
Sudharshan S. Vazhkudai

Oak Ridge National Laboratory

vazhkudaiss@ornl.gov, <http://users.nccs.gov/~vazhkuda>, Ph : 865-258-8986

1. CURRENT POSITION

Group Leader, Technology Integration, Oak Ridge National Laboratory (ORNL <http://www.ornl.gov>)

- *Occupation: R&D Management*

2. PROFILE

- **Experience:** 15 years of experience in the US Department Of Energy's (DOE) national lab system, working in the supercomputing and extreme-scale data center programs.
- **Leadership:** Lead a group of 25 R&D staff members to build and deploy solutions for the nation's premier supercomputing center, the Oak Ridge Leadership Computing Facility (OLCF <http://www.olcf.ornl.gov>). Leading projects in several areas such as high-performance computing (HPC), file and storage systems, non-volatile memory (NVRAM), data management, analytics appliance, system architecture, and distributed computing. OLCF comprises of some of the world's fastest supercomputers, providing billions of core hours to a national scientific user base from academia, government and industry, to perform breakthrough research in climate, materials, alternative energy sources and energy storage, chemistry, nuclear physics, astrophysics, quantum mechanics, and the gamut of scientific inquiry.
- **Technology Innovation and Creativity:**
 - Lead the Technology Integration (TechInt <http://techint.nccs.gov>) group in the *development and deployment* of the following for OLCF:
 - ✓ *Systems Software* for the world's No. 2 supercomputer, Titan (27 petaflop heterogeneous CPU/GPU machine with 560,640 cores).
 - ✓ *Storage Systems* such as the extreme-scale Lustre-based parallel file system (PFS), Spider, which is one of the world's fastest PFS (1 TB/s I/O throughput, 40 PB disk storage capacity and 600 million files) and the large-scale disk-cache/tape-based archival storage system software, HPSS (stores 50 PB, with 60+ million files).
 - ✓ *Technologies for future supercomputers*, e.g., 150 petaflop Summit machine in 2018, and *advanced data hierarchies*, comprising of node-local byte-addressable NVRAM, SSD-based burst buffer tiers, PFS, object storage for data analysis and long-term archival storage.
 - ✓ *Data/Metadata Management and Analytics* by federating metadata from a wealth of resources such as the millions of supercomputer jobs, hundreds of millions of data products, publications and user-specified metadata, using a graph network infrastructure, to answer numerous data disposition questions and discover new knowledge pathways.
 - Led the development of a distributed data management solution for the US DOE's Spallation Neutron Source (SNS), a billion dollar national infrastructure, producing O(100 TB) of data.
 - 60+ research papers in highly selective peer-reviewed conferences/journals, with around 1555 citations and an H-index of 17.
- **Program Development:** Initiated multi-institutional projects worth several million dollars, funded by federal agencies such as the National Science Foundation (NSF), the Department of Energy, and the National Institutes of Health (NIH).
- **Budget Management:** Manage a multi-million dollar annual budget (~ \$6M) and a large-scale storage acquisition budget every four years (> \$12M).
- **Education:** Doctorate in computer science with a focus on distributed storage and data management.
- **Results-oriented:** Ability to take research ideas to products. Experience delivering innovative products under tight time and budget constraints. Software products deployed on national infrastructure.
- **Mentoring & Team Building:** Excellent people management and communication skills. Ability to attract and retain talent. Successful in building dynamic teams to achieve both short/long-term goals. An "*Outstanding Mentor Award*" from the Oak Ridge Institute for Science and Education.
- Experience dealing with sensitive vendor and customer relationships.

3. EDUCATIONAL EXPERIENCE

Doctor of Philosophy in Computer Science, May 2003

Institution: University of Mississippi/Argonne National Laboratory (ANL)

Dissertation: Bulk Data Transfer Forecasts and the Implications to Grid Scheduling

Highlight: Wallace Givens fellowship from ANL; Thesis work on the Globus grid toolkit.

