PARALLEL DATA LABORATORY
CARNEGIE MELLON UNIVERSITY

PDL Packet

NEWSLETTER ON PDL ACTIVITIES AND EVENTS • FALL 2008
http://www.pdl.cmu.edu/

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

CONTENTS

PDL CONSORTIUM MEMBERS

American Power Corporation
Data Domain, Inc.
EMC Corporation
Facebook
Google
Hewlett-Packard Labs
Hitachi, Ltd.
IBM Corporation
Intel Corporation
LSI Corporation
Microsoft Research
NetApp, Inc.
Oracle Corporation
Seagate Technology
Symantec Corporation
VMware, Inc.

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

RECENT PUBLICATIONS

GIGA+ : Scalable Directories for Shared File Systems

Patil & Gibson

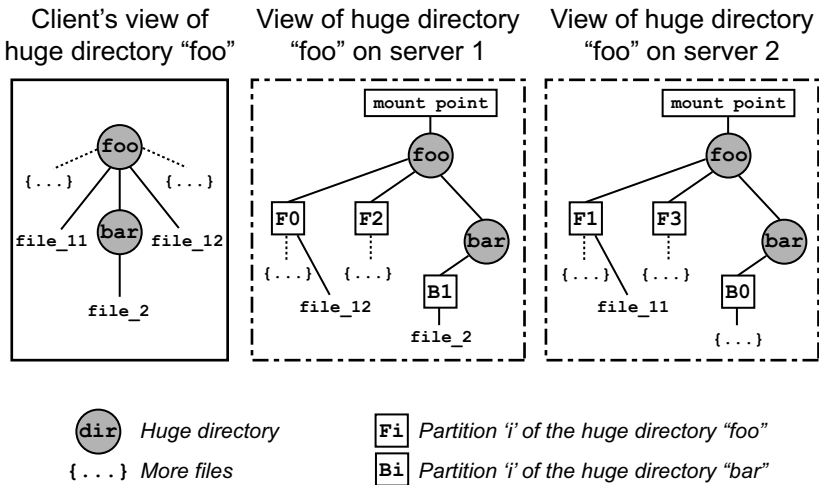
Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-08-110, August 2008.

Traditionally file system designs have envisioned directories as a means of organizing files for human viewing; that is, directories typically contain a few tens to thousands of files. Users of large, fast file systems have begun to put millions of files into single directories, for example, as simple databases. Furthermore, large-scale applications running on clusters with tens to hundreds of thousands of cores can burstily create files using all compute cores, amassing bursts of hundreds of thousands of creates or more.

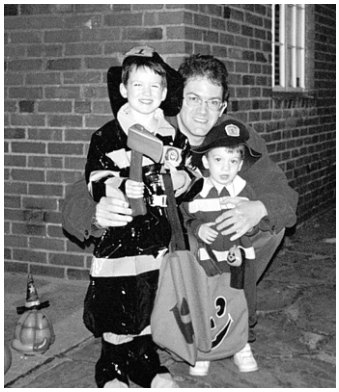
In this paper, we revisit data-structures to build large file system directories that contain millions to billions of files and to quickly grow the number of files when many nodes are creating concurrently. We extend classic ideas of efficient resizeable hash-tables and inconsistent client hints to a highly concurrent distributed directory service. Our techniques use a dense bitmap encoding to indicate which of the possibly created hash partitions really exist, to allow all partitions to split independently, and to correct stale client hints with multiple changes per update.

We implement our technique, GIGA+, using the FUSE user-level file system API

continued on page 5



Local representation of huge directory in Giga+ prototype layered on Ext3 — Currently, we replicate the huge directory tree structure (not the files or partitions) for ease of namespace management through common path-name manipulation across all servers.



FROM THE DIRECTOR'S CHAIR

Greg Ganger

Hello from fabulous Pittsburgh!

It has been another great year for the Parallel Data Lab. Some highlights include expansion of the Data Center Observatory (DCO), new research efforts initiated in energy efficiency and cloud computing, and great progress in ongoing research. Along the way, many students graduated and joined PDL Consortium companies, new students and faculty have joined PDL, and many papers have been published. Let me highlight a few things.

A lot of exciting activity is centered around the DCO, initiated over 5 years ago with a vision of enabling CMU research into automated cloud computing (we didn't have that name then), scalable storage, and data center energy efficiency. For example, APC has helped us to install a second power and cooling "zone", adding capacity for another 400 servers. Among other things, the new servers will be used as part of two open cloud computing testbeds in which PDL is participating: the OpenCloud testbed (led by the Open Cloud Consortium) and the OpenCirrus testbed (led by Intel, Yahoo!, and HP). These testbeds will greatly enhance CMU researchers' ability to explore new computing models, such as data-intensive scalable computing (DISC), and system support for them. The former received a great jumpstart from the year (and counting) of access to Yahoo!'s large-scale storage-heavy cluster called M45. The latter focus has spawned a new open source effort to create widely available software for cloud computing, initiated jointly by PDL, Intel, and Yahoo!.

A number of PDL research efforts focused on energy efficiency have emerged over the last year. One is exploiting data available from the DCO's deep instrumentation to develop and evaluate models for dynamic measurement and adaptive control of power and cooling in data centers. A second is exploring cluster scheduling and the use of various server energy levels (including off) to reduce energy usage. A third is exploring alternate hardware architectures for large-scale data-intensive computations, finding that large arrays of low-power nodes (based on embedded CPUs and Flash) can provide performance comparable to more conventional clusters at over an order of magnitude lower energy.

The Self-* Storage project continues to produce exciting new results as well as new approaches to building more automated scalable storage systems. Our ongoing effort to build a usable system, despite limited resources, has yielded a number of interesting schemes for creating and maintaining cluster-based storage systems with less effort. For example, we are exploring the use of virtual machines to address the porting problems associated with the client-side component of most cluster-based designs (including ours); we call the approach File System Virtual Appliances (FSVAs). We are exploring dynamic metadata server scaling without complex consistency protocols. We are exploring tools for exploiting end-to-end request flow tracing to simplify performance debugging of the complex distributed systems that are cluster-based storage systems. In addition, we continue to make progress on automation challenges, such as producing robust performance predictions in the face of system changes, insulating performance among clients sharing a storage cluster, and balancing different concerns (e.g., reliability, performance, and cost) in making provisioning and tuning decisions.

The Petascale Data Storage Institute (PDSI), led by Garth Gibson, continues to develop the community of academic, industry, and national lab experts focused

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.

