PDL Packet Fall Update

PDL CONSORTIUM MEMBERS

American Power Conversion
EMC Corporation
Facebook
Google
Hewlett-Packard Labs
Hitachi
IBM
Intel Corporation
LSI
Microsoft Research
NEC Laboratories
NetApp, Inc.
Oracle Corporation
Riverbed Technology
Samsung Information Systems America
Seagate Technology
STEC, Inc.
Symantec Corporation
VMware, Inc.
Yahoo! Labs

CONTENTS

THE PDL PACKET

SELECTED RECENT PUBLICATIONS

To Upgrade or Not to Upgrade: Impact of Online Upgrades across Multiple Administrative Domains

Dumitraş, Tilevich & Narasimhan

ACM Onward! Conference, Oct. 2010, Reno, NV.

Online software upgrades are often plagued by runtime behaviors that are poorly understood and difficult to ascertain. For example, the interactions among multiple versions of the software expose the system to race conditions that can introduce latent errors or data corruption. Moreover, industry trends suggest that online upgrades are currently needed in large-scale enterprise systems, which often span multiple administrative domains (e.g., Web 2.0 applications that rely on AJAX client-side code or systems that lease cloud computing resources). In such systems, the enterprise does not control all the tiers of the system and cannot coordinate the upgrade process, making existing techniques inadequate to prevent mixed-version races. In this paper, we present an analytical framework for impact assessment, which allows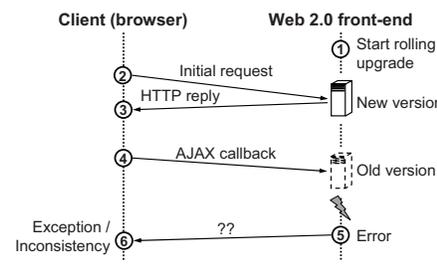 system administrators to directly compare the risk of following an online-upgrade plan with the risk of delaying or canceling the upgrade. We also describe an executable model that implements our formal impact assessment and enables a systematic approach for deciding whether an online upgrade is appropriate. Our model provides a method of last resort for avoiding undesirable program behaviors, in situations where mixed-version races cannot be avoided through other technical means.

Token Attempt: The Misrepresentation of Website Privacy Policies through the Misuse of P3P Compact Policy Tokens

Leon, L. Cranor, McDonald & McGuire

Cylab Technical Report CMU-CyLab-10-014, September 10, 2010.

Platform for Privacy Preferences (P3P) compact policies (CPs) are a collection of three-character and four-character tokens that summarize a website's privacy policy pertaining to cookies. User agents, including Microsoft's Internet Explorer (IE) web browser, use CPs to evaluate websites' data collection practices and allow, reject, or modify cookies based on sites' privacy practices. CPs can provide a technical means to enforce users' privacy preferences if CPs accurately reflect websites' practices. Through automated analysis we can identify CPs that are erroneous due to syntax errors or semantic conflicts. We collected CPs from 33,139 websites and detected errors in 11,176 of them,

Mixed version race.

October 2010

Satya Wins 2010 SIGMOBILE Award

Congratulations to M. Satyanarayanan (Satya), who has been honored with the SIGMOBILE 2010 Outstanding Contributions Award "for his pioneering a wide spectrum of technologies in support of disconnected and weakly connected mobile clients." He joins an illustrious group of previous winners: http://www.sigmobile.org/awards/oca.html The SIGMOBILE Outstanding Contribution Award is given for significant and lasting contributions to the research on mobile computing and communications and wireless networking.

September 2010

NSF Project to Make Internet Secure and Smart

Carnegie Mellon Computer Science and Electrical and Computer Engineering Professor Peter Steenkiste is leading a three-year, $7.1 million effort sponsored by the National Science Foundation (NSF) to develop a next-generation network architecture that fixes security and reliability deficiencies now threatening the viability of the Internet. The eXpressive Internet Architecture (XIA) Project, one of four new projects funded through the Future Internet Architecture Program of the NSF's Computer and Information Science and Engineering (CISE) Directorate, will include intrinsic security features so that users can be assured that the websites they access and the documents they download are legitimate.