Master of Science in Computer Science, December 1998

Institution: University of Mississippi

Thesis: Performance Oriented Distributed OS-Evolutionary Steps towards a Distributed Linux

Bachelor of Engineering in Computer Science, June 1996

Institution: Karnatak University, India

4. WORK EXPERIENCE

OAK RIDGE NATIONAL LABORATORY (US DOE Lab)

Oak Ridge, TN

2012 – present *Group Leader*, Technology Integration, National Center of Computational Sciences (NCCS)

2003 – 2012 *Research Scientist*, Computer Science and Mathematics Division (CSMD)

THE UNIVERSITY OF TENNESSEE

Knoxville, TN

2010 – present *Joint Faculty*, Joint Institute of Computational Sciences (JICS)

2005 – 2006 *Adjunct Assistant Professor*, Computer Science

ARGONNE NATIONAL LABORATORY (US DOE Lab)

Argonne, IL

2000 – 2002 *Givens Fellow/Doctoral Fellowship*, Math and Computer Science Division

THE UNIVERSITY OF MISSISSIPPI

Oxford, MS

1997 – 2000 *Instructor/Research Assistant*, Department of Computer Science

WIPRO INFOTECH

Bangalore, India

1996 – 1997 *Design Engineer, R&D*

5. R&D MANAGEMENT

• **Group Leader, Technology Integration, Oak Ridge National Laboratory**

Mission: The Technology Integration (TechInt <http://techint.nccs.gov>) group is charged with delivering new technologies in supercomputing and extreme-scale storage systems for the Oak Ridge Leadership Computing Facility (OLCF). The group's technology scope includes parallel file systems, non-volatile memory, architecture, archival storage, data management, and networking. The group is responsible for delivering system software solutions for the Titan supercomputer (world's No. 2 machine: 27 petaflops, 560,640 cores spread across 18,688 compute nodes with 16 AMD cores and a GPU per node and 710 TB DRAM), deploying one of the world's fastest Lustre PFS, Spider, from the block storage up, building the HPSS archival storage software, and technology development for future O(100 petaflop) acquisitions. We also research and evaluate emerging technologies in the aforementioned areas, and provide the systems programming to seamlessly integrate technologies and tools into the infrastructure as they are adopted.

Composition: 25 research staff, with experience ranging from 5-30+ years in systems R&D.

Role: Involves hands-on architecting of technology solutions, program and people management, and maintaining vendor/customer relationships.

- Set a vision for the group, identify R&D directions to pursue by ascertaining gaps in the systems software/hardware stack, conceive project ideas, and conceptualize the basic high-level system design.
- Work with cross-functional teams such as HPC operations, scientific user base and user assistance to gather requirements for products, and develop specifications.
- Build teams for projects, prioritize and assign tasks to staff, monitor progress of deliverables, and mentor staff on technical roadblocks.
- Architect solutions along with the teams, and participate closely on design deep-dives and reviews.
- Responsible for the successful deployment of products, and engage in outreach to improve usage.
- Employee performance management, appraisals, hiring, retention and compensation planning.
- Responsible for relationship management with vendors (e.g., large-scale storage acquisition involves disk, networking, and computer hardware companies), customers (diverse user base that uses our tools and systems), and funding agencies.
- Publish high-quality peer-reviewed papers where appropriate, contribute to program reviews, and develop project proposals to secure funding from sponsor agencies where appropriate.

Growth: In the years as the group lead, I have expanded the group to include more areas other than parallel file systems and archival storage. This includes a significant R&D investment into NVRAM, system architecture, and data management, which involved successfully lobbying for a need to grow in these areas to senior management as well as writing grant proposals to funding agencies.

Impact of group's R&D:

- Software products have been deployed in the OLCF environment (Titan, Spider and other clusters), other DOE labs, industry and several sites worldwide, and being used by thousands of users.
- Our tools have resulted in the efficient use of supercomputer time, which is a precious, rigorously peer-reviewed commodity.
- Software contributions have been adopted into the mainstream codebase by several vendors such as IBM and Intel.
- Our processes on storage system acquisition have been adopted by other labs as best practices.
- Our work has been published in high-quality venues, and cited in industry blogs on storage systems.

Thrust Areas: Leading projects in several thrust areas, towards a robust supercomputing infrastructure (<http://techint.nccs.gov/projects.html>).

