PDL Packet Spring Update

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

PDL CONSORTIUM MEMBERS

American Power Conversion
Cisco Systems
EMC
Google
Hewlett-Packard Labs
Hitachi
IBM
Intel
LSI
Microsoft Research
NetApp
Oracle
Seagate Technology
Symantec
VMware

CONTENTS

THE PDL PACKET

EDITOR
Joan Digney

CONTACTS
Greg Ganger
PDL Director

Bill Courtright
PDL Executive Director

Karen Lindenfelser
PDL Business Administrator

The Parallel Data Laboratory
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213-3891

TEL 412-268-6716

FAX 412-268-3010

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

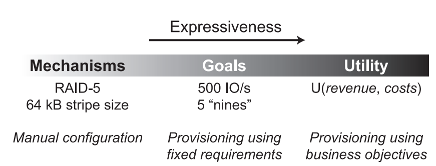
SELECTED RECENT PUBLICATIONS

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.



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

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.

continued on page 2

PDL NEWS & AWARDS

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 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, one of the top data mining conferences. 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. Congratulations, Hanghang, Spiros and Christos!

January 2008

Evan Hoke a Finalist for CRA Outstanding Undergraduate Award

Congratulations to Evan Hoke who was nominated as a finalist for the Computing Research Association

continued on page 6

RECENT PUBLICATIONS

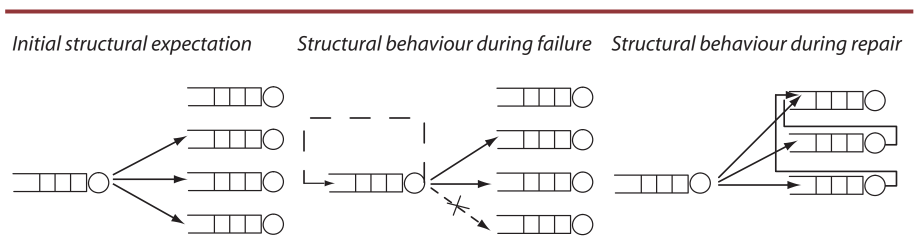
continued from page 2

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 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.



Structural deviations during failure and repair for 3-way replication. The initial structural expectation for write requests indicates that a write should contact three storage-nodes. If one of the storage-node fails, one of the writes times out and reenters the sending service center. During repair, a spare storage-node takes over from the failed one and is populated by the two remaining storage-nodes. Three graphs can be created and contrasted for each of these scenarios (nodes in such graphs would be entry and exit points from the service centers above and edges would represent the service center's response times).

continued on page 3

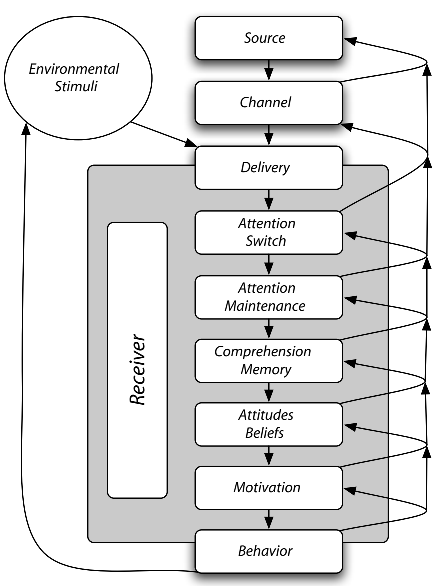
continued from page 2

You've Been Warned: An Empirical Study of the Effectiveness of Web Browser Phishing Warnings

Egelman, L. Cranor & Hong

The 26th Annual CHI Conference on Human Factors in Computing Systems (CHI 2008). April 5-10, 2008 in Florence, Italy. (CHI2008 Honorable Mention Paper)

Many popular web browsers now include active phishing warnings since research has shown that passive warnings are often ignored. In this laboratory study we examine the effectiveness of these warnings and examine if, how, and why they fail users. We simulated a spear phishing attack to expose users to browser warnings. We found that 97% of our sixty participants fell for at least one of the phishing messages that we sent them. However, we also found that when presented with the active warnings, 79% of participants heeded them, which was not the case for the passive warning that we tested—where only one participant heeded the warnings. Using a model from the warning sciences we analyzed how users perceive warning messages and offer suggestions for creating more effective phishing warnings.