FACULTY

Greg Ganger (pdl director)
412•268•1297
ganger@ece.cmu.edu

Anastasia Ailamaki
David Andersen
Lujo Bauer
Chuck Cranor
Lorrie Cranor
Christos Faloutsos
Rajeev Gandhi
Garth Gibson
Seth Copen Goldstein
Carlos Guestrin
Mor Harchol-Balter
Bruce Krogh
Julio López
Todd Mowry
David Nagle
Priya Narasimhan
David O'Hallaron
Adrian Perrig
Mike Reiter
Mahadev Satyanarayanan
Srinivasan Seshan
Bruno Sinopoli
Hui Zhang

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
Joan Digney
Adam Goode
James Moss
Doug Needham
Manish Prasad
Michael Stevens
Michael Stroucken
Spncer Whitman

GRADUATE STUDENTS

Michael Abd-El-Malek
Mukesh Agrawal
Jim Cipar
Debabrata Dash
Tudor Dumitraş
Anshul Gandhi
Varun Gupta
Nikos Hardavellas
James Hendricks
Shailesh Jain
Wesley Jin
Ryan Johnson
Christina Johns
Mike Kasick
Andrew Klosterman
Elie Krevat
Patrick Lanigan
Jure Leskovec
Eugene Marinelli
Michelle Mazurek
Iulian Moraru
Jim Newsome
Xinghao Pan
Ippokratis Pandis
Abdur Rehman Pathan
Swapnil Patil
Adam Pennington
Soila Pertet
Amar Phanishayee
Milo Polte
Rob Reeder
Dmitriy Ryaboy
Brandon Salmon
Raja Sambasivan
Hiral Shah
Faraz Shaikh
Tomer Shiran
Zoheb Shivani
Geeta Shroff
Jiri Simsa
Shafeeq Sinnamohideen
Joseph Slember
Ajay Surie
Atul Talesara
Jiaqi Tan
Wittawat Tantisiriroj
Vijay Vasudevan
Gaurav Veda
Matthew Wachs
Adam Wolbach
Lin Xiao

FROM THE DIRECTOR'S CHAIR

on the technology challenges faced in scaling storage systems to petascale sizes. Among other things, the third PDS Workshop, which will be held on November 17, 2008, at Supercomputing '08, brings together this community, and will again include two papers from PDL. PDL's research in this space continues along many directions. For example, we continue to explore approaches to mitigating the "incast" problem creating by data striping in high-performance networked storage. We are also developing new approaches to supporting very large-scale directories. The computer failure data repository continues to grow, enabling research into the characteristics of failures in various computing environments, and PDSI has begun a new repository for file system statistics from across a broad range of environments.

We have begun early deployments and user studies of the Perspective system that we have developed based on a new architecture for consumer storage in the home. The focus is on dramatically simplifying data management and sharing among the many storage-enhanced devices (e.g., DVRs, iPODs, laptops). Building on the concept of "views", which are queries against the attributes of objects, Perspective simplifies management of reliability and situations like device addition, travel with devices, and capacity exhaustion. We are also beginning to explore security mechanisms and access control policy management for this challenging environment.

Of course, many other ongoing PDL projects are also producing cool results. We have developed a new protocol (called Zzyzx) that provides unprecedented efficiency and scalability for Byzantine fault-tolerant services, providing a scheme for metadata to complement the fault-tolerant storage scheme developed last year. We are developing a new approach to upgrades in complex distributed systems, based on extensive use of virtual machines and storage. Extensive explorations into automated problem diagnosis in distributed systems include study of both model-based and learning-based schemes, centered around a new general architecture for instrumentation data of various types. This newsletter and the PDL website offer more details and additional research highlights.

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

NEW PDL FACES

Bruno Sinopoli

Bruno Sinopoli is an assistant professor in the department of Electrical and Computer Engineering at Carnegie Mellon University. Previously he was a postdoctoral scholar both at Stanford University and the University of California at Berkeley. Bruno Sinopoli received his M.S. and Ph.D. in Electrical Engineering from the University of California at Berkeley, in 2003 and 2005 respectively. Previously he received his Dr. Eng. degree from the University of Padova in Italy.

Bruno's research interests are in the design and analysis of networked embedded systems, with particular focus on wireless sensor actuator networks, distributed estima-

continued on page 16

YEAR IN REVIEW

November 2008

- 16th Annual PDL Retreat and Workshop.

October 2008

- Second zone installed in Data Center Observatory (DCO) providing power, cooling, and rack space for another 400 servers.
- Christos Faloutsos gave the keynote speech at the 2008 Internet Measurement Conference on "Graph Mining: Laws, Generators and Tools."

September 2008

- Julio López gave a half-day tutorial talk on "Data-Intensive Scalable Computing Systems for Science (DISCS)" at the Mass Storage Systems and Technology conference MSST'08 in Baltimore.
- Jiaqi worked with Steve Schlosser and Lily Mummert during his internship with Intel Research Pittsburgh over the summer.
- Abdur Rehman interned at Seagate Research, Pittsburgh on their Terabyte Home Project led by Erik Riedel (PDL Alumni) and Sami Iren.
- Jim Cipar worked with Intel Research Pittsburgh on a fellowship researching "Tashi: Dynamic VM Placement."
- Jure Leskovec interned at Microsoft studying the six degrees of separation theory with regard to internet instant messaging services.
- Elie Krevat spent the summer at VMware helping to build a prototype of a virtual data center as part of VMware's cloud computing efforts.
- Adam Wolbach completed his M.S. degree and presented his thesis research on "Improving the Deployability of Diamond."

August 2008

- Greg Ganger presented "Performance Insulation for Shared Cluster Storage" at the HECURA/FSIO Workshop in Arlington, VA.
- Garth Gibson presented "GIGA+: Scalable Directories for Shared File Systems" at the HECURA/FSIO Workshop in Arlington, VA.

July 2008

- Rob Reeder successfully defended his Ph.D. research titled "Expandable Grids: A user interface visualization technique and a policy semantics to support fast, accurate security and privacy policy authoring."
- Tudor Dumitraş proposed his Ph.D. research titled "Dependency-Agnostic Online Upgrade in Distributed Systems."
- Julio López visited LANL and gave a talk on "Data-Intensive Scalable Computing Systems for Science (DISCS)."

June 2008

- Michael Abd-El-Malek proposed his Ph.D. research on File System Virtual Appliances."
- Swapnil Patil gave an invited talk on "GIGA+: Scalable Directories for Shared File Systems" at the Conference on Scalability 2008 organized by Google in Seattle WA.

May 2008

- 10th Annual PDL Spring Industry Visit Day.
- Christos Faloutsos was the Keynote speaker at PAKDD 2008 in Osaka Japan, presenting "Graph Mining: Laws, Generators and Tools."

April 2008

- John Strunk completed his Ph.D. with the successful defense of his research on "Using Utility Functions to Control a Distributed Storage System."
- Congratulations to Hanghang Tong, Spiros Papadimitriou, Philip Yu and Christos Faloutsos on winning the best paper award for their work on "Proximity Tracking on Time-Evolving Bipartite Graphs" presented at SDM 2008 in Atlanta, GA.

February 2008

- John Strunk presented "Using Utility to Provision Storage Systems" at the 6th USENIX Conference on File and Storage Technologies (FAST '08) in San Jose, CA.
- Amar Phanishayee presented "Measurement and Analysis of TCP Throughput Collapse in Cluster-based Storage Systems" at the 6th USENIX Conference on File and Storage Technologies (FAST '08) in San Jose, CA.

January 2008

- James Hendricks proposed his Ph.D. research on "Low-overhead Byzantine Fault-Tolerant Storage."
- Swapnil Patil gave an invited talk on "GIGA+: Scalable Directories for Shared File Systems" at the Mathematics & Computer Science Division at Argonne National Laboratory in Argonne IL.

December 2007

- Michael Mesnier successfully defended his Ph.D. dissertation titled "On Modeling the Relative Fitness of Storage" and returned to work with Intel in Hillsboro, OR.

