PDL Packet Spring Update

NEWSLETTER ON THE PARALLEL DATA LABORATORY • SPRING 2006

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

CONSORTIUM MEMBERS

American Power Conversion
EMC
EqualLogic
Hewlett-Packard
Hitachi
IBM
Intel
Microsoft Research
Network Appliance
Oracle
Panasas
Seagate
Sun Microsystems
Symantec

CONTENTS

THE PDL PACKET

EDITOR
Joan Digney

CONTACT
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

Informed Data Distribution Selection in a Self-predicting Storage System

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

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

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

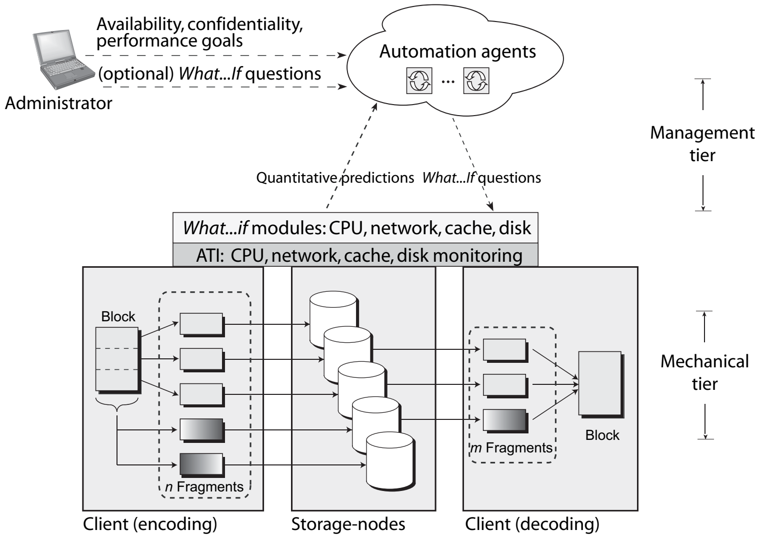
A Large-scale Study of Failures in High-performance-computing Systems

Schroeder & Gibson

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-05-112, December, 2005.

Designing highly dependable systems requires a good understanding of failure characteristics. Unfortunately little raw data on failures in large IT installations is publicly available, due to the confidential nature of this data. This paper analyzes soon-to-be public failure data covering systems at a large high-performance-computing site. The data has been collected over the past 9 years at Los Alamos National Laboratory and includes 23000 failures recorded on more than 20 differ-

... continued on pg. 2



High-level architecture of Ursa Minor. The mechanical tier, on the bottom, services I/O requests for clients. The management tier, on the top, provides automation. It makes use of the self-predicting capabilities of the individual system components to get answers to various *What...if* explorations.

PDL NEWS

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

April 2006
Brandon Salmon Awarded Intel Foundation Ph.D. Fellowship

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

April 2006
Best Demo at ICDE 2006

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

March 2006
Adrian Perrig Keynote Speaker at IPSN

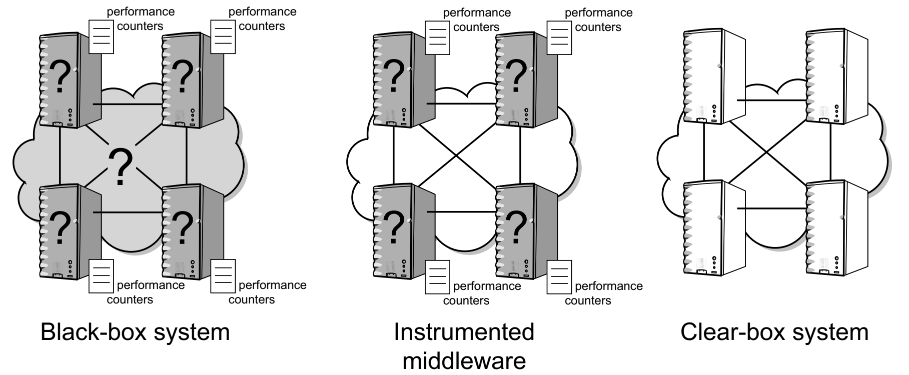
Adrian Perrig, assistant professor of electrical and computer engineering, engineering and public policy, and computer science, delivered the keynote presentation at the International Conference on Information Processing in Sensor Networks (IPSN) this April in Nashville, where he addressed his vision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

… continued on pg. 4

RECENT PUBLICATIONS

… continued from pg. 1

ent systems, mostly large clusters of SMP and NUMA nodes. We study the statistics of the data, including the root cause of failures, the mean time between failures, and the mean time to repair. We find for example that average failure rates differ wildly across systems, ranging from 20-1000 failures per year, and that time between failures is modeled well by a Weibull distribution with decreasing hazard rate. From one system to another, mean repair time varies from less than an hour to more than a day, and repair times are well modeled by a lognormal distribution.



The instrumentation framework depends on the system under consideration. Black-box systems are made of components that work together through well-defined interfaces but are closed-source. Instrumented middleware systems consist of the black box components running on top of a well-known middleware that provides resource multiplexing, management and accounting. Clear-box systems are a term we use for systems whose internals are completely known, either because the system is being built from the start or because its source code is available. Such systems offer the opportunity to have the necessary instrumentation built-in from the start.