The different phases of the Communication-Human Information Processing Model (C-HIP) for structuring warning research.

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.

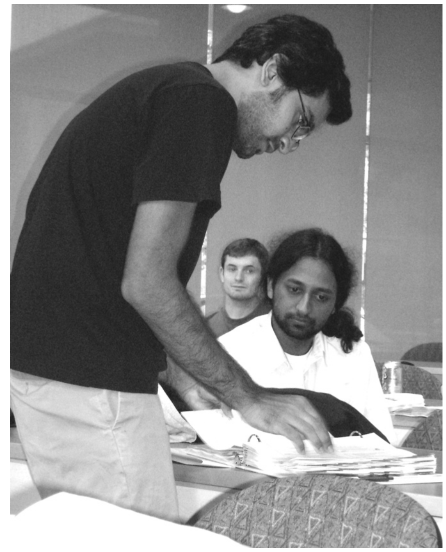
TCP Incast plagues scalable cluster-based storage built atop standard TCP/IP-over-Ethernet, often resulting in much lower client read bandwidth than can be provided by the available network links. This paper reviews the Incast problem and discusses potential application-level approaches to avoiding it.

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 novel interaction technique for creating, editing, and viewing many types of security policies. Security policies, such as file permissions policies, have traditionally been 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 multiple rules interact, because the interfaces have no means of conveying the interactions amongst rules 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.



Raja Sambasivan and Amar Phanishayee prepare resource materials for the 2007 PDL Workshop & Retreat.

Proximity Tracking on Time-Evolving Bipartite Graphs

Tong, Papadimitriou, Yu & Faloutsos

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

continued on page 4

continued from page 3

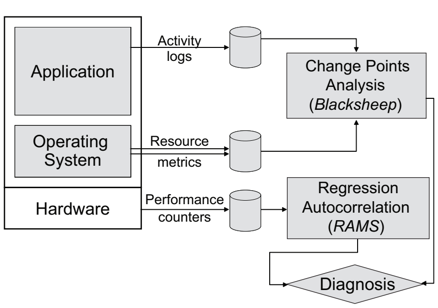
is to monitor (i) the centrality of an individual node (e.g., who are the most important authors?); and (ii) the proximity of two nodes or sets of nodes (e.g., who are the most important authors with respect to a particular conference?) Moreover, we want to do this efficiently and incrementally, and 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.

RAMS and BlackSheep: Inferring White-box Application Behavior using Black-box Techniques

Tan

CS Honors Thesis and Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-08-103.

A significant challenge in developing automated problem-diagnosis tools for distributed systems is the ability of these tools to differentiate between changes in system behavior due to workload changes from those due to faults. To address this challenge, current, typically white-box, techniques extract semantically-rich knowledge about the target application through fairly invasive, high-overhead instrumentation. We propose and explore two scalable, low-overhead, non-invasive techniques to infer semantics about target distributed systems, in a black-box manner, to facilitate problem diagnosis. RAMS applies statistical analysis on hardware performance counters to predict whether a given node in a distributed system is faulty, while BlackSheep corroborates multiple system metrics with application-level logs to determine whether a given node is faulty. In addition, we have developed and demonstrated a novel technique to extract, from existing application-level logs, semantically-rich behavior that is immediately amenable to analysis and synthesis with other numerical, black-box metrics. We have evaluated the efficacy of RAMS and BlackSheep in diagnosing real-world problems in the Hadoop distributed parallel programming system.



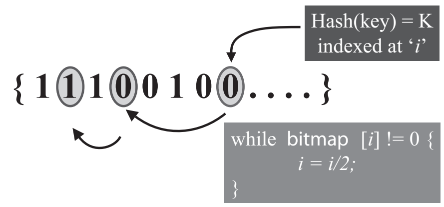
Overview of the *RAMS* and *BlackSheep* techniques for intra-node diagnosis by synthesizing multiple data sources.

Matthew Wachs discusses his research on "Argon: Performance Insulation for Shared Storage Servers" at the 2007 PDL Workshop & Retreat.

