



Seattle Conference on Scalability Saturday, June 23, 2007

Conference Fast Facts

Date & Time:

Saturday, June 23, 2007
8:30AM – 7:30PM

NOTE: New Conference Location!

The Westin Bellevue
600 Bellevue Way NE
Bellevue, Washington 98004
Tel 425.638.1000

Hotel Recommendations:

The Westin Bellevue (conference hotel - recommended)
600 Bellevue Way NE
Bellevue, Washington 98004
Tel 425.638.1000

<http://www.starwoodhotels.com>

Bellevue Club Hotel
11200 SE 6th St
Bellevue, WA 98004
Tel 425.454.4424

<http://www.bellevueclub.com>

Woodmark Hotel on Lake Washington
1200 Carillon Point
Kirkland, WA 98033
Tel 425.822.3700 or 800.822.3700

<http://www.woodmarkhotel.com>

CONFERENCE AT A GLANCE

Time	Session Description	Location
8:30AM – 9:00AM	Registration	3 rd Floor Lobby
9:00AM – 10:00AM	Keynote I: MapReduce, BigTable, and Other Distributed System Abstractions for Handling Large Datasets	Grand Ballroom B
10:15AM – 11:15AM	Breakout I: Lustre File System	Grand Ballroom A
	Breakout I: Building A Scalable Resource Management Layer for Grid Computing	Grand Ballroom B
11:30AM – 12:30PM	Breakout II: VeriSign's Global DNS Infrastructure	Grand Ballroom A
	Breakout II: Using MapReduce on Large Geographic Datasets & Google Talk: Lessons in Building Scalable Systems	Grand Ballroom B
12:30PM – 1:45PM	Lunch	Grand Ballroom C
1:45PM – 2:45PM	Keynote II: Description TBD	Grand Ballroom B
3:00PM – 4:00PM	Breakout III: SCTP's Additional Reliability and Fault Tolerance	Grand Ballroom A
	Breakout III: Scalable Test Selection Using Source Code Deltas	Grand Ballroom B
4:15PM – 5:15PM	Breakout IV: YouTube Scalability	Grand Ballroom A
	Breakout IV: Challenges in Building an Infinite Scalable Datastore	Grand Ballroom B
5:30PM – 7:30PM	Reception & Networking	Grand Ballroom C

DETAILED CONFERENCE SCHEDULE

TIME & LOCATION	SESSION ABSTRACT & SPEAKER BIO
<p>KEYNOTE SPEAKER: Jeff Dean 9:00AM – 10:00AM Grand Ballroom B</p> 	<p>MapReduce, BigTable, and Other Distributed System Abstractions for Handling Large Datasets by Jeff Dean, Google Inc.</p> <p>Search is one of the most important applications used on the internet, but it also poses some of the most interesting challenges in computer science. Providing high-quality search requires understanding across a wide range of computer science disciplines, from lower-level systems issues like computer architecture and distributed systems to applied areas like information retrieval, machine learning, data mining, and user interface design. In this talk, I'll highlight some of the behind-the-scenes pieces of infrastructure that we've built in order to operate Google's services.</p> <p>Jeff Dean joined Google in 1999 and is currently a Google Fellow in Google's Systems Infrastructure Group. While at Google he has worked on Google's crawling, indexing, query serving, and advertising systems, implemented several search quality improvements, and built several major pieces of Google's distributed computing infrastructure. He received a Ph.D. from the University of Washington in 1996 working with Craig Chambers on compiler optimization techniques for object-oriented languages.</p>
<p>BREAKOUT I 10:15AM – 11:15AM Grand Ballroom A</p> 	<p>Lustre File System by Peter Braam, Cluster File Systems, Inc.</p> <p>Lustre is a scalable open source Linux cluster file system that powers 6 of the top 10 computers in the world. It is resold by HP, SUN, Dell and many other OEM and storage companies, yet produced by a small powerful technology company, Cluster File Systems, Inc. This lecture will explain the Lustre architecture and then focus on how scalability was achieved. We will address many aspects of scalability mostly from the field and some from future requirements, from having 25,000 clients in the Red Storm computer to offering exabytes of storage. Performance is an important focus and we will discuss how Lustre serves up over 100GB/sec today going to 100TB/sec in the coming years. It will deliver millions of metadata operations per second in a cluster and, write 10's of thousands of small files per second on a single node. If you like big numbers (but less than a Gogol) please come to this talk.</p>