A. The Deep Storage Hierarchy: *Developed short-term scratch PFS and associated tools/services, mid-term object storage tier for data analysis and long-term archival storage for curation. The storage solutions developed and deployed are being used by thousands of users for their day-to-day activities at the OLCF HPC center. The solutions address extreme-scale needs in terms of capacity, performance and reliability.*

- **The Spider Lustre-based Parallel File System:**
 - ✓ Design, deployment, tuning and operating one of the world's fastest, HPC center-wide Lustre storage system, *Spider* (1 TB/s, 40 PB), from the block storage up (36 storage system units, SSUs, with 20,160 2 TB near-line SAS disks, organized as 2016 object storage targets, OST, and 288 object storage servers, OSS) and its integration into the Titan supercomputer's 18,688 compute nodes via a scalable I/O routing network infrastructure that comprises of strategically placed 432 I/O routers on the 3D-Torus interconnect fabric, which route the I/O traffic from the compute nodes to the backend storage system.
 - ✓ *I/O workload characterization* to optimize the file system design to serve both the Titan machine as well as other clusters, thereby accommodating a mixed I/O workload (bulk I/O writes from checkpoints and random I/O from data analysis clusters).
 - ✓ *Tools for end-to-end performance tuning* of the various storage system layers (block storage, OST/OSS layer, Lustre PFS layer and I/O routing layer) and performance QA to ensure continued high performance during operations.
 - ✓ *Lustre open-source PFS* code contributions in areas such as dynamic striping, balanced I/O placement based on OST load, checksums, temporal replication, novel data recovery by reconstructing from source in case of unrecoverable failure and use of GPUs for cost-effective distributed RAID. In addition, we perform testing and bug fixing of features for Lustre to ensure that it is viable at Titan's extreme-scale. Our group is the largest code contributor to Lustre outside of Intel.
 - ✓ *Highly parallel tools development* for file system operations (e.g., parallel cp, tar, find, checksums). These tools are being used to operate on petabytes of data and 600 million files.
- **I/O Monitoring, I/O-aware Scheduling and Reliability Modeling for better Provisioning:**
 - ✓ *Scalable I/O Monitoring:* A large-scale I/O monitoring infrastructure by tracking the workload on the backend storage controllers (96) of the Spider storage, and exposing the throughput and load statistics via databases for higher-level services such as I/O-aware tools.
 - ✓ *I/O-aware Scheduling:* Automatic extraction of application I/O signatures from noisy storage server-side logs using statistics and data mining. Use of I/O signatures along with the job scheduler to interleave applications to avoid I/O contention; use of storage controller load to select file system partition for a given job; and a coordinated scheduling mechanism to coincide data staging, computation and data offloading in an attempt to view the PFS as a cache.
 - ✓ *Modeling Storage System Reliability and Provisioning:* An end-to-end simulator to model the Spider PFS reliability and availability. The framework accommodates the various components and subsystems, their interconnections, failure patterns and propagation, and performs dependency analysis to capture a wide-range of failure cases, and predict data unavailability. Instead of the traditional "back of the envelope" calculations, the above framework is used as a tool to provision

