Vijay Vasudevan

Email: vrv@cs.cmu.edu
Web: http://www.cs.cmu.edu/∼vrv
Phone: (650) 387 3576

Interests

Broadly, my interests span topics in networked and large-scale storage systems. Specific topics include datacenter energy efficiency, non-volatile memory support in operating systems, datacenter networking, and systems at the intersection of databases and "cloud" storage.

Education

Carnegie Mellon University (2006 - 2011)
Ph.D. in Computer Science, Computer Science Department (2011)
Masters in Computer Science, Computer Science Department (2010)
Advisor: David G. Andersen
Ph.D Thesis: Energy-efficient Data-intensive Computing with a Fast Array of Wimpy Nodes

University of California, Berkeley (2002 - 2006)
Bachelor of Science with Honors, Electrical Engineering and
Computer Science Department, May 2006.

Refereed Publications

1. Vijay Vasudevan, Michael Kaminsky, and David G. Andersen. Using vector interfaces to deliver millions of IOPS from a networked key-value storage server. In *Proc. 3rd ACM Symposium on Cloud Computing (SOCC)*, San Jose, CA, October 2012.
2. Vijay Vasudevan, David G. Andersen, and Michael Kaminsky. The case for VOS: The vector operating system. In *Proc. HotOS XIII*, Napa, CA, May 2011.
3. David G. Andersen, Jason Franklin, Michael Kaminsky, Amar Phanishayee, Lawrence Tan, and Vijay Vasudevan. FAWN: A fast array of wimpy nodes. In *Proc. 22nd ACM Symposium on Operating Systems Principles (SOSP)*, Big Sky, MT, October 2009. (Won Best Paper Award)
4. Vijay Vasudevan, Amar Phanishayee, Hiral Shah, Elie Krevat, David G. Andersen, Gregory R. Ganger, Garth A. Gibson, and Brian Mueller. Safe and effective fine-grained TCP retransmissions for datacenter communication. In *Proc. ACM SIGCOMM*, Barcelona, Spain, August 2009.
5. Vijay Vasudevan, Jason Franklin, David Andersen, Amar Phanishayee, Lawrence Tan, Michael Kaminsky, and Iulian Moraru. FAWNdamentally power-efficient clusters. In *Proc. HotOS XII*, Monte Verita, Switzerland, May 2009.
6. Vijay Vasudevan, Sudipta Sengupta, and Jin Li. A first look at media conferencing traffic in the global enterprise. In *Passive & Active Measurement (PAM)*, Seoul, South Korea, April 2009.
7. Amar Phanishayee, Elie Krevat, Vijay Vasudevan, David G. Andersen, Gregory R. Ganger, Garth A. Gibson, and Srinivasan Seshan. Measurement and analysis of TCP throughput collapse in cluster-based storage systems. In *Proc. USENIX Conference on File and Storage Technologies (FAST)*, San Jose, CA, February 2008.
8. Elie Krevat, Vijay Vasudevan, Amar Phanishayee, David G. Andersen, Gregory R. Ganger, Garth A. Gibson, and Srinivasan Seshan. On application-level approaches to avoiding TCP throughput collapse in cluster-based storage systems. In *Proc. Petascale Data Storage Workshop at Supercomputing'07*, November 2007.

Other Articles

9. Vijay Vasudevan. Energy-efficient data-intensive computing with a fast array of wimpy nodes, October 2011. Ph.D Thesis, Department of Computer Science, Carnegie Mellon University.
10. David G. Andersen, Jason Franklin, Michael Kaminsky, Amar Phanishayee, Lawrence Tan, and Vijay Vasudevan. FAWN: A fast array of wimpy nodes. *Communications of the ACM*, 54(7):101–109, July 2011.
11. Vijay Vasudevan, David G. Andersen, Michael Kaminsky, Lawrence Tan, Jason Franklin, and Iulian Moraru. Energy-efficient cluster computing with FAWN: Workloads and implications. In *Proc. e-Energy 2010*, Passau, Germany, April 2010. (invited paper).
12. Vijay Vasudevan, David G. Andersen, and Hui Zhang. On Internet availability: Where does path choice matter? Technical Report CMU-CS-TR-114, Carnegie Mellon University, March 2009.
13. Vijay Vasudevan, David G. Andersen, and Hui Zhang. Understanding the AS-level path disjointness provided by multi-homing. Technical Report CMU-CS-TR-141, Carnegie Mellon University, July 2007.