In addition to Steenkiste who is the principal investigator, other CMU faculty members working on the project, including several members of the PDL, are David Andersen, David Feinberg, Srinivasan Seshan and Hui Zhang of the Computer Science Department; CyLab technical director Adrian Perrig; Sara Kiesler of the Human-Computer Interaction Institute; and Jon Peha and Marvin Sirbu of the Engineering and Public Policy Department.

--CMU 8.5x11 News Sept. 2, 2010

September 2010

Christos Faloutsos Wins SIGCOMM 2010 Test of Time Award

Congratulations to Christos and his co-authors (brothers Michalis Faloutsos and Petros Faloutsos) for winning the SIGCOMM Test of Time award for their paper "On the Power Law Relationships of the Internet Topology." The ACM SIGCOMM Test of Time Award recognizes papers published 10 to 12 years in the past in Computer Communication Review or any SIGCOMM sponsored or co-sponsored conference that is deemed to be an outstanding paper whose contents are still a vibrant and useful contribution today. Here is a link to the abstract of the 1999 paper.

July 2010

Best Paper Award at PAKDD 2010

School of Computer Science Ph.D. students Leman Akoglu and Mary McGlohon received the "best paper" award in late June at the 14th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2010). The paper, "OddBall: Spotting Anomalies in Weighted Graphs," by Akoglu, McGlohon and Professor Christos Faloutsos, gives fast algorithms to spot strange nodes in large social networks. The paper was selected among 412 submissions, and 42 accepted papers.

--CMU 8.5x11 News July 15, 2010

July 2010

Gregory Ganger Earns 2010 HP Innovation Research Award

CMU's Gregory Ganger was one of more than 60 recipients worldwide to receive the 2010 HP Innovation 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 (PDL) at Carnegie Mellon, will collaborate with HP labs on a research initiative focused on cloud computing issues. This is Ganger's second HP Innovation Award. He received his first HP Innovation Award in 2008 for research involving scalable and self-managing data storage systems.

HP reviewed more than 300 submissions from individuals at 202 universities in 36 countries. Ganger said the award will deepen and strengthen the PDL's long-standing ties with HP and with outstanding researchers globally. The PDL continues to work on solutions to critical problems of storage system design, implementation and evaluation.

"This is a wonderful award for Greg and his team because it recognizes the innovative, collaborative research excellence so endemic to the Parallel Data Lab," said Mark S. Kamlet, executive vice president and provost at Carnegie Mellon. "We applaud their dedication and energy in streamlining ubiquitous cloud computing use."

"The annual HP Labs Innovation Program is an ideal platform for HP to initiate highly innovative projects with leading researchers in universities

including 134 TRUSTe-certified websites and 21 of the top 100 most-visited sites. Our work identifies potentially misleading practices by web administrators, as well as common accidental mistakes. We found thousands of sites using identical invalid CPs that had been recommended as workarounds for IE cookie blocking. Other sites had CPs with typos in their tokens, or other errors. 98% of invalid CPs resulted in cookies remaining unblocked by IE under it's default cookie settings. It appears that large numbers of websites that use CPs are misrepresenting their privacy practices, thus misleading users and rendering privacy protection tools ineffective. Unless regulators use their authority to take action against companies that provide erroneous machine-readable policies, users will be unable to rely on these policies.

Behavior-Based Problem Localization for Parallel File Systems

Kasick, Gandhi & Narasimhan

HotDep '10. October 3, 2010, Vancouver, BC, Canada.

We present a behavior-based problem-diagnosis approach for PVFS that analyzes a novel source of instrumentation—CPU instruction-pointer samples and function-call traces—to localize the faulty server and to enable root-cause analysis of the resource at fault. We validate our approach by injecting realistic storage and network problems into three different workloads (dd, IO-zone, and PostMark) on a PVFS cluster.

Parsimonious Linear Fingerprinting for Time Series

Li, Prakash & Faloutsos

Proceedings of the VLDB Endowment, Vol. 3, No. 1, September 2010.

