

In addition to getting the DCO built and a number of servers up and running, we have made good initial progress on data collection about data center management. Instrumentation and processes are now in place and we are collecting information on how administrators spend their time, how much power and cooling is required, and where those resources go (right down to the individual machine level).

Looking forward, the coming year will see the addition of a second InfraStruXure® zone, which will bring the DCO's capacity to 23 racks of computers. Additionally, we will add a job control system for compute machines, which will enable users from distinct research groups to share them. This will be a first step towards a virtualized compute utility. We will also begin to move customers to storage based on our experimental storage system, Ursa Minor. Finally, we expect our data collection work to make progress as we continue to gather information from an increasing array of sources (e.g., failure data, resource utilization, etc.), and commence correlation and analysis studies.

Piping, redundant pumps and heat exchangers that provide chilled water to the DCO.

RECENT PUBLICATIONS

continued from page 17

of features for future database designs and outline a preliminary design and implementation of an adaptable multi-core optimized database engine, called Cordoba.

Nondeterminism in ORBs: The Perception and the Reality

Slember & Narasimhan

Workshop on High Availability of Distributed Systems, Krakow, Poland, September 2006

Nondeterminism is a source of problems for distributed replication because it makes it difficult to keep replicas consistent as they execute, process invocations and modify their internal states. Even if a middleware application is completely deterministic, the underlying middleware, e.g., the ORB, can continue to remain a source of nondeterminism. The paper presents our analysis of an open-source ORB from the viewpoint of nondeterminism. Our approach identifies the various sources of nondeterminism within the ORB. Our results demonstrate that while ORBs can contain several apparently nondeterministic system calls and functions, only a fraction of them manifest as actual nondeterminism and pose a threat to replica consistency.

Design Tradeoffs in Applying Content Addressable Storage to Enterprise-scale Systems Based on Virtual Machines

Nath, Kozuch, O'Hallaron, Harkes, Satyanarayanan, Tolia & Toups

Proceedings of the 2006 USENIX Annual Technical Conference (USENIX '06), Boston, Massachusetts, May-June 2006.

This paper analyzes the usage data from a live deployment of an enterprise client management system based on virtual machine (VM) technology. Over a period of seven months, twenty-three volunteers used VM-based computing environments hosted by the system and created over 800 checkpoints of VM state, where each checkpoint included the virtual memory and disk states. Using this data, we study the design tradeoffs in applying content addressable storage (CAS) to such VM-based systems. In particular, we explore the impact on storage requirements and network load of different privacy properties and data granularities in the design of the underlying CAS system. The study clearly demonstrates that relaxing privacy can reduce the resource requirements of the system, and identifies designs that provide reasonable compromises between privacy and resource demands.

continued on page 23

YEAR IN REVIEW

continued from page 4

December 2005

- Researchers from Carnegie Mellon's Parallel Data Lab (PDL) received both Best Paper awards at the 2005 File and Storage Technologies (FAST) conference in San Francisco, CA. "Ursa Minor: Versatile Cluster-based Storage," was presented by John Strunk and "On Multidimensional Data and Modern Disks" by Steve Schlosser.
- Stan Bielski received his M.S. for his work on "Design and Implementation of Self-Securing Network Interface Applications".
- Minglong Shao proposed her Ph.D. research, titled "Efficient Data Organization and Management on Heterogeneous Storage Hierarchies".
- Greg, Brandon and Raja attended the 2nd JST CREST Workshop on Advanced Storage Systems, sponsored by the Japan Science and Technology Agency in San Francisco. Greg gave an invited talk on Self-* Storage. Raja presented "Preserving Namespace Locality in Object-based Storage" and Brandon spoke on "Relative Fitness Models for Storage".