BREAKOUT I
10:15AM – 11:15AM
Grand Ballroom B



Peter Braam is founder and president of Cluster File Systems, Inc which has developed the Lustre file system. Lustre powers many of the world's largest super computers and has achieved outstanding bandwidth and scalability. Lustre has won several awards, as did Peter's earlier work on InterMezzo and Coda. Peter is a specialist in distributed file systems. Before founding Cluster File Systems, Peter was an academic and held senior faculty positions at Oxford and Carnegie Mellon University where the Lustre project was started. Peter was also Chief Architect at TurboLinux and a Cluster Architect for Red Hat, Inc and founded and ran Stelias Computing where much of the initial thinking about Lustre and InterMezzo was done.

Building A Scalable Resource Management Layer for Grid Computing by Khalid Ahmed, Platform Computing.

This talk will describe the architecture and implementation details for building a highly scalable resource management layer that can support a variety of applications and workloads. This technology has evolved from large scale computing grids deployed in production at customers such as Texas Instruments, AMD, JP Morgan, and various government labs. We will show how to build a centralized dynamic load information collection service that can handle up to 5000 nodes/20,000 cpus in a single cluster. The service is able to gather a variety of system level metrics and is extensible to collect up to 256 dynamic or static attributes of a node and actively feed them to a centralized master. A built-in election algorithm ensures timely failover of the master service ensuring high-availability without the need for specialized interconnects.

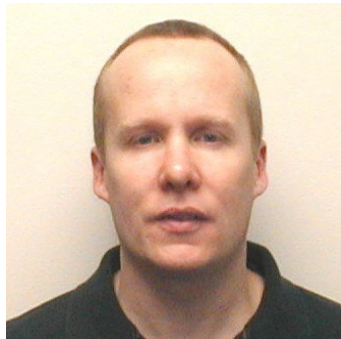
This building block is extended to multiple clusters that can be organized hierarchically to support a single resource management domain that can span multiple data centers. We believe the current architecture could scale to 100,000 nodes/400,000 cpus. Additional services such as a distributed process execution service, and a policy-based resource allocation engine which leverage this core scale-out clustering service are described. The protocols, communication overheads, and various design tradeoffs that were made the development of these services will be presented along with experimental results from various tests, simulations and production environments.

Khalid Ahmed works as the Chief Architect and Director of Technology Research at Platform Computing. In over 12 years at Platform he worked in a number of roles including development, product management and architecture. His work on distributed scheduling, wide-area resource sharing, workload management, system automation, virtualization management, and high availability

have been incorporated in Platform offerings including Platform LSF, Platform Symphony and more recently in the Platform Enterprise Grid Orchestrator (EGO) technology platform. Khalid oversees a team of architects and is responsible for managing Platform's technology strategy for both high performance computing and enterprise IT. Khalid holds a MA.Sc. in EE from the University of Toronto.

BREAKOUT II
11:30AM – 12:30PM
Grand Ballroom A

VeriSign's Global DNS Infrastructure by Patrick Quaid and Scott Courtney, VeriSign.



VeriSign's global network of nameservers for the .com and .net domains sees 500,000 DNS queries per second during its daily peak, and ten times that or more during attacks. By adding new servers and bandwidth, we've recently increased capacity to handle many times that query volume. Name and address changes are distributed to these nameservers every 15 seconds -- from a provisioning system that routinely receives one million domain updates in an hour. In this presentation we describe VeriSign's production DNS implementation as a context for discussing our approach to highly scalable, highly reliable architectures. We will talk about the underlying Advanced Transactional Lookup and Signaling software, which is used to handle database extraction, validation, distribution and name resolution. We also will show the central heads-up display that rolls up statistics reported from each component in the infrastructure.

Patrick Quaid is the technical director of Verisign's R&D group, where he is responsible for design and direction but still manages to write some code. The R&D group develops VeriSign's high capacity, highly available infrastructure services, including DNS and Whois for .com and .net as well as applications providing database, storage, network and monitoring services. Prior to his current role he led the development of ATLAS, the framework supporting many of VeriSign's resolution services.