We study the problem of mining and summarizing multiple time series effectively and efficiently. We propose PLiF, a novel method to discover essential characteristics ("fingerprints"), by exploiting the joint dynamics in numerical sequences. Our ̄fingerprinting method has the following benefits: (a) it leads to interpretable features; (b) it is versatile: PLiF enables numerous mining tasks, including clustering, compression, visualization, forecasting, and segmentation, matching top competitors in each task; and (c) it is fast and scalable, with linear complexity on the length of the sequences.
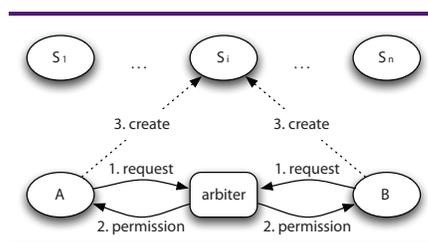
We did experiments on both synthetic and real datasets, including human motion capture data (17MB of human motions), sensor data (166 sensors), and network router traffic data (18 million raw updates over 2 years). Despite its generality, PLiF outperforms the top clustering methods on clustering; the top compression methods on compression (3 times better reconstruction error, for the same compression ratio); it gives meaningful visualization and at the same time, enjoys a linear scale-up.

dBug: Systematic Evaluation of Distributed Systems

Simsa, Bryant & Gibson

5th Int. Workshop on Systems Software Verification (SSV'10), co-located with 9th USENIX Symp. On Operating Systems Design and Implementation (OSDI'10), Vancouver BC, October 2010.

This paper presents the design, implementation and evaluation of "dBug" – a tool that leverages manual instrumentation for systematic evaluation of distributed and concurrent systems. Specifically, for a given distributed concurrent system, its initial state and a workload, the dBug tool systematically explores possible orders in which concurrent events triggered by the workload can happen. Further, dBug optionally uses the partial order reduction mechanism to avoid exploration of equivalent orders. Provided with a correctness check, the dBug tool is able

Steps taken to send a message: 1) An agent requests permission from the arbiter, 2) The arbiter grants the permission, 3) The agent sends the message.

to verify that all possible serializations of a given concurrent workload execute correctly. Upon encountering an error, the tool produces a trace that can be replayed to investigate the error.

We applied the dBug tool to two distributed systems – the Parallel Virtual File System (PVFS) implemented in C and the FAWN-based key-value storage (FAWN-KV) implemented in C++. In particular, we integrated both systems with dBug to expose the non-determinism due to concurrency. This mechanism was used to verify that the result of concurrent execution of a number of basic operations from a fixed initial state meets the high-level specification of PVFS and FAWN-KV. The experimental evidence shows that the dBug tool is capable of systematically exploring behaviors of a distributed system in a modular, practical, and effective manner.

OddBall: Spotting Anomalies in Weighted Graphs

Akoglu, McGlohon & Faloutsos

PAKDD 2010, Hyderabad, India, 21-24 June 2010. Best Paper Award.

Given a large, weighted graph, how can we find anomalies? Which rules should be violated, before we label a node as an anomaly? We propose the OddBall algorithm, to find such nodes. The contributions are the following: (a) we discover several new rules (power laws) in density, weights, ranks and eigenvalues that seem to govern the

RECENT PUBLICATIONS

so-called "neighborhood sub-graphs" and we show how to use these rules for anomaly detection; (b) we carefully choose features, and design OddBall, so that it is scalable and it can work unsupervised (no user-defined constants) and (c) we report experiments on many real graphs with up to 1.6 million nodes, where OddBall indeed spots unusual nodes that agree with intuition.

DiskReduce: Replication as a Prelude to Erasure Coding in Data-Intensive Scalable Computing

Fan, Tantisiriroj, Xiao & Gibson

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-10-111.