October 2005

- 13th Annual PDL Retreat & Workshop.
- Craig Soules presented "Connections: Using Context to Enhance File Search" at SOSP 2005 in Brighton, UK.
- Jay Wylie presented "Fault-Scalable Byzantine Fault-Tolerant Services" at SOSP 2005 in Brighton, UK.
- Michael Abd-El-Malek spoke on "Lazy Verification in Fault-Tolerant Distributed Storage Systems" at SRDS 2005 in Orlando, FL.
- Priya Narasimhan has been elected to membership in the International Federation for Information Processing (IFIP) Working Group 10.4 on Dependable Computing and Fault-Tolerance.
- Priya Narasimhan won an IBM Faculty Partnership Award for her recent proposed research on developing coordinated upgrades for distributed systems.

Bill points out DCO power management features to Dan Cassiday of Sun at the 2005 Spring Industry Visit Day.

continued from page 9

on usable privacy and security. The course is offered for the first time this semester in the School of Computer Science (http://cups.cs.cmu.edu/courses/ups.html). It is designed to introduce students to a variety of usability and user- interface problems related to privacy and security and give them experience in designing studies aimed at helping to evaluate usability issues in security and privacy systems.

-- CMU's 8 1/2 x 11 News

February 27, 2006
Song Selected for IBM Faculty Award

Dawn Song, Assistant Professor of ECE and CS, received an IBM Faculty Award, which recognizes and fosters novel, creative work as well as strengthens the relationships between universities and the IBM research and development community. Song's research offers fundamentally new techniques for defending against large-scale Internet attacks, including as worms and Distributed Denial-of-Service (DDoS) attacks, and supplies a foundation for building an attack-resilient communication infrastructure for both the current and next generation Internet.

-- ECE News Online

January 2006
James Newsome awarded Microsoft Research Fellowship

Congratulations to James for being awarded a Microsoft Research Fellowship! He won this award for his outstanding thesis work on "Sting: an automatic self-healing defense system against zero-day exploit attacks".

As stated by MSR, "[The] selection for this award is a tremendous honor and recognition of [the recipient's] accomplishments." This fellowship is one of the most prestigious fellowships for a PhD student, where each school is usually only allowed to submit up to three of their top candidates, and only less than 15% of these highly qualified candidates will be selected for the award.

January 2006
Jure Leskovec awarded Microsoft Research Fellowship

Congratulations to Jure Leskovec (PDL, CALD), who has been selected as Microsoft Research. A fellow for the next two years. Jure works with Christos Faloutsos, and is interested in link analysis and large graph mining.

The competition for these fellowships was extremely high. The award is one of only 10 specially funded MSR fellowships in "a major new Microsoft initiative that will be publicly announced in the coming weeks."

January 2006
Congratulations James and Anne!

James Hendricks and Anne Swift were married on January 8, 2006 at the First Unitarian Church of Pittsburgh in a private ceremony, but are planning a larger wedding ceremony for a little over a year from now to celebrate our marriage with friends. Anne is a first year doctoral student in the Department of Social and Decision Sciences here at Carnegie Mellon studying strategy, entrepreneurship, and technological change.

January 2006
PDL Researchers Receive Both Best Paper Awards at FAST 2005!

Researchers from Carnegie Mellon's Parallel Data Lab (PDL) received both Best Paper awards at the recent File and Storage Technologies (FAST) conference, the top forum for storage systems research.

"Ursa Minor: Versatile Cluster-based Storage," describes initial steps toward PDL's long-term target of storage systems that manage themselves (Self-* Storage). "On Multidimensional Data and Modern Disks" introduces a new approach to exploiting modern disk characteristics for better system performance. That research arises from collaboration between the PDL, Intel Pittsburgh, and EMC Corporation.

-- ECE News Online

December 2005
Jia-Yu Pan Receives Best Paper at ICDM

Jia-Yu (Tim) Pan, a doctoral student in computer science, won one of five best student paper awards at ICDM'05, one of the top data mining conferences. The paper is on mining biomedical images using a novel technique of visual vocabularies and independent component analysis.

-- CMU's 8 1/2 x 11 News.

December 2005
Hui Zhang Selected as ACM Fellow

Computer Science professor Hui Zhang has been selected as an ACM Fellow. Zhang's research interests are in computer networks, specifically on the scalability, robustness, dependability, security and manageability of broadband access networks, enterprise networks and the Internet. His end system multicast work has been used for the real-time broadcast of national events, including the John Kerry rally on campus during the 2004 presidential campaign.