November 2007

- 15th Annual PDL Retreat and Workshop.
- Elie Krevat presented "On Application-level Approaches to Avoiding TCP Throughput Collapse in Cluster-Based Storage Systems" at the 2nd international Petascale Data Storage Workshop (PDSW '07) held in conjunction with Supercomputing '07 in Reno, NV.
- Swapnil Patil presented "GIGA+: Scalable Directories for Shared File Systems" at the 2nd international Petascale Data Storage Workshop (PDSW '07) held in conjunction with Supercomputing '07 in Reno, NV.

continued from page 1

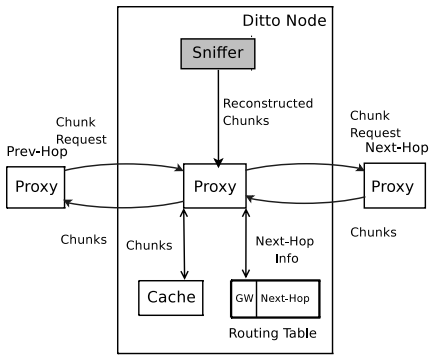
layered on Linux ext3. We measured our prototype on a 100-node cluster using the UCAR Metarates benchmark for concurrently creating a total of 12 million files in a single directory. In a configuration of 32 servers, GIGA+ delivers scalable throughput with a peak of 8,369 file creates/second, comparable to or better than the best current file system implementations.

Ditto - A System for Opportunistic Caching in Multi-hop Wireless Networks

Dogar, Phanishayee, Pucha, Ruwase & Andersen

The 14th Annual International Conference on Mobile Computing and Networking (Mobicom 2008). San Francisco, CA, Sept. 14-19, 2008.

This paper presents the design, implementation, and evaluation of Ditto, a system that opportunistically caches overheard data to improve subsequent transfer throughput in wireless mesh networks. While mesh networks have been proposed as a way to provide cheap, easily deployable Internet access, they must maintain high transfer throughput to be able to compete with other last-mile technologies. Unfortunately, doing so is difficult because multi-hop wireless transmissions interfere with each other, reducing the available capacity on the network. This problem is particularly severe in common gateway-based scenarios in which nearly all transmissions go through one or a few gateways from the mesh network to the Internet.



Ditto proxy design.

Ditto exploits on-path as well as opportunistic caching based on overhearing to improve the throughput of data transfers and to reduce load on the gateways. It uses content-based naming to provide application independent caching at the granularity of small chunks, a feature that is key to being able to cache partially overheard data transfers. Our evaluation of Ditto shows that it can achieve significant performance gains for cached data, increasing throughput by up to 7x over simpler on-path caching schemes, and by up to an order of magnitude over no caching.

On Application-level Approaches to Avoiding TCP Throughput Collapse in Cluster-based Storage Systems

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

Proceedings of the 2nd international Petascale Data Storage Workshop (PDSW '07) held in conjunction with Supercomputing '07. November 11, 2007, Reno, NV.

TCP Incast plagues scalable cluster-based storage built atop standard TCP/IP-over-Ethernet, often resulting in much lower client read bandwidth than can be provided by the available network links. This paper reviews the Incast problem and discusses potential application-level approaches to avoiding it.

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

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

6th USENIX Conference on File and Storage Technologies (FAST '08). Feb. 26-29, 2008. San Jose, CA.

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

Adaptive File Transfers for Diverse Environments

Pucha, Kaminsky, Andersen & Kozuch

2008 USENIX Annual Technical Conference (USENIX 2008), Boston, MA, June 22-27, 2008.

This paper presents dsync, a file transfer system that can dynamically adapt to a wide variety of environments. While many transfer systems work well in their specialized context, their performance comes at the cost of generality, and they perform poorly when used elsewhere. In contrast, dsync adapts to its environment by intelligently determining which of its available resources is the best to use at any given time. The resources sync can draw from include the sender, the local disk, and network peers. While combining these resources may appear easy, in practice it is difficult because these resources may have widely different performance or contend with each other. In particular, the paper presents a novel mechanism that enables dsync to aggressively search the receiver's local disk for useful data without interfering with concurrent network transfers. Our evaluation on several workloads in various network environments shows that dsync outperforms existing systems by a factor

continued on page 6

RECENT PUBLICATIONS

continued from page 5

of 1.4 to 5 in one-to-one and one-to-many transfers.

SALSA: Analyzing Logs as State Machines

Tan, Pan, Kavulya, Gandhi & Narasimhan

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-08-111, Sept. 2008.

SALSA examines system logs to derive state-machine views of the system's execution, along with control-flow, data-flow models and related statistics. Exploiting SALSA's derived views and statistics, we can effectively construct higher-level useful analyses. We demonstrate SALSA's approach by analyzing system logs generated in a Hadoop cluster, and then illustrate SALSA's value by developing visualization and failure-diagnosis techniques, for three different Hadoop workloads, based on our derived state-machine views and statistics.

Characterizing HEC Storage Systems at Rest

Dayal

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-08-109, July 2008.

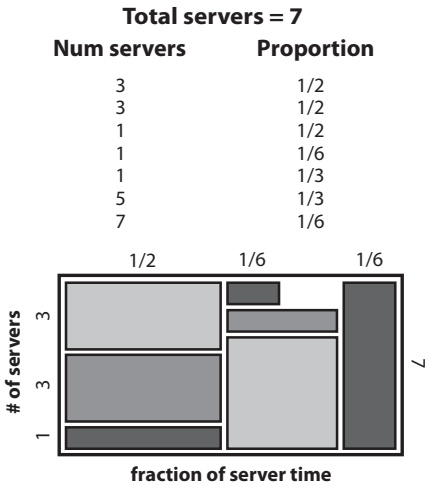
High-performance parallel file systems are a critical component of the largest computer systems, are primarily proprietary, and are specialized to high end computing systems that have many access patterns known to be unusual in enterprise and productivity workplaces. Yet little knowledge of even the basic distributions of file systems and file ages are publicly available, even though significant effort and importance is increasingly associated with small files, for example. In this paper we report on the statistics of supercomputing file systems at rest from a variety of national resource computing sites, contrast these to studies of the 80s and 90s of academic and software development campuses and observe the most interesting characteristics in this novel data.

Co-Scheduling of Disk Head Time in Cluster-based Storage

Wachs & Ganger

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-08-113, October 2008.

Disk timeslicing is a promising technique for storage performance insulation. To work with cluster-based storage, however, timeslices associated with striped data must be co-scheduled on the corresponding servers. This report describes algorithms for determining global timeslice schedules and mechanisms for coordinating the independent server activities. Experiments with a prototype show that, combined, they can provide performance insulation for workloads sharing a storage cluster—each workload realizes a configured minimum efficiency within its timeslices regardless of the activities of the other workloads.



Example problem instance and solution. Above is an example input to the scheduling algorithm; below is one possible solution. Rectangles correspond to workloads, with their height corresponding to the number of servers and their vertical location corresponding to which servers to use; their width corresponds to share of time and their horizontal location corresponds to what span of time during which their timeslices are scheduled. The enclosing rectangle represents a single round in the cluster; the schedule is repeated indefinitely.

Proximity Tracking on Time-Evolving Bipartite Graphs

Tong, Papadimitriou, Yu & Faloutsos

Bipartite Graphs. 5th VLDB Workshop on Secure Data Management (SDM 2008). Atlanta, GA, USA, April 2008. (Best Paper Award).

Given an author-conference network that evolves over time, which are the conferences that a given author is most closely related with, and how do they change over time? Large time-evolving bipartite graphs appear in many settings, such as social networks, co-citations, market-basket analysis, and collaborative filtering. Our goal is to monitor (i) the centrality of an individual node (e.g., who are the most important authors?); and (ii) the proximity of two nodes or sets of nodes (e.g., who are the most important authors with respect to a particular conference?) Moreover, we want to do this efficiently and incrementally, and to provide "any-time" answers. We propose pTrack and cTrack, which are based on random walk with restart, and use powerful matrix tools. Experiments on real data show that our methods are effective and efficient: the mining results agree with intuition; and we achieve up to 15–176 times speed-up, without any quality loss.