- spares, based on component failure rates and repair times, as well as analyze storage system capacity/bandwidth provisioning during initial procurement.
- **HPSS Archival Storage Development:**
 - ✓ Development of the HPSS disk-cache/tape storage system software as part of a consortium that comprises of multiple DOE labs and IBM. HPSS is widely used as the archival system by numerous sites worldwide. HPSS at ORNL stores over 50 PB and 60 million files.
 - ✓ Developed a scalable data quality assurance framework to verify the integrity of 60 million files.
 - ✓ Part of a core advanced technology working group (nine member panel) to shape the future of HPSS. The team will provide recommendations to the consortium on how HPSS should evolve in the next decade, to accommodate the new and emerging storage software and hardware technologies.
 - **Designing the deep storage hierarchy** for the 150 petaflop Summit supercomputer that will be deployed in 2018. This comprises of the design of the SSD-based burst buffer or caching tier, the 120 PB GPFS PFS as a scratch file system and a center-wide object storage solution for longer-term data analysis.
 - **Creating distributed index services** on the distributed file system, GlusterFS, storage servers using KV databases. Designing the ability to have user-specific views or tags within GlusterFS, enabling a science object view of the underlying file system.
 - B. SSDs and Non-Volatile Memory (NVRAM): Devising NVRAM solutions for next-generation supercomputers, from three perspectives, namely fault tolerance, memory extension, and active processing, using block-addressable flash on SATA, PCIe, NVMe, and byte-addressable NVRAM on DIMMs.**
 - **Fault Tolerance—SSD-based Burst Buffer Storage:** Design and development of a compute-node local, SSD as a burst buffer tier for fault tolerance of large-scale applications. The burst buffer absorbs the bulk data at high-speeds while seamlessly draining the data to a PFS. Developing a lightweight, relaxed POSIX distributed storage to be layered atop the node-local SSDs. This burst buffer solution will be deployed on the 2.5 PB of aggregate SSD space on the Summit supercomputer, distributed across the ~ 3500 computer nodes.
 - **Memory Extension—NVMalloc:** A runtime library for applications to use SSDs as a secondary memory partition by explicitly allocating variables on an NVRAM store. Built techniques to map a byte-addressable interface to a block store.
 - **Active Processing:**
 - ✓ *Active Flash:* Developed mechanisms to push computation into the flash device controller where the data already resides, facilitating in-situ data analysis, and minimizing energy consumed due to data movement. Developed several scheduling schemes in FTL to perform on-the-fly analysis. Developed an active storage target framework based on the SCSI T10 OSD-2 specification, and modified exofs in Linux to support active processing.
 - ✓ *Analysis-aware Storage System (AnalyzeThis):* Built an analytics workflow-aware storage appliance using an array of active flash devices (funded by the DOE's Scientific Data Management program). Developed a storage system with analysis awareness deeply embedded within every layer: an analysis object abstraction atop the active flash array to tie together the data, and the analytics to be (or was) performed on the data, using the OSD interface and SQLite databases; a workflow scheduler within the storage dispatches analysis jobs to the active flash fabric in a manner that minimizes data movement; a FUSE file system interface, *anFS*, with which users can submit analysis workflow jobs, write data, and monitor the status of the jobs (akin to `/proc`). Introducing AnalyzeThis concepts into a distributed file system, GlusterFS.
 - ✓ *Shared Memory for Processing in an array of byte-addressable NVRAM (AnalyzeThat):* Leading the development of a software shared memory infrastructure atop an array of byte-addressable NVRAM (e.g., PCM), capable of active processing. Building key-value and map-reduce programmable interfaces as well as a smart, lifetime and data movement aware scheduling runtime atop the shared memory infrastructure.
 - **Wear leveling:** Studying the impact on NVRAM lifetime based on multiple dimensions: when the device is used for all of the above workloads, placed either locally on the compute node or centrally on the I/O nodes to be shared by tens of thousands of compute nodes, and using different wear-management schemes in the FTL. Devising techniques to improve wear management based on the insights gained.