-- CMU's 8 1/2 x 11 News

continued from page 11

Prototype Design

Perspective has been designed in a modular fashion to allow alternate implementations of many of the components. Figure 2 shows the major components of Perspective. The view manager manages the control messages for Perspective by communicating with view managers on remote devices.

The transfer manager is responsible for transferring data between devices and may contain a set of data transfer protocols. The version manager compares two versions of an object and decides which version is newest, using the conflict manager to resolve problems. The object manager coordinates updates to objects, applying updates to the local store, passing messages to the frontends and the view manager, and processing frontend requests.

A frontend connects an application to Perspective, customizing the way in which objects are transferred between devices. For instance, a frontend could implement a standard file system interface by mapping object attributes into a file system hierarchy, or it could communicate directly with an application to customize the interface. Each frontend can be configured for callbacks on object modification. This allows a frontend to decide if updates should be ignored, maintain extra state if it chooses to do so, and perform extra operations, such as automatically transcoding objects on updates.

Finally, the local object store (LOS) stores object replicas and metadata. We have implemented our own LOS, but any semantic store would suffice.

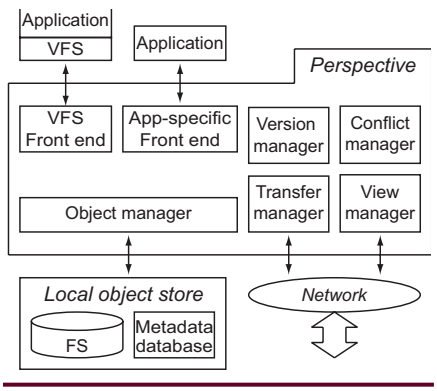


Figure 2: Major components of Perspective.

Some Early Results

Experiments with Perspective show that views provide significant performance benefits in creating device ensembles. They also show performance advantages over alternate methods of providing search and event notification in ad hoc ensembles, while automatically providing the same performance as a central server when similar resources exist.

The Perspective prototype is implemented in C++ as a user-level process. It currently runs on both Linux and Macintosh OS X. The object store is implemented by storing data in a backing filesystem and metadata in a custom XML database, which allows extensible semantic naming. The prototype currently allows data access from a client library, which communicates with the Perspective daemon through domain sockets and shared memory. It implements log-based sync, with a fall back of full object exchange to correctly propagate updates throughout the system. It also allows ensemble creation, disconnected operation, introduction of new views, and the addition of new devices. It uses TCP sockets to transfer view manager messages and data.

To evaluate Perspective, we connected two Mac-Book Pros, with 1.87 GHz processors and 1 and 2 GB RAM, one Linux laptop with a 1.6 GHz processor and 1 GB RAM, and a Linux desktop with a 3 GHz processor and 2 GB RAM to a 10Mbps half duplex wired hub to approximate home wireless bandwidth conditions.

Performance: Comparing the performance of Perspective with a local filesystem showed Perspective to have a significant overhead for a pure metadata workload due to IPC calls and a user-level implementation. For more typical data workloads, Perspective introduces a reasonable 4% overhead. We also found that the cost of a sync operation is a very small fraction of the overhead, meaning that most of the overhead is simply in the prototype's less efficient data transfer mechanism. Evaluating complex views, even up 200 views — considerably more than the expected number of views in a home deployment — imposed a negligible overhead on performance.

Search and event notification: In comparison with several approaches to implementing search and event notification, views matched or beat all approaches, even the ideal case with a central server, without requiring centralization. Perspective's use of views provided efficient search capabilities without the cost of building a central metadata store and without the extra cost of sending queries to unneeded power-limited devices. If there is a complete view currently accessible that covers a given query, it acts just like the central server case by noticing that the query only needs to go to that device. Perspective thus obtains the same benefits without requiring a centralized server and without requiring any device to store all metadata in the system.