The first generation of Data-Intensive Scalable Computing file systems employed only replication for reliability, typically delivering users with only about a third of the storage capacity of the raw disks. This paper presents DiskReduce, a framework for integrating RAID into these replicated storage systems to lower storage capacity overhead, for example, from 200% to 25% when triplicated data is dynamically replaced with 8+2 RAID 6 encoding. Based on data collected from Yahoo! and Facebook, we model the capacity effectiveness of simple and not so simple strategies for grouping data blocks into RAID sets; the most capacity efficient strategies suffer from "small write penalties" we ameliorate with deferred deletion. Because replication is intuitively stronger than common RAID erasure codes, we construct a data reliability model, apply it to scales similar to our collected data and explore the tradeoff between capacity and reliability. Failure detection effectiveness turns out to be a key parameter, either limiting reliability or enabling less invasive background reconstruction to not penalize reliability, depending on your perspective. Finally, an implementation of DiskRe-

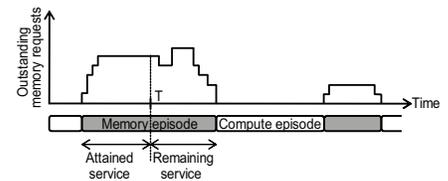duce has been built and submitted into the Apache Hadoop project.

ATLAS: A Scalable and High-Performance Scheduling Algorithm for Multiple Memory Controllers

Kim, Han, Mutlu & Harchol-Balter

Proceedings of the 16th International Symposium on High-Performance Computer Architecture (HPCA), Bangalore, India, January 2010.

Modern chip multiprocessor (CMP) systems employ multiple memory controllers to control access to main memory. The scheduling algorithm employed by these memory controllers has a significant effect on system throughput, so choosing an efficient scheduling algorithm is important. The scheduling algorithm also needs to be scalable – as the number of cores increases, the number of memory controllers shared by the cores should also increase to provide sufficient bandwidth to feed the cores. Unfortunately, previous memory scheduling algorithms are inefficient with respect to system throughput and/or are designed for a single memory controller and do not scale well to multiple memory controllers, requiring significant fine-grained coordination among controllers.

This paper proposes ATLAS (Adaptive per-Thread Least- Attained-Service memory scheduling), a fundamentally new memory scheduling technique that improves system throughput without requiring significant coordination among memory controllers. The key idea is to periodically order threads based on the service they have attained from the memory controllers so far, and prioritize those threads that have attained the least service over others in each period. The idea of favoring threads with least-attained-service is borrowed from the queueing theory literature, where, in the context of a single-server queue it is known that least-attained-service optimally schedules jobs, assuming a Pareto (or

Memory vs. compute episodes in a thread's execution time.

any decreasing hazard rate) workload distribution. After verifying that our workloads have this characteristic, we show that our implementation of least-attained service thread prioritization reduces the time the cores spend stalling and significantly improves system throughput. Furthermore, since the periods over which we accumulate the attained service are long, the controllers coordinate very infrequently to form the ordering of threads, thereby making ATLAS scalable to many controllers. We evaluate ATLAS on a wide variety of multiprogrammed SPEC 2006 workloads and systems with 4-32 cores and 1-16 memory controllers, and compare its performance to five previously proposed scheduling algorithms. Averaged over 32 workloads on a 24-core system with 4 controllers, ATLAS improves instruction throughput by 10.8%, and system throughput by 8.4%, compared to PAR-BS, the best previous CMP memory scheduling algorithm. ATLAS's performance benefit increases as the number of cores increases.

Phase Change Technology and the Future of Main Memory

Lee, Zhou, Yang, Zhang, Zhao, Ipek, Mutlu & Burger

IEEE Micro, Special Issue: Micro's Top Picks from 2009 Computer Architecture Conferences (MICRO TOP PICKS), Vol. 30, No. 1, p. 60-70, January/February 2010.

Phase-change memory may enable continued scaling of main memories, but PCM has higher access latencies, incurs higher power costs, and wears

out more quickly than DRAM. This article discusses how to mitigate these limitations through buffer sizing, row caching, write reduction, and wear leveling, to make PCM a viable DRAM alternative for scalable main memories.

Speeding Up Finite Element Wave Propagation for Large-Scale Earthquake Simulations

Taborda, López, Karaoglu, Urbanic & Bielak

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-10-109.