Ganesha: Black-Box Fault Diagnosis for MapReduce Systems

Pan, Tan, Kavulya, Gandhi & Narasimhan

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-08-112, October 2008.

Ganesha aims to diagnose faults transparently in MapReduce systems, by analyzing OS-level metrics alone. Ganesha's approach is based on peer-symmetry under fault-free conditions, and can diagnose faults that manifest asymmetrically at nodes within a MapReduce system. While our training is performed on smaller Hadoop

continued on page 7

continued from page 6

clusters and for specific workloads, our approach allows us to diagnose faults in larger Hadoop clusters and for unencountered workloads. We also candidly highlight faults that escape Ganesha's black-box diagnosis.

FAWN: A Fast Array of Wimpy Nodes

Andersen, Franklin, Phanishayee, Tan & Vasudevan

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-08-108, May 2008.

This paper introduces the FAWN—Fast Array of Wimpy Nodes—cluster architecture for providing fast, scalable, and power-efficient key-value storage. A FAWN links together a large number of tiny nodes built using embedded processors and small amounts (2–16GB) of flash memory into an ensemble capable of handling 700 queries per second per node, while consuming fewer than 6 watts of power per node. We have designed and implemented a clustered key-value storage system, FAWN-DHT, that runs atop these nodes. Nodes in FAWN-DHT use a specialized log-like back-end hash-based database to ensure that the system can absorb the large write workload imposed by frequent node arrivals and departures. FAWN uses a two-level cache hierarchy to ensure that imbalanced workloads cannot create hot-spots on one or a few wimpy nodes that impair the system's ability to service queries at its guaranteed rate. Our evaluation of a small-scale FAWN cluster and several candidate FAWN node systems suggest that FAWN can be a practical approach to building large-scale storage for seek-intensive workloads. Our further analysis indicates that a FAWN cluster is cost-competitive with other approaches (e.g., DRAM, multitudes of magnetic disks, solid-state disk) to providing high query rates, while consuming 3-10x less power.

User Level Implementation of Scalable Directories (GIGA+)

Hase, Jayaraman, Perneti, Sridharan, Patil, Polte & Gibson

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-08-107, May 2008.

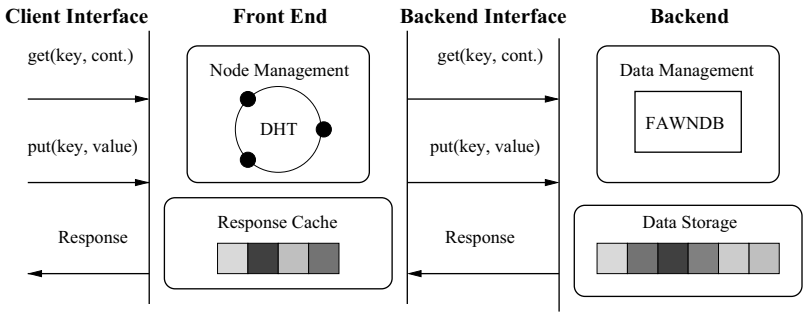
High performance computing applications are becoming increasingly widespread in a large number of fields. However the performance of I/O subsystems within such HPC computing environments has not kept pace with extreme processing and communication speeds of such computing clusters. The problem of high performance is tackled by system architects by employing a variety of storage technologies such as Parallel File Systems. While such solutions serve to significantly alleviate the problem of I/O performance scaling they still a lot of roam for improvement because they not endeavor to significantly scale the performance of meta-data operations. Such a situation can easily arise in database or telecommunication applications that create thousands of files per second in a single directory. When this happens, the consequent performance degradation effected by the slowdown of meta-data operations can severely slowdown the performance of the overall I/O system. GIGA+ affords a potential solution to this crucial issue of meta-data performance scaling in I/O sub-systems. GIGA+ is a scalable directory service that aims to scale and parallelize meta-data operations.

IRONModel: Robust Performance Models in the Wild

Thereska & Ganger

SIGMETRICS'08, June 2–6, 2008, Annapolis, Maryland, USA.

Traditional performance models are too brittle to be relied on for continuous capacity planning and performance debugging in many computer systems. Simply put, a brittle model is often inaccurate and incorrect. We find two types of reasons why a model's prediction might diverge from the reality: (1) the underlying system might be misconfigured or buggy or (2) the model's assumptions might be incorrect. The extra effort of manually finding and fixing the source of these discrepancies, continuously, in both the system and model, is one reason why many system designers and administrators avoid using mathematical models altogether. Instead, they opt for simple, but often inaccurate, "rules-of-thumb". This paper describes IRONModel, a robust performance modeling architecture. Through studying performance anomalies encountered in an experimental cluster-based storage system, we analyze why and how models and actual system implementations get out-of-sync. Lessons learned from that study are incorporated into IRONModel. IRONModel leverages the redundancy of high-level system specifications described through models and low-level system implementation to localize



The FAWN API and division of responsibilities.

continued on page12

October 2008

Carlos Guestrin Among Popular Science's "Brilliant 10"

Carlos Guestrin, assistant professor of machine learning and computer science, has been named as one of Popular Science's "Brilliant 10," the magazine's annual list of top young scientists. Dubbed "the Information Wrangler" by the magazine, Guestrin was cited for developing the Cascades algorithm, which obtains the most information with the least amount of effort. The algorithm works regardless of whether you want to determine the optimal number and placement of sensors in a water distribution system or simply the best blogs to read to get news as quickly as possible.

-- CMU 8.5x11 News Oct 16, 2008

October 2008

Jim and Melanie Wed!

Jim Cipar and Melanie Wilson, married October 4, 2008 in Worcester, Massachusetts. They honeymooned in Sonoma Valley and now Jim is back hard at work preparing for the retreat. Congratulations!

September 2008

Perspectives Developed to Thwart Internet Eavesdropping

The growth of shared Wi-Fi and other wireless computer networks has increased the risk of eavesdropping on Internet communications, but researchers at the School of Computer Science and College of Engineering have devised a low-cost system that can thwart these "Man-in-the-Middle" attacks.

The researchers - David Andersen, assistant professor of computer science, Adrian Perrig, associate professor of electrical and computer engineering and public policy, and Dan Wendlandt, a Ph.D. student in computer science - have incorporated Perspectives into an extension for the popular Mozilla Firefox v3 browser that can be downloaded free of charge at www.cs.cmu.edu/~perspectives/firefox.html.

"Perspectives provides an additional level of safety to browse the Internet," Perrig said. "To the security conscious user, that is a significant comfort."

-- CMU 8.5x11 News Sept 4, 2008

September 2008

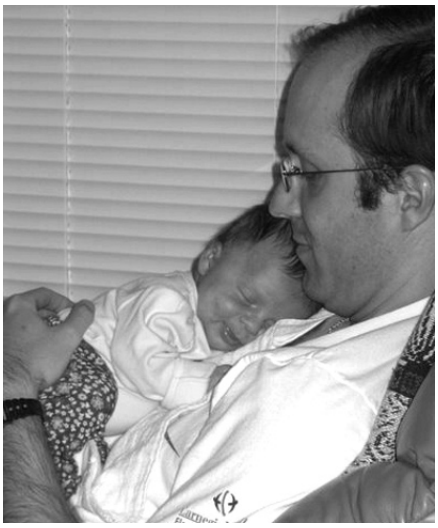
Welcome Morgan!

Joan, Bruce and big brother Evan welcomed Morgan Mary Georgia Digney on September 17 at 4:21 am. She weighed 6 lbs 13.5 oz and was 19.5 inches long. Looks like she came out fighting.