- C. Data and Metadata Management and Mining:** *Developed and deployed solutions for large-scale data search and analytics, metadata extraction, derivation and indexing, data curation and log analysis.*
- **Constellation Science Graph Network:** The overarching goal is the creation of a transformative “*science graph network*” that federates metadata from resources (e.g., systems, users, data, processes and lifecycle artifacts) in a scientific supercomputing collaboration, and builds complex associations between them by exploiting graph properties and performing deep mining and graph data analytics, to enable the scalable discovery of data and new knowledge pathways.
 - ✓ Metadata collected from the supercomputing resource infrastructure (600+ millions files from file systems, millions of jobs from schedulers, publications, users, “tags”, DOIs, etc.), and capturing it in a graph database, and deriving more metadata from the base metadata to create hierarchical databases indexes (SOLR). For example, metadata from within scientific data (HDF, NetCDF) is extracted, and then composite metadata (e.g., average, mean median, ranges and histograms) is automatically created across collections of datasets. Such indexes can be used to answer sophisticated queries on datasets, without human intervention.
 - ✓ Data mining and graph topological and temporal analytics are performed on the metadata to infer complex relationships between the resources, to make the associations in the science graph (e.g., are these data products related to this dataset? These datasets and jobs seem to overlap in time; are they related?).
 - ✓ Based on this graph engine, we can answer science queries such as “how did the temperature of the ice sheet vary between six months worth of job runs?” and even predict future data products of interest, e.g., a climate user who discovered a “hurricane” in one datasets from among the millions can be presented with other related datasets that may also potentially contain hurricanes?
 - **DOIs for HPC Data:** Developed a data curation process for long-term storage of HPC data. To this end, developed a Digital Object Identifier workflow as a means to identify and curate important data products. DOI provides a way to cite data products and the associated processes.
 - **Data Analytics on large-scale Logs to Optimize HPC Data center Operations:** Developed a scalable infrastructure to aggregate and analyze huge amounts of storage and supercomputer logs to present a “*window into and optimize HPC data center operations*” for OLCF. Logs include the Spider storage system I/O bandwidth information from the controllers, Titan supercomputer CPU/GPU RAS data, Moab job scheduler logs and job node allocation data. Created a Splunk infrastructure to aggregate the logs, and developed higher-level services to analyze and present them, e.g., a visual analytics tools to view storage system performance, query reliability information on compute node failure over time, visualize node allocation partitions for jobs, and perform data mining to decipher how fragmented the node allocations are. The tools are being used to identify hotspots and debug/resolve performance bottlenecks, e.g., the visual analytics of the Spider I/O bandwidth was used to identify and resolve the suboptimal use of the OSTs by a “hero” fusion application job that used 256,000 cores and produced 12 PB in 12 hours, improving the I/O performance by 15%.
- D. Computer System Architecture and Resilience:** *Developed software systems for the efficient use of the Titan supercomputer’s 560,640 CPU/GPU cores in the areas of runtime, scheduling and resilience.*
- **Core Pinning:** A Titan node has 16 cores and 8 FPUs, resulting in FPU sharing. The FPU sharing cannot be avoided if a job uses all the cores/node, however, when all the cores are not used we can use the alternate cores to avoid FPU contention. Developed a tool to help the multiple MPI processes of a job not compete with each other in sharing the FPUs on the AMD compute node on the Titan supercomputer. This tool is helping tens of thousands of user jobs on a day-to-day basis.
 - **Functional Partitioning Runtime for Many-core nodes:** On supercomputers, there is no timesharing of nodes due to jitter on the jobs. We have developed a runtime environment, Functional Partitioning (FP), to partition a many-core node such that an end-to-end application (simulation + data analysis tasks) data analyses can be scheduled on the same node, in-situ, alongside the application’s simulation job for better end-to-end performance. We exploit the CPU/GPU cores and memory that are under-utilized by the application’s simulation, and puts them to use for the application’s own end-to-end data analysis tasks. We can also perform I/O aggregation as a runtime service to streamline I/O from the many-core nodes, assign FP services closer to hardware resources using the hwloc. Since data analysis is an integer operation, it does not compete for the FPU of the node that is used by the FLOP-intensive simulation job. Runtime tool deployed on Titan, and used by large-scale Climate application jobs.

- **Dual Window scheduling policy** deployed on the Moab scheduler on Titan. Traditional policy schedules large and small jobs top down (1 to 18,688 nodes), which causes fragmentation of the large node jobs due to small node jobs coming and going. We have devised an algorithm that schedules large jobs top down and small jobs bottom up, thereby minimizing the fragmentation for large, long-running jobs that is caused by small, short-lived jobs. The policy has reduced the average hop count for larger jobs (1K, 2K, 4K, 8K, etc), without impacting the utilization of Titan. Thousands of jobs on Titan are benefitting due to this scheduling technique.
- **Supercomputer Failure Analysis:** Analysis of temporal/spatial failure characteristics of Titan's 560,640 CPU/GPU cores to understand trends in machine failure and MTBF. Studied SBEs, DBEs, temperature correlation to failure, etc. Based on the insights, devised novel techniques on when to do defensive checkpoint I/O and backfilling jobs. For example, we have proposed that applications relax their hourly checkpoint requirement, and instead perform checkpoints based on the machine failure rate and I/O rate, which can drastically reduce the data written to a PFS, and allow an application to spend more time computing instead of defensive I/O. Based on the GPU failure analysis, we found that there is a temperature correlation to GPU failures. To this end, we have devised an energy-based scheduling algorithm to schedule large node-count GPU jobs that stress the GPUs more to be scheduled at the bottom of the rack where it is much cooler than the top of the rack.
- **Heterogeneous Processor Architectures and End-to-end Computing:** Studying how end-to-end workflows (large-scale simulation jobs followed by data analysis) perform on heterogeneous supercomputer designs: HPC center-level (e.g., application simulations on Titan and data analyses on clusters in the HPC center), machine-level (separate partition of nodes for simulation and data analysis on the same machine, e.g., Cori machine at NERSC) and node-level heterogeneous architectures (CPU/GPU nodes conducting simulation and data analysis in situ on the Titan node itself) using a variety of processor architectures such as big processors (AMD Opterons, XEONs), Coprocessors (GPUs, MIC) and little processors (ARM, ATOM). Our study considers several competing dimensions such as performance, on-chip scalability, energy-efficiency, and error resiliency to provide insights on the design of future heterogeneous supercomputers for end-to-end workflow execution.