Selected Honors and Awards

- Yahoo! Key Scientific Challenges 2010 Award Winner – Green Computing
- JouleSort 10GB 2010 Winner (Indy and Daytona Categories)
- Facebook Fellowship Finalist 2010
- Best Paper, 22nd ACM Symposium on Operating System Principles (SOSP 2009)
- APC Fellowship 2009-2010
- SIGCOMM 2007, FAST 2008, FAST 2009, PAM 2009, SOSP 2009 Student Travel Grant Award Winner
- Member of Eta Kappa Nu, Berkeley Mu Chapter, Electrical Engineering and Computer Science Honor Society.
- Golden State Scholarship and Golden State for Maths and Sciences Scholarship, 2001.

Professional Experience

Google, Inc. Nov. 2011 - Present
Software Engineer in Storage Infrastructure

Carnegie Mellon University Computer Science Department Aug. 2006 - Oct. 2011
Research assistant under advisement of Professor David Andersen

- Conducted research on energy-efficient cluster computing.
- Investigated role of low-power compute nodes to datacenter workloads.
- Identified operating system support for high-performance non-volatile memories.
- A description of research projects begins on Page 3.

Intel Corporation September 2010 - September 2011
Consultant

- Worked with Intel Datacenter Group and Flash SSD Group to help identify supporting OS and interface improvements to high-performance flash storage for datacenter and data-intensive applications.

Research Intern, Microsoft Research Communications and Collaboration Systems Group Summer 2008

- Worked on characterizing media conferencing traffic in a large enterprise using dense call logs of audio/video deployments over a 6-month period.
- Designed a protocol for data source discovery for enterprise networks, aimed at building caching natively into a distributed enterprise network.

Research Assistant in ALPHA/Automation Lab, U.C. Berkeley 2005-2006
Led by Professor Ken Goldberg, IEOR/EECS Department.

- Worked on 3D Finite Element Modeled needle surgery simulation for use in medical operations such as Brachytherapy and for physician planning and training.
- Network Designer for production of Ballet Mori, an improvisational ballet performed at the San Francisco Opera House on April 4, 2006, using live streaming seismic activity data to generate music and lighting.

Xangati Inc., Palo Alto, California USA 2003-2004
Performance and Network Tester

- Tested Xangati's DDoS mitigation system by unleashing worms on a LAN and analyzing network and system statistics to determine the effectiveness of the solution.

Teaching Experience

15-440 Distributed Systems Teaching Assistant Aug. 2009 - Dec. 2009
Teaching assistant under Professor David Andersen and Charlie Garrod

- Helped revamp the Distributed Systems course to serve as a systems course requirement.
- Designed MapReduce project to data mine a crawled Twitter dataset.
- Designed test questions and homework problems, performed grading, held office hours.
- Ran two recitations per week on extra-curricular material (how to do systems programming).

15-744 Computer Networking Teaching Assistant Jan. 2007 - May 2007
Teaching assistant under Professor Srinivasan Seshan

- Maintained paper reading database software.
- Ran recitations, held office hours, designed test questions, performed grading.
- Helped run and supervise group projects.

Research Experience – Efficient Datacenter Computing

FAWN: A Fast Array of Wimpy Nodes September 2007 - October 2011

The majority of my dissertation research is based on FAWN, or a "Fast Array of Wimpy Nodes". In this project, we propose an approach to building clusters for low-power data-intensive computing [5, 3, 11]. The FAWN approach argues for energy-efficient, balanced computing platforms as the primitive building block for large-scale data-intensive cluster computing. FAWN achieves balance by pairing low-power processors to small amounts of local flash storage, choosing a particular pairing of computation and I/O that optimizes for energy efficiency rather than for performance alone.

To demonstrate the energy efficiency benefits of the FAWN approach, we developed FAWN-KV—a consistent, replicated, highly available, and high-performance key-value storage system built on a FAWN prototype. Our design centers around purely log-structured datastores that provide the basis for high performance on flash storage, as well as for replication and consistency obtained using chain replication on a consistent hashing ring. Our evaluation demonstrates that FAWN clusters can handle up to two orders of magnitude more *queries per Joule* than a disk-based system, and can be several times more efficient than balanced, flash-based server platforms.