August 2008

Julia Alexis Arrives!

Julia Alexis was born to John and Corley Strunk on August 17 at 3:15 pm. She weighed 6 lbs 4 oz and was 20 inches long. Good luck to John and his family as he settles into his new job with NetApp.



August 2008

Greg Ganger Earns HP Innovation Research Award

CMU's Greg Ganger was one of 33 recipients worldwide to receive a 2008 HP Innovation Research Award, which is designed to encourage open collaboration with HP labs resulting in mutually beneficial, high-impact research.

Ganger, a professor of electrical and computer engineering and director of the Parallel Data Lab at Carnegie Mellon, will collaborate with HP Labs on the research initiative titled "Toward Scalable Self-Storage."

HP reviewed more than 450 proposals from individuals at 200 universities in 28 countries on a range of topics including intelligent infrastructure, sustainability, information explosion, dynamic cloud services and content transformation. A key element of each award will be on-campus support for one graduate student researcher.

Ganger said the award will serve to strengthen and deepen the long-

continued on page 9

continued from page 8

standing relationship between HP Labs' scalable storage researchers and Carnegie Mellon's Parallel Data Lab.

"We will be collaborating on our common interests in scalable, self-managing storage to tackle key challenges, including performance insulation between tenants sharing a common infrastructure and tenants with different requirements," Ganger said.

The research will also generate a prototype that can be tested in both commercial and academic venues. "HP partners with the best and brightest in the industry and academia to drive open innovation and set the agenda for breakthrough technologies that are designed to change the world," said Prith Banerjee, senior vice president of research at HP and director of HP Labs.

-- CMU Press Release Aug 21, 2008

July 2008

Jimeng Sun Runner-up for Best SIGKDD Dissertation Award

Dr. Jimeng Sun (Ph.D. CMU-CSD 2007), received the runner-up award for the best SIGKDD, which is the premier community for data mining research. Jimeng's dissertation is on tensor and stream mining, proposing novel and efficient methods to handle streams of numerical data, as well as streams of graphs. He applied his methods on chlorine monitoring in the drinking water (joint project with Prof. Jeanne VanBriesen of CIT/CMU), on monitoring the self-star data center of PDL/CMU (with Prof. Greg Ganger and his group), and also on monitoring computer traffic (with Prof. Hui Zhang, SCS/CMU).

June 2008

Welcome Reut!

After a "long wait" Reut Shiran made her entrance into the world on June 19. She is a healthy baby girl, and weighed in at 8.3 lbs and was 20.5 inches long. Her name is pronounced Re-oot (the first syllable sounds like the "re" in the color "red"). The literal meaning of the name (in Hebrew) is "friendship".

June 2008

Priya Narasimhan Receives Teaching Award

ECE Associate Professor Priya Narasimhan has been presented an award recognizing her teaching excellence by Eta Kappa Nu (Sigma Chapter, Carnegie Mellon). Congratulations Priya!

May 2008

Penguins Fans Get a New View

In the midst of hockey fever came news that software developed at Carnegie Mellon could offer up new benefits for Pittsburgh Penguins fans in coming seasons.

Called the "Yinz Cam," the tool could let spectators watch the game from any vantage point in the new arena on their cell phones. It could even tell them the best times to head for the refreshment line (or yes, even the bathroom line).

"Hockey moves really, really fast, and you want to catch every second of the game," explained Carnegie Mellon Professor Priya Narasimhan. "Even if you have excellent seats right up against the glass, when the action is on the other side of the arena you no longer have the best seats in the house."

Narasimhan and her students began working on a solution when they heard the Penguins were looking for ideas for the new arena, which is set to open in the Fall of 2010. With the Yinz Cam, spectators could download a widget onto their cell phones prior to the game. They can then choose from a variety of camera views. It also allows them to replay a goal, a fight, or any other action that happens on the ice.

Narasimhan said her students' passion for the Pens was her inspiration for the project. "The students are amazing," she said. "And they and I are big fans, so this is a really big deal for them."

-- Carnegie Mellon News

May 2008

PDL and Intel Welcome Henry Liam Schlosser!

Henry was born on Tuesday, May 27, at 11:40am to PDL Alumni Steve Schlosser and his wife Rachel. He was 6 lbs 7 oz, and is 19 inches long.

April 2008

Carlos Guestrin among Office of Naval Research 2008 Young Investigators Awardees

The Young Investigator Program (YIP) aims to attract to naval research those outstanding new faculty members at institutions of higher education. As part of the program, the Office of Naval Research (ONR) grants monetary support to award recipients for research and encourages their promising teaching and research careers. This year's YIP recipients showed exceptional talent in the following naval

continued on page 15

DISSERTATION ABSTRACT:
On Modeling the Relative Fitness of Storage

Michael P. Mesnier

Carnegie Mellon University ECE
Ph.D. Dissertation, Dec.19, 2007.

Storage management is usually handled by skilled system administrators. The specific task of configuring and allocating disk space for applications, often referred to as storage system design, is especially time-consuming and error-prone. Automated storage system design, a solution proposed by many, relies on fast and accurate performance predictions. However, challenges with conventional performance modeling have prevented such automation from being fully realized in practice.

Relative fitness is a new approach to modeling the performance of storage systems. In contrast to conventional models that predict the performance of storage systems based on the characteristics of workloads, referred to in this dissertation as absolute models, relative fitness models predict performance differences as workloads are moved across storage systems. There are two primary advantages. First, because relative fitness models are constructed for each pair of storage systems, the feedback of a closed workload can be captured (e.g., how the I/O arrival rate changes as the workload moves from storage system A to storage system B). Second, relative fitness models allow performance and resource utilization to be used in addition to workload characteristics. This is beneficial when workload characteristics are difficult to obtain or concisely express. For example, rather than trying to describe the spatio-temporal characteristics of a workload, one could use the observed performance and cache hit rate of storage system A to help predict the performance of storage system B.

This dissertation describes the steps necessary to build a relative fitness model, with an approach that is general enough to be used with any black-box modeling technique. Relative fitness models and absolute models are compared across a variety of workloads and disk arrays (RAID). When compared to absolute models, relative fitness models reduce the bandwidth prediction error up to 53%, throughput up to 23%, and latency up to 20%. In general, the best predictors of the relative fitness models are performance observations, followed by conventional workload characteristics.

Relative fitness models can be used in automated storage system design in a similar way that absolute models are used. Specifically, workloads can be observed on the storage systems that they are initially assigned to, relative fitness models can use these observations to predict the performance of different assignments, and optimization techniques can be used to select an assignment that optimizes performance.

DISSERTATION ABSTRACT:
Using Utility Functions to Control a Distributed Storage System

John D. Strunk

Carnegie Mellon University ECE
Ph.D. Dissertation, April 18, 2008.

Provisioning, and later optimizing, a storage system involves an extensive set of trade-offs between system metrics, including purchase cost, performance, reliability, availability, and power. Previous work has tried to simplify provisioning and tuning tasks by allowing a system administrator to specify goals for various storage metrics. While this helps by raising the level of specification from low-level mechanisms to high-level storage system metrics, it does not permit trade-offs between those metrics.

This dissertation goes beyond goal-based requirements by allowing the system administrator to use a utility function to specify his objectives. Using utility, both the costs and benefits of configuration and tuning decisions can be examined within a single framework. This permits a provisioning system to make automated trade-offs across system metrics, such as performance, data protection and power consumption. It also allows an automated optimization system to properly balance the cost of data migration with its expected benefits.