- **Research Scientist, Computer Science Research, Oak Ridge National Laboratory**

In this role, served as the principal investigator on several storage, and data management projects. Led several multi-institutional projects comprising of several graduate students, post-masters, postdoc, and university faculty researchers. I have been responsible for setting goals, software development, mentoring, monitoring progress, and publishing papers. Served on several masters and doctoral committees of graduate students from North Carolina State University, Virginia Tech, Pennsylvania State University, and Northeastern University. For my mentoring efforts, I was awarded the “*Outstanding Mentor Award*” by the Oak Ridge Institute for Science and Education in 2008. Selected projects are listed below.

- A. **FreeLoader Cache using Distributed Storage Scavenging:** Led the design and development of an aggregated, distributed storage infrastructure as a client-side cache by data-intensive applications. The basic idea is the aggregation of space and I/O bandwidth contributions from commodity desktops within a domain to provide a mountable (through FUSE), highly-available, shared cache/scratch space for large, immutable scientific data sets that can be accessed in parallel. Built novel techniques such as asymmetric striping, prefix caching with suffix patching from remote sources, and cost-of-recovery of a dataset to determine its redundancy scheme.
- B. **Scientific Data Management for SNS:** Lead architect for a scientific gateway, a distributed computation solution and data management for O(100 TB) from the Spallation Neutron Source (SNS), a billion dollar infrastructure, catering to a large user community.
- C. **Networking: Staged Data Transfer.** Expedite the end-user data delivery between HPC centers and end-user locations, thereby alleviating the last-mile problem. Developed decentralized data-offloading and just-in-time staging schemes to move job output and input data by reconciling center purge policies, user delivery and job startup deadlines. Combined point-to-point transfer tools (e.g., GridFTP) and decentralized schemes (e.g., BitTorrent) along with NWS to enable timely data delivery. Using such a scheme, users can exploit advanced networks (100Gb/s) and commodity networks in a seamless fashion to deliver data in time.
- D. **Cloud Storage:** Led the development of a FUSE file system atop Azure cloud storage. Developed ways to use the cloud storage as intermediate nodes for data delivery from HPC centers to end-users.

- **Joint Faculty Associate Professor, The University of Tennessee**

As an Associate Professor, I work with students and postdocs on several problems in storage and data management. The National Science Foundation and the National Institutes of Health fund these projects.

- A. **Distributed NVRAM/DRAM-based staging Storage:** Led a multi-institutional National Science Foundation's High-end Computing File systems and I/O project. Built a distributed, lightweight, intermediate staging storage system for supercomputers, using SSDs and DRAM tiers, in order to alleviate the I/O bottleneck in checkpointing and data analysis. Built novel incremental checkpointing support using checksum comparisons of chunks (de-dup).
- B. Technical lead and Co-PI on an NIH effort to build a data repository so that National Cancer Institute (NCI) funded centers can aggregate, annotate, search, index, and download data.

- **Doctoral Thesis at Argonne National Laboratory:** Developed middleware for the orchestration of bulk data transfers in the Globus Data Grid. The middleware comprised of a scalable storage broker and selection heuristics for locating widely replicated data, statistical models and tools to predict the performance of wide-area data transfer times, and techniques for co-allocating transfers.

- **Masters Thesis:** Developed a distributed OS for Linux. Built a high-speed communication protocol for a Linux cluster by short-circuiting the protocol stack; a networked file system, global IPC mechanism, group communication and remote process execution environment using the communication scheme. Patches for the 2.0 kernel. *Team:* Led several graduate students.

