Lavanya Subramanian

Electrical & Computer Engineering Department
Phone: (412) 889 2792
Email: lsubrama@andrew.cmu.edu
Web: www.ece.cmu.edu/~lsubrama/

Carnegie Mellon University
5000 Forbes Avenue
Collaborative Innovation Center 4th Floor
Pittsburgh PA 15213

Research Interests

Computer architecture, memory systems, quality of service (QoS), heterogeneous systems

Education

Carnegie Mellon University [Aug. 2009 - Aug. 2015 (expected)]
Ph.D., Electrical and Computer Engineering
Thesis: Achieving high and predictable performance in multicore systems through shared resource management
Advisor: Prof. Onur Mutlu

Carnegie Mellon University [Aug. 2009 - Aug. 2012]
M.S., Electrical and Computer Engineering

Madras Institute of Technology, Anna University [July 2003 - May 2007]
Bachelor of Engineering, Electronics & Communication Engineering

Research Experience

Graduate Student Researcher, Carnegie Mellon University [Aug. 2009 - present]
High and predictable performance in the presence of shared resources
During my PhD, I have explored different approaches to tackle the problem of shared resource interference in multicore systems, as part of my thesis. I have worked on different solution directions such as memory channel partitioning, low-complexity memory scheduling with the goals of mitigating memory interference and improving performance/fairness. Most recently, I have been focusing on the problem of achieving predictable performance in the presence of shared resource interference. Towards this end, we have developed a model to estimate application slowdowns due to main memory and shared cache interference and mechanisms that leverage this model to bound an application's slowdown.
Besides my work on such homogeneous multicore systems, I have also explored the memory interference problem in heterogeneous systems with CPUs, GPUs and hardware accelerators, in collaboration with other members of my research group. We have designed a memory scheduler that mitigates interference between CPU and GPU requests. Most recently, we have been exploring the problem of meeting deadlines for hardware accelerators and GPUs, while still achieving high CPU performance.

Work Experience

Intel Labs, Intern [Summer 2011]
Analyzed the impact of sharing between different agents in a multicore system. The goal of the project was to study the sensitivity of different cores to shared cache capacity. We performed experiments on a CPU-GPU system to understand the sensitivity of CPU cores and the GPU to cache capacity allocation. Based on our studies, we proposed general guidelines for partitioning the shared cache capacity between the CPU and the GPU. This work resulted in an internal Intel publication.

Microsoft Research, Intern [Summer 2010]

Worked on architecture level performance analysis and optimization of web search applications. The goal of this project was to understand the memory access characteristics of search queries and optimize the design of the memory system to achieve high performance and low energy. We observed that there were hot regions of memory that constituted a large percent of memory accesses and proposed to keep only such data in always-on memory partitions.

SanDisk, ASIC Design Engineer [July 2007 - June 2009]
Worked on the design of flash memory controller chips used in SD cards, memory sticks, SSDs and USB drives. I was primarily involved in RTL design, verification, design for testability (DFT) and test vector generation of 3 chips in TSMC and UMC 130 and 90 nm processes. The ASIC design team in Sandisk India was small and young. This gave me the opportunity to work on different aspects of the ASIC design flow.

Publications

A-DRM: Architecture-aware Distributed Resource Management of Virtualized Clusters [VEE'15]
Hui Wang, Canturk Isci, **Lavanya Subramanian**, Jongmoo Choi, Depei Qian, Onur Mutlu

The Blacklisting Memory Scheduler: Achieving High Performance and Fairness at Low Cost [ICCD'14]
Lavanya Subramanian, Donghyuk Lee, Vivek Seshadri, Harsha Rastogi, Onur Mutlu

MISE: Providing Performance Predictability and Improving Fairness in Shared Main Memory Systems [HPCA'13]
Lavanya Subramanian, Vivek Seshadri, Yoongu Kim, Ben Jaiyen, Onur Mutlu

Tiered-Latency DRAM: A Low Latency and Low Cost DRAM Architecture [HPCA'13]
Donghyuk Lee, Yoongu Kim, Vivek Seshadri, Jamie Liu, **Lavanya Subramanian**, Onur Mutlu

Staged Memory Scheduling: Achieving High Performance and Scalability in Heterogeneous Systems [ISCA'12]
Rachata Ausavarungnirun, Kevin Chang, **Lavanya Subramanian**, Gabriel Loh, Onur Mutlu

Reducing Memory Interference in Multicore Systems via Memory Channel Partitioning [MICRO'11]
Sai Prashanth Muralidhara, **Lavanya Subramanian**, Onur Mutlu, Mahmut Kandemir, Thomas Moscibroda

The Application Slowdown Model: Quantifying and Mitigating the Impact of Inter-Application Interference at Shared Caches and Main Memory [Under submission]

SQUASH: Simple, QoS-Aware, High-Performance Memory Scheduler For Heterogeneous Systems with Hardware Accelerators [Under submission]

Decoupled Direct Memory Access: Isolating CPU and IO Traffic by Leveraging a Dual-Port DRAM [Under submission]

Invited Papers

The Main Memory System: Challenges and Opportunities [Communications of the KIISE, 2015]
Onur Mutlu, Justin Meza, **Lavanya Subramanian**

Research Problems and Opportunities in Memory Systems [Superfri, 2014]
Onur Mutlu, **Lavanya Subramanian**