Prior to that, he participated in the usual assortment of start-ups and consulting companies. He lives in Northern Virginia with his wife and three kids.

Scott Courtney is a principal architect in VeriSign's Architecture and Technology Services group. He leads the Naming and Edge Services team, where he focuses on systems design for the Shared Registration System and related resolution services (root, TLDs, managed DNS). His background at several financial firms and at one tech startup/shutdown is in highly available systems design, failure analysis and systems administration ... which often seem intercorrelated. He received two engineering degrees from Virginia Tech, and lives in Virginia with his wife and three children.

BREAKOUT II
11:30AM – 12:30PM
Grand Ballroom B

Using MapReduce on Large Geographic Datasets by Barry Brumitt, Google, Inc.

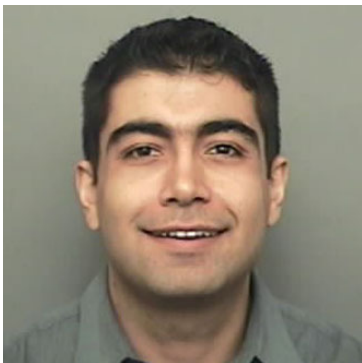
MapReduce is a programming model and library designed to simplify distributed processing of huge datasets on large clusters of computers. This is achieved by providing a general mechanism which largely relieves the programmer from having to handle challenging distributed computing problems such as data distribution, process coordination, fault tolerance, and scaling. While working on Google maps, I've used MapReduce extensively to process and transform datasets which describe the earth's geography. In this talk, I'll introduce MapReduce, demonstrating its broad applicability through example problems ranging from basic data transformation to complex graph processing, all the in the context of geographic data.


Barry Brumitt is a software engineer with Google, Inc. who has been working on maps-related applications since joining the Kirkland, WA office in November 2005. Prior to coming to Google, he was at Microsoft Corp for 8 years, working in both MS Games Studios and MS Research. At MGS, he was responsible for the Artificial Intelligence in Forza Motorsport, a simulation-style XBox racing game. Previously at MSR, he worked in Ubiquitous Computing, exploring location-based services, geometric models, and multi-modal interfaces for smart environments, and publishing over a dozen peer-reviewed papers. He received his Ph.D. in Robotics from Carnegie Mellon in December 1997, and two B.Sc.s' in Computer Engineering and Physics from the same institution in 1991. Offline, he's currently fascinated by yoga, downhill skiing, cycling, bridge, juggling, electronic music, and community building.

Google Talk: Lessons in Building Scalable Systems by Reza Behforooz, Google, Inc.

Since launching Google Talk in the summer of 2005, we have integrated the service with two large existing products: Gmail and orkut. Each of these integrations provided unique scalability challenges as we had to handle a sudden big increase in the number of users. Today, Google Talk supports millions of users and handles billions of packets per day. I will discuss several practical lessons and key insights from our experience that can be used for any project. These lessons will cover both engineering and operational areas.

Reza Behforooz is a Staff Engineer at Google and is currently the technical lead for the Google Talk servers. He's passionate about building large systems and working on communication products in an attempt to make the world a smaller place. While at Google, he