GIGA+ : Scalable Directories for Shared File Systems

Patil, Gibson, Lang & Polte

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



This figure shows how we use the BITMAP representation to lookup the presence or absence of a partition on any server. A bit-value of "1" indicates the presence of a partition on a server, and value "0" indicates the absence of the partition on a server. If the bit-value of "0", GIGA+ indexing techniques halves the index and checks the status of the parent partition.

Demand for scalable storage I/O continues to grow rapidly as more applications begin to harness the parallelism provided by massive compute clusters. Much of the research in storage systems has focused on improving the scale and performance of the "data-path". Large-scale file systems do a good job at handling large files by scaling the storage bandwidth by either striping or sharing I/O resources across many servers or disks. Some parallel file systems, like IBM's GPFS and Sun's LustreFS, enable highly concurrent access to the file data using sophisticated locking disciplines. While today's storage systems have been on pace to scale-up the data-path, the same cannot be said about scaling file metadata operations, which is becoming a growing concern for file system vendors and users.

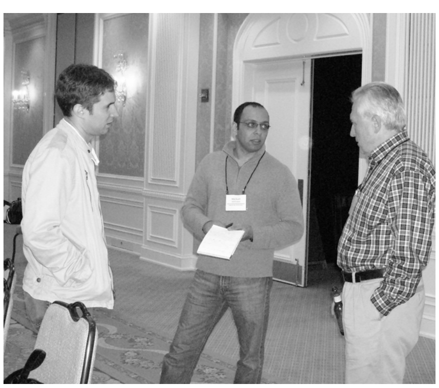
Two trends motivate the need for scalable metadata services in shared file systems. First, there's a burgeoning set of applications that demand a scalable and fast metadata service, like large distributed directories. These applications use the file system as a fast, lightweight "database" by creating large number of small files at high speeds. Such workloads are often seen in scientific computing applications and Internet services. Furthermore, by running on large clusters, these ap-

continued on page 5

continued from page 4

plications end up executing on thousands of concurrent threads (and this number will grow with the adoption of multi-core machines). This increasing application-level parallelism will impose additional burdens on the underlying metadata service. Second, most file systems store metadata on a single metadata server (MDS), thus limiting the overall scalability of the system. Few storage solutions use multiple MDSs either for additional capacity or for fault tolerance. However, additional MDSs can scale performance only if the file system uses them in a distributed and concurrent manner.

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).



Garth Goodson, PDL Alum, now with NetApp (l), Michael Abd-El-Malek and David Ford of NetApp (r) discuss PDL research directions at the retreat.

John Strunk delivers his final Retreat talk on "Using Utility for Storage Provisioning and Tuning." John is now working with NetApp as a member of their Advanced Technology Group.

A User Study of Policy Creation in a Flexible Access-Control System

Bauer, L. Cranor, Reeder, Reiter & Vaniea

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

Significant effort has been invested in developing expressive and flexible access-control languages and systems. However, little has been done to evaluate these systems in practical situations with real users, and few attempts have been made to discover and analyze the access-control policies that users actually want to implement. We report on a user study in which we derive the ideal access policies desired by a group of users for physical security in an office environment. We compare these ideal policies to the policies the users actually implemented with keys and with a smartphone-based distributed access-control system. We develop a methodology that allows us to show quantitatively that the smartphone system allowed our users to implement their ideal policies more accurately and securely than they could with keys, and we describe where each system fell short.

Ideal Access Conditions
I1. True (can access anytime)
I2. Logged
I3. Owner notified
I4. Owner gives real-time approval
I5. Owner gives real-time approval and witness present
I6. Trusted person gives real-time approval and is present
I7. False (no access)

Physical Key Access Conditions
K1. True (has a key)
K2. Ask trusted person with key access
K3. Know location of hidden key
K4. Ask owner who contacts witness
K5. False (no access)

Grey Access Conditions
G1. True (has Grey access)
G2. Ask trusted person with Grey access
G3. Ask owner via Grey
G4. Ask owner who contacts witness
G5. False (no access)

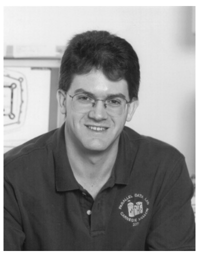
Conditions for access rules in ideal policies, as well as in actual policies implemented with physical keys or Grey, a distributed access-control system that uses off-the-shelf smartphones to allow users to access and manage resources.

continued from page 2

(CRA) Outstanding Undergraduate Award for 2008. The CRA award is extremely competitive and prestigious, with fierce competition from the top undergraduates of all the schools in the nation. Evan has worked on the InteMon and SPIRIT projects. His advisor is Christos Faloutsos.

