Sumanth Umesh

Computer Science PhD Student
University of Michigan, Ann Arbor
sumanthu@umich.edu | +1 (734) 834-6197
Google Scholar | Linkedin | Personal Website

Summary

My primary research focus is on optimizing memory architecture for improving the performance of memory bandwidth-bound or latency-bound workloads. My current work involves using Compute Express Link (CXL) for memory expansion and host-accelerator/accelerator-accelerator interfaces. My goal is to uncover methodologies and strategies that tackle issues such as memory shortage and bandwidth bottlenecks, ultimately contributing to more efficient and effective computing systems

Education

University of Michigan, Ann Arbor
PhD in Computer Science **[Aug 2022 - Present]**
Advisor: Prof. Reetuparna Das
GPA: 3.72/4.0

Indian Institue of Technology, Jodhpur
BTech in Electrical Engineering **[Aug 2016 - May 2020]**
Advisor: Prof. Shree Prakash Tiwari
GPA: 9.57/10.0

Research Experience

Graduate Student Research Assistant
University of Michigan, Ann Arbor **[Aug 2022 - Present]**
Advisor: **Prof. Reetuparna Das**

Project: CXL Based Memory Expansion for Databases
- Working on a CXL-based memory expanded system to accelerate queries on large main-memory databases that overflow to storage on conventional systems
- Evaluated strategies to take advantage of device level parallelism among CXL devices and migrate data chunks between CXL attached memory and DRAM based on query workload analysis
- Implemented an extensible CXL.mem simulator (using Ramulator for DRAM simulations) to evaluate the above strategies

Project: CXL-enabled Large Language Model Accelerator
- Designed processing in memory (PIM) platform for LLM inference as an alternative to GPU-based platforms
- Developed the CXL topology to connect multiple PIM devices to accommodate sufficient memory for LLMs and to enable inter-PIM device communication
- Proposed modifications and workarounds over base CXL.mem protocol to support broadcast and multicast transactions to reduce network traffic and improve latency

Work Experience

ASIC Engineer
NVIDIA, Bangalore, India **[Aug 2021 - Jul 2022]**
- Designed microarchitecture and RTL of forward error correction modules in the datalink layer of PCIe 6.0
- Resolved datalink layer timing and synthesis issues for PCIe - Gen 5 controller

Design Engineer
Silicon Labs, Hyderabad, India **[Aug 2020 - Aug 2021]**
- Designed microarchitecture and RTL of security accelerators (SHA3, Poly1305 and ChaCha20) for low-power wireless SoC
- Handled design quality, synthesis and timing checks for security sub-module
- Wrote firmware for the host ThreadArch domain-specific processor

Research Intern

Bosch Corporate Research, Bangalore, India **[May 2019 - Jul 2019]**

- Designed parameterized and scalable arithmetic modules (adder, subtractor and multiplier) for posit numbers as an alternative to floating point
- Optimized the design to deliver performance comparable to floating point with lower resource utilization on an FPGA

Publications

- A. Khadem, Y. Gu, **S. Umesh**, N. Liang, X. Servot, O. Mutlu, R. Iyer, R. Das. "Cellar: CXL-enabled Large Language Model Accelerator". *International Symposium on Computer Architecture (ISCA) 2024* *Under review
- **S. Umesh**, and S. Mittal. "A survey of techniques for intermittent computing." Journal of Systems Architecture 112 (2021): 101859
- **S. Umesh**, and S. Mittal. "A survey of spintronic architectures for processing-in-memory and neural networks." Journal of Systems Architecture 97 (2019): 349-372
- S. Mittal, and **S. Umesh**. "A survey on hardware accelerators and optimization techniques for RNNs." Journal of Systems Architecture 112 (2021): 101839

Skills

Programming Languages	:	C • C++ • Python
HDL	:	Verilog • Systemverilog
Software	:	Synopsys VCS/Design Compiler • Xilinx Vivado
Simulator	:	Ramulator • ZSim
Profiling	:	VTune • PIN • Nsight Compute

Projects

N-Way Superscalar RISC-V Core

[Aug 2022 - Dec 2022]

EECS 470 | Prof Ronald Dreslinski

- Designed microarchitecture and RTL for a N-Way superscalar RISC-V core based on MIPS R10K architecture
- Implemented the memory subsystem with a non-blocking data cache, victim cache, multi-ported instruction cache with prefetching and branch predictor

MESI Cache Coherence Protocol

[Jan 2023 - Apr 2023]

EECS 570 | Prof Yatin Manerkar

- Developed and verified a directory-based cache coherence protocol including Modified, Exclusive, Shared, and Invalid states in Murphi
- Optimized the protocol to a Nack-Free coherency model featuring silent Exclusive to Modified state transitions

Whether and How to In-Cache Compute

[Aug 2022 - Dec 2022]

EECS 583 | Prof Scott Mahlke

- Developed an auto-vectorization compiler pass to target an in-cache compute accelerator in LLVM with the help of polyhedral compilation
- Designed a cost model to automatically decide if a task should be offloaded to in-cache compute by estimating the cost and benefit of such an offlaod

Academic Achievements

- **Gold medal for the best all-round performance in the class of 2020, IIT Jodhpur**
- **Silver medal for best academic performance in Electrical Engineering, IIT Jodhpur**
- **National Talent Search Examination (NTSE) Scholarship 2014 (awarded to 1000 high school students in India)**

Relevant Coursework

- Computer Architecture (470)
- Advanced Databases (584)
- Parallel Computer Architecture (570)
- Advanced Compilers (583)