	<p>has primarily worked on Google Talk, Gmail, orkut, Google Groups, and shared infrastructure used by several Google applications. Reza holds a BS from Cornell and a MS from Stanford in Computer Science. Prior to Google, he held various engineering and management positions at Microsoft and two startups, Zaplet and Epiphany.</p>
<p>12:30PM – 1:45PM Lunch Grand Ballroom C</p>	
<p>KEYNOTE SPEAKER: Marissa Mayer 1:45PM – 2:45PM Grand Ballroom B</p> 	<p>TOPIC TBD</p> <p>Marissa Mayer, Vice President, Search Products & User Experience, leads the product management efforts on Google's search products – web search, images, groups, news, Froogle, the Google Toolbar, Google Desktop, Google Labs, and more. She joined Google in 1999 as Google's first female engineer and led the user interface and webserver teams at that time. Her efforts have included designing and developing Google's search interface, internationalizing the site to more than 100 languages, defining Google News, Gmail, and Orkut, and launching more than 100 features and products on Google.com. Several patents have been filed on her work in artificial intelligence and interface design. In her spare time, Marissa also organizes Google Movies – outings a few times a year to see the latest blockbusters – for 6,000+ people (employees plus family members and friends).</p> <p>Concurrently with her full-time work at Google, Marissa has taught introductory computer programming classes at Stanford to over 3,000 students. Stanford has recognized her with the Centennial Teaching Award and the Forsythe Award for her outstanding contribution to undergraduate education.</p> <p>Prior to joining Google, Marissa worked at the UBS research lab (Ubilab) in Zurich, Switzerland and at SRI International in Menlo Park, California. Marissa has been featured in various publications, including <i>Newsweek</i> ("10 Tech Leaders of the Future"), <i>Red Herring</i> ("15 Women to Watch"), <i>Business 2.0</i> ("Silicon Valley Dream Team"), <i>BusinessWeek</i>, <i>Fortune</i>, and <i>Fast Company</i>. Graduating with honors, Marissa received her B.S. in Symbolic Systems and her M.S. in Computer Science from Stanford University. For both degrees, she specialized in artificial intelligence.</p>

BREAKOUT III
3:00PM – 4:00PM
Grand Ballroom A

SCTP's Additional Reliability and Fault Tolerance by Brad Penoff, Mike Tsai, and Alan Wagner, The University of British Columbia.

Low cost clusters are usually built from commodity parts and use standard transport protocols like TCP/IP. Once systems become large enough, reliability and fault tolerance become an important issue and TCP/IP often requires additional mechanisms to ensure reliability of the application. The Stream Control Transmission Protocol (SCTP) is a newly standardized transport protocol that provides additional mechanisms for reliability beyond that of TCP. The added reliability and fault tolerance of SCTP may function better for MapReduce-like distributed applications on large commodity clusters.

SCTP has the following features that provide additional levels of reliability and fault tolerance. Selective acknowledgment (SACK) is built-in to the protocol with the ability to express larger gaps than TCP; as a result, SCTP outperforms TCP under loss. For cluster nodes with multiple interfaces, SCTP supports multihoming, which transparently provides failover in the event of network path failure. SCTP has the stronger CRC32c checksum which is necessary with high data rates and large scale systems. SCTP also allows multiple streams within a single connection, providing a solution to the head-of-line blocking problem present in TCP-based farming applications like Google's MapReduce. Like TCP, SCTP provides a reliable data stream by default, but unlike TCP, messages can optionally age or reliability can be disabled altogether. The SCTP API provides both a one-to-one (like TCP) and a one-to-many (like UDP) socket style; use of a one-to-many style socket can reduce the number of file descriptors required by an application, making it more scalable.

The additional scalability and fault tolerance come at a cost. The CRC32c checksum calculation currently is not off-loaded to any NIC available on the market, so it must be performed by the host CPU. In high bandwidth environments with no loss, SACK processing may become a burden on the host CPU. Understanding the trade-offs between the benefits of SCTP and these additional CPU costs is an interesting research question. We have experimented and developed middleware and applications using SCTP underneath a popular programming model used for parallel programs called the Message Passing Interface (MPI). Argonne National Laboratory releases a widely-used open-source version of MPI called MPICH2; we have incorporated our work into their most recent release of MPICH2. Although relatively new, SCTP is maturing and stacks exist for most major operating systems.

Brad Penoff received a B.Sc. degree from Ohio State University. He did two summer internships at Sun Microsystems and after graduation worked for Sun Microsystems Ireland, for two years. Brad is currently a PhD student in CS at UBC under the supervision of Alan Wagner after having completed his M.Sc. under Alan at UBC as well. Brad has been a research assistant the past two summers at Argonne National Labs working in their MPI group.

Mike Tsai received a B.Sc. degree from UBC in 2005. He is currently in the M.Sc. program also under Alan Wagner's supervision. Mike has been a full-time research assistant with our team at UBC for the past year and a half, working with MPI and SCTP.

Alan Wagner received his Ph.D. from the University of Toronto in 1987 under the supervision of Derek Corneil. He joined the CS Department at the University of British Columbia in 1987, and is currently an Associate Professor. In 2000 he spent one year with Terabeam Networks in Seattle, Washington in the distributed networking group headed by Henry Sowizral.