December 2007

Greg Ganger Recognized by ACM as a Distinguished Engineer



ACM (the Association for Computing Machinery) has named 20 of its members as recipients of a recently created recognition program for their contributions to both the practical and theoretical aspects of computing and information technology. The new ACM Distinguished Members include computer scientists and engineers from some of the world's leading corporations, research labs, and universities who made significant advances in technology that are having lasting impacts on the lives of people across the globe.

"These prominent scientists and engineers have contributed breakthroughs in computing that drive the technologies which benefit our world," said Stuart Feldman, president of ACM. "Their computing innovations address problems in virtually every industry, and make possible advances in communications, health care, finance, entertainment, environmental control, computer security, and many other real life applications. We are proud to recognize these dedicated men and women and to raise their profile in the computing community."

Greg Ganger is one of 12 recipients who conducts his research at a university and received the honor in recognition of his work in data file and storage systems. Eight other recipients are from the industrial sector.

Michael Stroucken, Sr. Research Programmer with the PDL, discusses the DCO with James Nunez of Los Alamos National Laboratory at a PDL Retreat poster session.

For more information about the selection criteria and a complete list of 2007 Distinguished Members and their citation, click on http://distinguished.acm.org.

-- ACM Press Release, Dec. 5, 2007

December 2007

Carlos Guestrin Awarded a 2007 IBM Faculty Fellowship

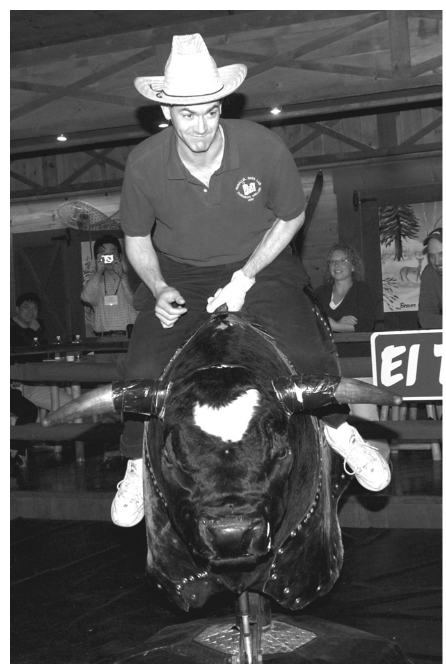
Congratulations to Carlos Guestrin, assistant professor in the Machine Learning Department and in the Computer Science Department at CMU on his receipt of an IBM Faculty Award. The award is part of a competitive worldwide program intended to foster collaboration between researchers at leading universities worldwide and those in IBM research, development and services organizations and promote courseware and curriculum innovation to stimulate growth in disciplines and geographies that are strategic to IBM.

November 2007

Scientists First to Use New Yahoo! Supercomputing Cluster

Yahoo! has launched a new program that will give university scientists an opportunity to advance systems software for distributed computing while using a 4,000-processor supercomputer cluster that the company calls M45. Carnegie Mellon scientists will be the first to take advantage of the M45, which is capable of more than 27 trillion calculations per second and boasts 3 trillion bytes of memory. Carnegie Mellon researchers Garth Gibson and Greg Ganger will instrument the system and evaluate its performance; computer science professors Jamie Callan and Christos Faloutsos will use M45 to solve information retrieval and large-scale graph problems; and faculty members Alexei Efros, Noah Smith and Stephan Vogel will tackle large-scale computer graphics, natural language processing and machine translation problems. "We are excited about collaborating with Yahoo! on system software research, helping to advance the state of the art and creating new research possibilities in a critical area," said Randal E. Bryant, dean of the School of Computer Science.

-- CMU 8 1/2 x 11 News



Our fearless leader takes the first turn on the mechanical bull at Nemacolin Woodlands Resort, site of the annual PDL Workshop and Retreat, to show the rest of us how it is done.