David (Shuangliang) Chen

Website: https://davidchen.page/
Address: 1007 W University Ave, Unit 309, Urbana, IL, 61801
Email: sc60@illinois.edu, Phone: (414) 439-5080

RESEARCH INTERESTS

Computer Architecture, Large-scale Chiplet-based Architecture, Waferscale Integration, Network-on-Chip, Network Topology, Network Switches, Hardware System for Efficient LLM Inference.

EDUCATION

University of Illinois at Urbana-Champaign, Champaign August 2023-Now

Ph.D in Electrical and Computer Engineering

- Advisor: Prof. Rakesh Kumar

University of Illinois at Urbana-Champaign, Champaign August 2019-June 2023

Bachelor of Science in Computer Engineering

- Major GPA 3.94/4.00

RESEARCH SUMMARY

Waferscale Network Switches

S. Chen, S. Pal and R. Kumar ISCA-51st, 2024

- We are the first to propose using chiplet-based waferscale integration technology to build a waferscale network switch that has 32x higher radix than state-of-the-art switch ASIC.
- We show that the actual radix of a waferscale network switch is not area-limited. Rather, it is limited by a combination of internal bandwidth, external bandwidth, and power density.
- We propose a heterogeneous network switch design that reduces switch power by 30.8%-33.5% which, in turn, allows an increase in radix (by up to 4x) by increasing internal I/O bandwidth at the expense of energy efficiency.
- We propose subswitch deradixing that increases the overall radix by 2x by decreasing the radix of the subswitches to alleviate the internal I/O bottleneck.
- We present a system architecture for a waferscale network switch that supports its port count, power delivery, and cooling requirements in a compact form factor.

Distributed Scan: An Architectural Approach to Improve Utilization of Hardware Firewalls

S. Chen, S. Pal and R. Kumar Under submission

- We identify that packets from a single high-bandwidth session will be mapped to the same data processing card to ensure that state access and updates are sequentially consistent, lowering overall utilization
- We observe that only the stateful inspection phase of packet processing truly needs to be serialized – the content inspection phase of packet processing can be parallelized across multiple DPCs without impacting correctness.
- We propose a distributed scan architecture of hardware NGFWs where the stateful inspection for all packets in a session is first performed sequentially on a dedicated processor before the packets are sent to DPCs for distributed content inspection.
- Our evaluations demonstrate that the proposed distributed scan architecture improves the average firewall throughput by 1.08x for mid-end firewalls, 4.29x for high-end firewalls, and 14.3x when using an optical backplane.

A Waferscale Memory Architecture for Low Arithmetic Intensity High Throughput Applications

S. Chen, P. Hanumolu, S. Pal and R. Kumar Under submission

- We show that 1000s of high bandwidth memory devices may be required to maximize performance and energy efficiency for low arithmetic intensity applications, including LLM inference applications.
- We propose a waferscale memory architecture that consists of a compute die placed in the middle of a wafer and directly connected to over 250 high bandwidth memory devices that fill the rest of the wafer.
- We show that a naive implementation of waferscale memory architecture is infeasible due to the high overhead of memory controllers and crossbar between the compute die and the memory devices, and the short reach and low shoreline bandwidth density of conventional interconnects.
- We present a feasible implementation of a waferscale memory architecture that uses high data rate modulation schemes to address shoreline bandwidth constraints, retimer-based interconnects to address the reach constraints, and ultra-wide channels implemented using a multicast retimer design to address memory controller and memory crossbar overheads.
- We present a hybrid interposer design that can be used to increase the number of high bandwidth devices connected to the compute die to over 450 and a massively-banked memory design that can be used to reduce the cost of a waferscale memory system for applications that leave some on-wafer memory capacity un-utilized.
- We quantify the benefits of a waferscale memory architecture. The architecture improves cost efficiency by up to 3.8X and energy efficiency by up to 5.33X over a GPU cluster. The cost efficiency benefits increase to 7.74x when optimizations such as

hybrid interposers and massively-banked memory are considered.

Waferscale Silicon Photonics Systems: A Feasibility Study
R. Bao, F. Cai, S. Chen, A. Joshi, D. Bunandar, and R. Kumar Under submission

- We perform the first characterization of power overheads of implementing silicon photonics at waferscale. This characterization was performed with and without reconfigurability support.
- We show that it is not possible to build some topologies (hypercube, Clos, A2A), at least naively, at waferscale using silicon photonics because the maximum power limit for silicon waveguides is violated. Overall power overheads are also excessively high and get higher when support for reconfigurability is added.
- We propose several optimizations - waveguide disaggregation, waveguide power-optimized floorplanning, and bypass waveguides that provide a significant reduction in overall power and allow each waveguide's power to stay within the safety threshold.
- We perform the first characterization of application-level performance benefits of a silicon photonics-based waferscale HPC node versus a system based on conventional electrical interconnections. We show that an over 4x speedup can be achieved by switching to optical connections. However, our optimizations are needed to make the implementation feasible.
- We make a set of additional recommendations - link splitting, multilayer routing, and alternative waveguide network layouts - for further improving the feasibility and effectiveness of implementing silicon photonics systems at waferscale. Overall, this is the first paper that analyzes and optimizes the feasibility of implementing waferscale silicon photonics systems.

WORK EXPERIENCE

Meta Platforms Menlo Park, CA
Software Engineer Intern May 2022-Aug 2022

- Developed a PyTorch profiler trace analysis utility that can identify users' suboptimal code
- Worked with stakeholders and users to come up with common antipatterns in user's code and implement efficient call tree pattern matchers to flag offending code patterns, including vectorizable for loop, redundant memory copy, not using efficient memory format for tensors, etc.
- Significantly improved model performance in Torch Benchmark by 20-30% after applying fixes suggested by this utility.
- Extended PyTorch Profiler to collect extensive tensor information on torch operations that can enable more complex analysis of the execution graph.

MetaX Integrated Circuits Shanghai, China
Software Engineer Intern (Compiler Team) Feb 2021-Aug 2021

- Involved in building an LLVM compiler backend for a GPGPU, and implemented part of both the assembler and instruction lowering functionality.
- Helped set up a Continuous Integration & Continuous Development workflow for the compiler team with Gerrit+Jenkins to allow automated testing upon 10k+ lines of code submissions.
- Developed a suite of Python-based testing scripts to unit test the functionality of the assembler.
- Received the “Best Intern of the Season” award.

AWARDS

🏆**Qualcomm Graduate Award for the Spring 2025 semester**
- $5000 awarded to 6 graduate students in the Electrical & Computer Engineering Department at UIUC

🏆**UIUC ECE High Honor**
- GPA 3.94/4.00 in the Electrical & Computer Engineering Department at UIUC

🏆**James Scholar Honors**

TECHNICAL SKILLS

Programming Languages:
- C++, Python, Java, Rust, Javascript, Verilog, SystemVerilog, CUDA

Software:
- PyTorch, LLVM, Gem5, Accel-Sim, Booksim2.0, Synopsys-VCS, Cadence-Virtuoso, Matlab, LaTeX, Docker, K8s