This paper describes the implementation and performance of a new approach to finite element earthquake simulations that represents a speedup factor of 3x in the total solving time employed by Hercules--the octree-based earthquake simulator developed by the Quake Group at Carnegie Mellon University. This gain derives from applying an efficient method for computing the stiffness contribution at the core of the solving algorithm for the discretized equations of motion. This efficient method is about 5 times faster than our previous conventional implementation. We evaluate the performance and scalability of the new implementation through numerical experiments with the 2008 Chino Hills earthquake under various problem sizes and resource conditions on up to 98K CPU cores, obtaining excellent results. These experiments required simulations with up to 11.6 billion mesh elements. The newly obtained efficiency reveals that other areas in Hercules, such as inter-processor communication, waiting time, and additional computing processes become more critical, and that improvements in these areas will result in significant enhancement in overall performance. This latest advance has enormous implications for saving CPU hours and catapults the potential of Hercules to target larger and more realistic problems, taking full advan-

tage of the new generation of petascale supercomputers.

FAWNSort: Energy-efficient Sorting of 10GB

Vasudevan, Tan, Kaminsky, Kozuch, Andersen & Pillai

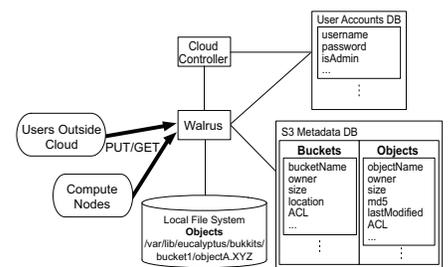Winner of 2010 10GB Joulesort, Daytona and Indy categories. http://sortbenchmark.org/

This document describes our submission for the 2010 10GB JouleSort competition. Our system consists of a machine with a low-power server processor and five flash drives, sorting the 10GB dataset in 21.2 seconds (±0.227s) seconds with an average power of 104.9W (±0.8W). The system sorts the 10GB dataset using only 2228 Joules (±12 J), providing 44884 (±248) sorted records per Joule. Our entry tried to use the most energy-efficient platform we could find that could hold the dataset in memory to enable a one-pass sort. We decided to use a one-pass sort on this hardware over a two-pass sort on more energy efficient hardware (such as Intel Atombased boards) after experimenting with several energy efficient hardware platforms that were unable to address enough memory to hold the 10GB dataset in memory. The lowpower platforms we tested suffered from either a lack of I/O capability or high, relative fixed power costs, both stemming from design decisions made by hardware vendors rather than being informed by fundamental properties of energy and computing.

pWalrus: Towards Better Integration of Parallel File Systems into Cloud Storage

Abe & Gibson

Workshop on Interfaces and Abstractions for Scientific Data Storage (IASDS10), Heraklion, Greece, September 2010.

Amazon S3-style storage is an attractive option for clouds that provides data ac-

Architecture of Walrus (in Eucalyptus 1.6.2).

cess over HTTP/HTTPS. At the same time, parallel file systems are an essential component in privately owned clusters that enable highly scalable dataintensive computing. In this work, we take advantage of both of those storage options, and propose pWalrus, a storage service layer that integrates parallel file systems effectively into cloud storage. Essentially, it exposes the mapping between S3 objects and backing files stored in an underlying parallel file system, and allows users to selectively use the S3 interface and direct access to the files. We describe the architecture of pWalrus, and present preliminary results showing its potential to exploit the performance and scalability of parallel file systems.

Thread Cluster Memory Scheduling: Exploiting Differences in Memory Access Behavior

Kim, Papamichael, Mutlu & Harchol-Balter

Proceedings of the 43rd International Symposium on Microarchitecture (MICRO), Atlanta, GA, December 2010.

In a modern chip-multiprocessor system, memory is a shared resource among multiple concurrently executing threads. The memory scheduling algorithm should resolve memory contention by arbitrating memory access in such a way that competing threads progress at a relatively fast and even pace, resulting in high system throughput and fairness. Previously

proposed memory scheduling algorithms are predominantly optimized for only one of these objectives: no scheduling algorithm provides the best system throughput and best fairness at the same time.