If an appropriate complete view is not available, views allow Perspective to query only devices that could contain matching objects. This allows a power-limited device to participate in an ensemble without having to see most of the queries in the system (a power and CPU limited device like a cell phone would have little chance of keeping up with the other devices if all devices are searched on all queries).

Summary

We believe views are a powerful structuring tool for home and personal data management. By concisely describing the data objects that a device may store and access, views enable a group of devices to efficiently coordinate for consistency, search, reliability, and ad hoc ensemble creation without a central server. We believe that Perspective's use of views provides a strong foundation on which to build automated data management for home storage.

continued from page 20

Static Analysis Meets Distributed Fault-Tolerance: Enabling State-Machine Replication with Nondeterminism

Slember & Narasimhan

HotDep, Seattle, WA, Nov. 2006.

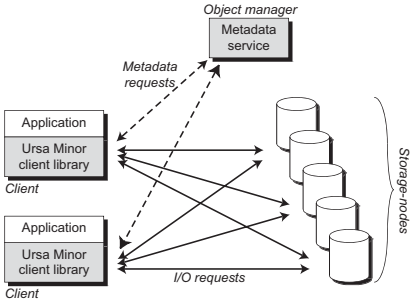
Midas is an inter-disciplinary approach to supporting state-machine replication for nondeterministic distributed applications. The approach exploits compile-time static analysis to identify both first-hand and second-hand sources of nondeterminism. Subsequent runtime compensation occurs through either the transfer of nondeterministic checkpoints or the re-execution of inserted code, and restores consistency among replicas before each new client request. The approach avoids the need for lock-step synchronization and leverages application-level insight to address only the nondeterminism that matters. Our preliminary evaluation demonstrates Midas' feasibility and current performance overheads.

Ursa Minor: Versatile Cluster-based Storage

Abd-El-Malek, Courtright, Cranor, Ganger, Hendricks, Klosterman, Mesnier, Prasad, Salmon, Sambasivan, Sinnamohideen, Strunk, Thereska, Wachs & Wylie

The 4th USENIX Conference on File and Storage Technologies (FAST '05). San Francisco, CA. Dec., 2005.

No single encoding scheme or fault model is right for all data. A versatile storage system allows them to be matched to access patterns, reliability requirements, and cost goals on a per-data item basis. Ursa Minor is a cluster-based storage system that allows data-specific selection of, and on-line changes to, encoding schemes and fault models. Thus, different data types can share a scalable storage infrastructure and still enjoy specialized choices, rather than suffering from "one size fits all." Experiments with Ursa Minor show performance benefits of 2-3x when using specialized choices as opposed to a single, more general, configuration. Experiments also show that a single cluster supporting multiple workloads simultaneously is much more efficient when the choices are specialized for each distribution rather than forced to use a "one size fits all" configuration. When using the specialized distributions, aggregate cluster throughput increased by 130%.



Ursa Minor high-level architecture.

Living with Nondeterminism in Replicated Middleware Applications

Slember & Narasimhan

International Middleware Conference, Melbourne, Australia, Nov. 2006

Application-level nondeterminism can lead to inconsistent state that defeats the purpose of replication as a fault-tolerance strategy. We present Midas, a new approach for living with nondeterminism in distributed, replicated, middleware applications. Midas exploits (i) the static program analysis of the application's source code prior to replica deployment and (ii) the online compensation of replica divergence even as replicas execute. We identify the sources of nondeterminism within the application, discriminate between actual and superficial nondeterminism, and track the propagation of actual nondeterminism. We evaluate our techniques for the active replication of servers using micro-benchmarks that contain various sources (multi-threading, system calls and propagation) of nondeterminism.

On Multidimensional Data and Modern Disks

Schlosser, Schindler, Shao, Papadomanolakis, Ailamaki, Faloutsos & Ganger

The 4th USENIX Conference on File and Storage Technologies (FAST '05). San Francisco, CA. Dec., 2005.