6. **SELECTED PUBLICATIONS** (full list at <http://users.nccs.gov/~vazhkuda/publications.html>)

1. H. Sim, Y. Kim, S.S. Vazhkudai, D. Tiwari, A. Anwar, A.R. Butt, L. Ramakrishnan, "AnalyzeThis: An Analysis Workflow-Aware Storage System," Supercomputing 2015 (SC'15): 28th IEEE/ACM Int'l Conference on High Performance Computing, Networking, Storage and Analysis, Austin, Texas, November 2015.
2. L. Wan, F. Wang, S. Oral, D. Tiwari, S.S. Vazhkudai, Q. Cao, "A Practical Approach to Reconciling Availability, Performance, and Capacity in Provisioning Extreme-scale Storage Systems," Supercomputing 2015 (SC'15), Austin, Texas, November 2015.
3. D. Tiwari, et. al., "Understanding GPU Errors on Large-scale HPC Systems and the Implications for System Design and Operation", 21st IEEE High Performance Computer Architecture (HPCA'15), California, February, 2015.
4. S. Oral, et. al., "Best Practices and Lessons Learned from Deploying and Operating Large-Scale Data-Centric Parallel File Systems", Supercomputing 2014 (SC'14), New Orleans, Louisiana, November 2014.
5. Devesh Tiwari, Saurabh Gupta, Sudharshan S. Vazhkudai, "Lazy Checkpointing: Exploiting Temporal Locality in Failures to Mitigate Checkpointing Overheads on Extreme-Scale Systems," 44th Annual IEEE/IFIP Conference on Dependable Systems and Networks (DSN'14), Atlanta, Georgia, June 2014.
6. Y. Liu, R. Gunasekaran, X. Ma, S.S. Vazhkudai, "Automatic Identification of Applications I/O Signatures from Noisy Server-Side Traces", 12th USENIX Conference on File and Storage Technologies (FAST'14), Santa Clara, California, February 2014.
7. D. Tiwari, S. Boboila, S.S. Vazhkudai, Y. Kim, X. Ma, P. Desnoyers and Y. Solihin, "Active Flash: Towards Energy-Efficient, In-Situ Data Analytics on Extreme-Scale Machines," 11th USENIX Conference on File and Storage Technologies (FAST'13), San Jose, California, February 2013.
8. A. Khasymski, M.M. Rafique, A.R. Butt, S.S. Vazhkudai, D.S. Nikolopoulos, "On the Use of GPUs in Realizing Cost-Effective Distributed RAID," IEEE Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS'12), Washington, D.C., August 2012.
9. C. Wang, S.S. Vazhkudai, X. Ma, F. Meng, Y. Kim, C. Engelmann, "NVMalloc: Exposing an Aggregate SSD Store as a Memory Partition in Extreme-Scale Machines," 26th IEEE Int'l Parallel & Distributed Processing Symposium (IPDPS'12), Shanghai, China, May 2012.
10. S. Boboila, Y. Kim, S.S. Vazhkudai, P.J. Desnoyers, G. Shipman, "Active Flash: Out-of-core Data Analytics on Flash Storage," 28th IEEE Conference on Mass Storage Systems and Technologies (MSST'12), Monterey, CA, April 2012.
11. H. Monti, A.R. Butt, S.S. Vazhkudai, "Timely Result-Data Offloading for Improved HPC Center Scratch Provisioning and Serviceability," IEEE Transactions on Parallel and Distributed Systems (TPDS'11), Vol. 22, No. 8, pp. 1307-1322, August 2011.