This paper presents a new memory scheduling algorithm that addresses system throughput and fairness separately with the goal of achieving the best of both. The main idea is to divide threads into two separate clusters and employ different memory request scheduling policies in each cluster. Our proposal, Thread Cluster Memory scheduling (TCM), dynamically groups threads with similar memory access behavior into either the latency-sensitive (memory-non-intensive) or the bandwidth-sensitive (memory-intensive) cluster. TCM introduces three major ideas for prioritization: 1) we prioritize the latency-sensitive cluster over the bandwidth-sensitive cluster to improve system throughput; 2) we introduce a "niceness" metric that captures a thread's propensity to interfere with other threads; 3) we use niceness to periodically shuffle the priority order of the threads in the bandwidth-sensitive cluster to provide fair access to each thread in a way that reduces interthread interference. On the one hand, prioritizing memory-non-intensive threads significantly improves system throughput without degrading fairness, because such "light" threads only use a small fraction of the total available memory bandwidth. On the other hand, shuffling the priority order of memory-intensive threads improves fairness because it ensures no thread is disproportionately slowed down or starved.

We evaluate TCM on a wide variety of multiprogrammed workloads and compare its performance to four previously proposed scheduling algorithms, finding that TCM achieves both the best system throughput and fairness. Averaged over 96 workloads on a 24-core system with 4 memory channels, TCM improves system throughput
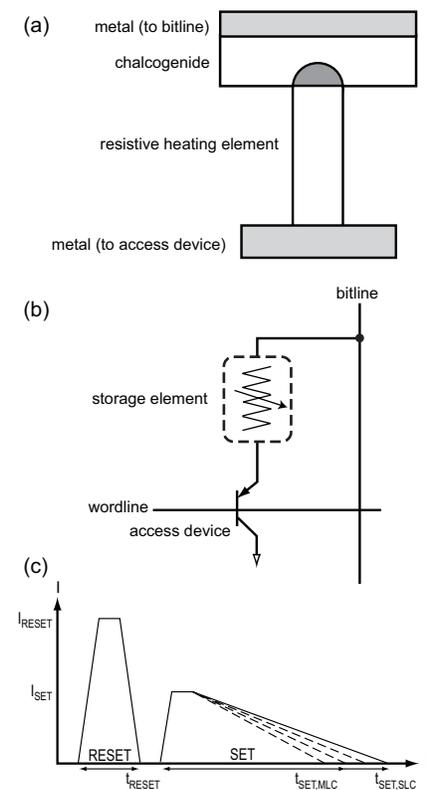
and reduces maximum slowdown by 4.6%/38.6% compared to ATLAS (previous work providing the best system throughput) and 7.6%/4.6% compared to PAR-BS (previous work providing the best fairness).

Phase Change Memory Architecture and the Quest for Scalability

Lee, Ipek, Mutlu & Burger

Communications of the ACM (CACM), Research Highlight, Vol. 53, No. 7, pages 99-106, July 2010.

Memory scaling is in jeopardy as charge storage and sensing mechanisms

Phase change memory. (a) Storage element with heating resistor and chalcogenide between electrodes. (b) Cell structure with storage element and BJT access device. (c) Reset to an amorphous, high resistance state with a high, short current pulse. Set to a crystalline, low resistance state with moderate, long current pulse. Slope of set current ramp down determines the state in MLC.

become less reliable for prevalent memory technologies, such as dynamic random access memory (DRAM). In contrast, phase change memory (PCM) relies on programmable resistances, as well as scalable current and thermal mechanisms. To deploy PCM as a DRAM alternative and to exploit its scalability, PCM must be architected to address relatively long latencies, high energy writes, and finite endurance.

We propose architectural enhancements that address these limitations and make PCM competitive with DRAM. A baseline PCM system is 1.6× slower and requires 2.2× more energy than a DRAM system. Buffer reorganizations reduce this delay and energy gap to 1.2× and 1.0×, using narrow rows to mitigate write energy as well as multiple rows to improve locality and write coalescing. Partial writes mitigate limited memory endurance to provide more than 10 years of lifetime. Process scaling will further reduce PCM energy costs and improve endurance.