With the well-ingrained notion that disks can efficiently access only one dimensional data, current approaches for mapping multidimensional data to disk blocks either allow efficient accesses in only one dimension, trading off the efficiency of accesses in other dimensions, or equally penalize access to all dimensions. Yet, existing technology and functions readily available inside disk firmware can identify non-contiguous logical blocks that preserve spatial locality of multidimensional datasets. These blocks, which span on the order of a hundred adjacent tracks, can be accessed with minimal positioning cost. This paper details these technologies, analyzes their trends, and shows how they can be exposed to applications while maintaining existing abstractions. The described approach can achieve the best possible access efficiency afforded by the disk technologies: sequential access along primary dimension and access with minimal positioning cost for all other dimensions. Experimental evaluation of a prototype implementation demonstrates a reduction of the overall I/O time between 30% and 50% for multidimensional data queries when compared to existing approaches.

Michael Stroucken begins assembly of the DCO computer storage racks.

continued from page 13

THE Open GROUP Making standards work® **High End Computing Extensions Working Group**

You are here: Platform Forum > HECEWG > Documents

Created	Title (see details)	Version (+ implies others)	Formats (download)
17-Aug-2006	Evaluation Criteria for Proposed High End Computing Extensions to the POSIX I/O API	1.2 +	PDF
30-Jun-2006	Manpage - readdirplus	1	PDF
30-Jun-2006	Manpage - lockg (group lock)	1	PDF
30-Jun-2006	Manpage - sutoc (convert file handle to file descriptor)	1	PDF
30-Jun-2006	Manpage - NFSV4acls	1	PDF
30-Jun-2006	Manpage - openg (group open)		PDF
30-Jun-2006	Manpage - statlite and family of light weight stat calls	1	PDF
30-Jun-2006	Manpage - open (O_LAZY flags)	1	PDF
30-Jun-2006	POSIX I/O High Performance Extensions presentation Panasas SC05	1	PDF
30-Jun-2006	POSIX I/O High Performance Computing Extensions ASC SC05 presentation	1	PDF
30-Jun-2006	High End Computing Early Goals for extesions to POSIX I/O API	1	PDF
30-Jun-2006	A Business Case for Extensions to the POSIX I/O API for High End, Clustered, and Highly Concurrent Computing	1	PDF

Figure 2: High End Computing Extentions Working Group web page at www.opengroup.org/platform/hecewg/.

agement technology such as is being developed in PDL's Self-* Storage project. Large-scale clusters are the computing systems both in greatest need of self-management and under greatest strain from scale. Instrumentation and diagnosis, data placement and load balancing, and configuration and tuning are all targets of automation for large scale clusters.

Other innovation projects under consideration for PDSI include security mechanisms that scale with the speeds'n'feeds of HPC clusters, and mechanisms to ensure that the result of sharing resources between significantly different apps will be predictable. With leadership from Institute partners U.C. Santa Cruz and their object storage system, Ceph, PDSI will develop mechanisms to scale rich metadata and its indexing. And with leadership from Institute partner the Pacific Northwest Nat. Lab., PDSI will explore file system integration with and exploitation of server virtualization technology. New, scalable interfaces for archive systems is another target for PDSI innovation, with leadership from Institute partner, Lawrence Berkeley Nat. Lab.

So ends a quick survey of plans in the Petascale Data Storage Institute. But, many of the readers of this newsletter don't work in supercomputing, and might wonder if PDSI applies to supercomputing only. Of course, supercomputing is on the minds of our funders and partners, but high-performance computing is far more than supercomputing. IDC expects the 2006 revenue in HPC systems to exceed $10 billion, with most of the pleasantly double digit growth in departmental sized systems, not in the supercomputing systems. It seems more and more companies are figuring out how to raise top line revenue by improving their analysis, analytics, simulation, and search computing with clusters.

Hopefully, a few of you will be intrigued by the Petascale Data Storage Institute. To that end, the Institute is open to collaboration and partnership with corporate research, product teams and other academic groups. All it takes is a project to engage in and interest in participating. Some PDL sponsors are already doing this, in that they are participating in the POSIX extensions, pNFS prototyping or data collection projects. For further information or to propose a partnership, talk to Garth who'll take it up with his co-PIs.

PDL Retreat 2005 attendee group photo.