Stardust: Tracking Activity in a Distributed Storage System

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

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

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

… continued on pg. 3

… continued from pg. 2

having other systems incorporate such an instrumentation framework.

Causes of Failure in Web Applications

Pertet & Narasimhan

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-05-109, December, 2005.

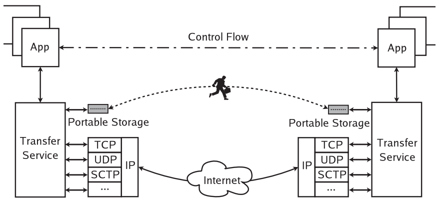
This report investigates the causes and prevalence of failure in Web applications. Data was collected by surveying case studies of system failures and by examining incidents of website outages listed on technology websites such as CNET.com and eweek.com. These studies suggest that software failures and human error account for about 80% of failures. The report also contains an appendix that serves as a quick reference for common failures observed in Web applications. This appendix lists over 40 incidents of real-world site outages, outlining how these failures were detected, the estimated downtime, and the subsequent recovery action.

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

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

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

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



Overview of DOT, a flexible architecture for data transfer.

An Architecture for Internet Data Transfer

Tolia, Kaminsky, Andersen & Patil

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

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

Chuck Cranor, PDL research scientist, and Michael Abd-El-Malek, graduate student, stand ready to discuss PDL research at one of the 2005 Retreat and Workshop poster sessions.

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

Scheduling Speculative Tasks in a Compute Farm

Petrou, Gibson & Ganger

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

Users often behave speculatively, submitting work that initially they do not know is needed. Farm computing often consists of single node speculative tasks issued by, e.g., bioinformaticists comparing DNA sequences and computer graphics artists rendering scenes who wish to reduce their time waiting for needed tasks and the amount they will be charged for unneeded speculation. Existing schedulers are not effective for such behavior. Our 'batchactive' scheduling exploits speculation: users submit explicitly labeled batches of speculative tasks, interactively request outputs when ready to process them, and cancel tasks found not to be needed. Users are encouraged to participate by

… continued on pg. 5

… continued from pg. 2

February 2006
Adrian Perrig Recipient of 2006 Sloan Award

Three CMU 2006 winners of a Sloan Research Fellowship in computer science have been announced: Carlos Guestrin, CALD and CSD, Doug James, CSD and RI, and Adrian Perrig, ECE and CSD. A Sloan Fellowship is a prestigious award intended to enhance the careers of the very best young faculty members in specified fields of science. Currently a total of 116 fellowships are awarded annually in seven fields: chemistry, computational and evolutionary molecular biology, computer science, economics, mathematics, neuroscience, and physics. Only 14 are given in computer science each year so CMU once again shines.

February 2006
Microsoft Fellowship Helps Develop New Security Course

Faculty members Lorrie Cranor, Institute for Software Research International, Jason Hong, Human-Computer Interaction Institute, and Michael Reiter, Electrical and Computer Engineering, have received a 2005 Microsoft Research Trustworthy Computing Curriculum Award to fund the development of a new course on usable privacy and security. The course is offered for the first time this semester in the School of Computer Science (http://cups.cs.cmu.edu/courses/ups.html). It is designed to introduce students to a variety of usability and user-interface problems related to privacy and security and give them experience in designing studies aimed at helping to evaluate usability issues in security and privacy systems.

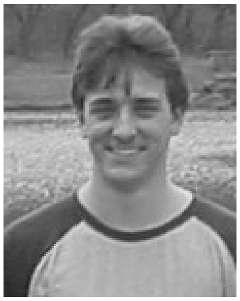
-- CMU 8 1/2 x 11 News

January 2006
Jure Leskovec awarded Microsoft Research Fellowship

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

The competition for these fellowships was extremely high. The award will be one of only 10 specially funded MSR fellowships to be funded by “a major new Microsoft initiative that will be publicly announced in the near future.”

January 2006
James Newsome awarded Microsoft Research Fellowship



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

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

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

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

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

One of academia’s premiere storage systems research centers, the PDL is an interdisciplinary group, bringing together graduate students and faculty mainly from the Computer Science and ECE departments. Both of the winning papers have authors from both departments.

-- ECE News Online

December 2005
Jia-Yu Pan Receives Best Paper at ICDM

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

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

December 2005
Hui Zhang Selected as ACM Fellow

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

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

… continued from pg. 3

a new pricing mechanism charging for only requested tasks no matter what ran. Over a range of simulated user and task characteristics, we show that: batchactive scheduling improves visible response time—a new metric for speculative domains—by at least 2X for 20% of the simulations; batchactive scheduling supports higher billable load at lower visible response time, encouraging adoption by resource providers; and a batchactive policy favoring users who use more of their speculative tasks provides additional performance and resists a denial-of-service.