This work develops a prototype storage provisioning tool that uses an administrator-specified utility function to generate cost-effective storage configurations. The tool is then used to provide examples of how utility can be used to balance competing objectives (e.g., performance and data protection) and to provide guidance in the presence of external constraints. A framework for using utility to evaluate data migration is also developed. This framework balances data migration costs (decreases to current system metrics) against the potential benefits by discounting future expected utility. Experiments show that, by looking at utility over time, it is possible to choose the migration speed as well as weigh alternate optimization choices to provide the proper balance of current and future levels of service.

DISSERTATION ABSTRACT:
Expandable Grids: A user interface visualization technique and a policy semantics to support fast, accurate security and privacy policy authoring

Rob Reeder

Carnegie Mellon University SCS
Ph.D. Dissertation, July 21, 2008.

This thesis addresses the problem of designing user interfaces to support creating, editing, and viewing security and privacy policies. Policies are declarations of who may access what under which conditions. Creating, editing, and viewing—in a word, authoring—accurate policies is essential to keep-

continued on page 11

continued from page 11

ing resources both available to those who are authorized to use them and secure from those who are not. User interfaces for policy authoring can greatly affect whether policies match their authors' intentions; a bad user interface can lead to policies with many errors, while a good user interface can ensure that a policy matches its author's intentions. Traditional methods of displaying security and privacy policies in user interfaces are deficient because they place an undue burden on policy authors to interpret nuanced rules or convoluted natural language.

We introduce the Expandable Grid, a novel technique for displaying policies in a user interface. An Expandable Grid is an interactive matrix visualization designed to address the problems that traditional policy-authoring interfaces have in conveying policies to users. This thesis describes the Expandable Grid concept, then presents three pieces of work centered on the concept:

- a design, implementation, and evaluation of a system using an Expandable Grid for setting file permissions in the Microsoft Windows XP operating system;
- a description and evaluation of a file-permissions policy semantics that complements the Expandable Grid particularly well for reducing policy-authoring errors; and
- a design, implementation, and evaluation of a system using an Expandable Grid for displaying website privacy policies to Web users.

The evaluations of the Expandable Grid system for setting file permissions and its associated policy semantics show that the Expandable Grid can greatly improve the speed and accuracy with which policy authors complete tasks compared to traditional policy-authoring interfaces. However, the evaluation of the Expandable Grid system for displaying website privacy policies suggest some limitations of the Grid concept. We conclude that the Expandable Grid is a beneficial promising approach to policy-authoring interface design, but that it must be applied with care and tailored to each domain to which it is applied.

THESIS ABSTRACT: Improving the Deployability of Diamond

Adam Wolbach

Carnegie Mellon University SCS M.S. Thesis, CMU-CS-08-158, September 2008.

This document describes three engineering contributions made to Diamond, a system for discard-based search, to improve its portability and maintainability, and add new functionality. First, core engineering work on Diamond's RPC and content management subsystems improves the system's maintainability. Secondly, a new mechanism supports "scoping" a Diamond search through the use of external metadata sources. Scoping selects a subset of objects to perform content-based search on by executing a query on an external metadata source related to the data. After query execution, the scope is set for all subsequent searches performed by Diamond applications. The final contribution is Kimberley, a system that enables mobile application use by leveraging virtual machine technology. Kimberley separates application state from a base virtual machine by differencing the VM before and after application customization. The much smaller application state can be carried with the user and quickly applied in a mobile setting to provision infrastructure hardware. Experiments confirm that the startup and teardown delays experienced by a Kimberley user are acceptable for mobile usage scenarios.

THESIS PROPOSAL: File System Virtual Appliances

Michael Abd-El-Malek, ECE
June 2008

File system virtual appliances (FSVAs) address a major headache faced by third-party FS developers: OS version compatibility. By packaging their FS implementation in a VM, separate from the VM that runs user applications, they can avoid the need to provide an FS port for every kernel version and OS distribution. A small FS-agnostic proxy, maintained by the core OS developers, connects the FSVA to whatever kernel version the user chooses. Evaluation of prototype FSVA support in Linux, using Xen as the VM platform, demonstrates that this separation can be efficient and maintain desired OS and virtualization features. Experiments with three existing FSs demonstrate that the FSVA architecture can insulate FS implementations from user OS differences that would otherwise require explicit porting changes.

THESIS PROPOSAL: Low-overhead Byzantine Fault-Tolerant Storage

James Hendricks, SCS
January 2008

As distributed storage systems grow in size and importance, they must tolerate faults other than crashes and simple corruptions. Ideally, storage systems should tolerate Byzantine faulty clients or servers, but Byzantine fault-tolerance is often believed to be too expensive to justify in practice. Previous Byzantine fault-tolerant block storage protocols have either relied upon replication, which is inefficient for large blocks of data when tolerating multiple faults, or have utilized erasure codes for efficient storage but required extra bandwidth and computation. This dissertation will present an erasure-coded Byzantine fault-tolerant block storage protocol that is nearly as efficient as protocols that tolerate only crashes. To accomplish this, I will first demonstrate a novel cryptographic primitive that I call homomorphic fingerprinting that

continued on page 14

RECENT PUBLICATIONS

continued from page 7

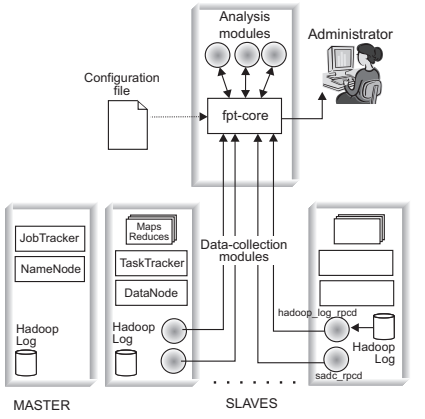
many types of system-model inconsistencies. IRONModel can guide designers to the potential source of the discrepancy, and, if appropriate, can semi-automatically evolve the models to handle unanticipated inputs.

ASDF: Automated, Online Fingerpointing for Hadoop

Bare, Kasick, Kavulya, Marinelli, Pan, Tan, Gandhi & Narasimhan

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-08-104. May 2008.

Localizing performance problems (or fingerpointing) is essential for distributed systems such as Hadoop that support long-running, parallelized, data-intensive computations over a large cluster of nodes. Manual fingerpointing does not scale in such environments because of the number of nodes and the number of performance metrics to be analyzed on each node. ASDF is an automated, online fingerpointing framework that transparently extracts and parses different time-varying data sources (e.g., sysstat, Hadoop logs) on each node, and implements multiple techniques (e.g., log analysis, correlation, clustering) to analyze these data sources jointly or in isolation. We demonstrate ASDF's online fingerpointing for documented performance problems in Hadoop, under different workloads; our results indicate that ASDF incurs an average monitoring overhead of 0.38% of CPU time, and exhibits average online fingerpointing latencies of less than 1 minute with false-positive rates of less than 1%.



Deploying ASDF and its modules for a Hadoop cluster. The fpt-core runs on the ASDF control node.

Data-intensive File Systems for Internet Services: A rose by any other name…

Tantisiriroj, Patil & Gibson

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-08-114, October 2008.

Data-intensive distributed file systems are emerging as a key component of large scale Internet services and cloud computing platforms. They are designed from the ground up and are tuned for specific application workloads. Leading examples, such as the Google File System, Hadoop distributed file system (HDFS) and Amazon S3, are defining this new purpose-built paradigm. It is tempting to classify file systems for large clusters into two disjoint categories, those for Internet services and those for high performance computing.

