# Vignesh Suresh

Email: vv15@illinois.edu
Website: https://publish.illinois.edu/vigneshsuresh

---

## Education

**Ph.D., Electrical and Computer Engineering** Jan 2021 - Present
University of Illinois, Urbana-Champaign
Advisor: Prof. Sarita Adve

**B.Tech., Electronics and Communication Engineering** Jul 2013 - Jun 2017
Vellore Institute of Technology, India

---

## Research Interests

I am broadly interested in computer system design. Specifically, I am interested in the design of heterogeneous systems-on-chip (SoC), including the design of efficient hardware accelerators, their interaction with general-purpose CPUs and the memory system. My current work aims to improve aspects of coherent shared memory and the role of the operating system (OS) in the context of heterogeneous SoCs. For my future work, I would like to co-design systems with emerging applications, like Extended Reality (XR) and deep learning techniques (e.g., NeRF, Generative models).

---

## Publications

- **Taming the Acceleration Tax: Enabling New Opportunities for Accelerator-Level Parallelism. Vignesh Suresh**, Bakshree Mishra, Zeran Zhu, Ying Jing, Naiyin Jin, Charles Block, Paolo Mantovani, Davide Giri, Joseph Zuckerman, Luca Carloni and Sarita Adve. *Under review.*
- **A 12nm Linux-SMP-Capable RISC-V SoC with 14 Accelerator Types, Distributed Hardware Power Management and Flexible NoC-Based Data Orchestration.**
  Maico Cassel dos Santos, Tianyu Jia, Joseph Zuckerman, Martin Cochet, Davide Giri, Erik Jens Loscalzo, Karthik Swaminathan, Thierry Tambe, Jeff Jun Zhang, Alper Buyuktosunoglu, Kuan-Lin Chiu, Giuseppe di Guglielmo, Paolo Mantovani, Luca Piccolboni, Gabriele Tombesi, David Trilla, John-David Wellman, En-Yu Yang, Aporva Amarnath, Ying Jing, Bakshree Mishra, Joshua Park, **Vignesh Suresh**, Sarita Adve, Pradip Bose, David Brooks, Luca P. Carloni, Kenneth L. Shepard, Gu-Yeon Wei. *Under review.*

---

## Experience

- **Research Assistant**
  University of Illinois, Urbana-Champaign
  Jan 2021 - Present
  - **Enabling Fine-grained Acceleration in Shared Memory SoCs:**
    * Emerging applications with high task diversity have no single dominant acceleration candidate, therefore, there is a need to enable efficient acceleration of composable fine-grained primitives.
    * Demonstrated that acceleration taxes (such as control and data movement) significantly impact the performance of fine-grained acceleration.
    * Implemented an interface for low-overhead shared memory synchronization to minimize control taxes, and Spandex to minimize data taxes through direct producer-consumer forwarding.
    * Evaluated the proposed architecture with several applications of different complexities on FPGA, including 3D spatial audio and mini-ERA, showing performance improvements over the baseline.

  - **EPOCHS:**
    * Collaboration between IBM, Columbia, Harvard and UIUC, under the DARPA DSSoC program.
    * Effort to enable rapid and scalable development of heterogeneous multi-accelerator SoCs.
    * Contributed to the development of Spandex caches and 3D spatial audio accelerators.
    * Optimized the data movement within mini-ERA using Spandex to show benefit over MESI.
  - **Spandex Caches:**
    * Implemented the Spandex coherence protocol in synthesizable RTL using the ESP framework.
    * Integrated with RISC-V CPUs and validated the stability of multi-core Linux boot on FPGA.
    * Evaluated the performance using heterogeneous systems showing benefit of up to 1.8X over MESI.

- **Silicon Validation Engineer**
  Intel Corporation, India
  Jan 2017 - Jan 2021
  - **Power and Performance (PnP) Validation:**
    * Worked alongside SoC performance architects to evaluate the PnP of the vision processing unit (VPU) for an edge vision SoC using use cases such as ResNet50 and MobileNet.
    * Identified and worked with architects to resolve a showstopper bug involving large voltage droops during compute intensive workloads, such as early layers of ResNet50.
  - **Post-Silicon Validation:**
    * Functional validation of the VPU and media (image and video) codecs using directed and end-to-end tests for an edge vision SoC.
  - **Pre-Silicon Validation:**
    * Functional validation (including end-to-end 5G NR use cases) for a 5G modem SoC.
    * Functional validation of an ARM subsystem, including boot, coherence and low power modes.

---

## Hardware/Software Contributions

- **Spandex Caches:**
  - Implementation of the Spandex coherence protocol in synthesizable RTL and SystemC.
  - Integrates with CPUs and accelerators through a seamless plug-and-play interface in ESP.
- **3D Spatial Audio:**
  - Hardware accelerated using accelerators available in the ESP and custom accelerators.
  - Implemented a bare-metal version to enable debugging using RTL waveforms.
- **3D Spatial Audio Accelerators:**
  - Implemented several variants of the spatial audio accelerators to evaluate monolithic and composable fine-grained acceleration.
- **GeMM Accelerator:**
  - Implemented in SystemC with high-level synthesis (HLS) within ESP.

---

## Talks

- **Towards a Flexible and Low-overhead Memory Hierarchy for Heterogeneous Systems**
  Applications Driving Architectures (ADA), University of Michigan
  Aug 2022

- **Cache Coherence Flexibility in Hardware with Spandex and ESP**
  Prof. Luca Carloni's Group, Columbia University
  Apr 2021

---

## Skills

- **Programming languages**: C, C++, SystemVerilog, Python, SystemC, VHDL
- **Tools**: EDA tools (RTL simulation, High-Level Synthesis, FPGA synthesis)
- **Frameworks**: PyTorch

---

## Relevant Coursework

- Computer System Organization (CS433)
- Parallel Computer Architecture (CS533)
  - **Project**: Optimizing a GeMM Accelerator with HLS optimizations for parallelism and pipelining.
- System-on-chip Design (ECE527)
  - **Project**: Designing a hardware accelerator for PageRank using the Xilinx Pynq platform.
- Architectures for Mobile and Edge Computing (CS598SG)
  - **Project**: Algorithmic and hardware optimizations for scene reconstruction in InfiniTAM.
- Advanced Topics in Computer Architecture: Immersive Computing Systems (CS534)
  - **Project**: Hardware acceleration for NeRF with an MLP accelerator.
- Compiler Construction (CS426)
- Real-World Algorithms for IoT and Data Science (ECE434)
- Deep Learning for Computer Vision (CS444)
  - **Project**: Designing a simpler approach for multi-view consistent instance segmentation with NeRF.