Select Technical Reports

The Blacklisting Memory Scheduler: Balancing Performance, Fairness and Complexity
[SAFARI Technical Report No. 2015-004: March, 2015]
Lavanya Subramanian, Donghyuk Lee, Vivek Seshadri, Harsha Rastogi, Onur Mutlu

SQUASH: Simple QoS-Aware High-Performance Memory Scheduler for Heterogeneous Systems with Hardware Accelerators
[SAFARI Technical Report No. 2015-003: March, 2015]
Hiroyuki Usui, **Lavanya Subramanian**, Kevin Chang, Onur Mutlu

Select Projects

Predictable performance in multicore systems:
In a multicore system, multiple applications running on different cores interfere at shared resources, resulting in unpredictable application performance. Our goal, in this project, is to address this performance unpredictability. As a first step towards achieving this goal, we built a model to estimate application slowdowns due to main memory bandwidth interference and resource management schemes that leverage our model to appropriately allocate memory bandwidth to different applications. As a next step, we have extended this model to account for shared cache capacity contention as well. We are currently working on leveraging our combined memory/cache model to perform resource allocation.

Simple and high-performance memory management in multicore systems:
The DRAM main memory is a critical bottleneck and heavily contended resource in multicore systems. Prior work has proposed application-aware memory request scheduling to mitigate this contention. However, application-aware memory schedulers incur high hardware implementation cost. In this project, we tackle this problem and build a simple memory request scheduling technique to mitigate memory contention and achieve high performance. Our scheduler achieves performance and fairness improvements, while reducing critical path latency and area (from RTL implementations) significantly.

Application-aware memory channel partitioning:
This research project tackles the problem of contention for main memory, among applications in a multicore system too, but from an orthogonal perspective. The proposed solution is to partition the memory such that interfering applications'

data are allocated to different memory channels. This is a drastically different approach than previous proposals that aim to reorder memory requests to mitigate memory contention and improve performance. This approach achieves performance on par with the best previous memory schedulers, while requiring no modifications to the scheduler hardware.

QoS-aware memory scheduling in heterogeneous systems:
This research project tackles the problem of providing QoS to different agents such as hardware accelerators, GPUs, while still achieving high CPU performance, when different agents share and interfere at the main memory in a heterogeneous SoC setting. We propose to tackle this problem by prioritizing hardware accelerators' memory requests when they are not on track to meet their deadlines, while prioritizing CPU applications' requests when hardware accelerators are making sufficient progress to meet their deadlines. We also observe that different hardware accelerators have different access characteristics and appropriately prioritize between different hardware accelerators' requests. Our evaluations show that we are able to meet 99.9% of hardware accelerators' deadlines, while still achieving high CPU performance.

Data mapping in database systems:
This project was carried out as part of a course on big data systems. We, as a group of 3 students, observed that when data is mapped to physical memory (DRAM) in a database system, several key characteristics of the memory are not leveraged. We built a data mapping scheme for the Redis in-memory database that exploited specific access characteristics of DRAM main memory systems to achieve high performance. Our results showed that our proposed scheme was promising and significant performance benefits could be reaped by taking into account the access characteristics of the memory technology being used.

Architecture, design and technology co-optimization:
This project was part of a course on technology foundations for System On Chip (SoC) design. The project group consisted of 4 students from different specializations, namely, process technology development, circuit design and computer architecture. The goal of this project was to come up with the high-level design of a product for the server market, taking into account architecture, technology, cost considerations. Our recommendations, after thorough explorations, were well-received by the class and helped us gain significant insights into the different product design considerations.

Flash memory controller design:
During my two years at SanDisk India, I worked on different aspects of front end SoC design such as SoC performance analysis, RTL design, verification and some aspects of DFT/test, of 3 chips that were taped out in 90 nm and 130 nm TSMC and UMC processes. These designs were targeted towards USB drives, memory sticks and SD cards. The team in India was only a year old when I joined them and the job role was very flexible. This gave me the opportunity to get a good sense of the entire IC design flow, even though my main charter was front end design.

Relevant Graduate Course Work

Computer Architecture, Parallel Computer Architecture, Energy Aware Computing, Probabilistic Modeling, Performance Modeling, Big Data Systems, Technology Foundations For SoC Design

Select Talks

The Blacklisting Memory Scheduler: Achieving High Performance and Fairness at Low Cost	[ICCD'14]
MISE: Providing Performance Predictability in Shared Main Memory Systems	[HPCA'13]
MISE: Providing Performance Predictability in Shared Main Memory Systems	[PDL Retreat, 2012]
Towards Fairness and Performance Predictability in Multicore Memory Systems	[VMware, 2012]
Reducing Memory Interference in Multicore Systems via Memory Channel Partitioning	[MICRO'11]

Honors and Awards

John and Claire Bertucci Fellowship (2013)
Carnegie Institute of Technology Dean's Fellowship (2009)

Service and Leadership

Helped prepare SRC/NSF research proposal submissions
Reviewer for IEEE TC, ACM TACO and ACM TECS
Mentored several younger graduate students and undergraduate students
Served on the President's Student Advisory Council (PSAC) at CMU
Served as treasurer of the Society for Promotion of Indian Classical Music Among Youth (SPICMACAY), CMU Chapter