In this paper we compare and contrast parallel file systems, developed for high performance computing, and data-intensive distributed file systems, developed for Internet services. Using PVFS as a representative for parallel file systems and HDFS as a representative for Internet services file systems, we configure a parallel file system into a data-intensive Internet services stack, Hadoop, and test performance with microbenchmarks and macrobenchmarks running on a 4,000 core Internet services cluster, Yahoo!'s M45.

Once a number of configuration issues such as stripe unit sizes and application buffering sizes are dealt with, issues of replication, data layout and data-guided function shipping are found to be different, but supportable in parallel file systems. Performance of Hadoop applications storing data in an appropriately configured PVFS are comparable to those using a purpose built HDFS.

Proximity Tracking on Time-Evolving Bipartite Graphs

Tong, Papadimitriou, Yu & Faloutsos

Proceedings 2008 SIAM Conference on Data Mining, April 2008, Atlanta, GA.

Given an author-conference network that evolves over time, which are the conferences that a given author is most closely related with, and how do they change over time? Large time-evolving bipartite graphs appear in many settings, such as social networks, co-citations, market-basket analysis, and collaborative filtering. Our goal is to monitor (i) the centrality of an individual node (e.g., who are the most important authors?); and (ii) the proximity of two nodes or sets of nodes (e.g., who are the most important authors with respect to a particular conference?) Moreover, we want to do this efficiently and incrementally, and to provide "any-time" answers. We propose pTrack and cTrack, which are based on random walk with restart, and use powerful matrix tools. Experiments on real data show that our methods are effective and efficient: the mining results agree with intuition; and we achieve up to 15–176 times speed-up, without any quality loss.

Expandable Grids for Visualizing and Authoring Computer Security Policies

Reeder, Bauer, L. Cranor, Reiter, Bacon, How & Strong

The 26th Annual CHI Conference on Human Factors in Computing Systems (CHI 2008). April 5-10, 2008 in Florence, Italy.

We introduce the Expandable Grid, a

continued on page 13

continued from page 12

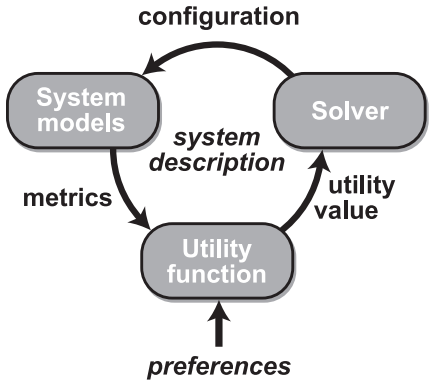
novel interaction technique for creating, editing, and viewing many types of security policies. Security policies, such as file permissions policies, are traditionally displayed and edited in user interfaces based on a list of rules, each of which can only be viewed or edited in isolation. These list-of-rules interfaces cause problems for users when rules interact, because the interfaces have no means of conveying the interactions to users. Instead, users are left to figure out these rule interactions themselves. An Expandable Grid is an interactive matrix visualization designed to address the problems that list-of-rules interfaces have in conveying policies to users. This paper describes the Expandable Grid concept, shows a system using an Expandable Grid for setting file permissions in the Microsoft Windows XP operating system, and gives results of a user study involving 36 participants in which the Expandable Grid approach vastly outperformed the native Windows XP file-permissions interface on a broad range of policy-authoring tasks.

Using Utility to Provision Storage Systems

Strunk, Thereska, Faloutsos & Ganger

6th USENIX Conference on File and Storage Technologies (FAST '08). Feb. 26-29, 2008. San Jose, CA.

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



Overview of a utility-based provisioning tool. The solver produces candidate system configurations. The system models annotate the configurations with system, workload, and dataset metrics. The utility function uses the administrator's preferences to rank the annotated configurations by assigning a utility value to each.

File System Virtual Appliances: Third-party File System Implementations without the Pain

Abd-El-Malek, Wachs, Cipar, Ganger, Gibson & Reiter

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-08-106, May 2008.

File system virtual appliances (FSVAs) address a major headache faced by third-party FS developers: OS version compatibility. By packaging their FS implementation in a VM, separate from the VM that runs user applications, they can avoid the need to provide an FS port for every kernel version and OS distribution. A small FS-agnostic proxy, maintained by the core OS developers, connects the FSVA to whatever kernel version the user chooses. Evaluation of prototype FSVA support in Linux, using Xen as the VM platform, demonstrates that this separation can be efficient and maintain desired OS and virtualization features. Using three existing file systems and a cooperative caching extension as a case study, we demonstrate that the FSVA architecture can insulate FS implementations from user OS differences that would otherwise require explicit porting changes.

Perspective: Semantic Data Management for the Home

Salmon, Schlosser, L. Cranor & Ganger

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-08-105, May 2008.

Perspective uses a new semantic file-system construct, the view, to simplify management of distributed storage in the home. A view is a semantic description of a set of files, specified as a query on file attributes. In Perspective, users can identify and control the files stored on a given device by examining and modifying the views associated with it. This approach allows them to reason about what is where in the same way (semantic naming) as they navigate their digital content. Thus, in serving as their own administrators, users do not have to deal with a second data organization scheme (hierarchical naming) to perform replica management tasks, such as specifying redundancy to provide reliability and data partitioning to address device capacity exhaustion. A set of extensive user studies confirm the difficulties created by current approaches and the efficacy of view-based data management.

continued on page 15

continued from page 11

can be used to efficiently verify the encoding of distributed erasure-coded data, avoiding some of the overheads required in previous protocols. I will also employ novel mechanisms to minimize computation and keep the number of servers minimal.

Unlike consensus protocols, which inherently require more servers to tolerate Byzantine rather than crash faults (a significant barrier to adoption), erasure-coded storage protocols can tolerate Byzantine faults with the same number of servers used to tolerate crashes. Given an erasure code where m fragments are required to reconstruct a block, tolerating f crash faults or Byzantine faults in an asynchronous environment requires writing fragments to m+f servers (assuming m>f) out of m+2f total servers. Real-world storage systems have recently begun to tolerate faults other than crashes, but it is unclear which faults such systems should tolerate. My dissertation may allow designers of such systems to choose Byzantine fault tolerance by default by demonstrating Byzantine fault tolerance without much overhead.

THESIS PROPOSAL: Dependency-Agnostic Online Upgrade in Distributed Systems

Tudor Dumitraş, ECE
July 2008

Online software-upgrades are unavoidable in enterprise systems. For example, business reasons sometimes mandate switching vendors; responding to customer expectations and conforming with government regulations can require new functionality. Moreover, many enterprises can no longer afford to incur the high cost of downtime and must perform such upgrades online, without stopping their systems. Most online-upgrade techniques developed over the past 40 years rely on tracking the dependencies among the components of the system-under-upgrade in order to ensure the correctness of the system, both during and after the upgrade. Today, the benefits of dependency-tracking are reaching their limit due to the increasing complexity of configuration dependencies and to the presence of dynamic dependencies that cannot always be discovered automatically. These fundamental limitations of dependency-tracking lead to frequent upgrade failures. A 2007 survey of 50 system administrators from multiple countries (82% of whom had more than five years of experience) identified broken dependencies and altered system-behavior as the leading causes of upgrade failure, followed by bugs in the new version, incompatibility with legacy configurations and improper packaging of the new components to be deployed. According to the survey, the average upgrade-failure rate was 8.6%, with some administrators reporting that up to 50% of upgrades had failed in their respective installations.