Scalable Test Selection Using Source Code Deltas by Ryan Gerard, Symantec Corporation.

As the number of automated regression tests increase, the ability to run all of them in a reasonable amount of time becomes more and more difficult, and simply doesn't scale. Since we are looking for regressions, it is useful to hone in on the parts of the code that have changed from the last run to help select a small subset of tests that are likely to find the regression. In this way we are only running the tests that need to be run as your system gets larger and the number of possible tests scales outward. We have devised a method to select a subset of tests from an existing test set for scalable regression testing based on source code changes, or deltas. The selection algorithm is a static data mining technique that establishes the relationship between source code deltas and test case execution results. Test selection is then based on the established correlation. In this talk, we will discuss the benefits and also the pitfalls involved in having such an infrastructure. Finally, we will talk about how best to add it to a nightly or continuous test automation infrastructure.

Ryan Gerard is currently an SQA Engineer at Symantec. He has a BS in Computer Science and Engineering from UCLA, and is currently pursuing his MS in Information Security. Ryan's particular specialties are in web technologies and security testing, although his interests span kernel-level technologies to process improvements to data analysis.

BREAKOUT III
3:00PM – 4:00PM
Grand Ballroom B



BREAKOUT IV
4:15PM – 5:15PM
Grand Ballroom A

YouTube Scalability by Cuong Do, Engineering Manager, YouTube.

This talk will discuss some of the scalability challenges that have arisen during YouTube's short but extraordinary history. YouTube has grown incredibly rapidly despite having had only a handful of people responsible for scaling the site. Topics of discussion will include hardware scalability, software scalability, and database scalability.

Cuong is currently an engineering manager at YouTube/Google. He was part of the engineering team that scaled the YouTube software and hardware infrastructure from its infancy to its current scale. Prior to YouTube/Google, he held various software development and management positions at PayPal and Inktomi.

BREAKOUT IV
4:15PM – 5:15PM
Grand Ballroom B

Challenges in Building an Infinite Scalable Datastore by Swami Sivasubramanian and Werner Vogels, Amazon.com.



Amazon.com runs a world-wide e-commerce platform that serves tens of millions customers at peak times using tens of thousands of servers located in many data centers around the globe. Reliability and scalability are the most important challenges in building our platform. The most important scalability challenge in our system is our ability to scale our persistent layer. Amazon.com has built several in-house datastores to meet the needs of its internal applications. In this talk, we will present the design of one of our internal datastores, HASS.



HASS is designed to be "always" available, i.e., it will always accept read/write requests even if disks are failing, routes are flapping or if datacenters are being destroyed by tornados. HASS is designed for incremental scalability where adding or removing nodes can be done easily and the load gets evenly distributed among the nodes uniformly without requiring any operator intervention. In this talk, we will focus on a single and one of the most crucial ideas in HASS's design: its ability to partition data. HASS uses consistent hashing to partition its data across its storage nodes. The basic consistent hashing algorithm is well understood in the academic literature and several research systems have been designed using it. In this talk, we will discuss our experiences with using the basic consistent hashing algorithm and the optimizations we performed to achieve more uniform load distribution and ease of operation.

Swaminathan Sivasubramanian is a distributed systems engineer in Amazon.com. His primary interests are in the design and implementation of scalable and reliable distributed systems. He has a Ph.D. in Computer Science from Vrije Universiteit in Amsterdam.

	<p>Dr. Werner Vogels is Vice President & Chief Technology Officer at Amazon.com where he is responsible for driving the company's technology vision, which is to continuously enhance the innovation on behalf of Amazon's customers at a global scale. Prior to joining Amazon, he worked as a research scientist at Cornell University where he was a principal investigator in several research projects that target the scalability and robustness of mission-critical enterprise computing systems. He has held positions of VP of Technology and CTO in companies that handled the transition of academic technology into industry. Vogels holds a Ph.D. from the Vrije Universiteit in Amsterdam and has authored close to 80 articles for journals and conferences, most of them on distributed systems technologies for enterprise computing.</p>
<p>RECEPTION 5:30PM – 7:30PM Grand Ballroom C</p>	<p>Join your fellow conference attendees for food, drinks and networking.</p>

Questions?
seattle-events@google.com