We have also explored a variety of data-intensive workloads, identifying under what conditions and parameters that a FAWN system can be more efficient, and when traditional systems win out [11]. This work demonstrates that achieving system balance depends on a variety of workload characteristics—the software challenges posed by the FAWN approach often require rethinking and redesigning software systems to account for FAWN's different balance of computation, capacity, and I/O capability.

Rethinking OS Support for High-Performance Flash Storage March 2010 - Present

Flash storage has recently emerged as a viable new tier between DRAM and disk. Its cost per byte exceeds DRAM, its random I/O performance obliterates that provided by rotating media, and its low power consumption makes it an attractive technology to help reduce power in datacenters. Our FAWN work leveraged these features of flash to save energy for random I/O workloads by reducing the CPU-I/O gap.

However, flash random I/O performance has increased by about a factor of 100 in just three years. Combined with the move to many-core CPUs where individual core speeds have plateaued, this hundred-factor increase in I/O performance has created a sudden I/O-gap reduction: random I/O performance from flash has become CPU-bound. The reasons are twofold: First, a device capable of hundreds of thousands of I/Os per second can create the same number of hardware interrupts per second, reducing the operating system's ability to perform useful work. Second, flash is currently exposed behind the same interface as much slower rotating media, and the I/O stack used for disk is too inefficient for the two orders of magnitude improvement in both random I/O throughput and latency that flash provides.

I explored several ways in which operating systems interfaces should evolve to improve handling of low-latency, high-throughput asynchronous I/O devices such as flash storage. Our proposal for, and implementation of using "vector" interfaces to storage and RPC systems provided a 14-factor improvement in the number of IOPS delivered using a prototype NVM device [1]. Our experience showed that rethinking the OS interface and implementations for future storage technologies is necessary to take advantage of this new tier in the storage hierarchy.

Research Experience – Networking

TCP Incast September 2006 - August 2009

In TCP-based high-bandwidth, low-latency storage area networks where one client synchronously requests data striped across many servers, TCP throughput sometimes drops by one or two orders of magnitude below the client's link bandwidth. This throughput collapse, termed *Incast*, poses a problem for large-scale networked services running on commodity hardware (Ethernet/IP), representing an important barrier to the scalability of future networked storage systems.

The fundamental cause of this behavior is a combination of three aspects of networking: small network switch buffers, barrier-synchronized request/response traffic, and small object transfers. This traffic pattern creates excessive, synchronized packet loss that TCP is unable to recover from without a TCP timeout lasting 200 milliseconds. These long timeouts lead to reduced TCP throughput and high request latency.

Our FAST2008 paper presented a full network-level analysis of Incast in simulation and real networks [7], introducing the Incast problem to a broader community. In our SIGCOMM2009 paper, we demonstrated that reducing the 200ms minimum retransmission timeout value, long held as an important lower bound, was both a safe and effective change to the TCP stack to combat Incast [4]. We developed a Linux patch to enable high resolution microsecond TCP retransmissions, allowing retransmissions in microseconds to alleviate the problem. Reducing the retransmission timeout is now being used by members in industry as well as follow-on work to datacenter transport protocols.

Media Traffic Characterization in the Enterprise May 2008 - August 2008

We analyzed and characterized media traffic in the Microsoft enterprise network using dense call logs of audio/video conferencing and IP-phone deployments over a 6-month period [6]. We also identified major causes of poor call quality and the impact of QoS on voice traffic in a real-world setting, suggesting specific improvements to networking infrastructure and software to the Microsoft product team.

Proximity-guided Data Source Discovery for Enterprise Networks May 2008 - August 2008

We designed a protocol for data source discovery for enterprise networks, building data caching infrastructure natively into enterprise hosts for improved scalability, robustness, performance, and content locality. We also quantified the benefit and importance of cross-subnet sharing through collaborations with the Microsoft P2P product group.

SWAMP: Simple Wide-Area Multi-Path August 2006 - May 2008

This work explores routing architectures to improve Internet availability and performance using multiple concurrent interdomain end-to-end paths. The goal of this work is to find which paths provide the highest degree of AS disjointness with low complexity. Our first design focuses on a simple, effective, and economically-friendly architecture to promote real-world deployment. We find that providing path choice at both a multi-homed source and destination can provide nearly as many AS disjoint paths as a policy-free source routing mechanism could [13]. Moreover, today's multi-homing severely underutilizes the diversity of network links due to single path BGP forwarding. We have also conducted real-world measurements gauging the benefits of SWAMP-style routing[12]. SWAMP is an immediately deployable service/architecture allowing multi-homed end-hosts and organizations to unlock the currently hidden benefits of multi-homing for improved availability and performance.