Applying Performance Models to Understand Data-intensive Computing Efficiency

Krevat, Shiran, Anderson, Tucek, J. Wylie & Ganger

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-10-108. May 2010.

New programming frameworks for scale-out parallel analysis, such as MapReduce and Hadoop, have become a cornerstone for exploiting large datasets. However, there has been little analysis of how these systems perform relative to the capabilities of the hardware on which they run. This paper describes a simple analytical model that predicts the optimal performance of a parallel dataflow system. The model exposes the inefficiency of popular scale-out systems, which take 3--13× longer to complete jobs than the hardware should allow, even in well-tuned systems used to achieve record-break-

ing benchmark results. To validate the sanity of our model, we present small-scale experiments with Hadoop and a simplified dataflow processing tool called Parallel DataSeries. Parallel DataSeries achieves performance close to the analytic optimal, showing that the model is realistic and that large improvements in the efficiency of parallel analytics are possible.

Diagnosing Performance Changes by Comparing System Behaviours

Sambasivan, Zheng, Krevat, Whitman, Stroucken, Wang, Xu & Ganger

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-10-107. July 2010.

The causes of performance changes in a distributed system often elude even its developers. We develop a new technique for gaining insight into such changes: comparing system behaviours from two executions (e.g., of two system versions or time periods). Building on end-to-end request flow tracing within and across components, algorithms are described for identifying and ranking changes in the flow and/or timing of request processing. The implementation of these algorithms in a tool called Spectroscope is described and evaluated. Five case studies are presented of using Spectroscope to diagnose performance changes in a distributed storage system caused by code changes and configuration modifications, demonstrating the value and efficacy of comparing system behaviours.

DiscFinder: A Data-Intensive Scalable Cluster Finder for Astrophysics

Fu, López, Fink, Ren & Gibson

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-10-104. October 2010.

DiscFinder is a scalable, distributed, data-intensive group finder for analyzing observation and simulation as-

trophysics datasets. Group finding is a form of clustering used in astrophysics for identifying large-scale structures such as galaxies and clusters of galaxies. DiscFinder runs on commodity compute clusters and scales to large datasets with billions of particles. It is designed to operate on datasets that are much larger than the aggregate memory available in the computers where it executes. As a proof-of-concept we have implemented DiscFinder as an application on top of the Hadoop framework. DiscFinder has been used to cluster the largest open-science cosmology simulation datasets containing as many as 14.7 billion particles. We evaluate its performance and scaling properties and describe the performed optimization.

Scale and Concurrency in GIGA+: File System Directories with Millions of Files

Patil & Gibson

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-10-110. October 2010.

We examine the problem of scalable file system directories, motivated by data-intensive applications requiring millions to billions of small files to be ingested in a single directory at rates of hundreds of thousands of file creates every second. We introduce a POSIX-compliant scalable directory design, GIGA+, that distributes directory entries over a cluster of server nodes that make only local, independent decisions about migration. GIGA+ uses two tenets, asynchrony and inconsistency, to: (1) partition the index among all servers without any synchronization or serialization, and (2) minimize stale and inconsistent mapping state at the clients. Applications are provided traditional strong data consistency semantics, and cluster growth requires minimal directory entry migration. We have built and demonstrated that the GIGA+ approach scales better than existing distributed directory implementations, delivers a sustained

throughput of more than 98,000 file creates per second on a 32-server cluster, and balances load more efficiently than consistent hashing.

SmartScan: Efficient Metadata Crawl for Storage Management Metadata Querying in Large File Systems

Liu, Xu, Wu, Yang & Ganger

Carnegie Mellon University Parallel Data Lab Technical Report CMU-PDL-10-112, October 2010.

SmartScan is a metadata crawl tool that exploits patterns in metadata changes to significantly improve the efficiency of support for file-system-wide metadata querying, an important tool for administrators. Usually, support for metadata queries is provided by databases populated and refreshed by calling stat() on every file in the file system. For large file systems, where such storage management tools are most needed, it can take many hours to complete each scan, even if only a small percentage of the files have changed. To address this issue, we identify patterns in metadata changes that can be exploited to restrict scanning to the small subsets of directories that have recently had modified files or that have high variation in file change times. Experiments with using SmartScan on production file systems show that exploiting metadata change patterns can reduce the time needed to refresh the metadata database by one or two orders with minimal loss of freshness.