Eliminating Cross-server Operations in Scalable File Systems

Hendricks, Sinnamohideen, Sambasivan & Ganger

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

Distributed file systems that scale by partitioning files and directories among a collection of servers inevitably encounter crossserver operations. A common example is a RENAME that moves a file from a directory managed by one server to a directory managed by another. Systems that provide the same semantics for cross-server operations as for those that do not span servers traditionally implement dedicated protocols for these rare operations. This paper suggests an alternate approach that exploits the existence of dynamic redistribution functionality (e.g., for load balancing, incorporation of new servers, and so on). When a client request would involve files on multiple servers, the system can redistribute files onto one server and have it service the request. Although such redistribution is more expensive than a dedicated cross-server protocol, the rareness of such operations makes the overall performance impact minimal. Analysis of NFS traces indicates that cross-server operations make up fewer than 0.001% of client requests, and experiments with a prototype implementation show that the performance impact is negligible when such operations make up as much as 0.01% of operations. Thus, when dynamic redistribution functionality exists in the system, cross-server operations can be handled with almost no additional implementation complexity.

Somehow, Bill Courtright, PDL Executive Director, has shanghaied Bob Wolfgang of APC into helping stuff binders in preparation for last fall's PDL Retreat.

Quantifying Interactive User Experience on Thin Clients

Tolia, Andersen & Satyanarayanan

IEEE Computer. March, 2006.

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

Improving Small File Performance in Object-based Storage

Hendricks, Sambasivan, Sinnamohideen & Ganger

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

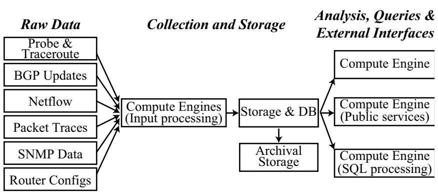
This paper proposes architectural refinements, server-driven metadata prefetching and namespace flattening, for improving the efficiency of small file workloads in object-based storage systems. Server-driven metadata prefetching consists of having the metadata server provide information and capabilities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

Challenges and Opportunities in Internet Data Mining

Andersen & Feamster

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

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



The datapository architecture.

… continued on pg. 6

… continued from pg. 5

Eliminating Cross-server Operations in Scalable File Systems

Hendricks, Sinnamohideen, Sambasivan & Ganger

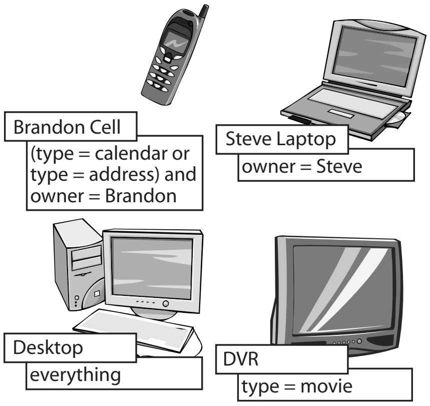
Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-06-105, May 2006.

Distributed file systems that scale by partitioning files and directories among a collection of servers inevitably encounter crossserver operations. A common example is a RENAME that moves a file from a directory managed by one server to a directory managed by another. Systems providing the same semantics for cross-server operations as for those that do not span servers traditionally implement dedicated protocols for these rare operations. We suggest an alternate approach that exploits the existence of dynamic redistribution functionality (e.g., for load balancing, incorporation of new servers, and so on). When a client request would involve files on multiple servers, the system can redistribute files onto one server and have it service the request. Although such redistribution is more expensive than a dedicated cross-server protocol, the rareness of such operations makes the overall performance impact minimal. Analysis of NFS traces indicates that cross-server operations make up fewer than 0.001% of client requests, and experiments with a prototype implementation show that the performance impact is negligible when such operations make up as much as 0.01% of operations. Thus, when dynamic redistribution functionality exists in the system, cross-server operations can be handled with almost no additional implementation complexity.

Towards Efficient Semantic Object Storage for the Home

Salmon, Schlosser & Ganger

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



Each device has a customized view. The cell phone is only interested in "calendar" and "address" objects belonging to Brandon. The laptop is interested in all objects owned by Steve, the house desktop is interested in all data, while the DVR is interested in all "movie" objects.

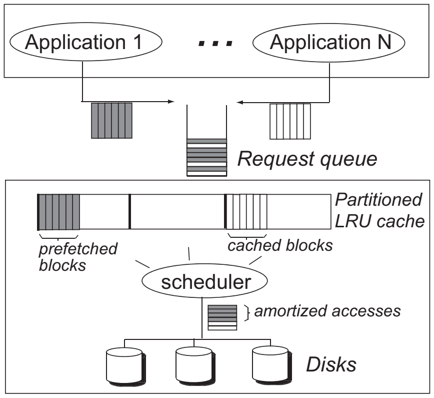
The home provides a new and challenging environment for data management. Devices in the home are extremely heterogeneous in terms of computational capability, capacity, and usage. Yet, ideally, information would be shared easily across them. Current volume-based filesystems do not provide the flexibility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

Argon: Performance Insulation for Shared Storage Servers

Wachs, Abd-El-Malek, Thereska & Ganger

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

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



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