To improve the dependability of online software-upgrades, I propose removing the leading cause of upgrade failures — broken dependencies — by presenting an alternative to dependency-tracking. While relying on knowledge of the planned changes in data-formats and observable system-behavior between the old and new versions, this approach treats the system-under-upgrade as a black box and is, by design, guaranteed not to modify the distributed dependencies within this system. The key to achieving such a dependency-agnostic upgrade is isolating the new version of the system from the old version, to avoid sharing dependencies. I enforce the dependency-isolation by installing the new version in a parallel universe — a logically distinct collection of resources, realized either using different hardware or through virtualization — that is isolated from the universe of the old version. With this approach, a complex distributed-system upgrade can be performed as an atomic action. While it cannot prevent all possible configuration errors or upgrade-failures due to external dependencies, this approach eliminates the internal single-points-of-failure for dependency-breaking faults. Experiments conducted with a prototype implementation indicate that the response time of the system-under-upgrade is not affected during the upgrade process. More specifically, using this prototype to upgrade a three-tier web application (RUBiS) reduces the risk of downtime due to broken dependencies by over 60%, compared with two widely-used alternative techniques for online upgrades in distributed systems.

THESIS PROPOSAL Managing Multi-core Resources in Database Engines

Ryan Johnson, SCS
June 2008

Multi-core computing causes fundamental changes in the hardware landscape with implications for all layers of the software stack. In the past, each processor generation brought significant improvements in single-thread performance, allowing existing software to benefit directly from Moore's Law. Recently, however, power constraints and diminishing returns have pushed chip manufacturers away from aggressive single-thread architecture designs. Designers now use growing transistor budgets to increase the number of hardware contexts available per chip. For the foreseeable future, Moore's Law provides software with twice as many cores each processor generation, with only modest improvements in single-thread performance. In order to achieve peak performance with these new architectures, software must provide abundant parallelism to keep an exponentially growing number of cores busy.

Software faces a second major challenge from chip multiprocessors. In contrast with the traditional shared-nothing cluster, each node of the

continued on page 15

AWARDS & OTHER PDL NEWS

continued from page 9

priority research areas: Command Control Communications, Computers, Intelligence, Surveillance and Reconnaissance. Congratulations to Carlos Guestrin on receiving an award to research "Novel Computational Paradigm for Integration of Uncertain Information in Adversarial Activity Recognition."

April 2008

Best Paper Award at the SIAM Data Mining 2008 Conference

Hanghang Tong (CMU), Spiros Papadimitriou (IBM; CMU Alumni), Philip Yu (IBM) and Christos Faloutsos (CMU) have received the best paper award for their paper titled "Proximity Tracking on Time-Evolving Bipartite Graphs" at the 2008 SIAM (Society for Industrial and Applied Mathematics) Data Mining Conference. The work focuses on social networks, and specifically on measuring the proximity of nodes, as the networks change over time. With careful design, the proposed methods achieve up to 2 orders of magnitude faster computation over straightforward competitors.

RECENT PUBLICATIONS

continued from page 13

GIGA+ : Scalable Directories for Shared File Systems

Patil, Gibson, Lang & Polte

Proceedings of the 2nd International Petascale Data Storage Workshop (PDSW '07) held during Supercomputing '07. Nov. 11, 2007, Reno, NV.

In this paper we describe a distributed metadata service that achieves high parallelism both in the way it stores the metadata and in the way it accesses the metadata. We will present the design and implementation of GIGA+, a POSIX-compliant directory implementation that can scale capacity (i.e., storing >1012 files) and performance (i.e., handling >100K operations/second). In contrast to several attractive "domain-specific" systems (like Google's BigTable and Amazon's Dynamo) that achieve similar scalability, GIGA+ builds file system directories that maintain UNIX file system semantics like no duplicates, no range queries, and unordered readdir() scans. The core of our design is an indexing technique that partitions a directory over a scalable number of servers in an incremental, load-balanced, and unsynchronized manner. GIGA+ achieves highly parallel growth by allowing the servers to grow their partitions independently, without synchronizing with the rest of the system. GIGA+ tolerates the use of stale partition-to-server mapping at the clients without affecting the correctness of their operations. Our system also handles operational realities like client and server failures, addition and removal of servers, and "request storms" that overload any server. We will show the evaluation results of our GIGA+ prototype implemented in an open-source cluster file system called Parallel Virtual File System (PVFS).

DISSERTATIONS & PROPOSALS

continued from page 14

"cluster on a chip" must share off-chip storage capacity and bandwidth, on-chip cache hierarchies, and even processor pipelines with its neighbors. Application designers must find ways to achieve parallelism without placing undue stress on shared resources. As the degree of sharing increases with the number of cores the importance of cooperation between threads will also grow. This thesis focuses on identifying and addressing the challenges for database engines which arise from this new hardware landscape. Many database engine designs date from an era dominated by I/O and few execution contexts, while modern machines feature huge main memories and hardware support for abundant parallelism. My work will analyze the impact of multi-core computing on database engine performance and develop approaches that extract the full potential from modern architectures. I will show that database engines must provide abundant parallelism while carefully managing shared resources such as on-chip cache hierarchies, then demonstrate ways in which this goal can be achieved. I first identify particular areas where multi-core architectures create different challenges than traditional systems, such as managing extreme parallelism and shared resources. I then explore how to best modify database algorithms and internal database engine design, in order to address these challenges.

continued from page 3

tion, control over unreliable networks and security. A common feature of these systems is the presence of significant communication delays and data loss across the network. From the point of view of control theory, significant delay is equivalent to loss, as data needs to arrive at its destination in time to be used for control. In short, communication and control become tightly coupled such that the two issues cannot be addressed independently. Applications for this technology include environmental monitoring, industrial automation, office and building automation, Supervisory Control and Data Acquisition (SCADA) systems, and the automotive industry.

Bruce Krogh

Bruce H. Krogh is a professor of electrical and computer engineering at Carnegie Mellon University. He received his MS and PhD in Electrical Engineering from the University of Illinois in 1978 and 1982 respectively, and his BS in Mathematics and Physics from Wheaton College in 1975.In the past he has served as an Associate Editor of the IEEE Transactions on Automatic Control and Discrete Event Dynamic Systems: Theory and Applications, and was the founding Editor-in-Chief of the IEEE Transactions on Control Systems Technology. Dr. Krogh is a Distinguished Member of the IEEE Control Systems Society and a Fellow of the IEEE. His current research interests include design and verification of embedded control systems, discrete event and hybrid dynamic systems, and information processing in wireless sensor networks.

Lujo Bauer

Lujo Bauer is a Research Scientist in CyLab and the Electrical and Computer Engineering Department at Carnegie Mellon University. He received his BS in Computer Science from Yale University and his PhD, also in Computer Science, from Princeton University.

Lujo Bauer's research interests are in computer security—he is particularly interested in building usable access-control systems with sound theoretical underpinnings, and generally in narrowing the gap between a formal model and a usable system. Topics that Lujo is actively studying include distributed access control, proof-carrying authorization, program monitors, security automata, and languages for specifying security policies. His current projects include:

- Grey—an experiment to create a universal and highly secure access-control device via software extensions to off-the-shelf "smart phones". Grey builds from formal techniques for proving authorization that assure sound access decisions and that permit virtually unlimited flexibility in the policies that can be implemented.
- Secure Digital Home—a project that explores an architecture, mechanisms, and interfaces for helping users manage access control in the digital home of the future.
- Polymer—a language and system for specifying and enforcing composable run-time security policies on Java programs. This project studies various facets of the theory, design, and implementation of software program monitors and monitor-specification languages.

PDL Workshop and Retreat 2007.