Matthew Wachs discusses his research with industry guests at the PDL Spring Industry Visit Day.

Milo Polte presents his "Table Software Benchmark" demonstration at the PDL Spring Industry Visit Day. His advisor, Garth Gibson, looks on.

worldwide. The collaborative effort between HP and these universities has delivered breakthroughs in areas such as cloud computing, optical computing and nano-materials — fundamental enablers of the next generation of products and services for communities around the globe," said Rich Friedrich, director of strategy and innovation at HP

--CMU Press Release July 8, 2010

July 2010

Gregory Ganger Testifies in Washington About Benefits and Risks of Using Cloud Computing

In testimony to the U.S. House Committee on Oversight and Government Reform and the Subcommittee on Government Management, Organization and Procurement, Gregory Ganger discussed the benefits and risks of using cloud computing.

Ganger, head of Carnegie Mellon's Parallel Data Lab and a professor in the Department of Electrical and Computer Engineering, said that cloud computing has the potential to provide large efficiency improvements for federal information technology (IT) functions. Cloud computing refers to computing that is based on the Internet, which allows computer users to share software, databases and other services that are provided or managed by other parties over the Web. This contrasts with personal computing,

where all data storage and processing occurs within the user's computer and uses software loaded onto that computer.

Ganger recommended to federal officials that the government support both standardization and research/experimentation efforts in the pursuit of cloud computing's potential. He also noted that moving federal IT "to the cloud" will require significant technical and change management training for IT staff and managers as well as explicit information and effort sharing across a broad swath of federal agencies considering the use of cloud computing.

"Cloud computing is an exciting realization of a long-sought concept: computing as a utility. Pursuing judicious use for federal IT functions is important, given the large potential benefits," Ganger said.

--Carnegie Mellon Media Notification, June 30, 2010

July 2010

FAWN Team Wins 2010 JouleSort Challenge

Congratulations to Vijay Vasudevan, Lawrence Tan, David Andersen of Carnegie Mellon University, and Michael Kaminsky, Michael A. Kozuch, Padmanabhan Pillai of Intel Labs Pittsburgh for winning the 2010 JouleSort (energy-efficient sort) for the 108 records category. Using FAWNSort on the following equipment (Intel Xeon L3426 1.86GHz, 12GB RAM, Nsort, Fusion-io ioDrive (80GB), 4 x Intel X25-E (3 x 32GB, 1 x 64GB)) , they achieved 44,900 records sorted/joule. Medals are awarded each year at ACM SIGMOD. More information on the challenge, including the rules, may be found at http://sortbenchmark.org/.

June 2010

Christos Faloutsos Receives 2010 ACM SIGKDD Innovation Award

Congratulations to Christos Faloutsos, who is the winner of the 2010

ACM SIGKDD Innovations Award. The Innovation Award recognizes one individual or one group of collaborators whose outstand-

ing technical innovations in the field of Knowledge Discovery and Data Mining have had a lasting impact in advancing the theory and practice of the field. The contributions must have significantly influenced the direction of research and development of the field or transferred to practice in significant and innovative ways and/or enabled the development of commercial systems.

Christos, a Professor of Computer Science and Electrical and Computer Engineering, focuses his research on data Mining for graphs and streams, fractals, self-similarity and power laws, indexing and data mining for video, biological and medical databases, and data base performance evaluation (data placement, workload characterization).

May 2010

Lorrie Cranor Expert on Privacy Issues in Advertising Panel

Lorrie Cranor, associate professor of computer science and engineering and public policy, discussed the privacy issues swirling around

the technical mechanics of online advertising as part of a panel of experts in Washington, D.C., sponsored by The Progress & Freedom Foundation. Read more about the discussion at http://www.cmu.edu/news/archive/2010/May/may21_onlineadvertisingprivacy.shtml.

-- CMU 8.5x11 News May 27, 2010