Personal References

Prof. David G. Andersen
Computer Science Department
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA 15213
dga@cs.cmu.edu

Prof. Garth A. Gibson
Computer Science Department
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA 15213
garth@cs.cmu.edu

Prof. Gregory R. Ganger
Electrical and Computer Engineering Dept.
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA 15213
ganger@ece.cmu.edu

Dr. Michael E. Kaminsky
Intel Labs Pittsburgh
4720 Forbes Ave, Suite 410
Pittsburgh, PA 15213
michael.e.kaminsky@intel.com

References

[1] Vijay Vasudevan, Michael Kaminsky, and David G. Andersen. Using vector interfaces to deliver millions of IOPS from a networked key-value storage server. In *Proc. 3rd ACM Symposium on Cloud Computing (SOCC)*, San Jose, CA, October 2012.

[2] Vijay Vasudevan, David G. Andersen, and Michael Kaminsky. The case for VOS: The vector operating system. In *Proc. HotOS XIII*, Napa, CA, May 2011.

[3] David G. Andersen, Jason Franklin, Michael Kaminsky, Amar Phanishayee, Lawrence Tan, and Vijay Vasudevan. FAWN: A fast array of wimpy nodes. In *Proc. 22nd ACM Symposium on Operating Systems Principles (SOSP)*, Big Sky, MT, October 2009.

[4] Vijay Vasudevan, Amar Phanishayee, Hiral Shah, Elie Krevat, David G. Andersen, Gregory R. Ganger, Garth A. Gibson, and Brian Mueller. Safe and effective fine-grained TCP retransmissions for datacenter communication. In *Proc. ACM SIGCOMM*, Barcelona, Spain, August 2009.

[5] Vijay Vasudevan, Jason Franklin, David Andersen, Amar Phanishayee, Lawrence Tan, Michael Kaminsky, and Iulian Moraru. FAWNdamentally power-efficient clusters. In *Proc. HotOS XII*, Monte Verita, Switzerland, May 2009.

[6] Vijay Vasudevan, Sudipta Sengupta, and Jin Li. A first look at media conferencing traffic in the global enterprise. In *Passive & Active Measurement (PAM)*, Seoul, South Korea, April 2009.

[7] Amar Phanishayee, Elie Krevat, Vijay Vasudevan, David G. Andersen, Gregory R. Ganger, Garth A. Gibson, and Srinivasan Seshan. Measurement and analysis of TCP throughput collapse in cluster-based storage systems. In *Proc. USENIX Conference on File and Storage Technologies (FAST)*, San Jose, CA, February 2008.

[8] Elie Krevat, Vijay Vasudevan, Amar Phanishayee, David G. Andersen, Gregory R. Ganger, Garth A. Gibson, and Srinivasan Seshan. On application-level approaches to avoiding TCP throughput collapse in cluster-based storage systems. In *Proc. Petascale Data Storage Workshop at Supercomputing'07*, November 2007.

[9] Vijay Vasudevan. Energy-efficient data-intensive computing with a fast array of wimpy nodes, October 2011. Ph.D Thesis, Department of Computer Science, Carnegie Mellon University.

[10] David G. Andersen, Jason Franklin, Michael Kaminsky, Amar Phanishayee, Lawrence Tan, and Vijay Vasudevan. FAWN: A fast array of wimpy nodes. *Communications of the ACM*, 54(7):101–109, July 2011.

[11] Vijay Vasudevan, David G. Andersen, Michael Kaminsky, Lawrence Tan, Jason Franklin, and Iulian Moraru. Energy-efficient cluster computing with FAWN: Workloads and implications. In *Proc. e-Energy 2010*, Passau, Germany, April 2010. (invited paper).

[12] Vijay Vasudevan, David G. Andersen, and Hui Zhang. On Internet availability: Where does path choice matter? Technical Report CMU-CS-TR-114, Carnegie Mellon University, March 2009.

[13] Vijay Vasudevan, David G. Andersen, and Hui Zhang. Understanding the AS-level path disjointness provided by multihoming. Technical Report CMU-CS-TR-141, Carnegie Mellon University, July 2007.