12. R. Prabhakar, S. S. Vazhkudai, Y. Kim, A.R. Butt, M. Li, M. Kandemir, "Provisioning a Multi-Tiered Data Staging Area for Extreme-Scale Machines," 31st Int'l Conference on Distributed Computing Systems (**ICDCS'11**), Minneapolis, MN, June 2011.
13. H. Monti, A.R. Butt, S.S. Vazhkudai, "CATCH: A Cloud-based Adaptive Data Transfer Service for HPC," 25th IEEE Int'l Parallel & Distributed Processing Symposium (**IPDPS'11**), Anchorage, AK, May 2011.
14. M. Li, S. S. Vazhkudai, A.R. Butt, F. Meng, X. Ma, Y. Kim, C. Engelmann, G. Shipman, "Functional Partitioning to Optimize End-to-End Performance on Many-core Architectures," Supercomputing 2010 (**SC'10**), New Orleans, LA, November 2010.
15. X. Ma, S.S. Vazhkudai, Z. Zhang, "Improving Data Availability for Better Access Performance: A Study on Caching Scientific Data on Distributed Desktop Workstations," **Journal of Grid Computing** - Special Issue on Volunteer Computing and Desktop Grids, Vol. 7, No. 4, pp. 419-438, December 2009.
16. H. Monti, A.R. Butt, S.S. Vazhkudai, "Scratch as a Cache: Rethinking HPC Center Scratch Storage," 23rd ACM Conference on Supercomputing (**ICS'09**), Yorktown Heights, NY, June 2009.
17. S.A. Kiswany, M. Ripeanu, S. S. Vazhkudai, A. Gharaibeh, "stdchk: A Checkpoint Storage System for Desktop Grid Computing," 28th Conference on Distributed Computing Systems (**ICDCS'08**), Beijing, China, June 2008.
18. Z. Zhang, C. Wang, S. S. Vazhkudai, X. Ma, G. Pike, F. Mueller, J.W. Cobb, "Optimizing Center Performance through Coordinated Data Staging, Scheduling and Recovery," Supercomputing 2007 (**SC'07**), Reno, NV, November 2007.
19. S. Vazhkudai, X. Ma, "Recovering Transient Data: Automated On-demand Data Reconstruction and Offloading on Supercomputers," **Operating Systems Review**: Special Issue on File and Storage Systems, Vol. 41, No. 1, pp. 14-18, January 2007.
20. S. Vazhkudai, X. Ma, V. Freeh, J. Strickland, N. Tammineedi, T.A. Simon, S.L. Scott, "Constructing Collaborative Desktop Storage Caches for Large Scientific Datasets," **ACM Transactions on Storage (TOS'06)**, Volume 2, No. 3, pp. 221-254, August 2006.
21. X. Ma, S. S. Vazhkudai, V. Freeh, T.A. Simon, T. Yang, S.L. Scott, "Coupling Prefix Caching and Collective Downloads for Remote Data Access," 20th ACM Conference on Supercomputing (**ICS'06**), pp. 229-238, Cairns, Australia, June 2006.
22. S. Vazhkudai, X. Ma, V. Freeh, J. Strickland, N. Tammineedi, S.L. Scott, "FreeLoader: Scavenging Desktop Storage Resources for Scientific Data," of Supercomputing 2005 (**SC'05**), Seattle, WA, November 2005.
23. S. Vazhkudai, J. Schopf, "Predicting Sporadic Grid Data Transfers," 11th IEEE Symposium on High Performance Distributed Computing (**HPDC'02**), Edinburgh, Scotland, July 2002.
24. S. Vazhkudai, J. Schopf, I. Foster, "Predicting the Performance of Wide-Area Data Transfers," 16th Parallel and Distributed Processing Symposium (**IPDPS'02**), Fort Lauderdale, Florida, April 2002.
25. S. Vazhkudai, S. Tuecke, I. Foster, "Replica Selection in the Globus Data Grid," IEEE Conference on Cluster Computing and the Grid (**CCGRID'01**), Brisbane, Australia, May 2001.

7. **SELECTED PROGRAM DEVELOPMENT ACTIVITIES**

Initiated several multi-institutional projects by obtaining external grants from DOE, NSF, and NIH.

1. L. Ramakrishnan, D. Agarwal, S.S. Vazhkudai, M. Franklin, C. Aragon, "Usable Data Abstractions for Next-Generation Scientific Workflows," DOE ASCR Scientific Data Management Program Announcement LAB 14-1043, 09/2014-08/2017 (Role: Co-PI).
2. S.S. Vazhkudai, X. Ma, D.K. Panda, "Dynamic Staging Architecture for Accelerating I/O Pipelines," *NSF High-End Computing Research Activity (HECURA)*, CCF-0937827, 04/2010-03/2013 (Role: PI).
3. X. Ma, Y. Zhou, V.W. Freeh, S. Vazhkudai, "Application-adaptive I/O Stack for High End Computing," *NSF HECURA*, CCF-0621470, FY 2007-2009 (Role: Co-PI).
4. S. Vazhkudai, X. Ma, J.W. Cobb, G. Pike, "Storage Virtualization: An Integrated Approach to Machine-Room Storage Management," *DOE ORNL LDRD*, FY 2007-2008 (Role: PI).

8. **AWARDS**

- Outstanding Mentor Award (Feb 2008) - from the Oak Ridge Institute for Science and Education.
- Ph.D. Dissertation Fellowship (2001 - 2002) - Argonne National Laboratory (*Data Grid Research*).
- Wallace Givens Fellowship (2000) - Argonne National Laboratory