Yang Yang

Rice Hall 530, University of Virginia, 85 Engineer's Way, Virginia, VA 22904

Keywords: Computer Architecture/GPU/Memory System/Confidential Computing/Performance

yangyang@virginia.edu linkedin.com/in/jasonyy01 github.com/Elio-yang elio-yang.github.io

EDUCATION

University of Virginia, Charlottesville, USA *Aug. 2023 – Present*
Ph.D. in Computer Science
GPA: 3.91/4.0
Topics: GPU × {Cryptography, Trusted-Computing, Memory, CXL}
Advisor: Prof. Adwait Jog

Jilin University, Changchun, China *Sept. 2019 – Jul. 2023*
B.S. in Computer Science
GPA: 3.69/4.0
Thesis: The Design and Implementation of Binary Code Analysis Framework for NVIDIA GPU.
Advisor: Prof. Jingweijia Tan

PUBLICATIONS

[C3] (In-Submission)
LÆGIS: Understand and Optimize Unified Virtual Memory in GPU-based Confidential Computing

Abstract: Building on our ISPASS'25 results, we examined the open-source GPU driver, focusing on the **/dev/nvidia-uvm** components. From this analysis, we propose **LÆGIS**, a set of novel designs to reduce encryption overhead in GPU-based CC and to improve UVM performance for both large DNN workloads and SSD-backed oversubscription. LÆGIS leverages heterogeneous memory to decouple encryption counters, enabling prediction to support pre-encryption, prefetching and eviction.

[C2] (ISCA'25) [PDF] [Slides.pdf] [Slides.pptx]
NetCrafter: Tailoring Network Traffic for Non-Uniform Bandwidth Multi-GPU Systems
Amel Fatima, **Yang Yang**, Yifan Sun, Rachata Ausavarungnirun, Adwait Jog
In the Proceedings of ACM International Symposium on Computer Architecture (ISCA), Tokyo, Japen, June 2025

Abstract: We present **NetCrafter**, a set of novel techniques to manage network traffic, especially across low-bandwidth links in multi-GPU systems. NetCrafter reduces the volume of flit traffic by (i) stitching compatible, partially filled flits, (ii) trimming unnecessary flits to avoid redundant transfers, and (iii) sequencing flits so that latency-sensitive ones arrive at their destinations faster.

[C1] (ISPASS'25) [PDF] [Talk@ISPASS] [Talk@CC-Summit] [Artifact Available]
Dissecting Performance Overheads of Confidential Computing on GPU-based Systems
Yang Yang, Mohammad Sonji, Adwait Jog
In the Proceedings of IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS), Ghent, Belgium, May 2025

Abstract: For the first time, we present a detailed event-level analysis of GPU applications, showing that kernels without unified virtual memory (UVM) are less affected in execution time but face higher launch and queuing delays. In contrast, UVM kernels experience severe slowdowns under CC. We further evaluate CNN training and LLM inference, where overheads grow at scale, and explore optimizations such as kernel fusion, overlapping and quantization to mitigate these costs.

RESEARCH EXPERIENCE

Insight Computer Architecture Lab *Aug. 2023 – Present*
University of Virginia, Charlottesville, Virginia, USA
Advisor: Prof. Adwait Jog
Focus:

- GPU memory and storage system (UVM across DDR/HBM/SSD via PCIe/CXL/RDMA).
- TDX-based confidential computing on GPUs.
- Cryptography (FHE/PIR) with GPU acceleration.
- Efficient encryption for GPU-based heterogeneous memory system.

State Key Laboratory of Processor *Jul. 2022 – Sept. 2023*
Institute of Computing Technology, Chinese Academy of Science, Beijing, P.R.China
Advisor: Prof. Guangli Li
Topics: Compiler & Programming Systems
Focus: Facilitating Profile-Guided Compiler Optimization with Graph Learning

- Proposed a branch predictor using XGBoost based on compile time static analysis.
- Utilize graph representations to build predictive profile-guided optimization framework and integrated it into LLVM.
- Released a new dataset for graph-related static analysis tasks.

Emerging Technology Enabled Computer Architecture Lab *Feb. 2022 – Jul. 2023*
Jilin University, Changchun, Jilin, P.R.China
Advisor: Prof. Jingweijia Tan
Topics: GPU × {PTX/SASS, Reliability, Energy Efficiency}
Focus:

- Process variation of FinFET and chiplet based MCM-GPUs.
- SASS level analysing and modeling framework for NVIDIA Ampere GPUs.
- Learning techniques for GPU power modeling.
- Instruction level under-voltage reliability of GPUs

TEACHING EXPERIENCE

25 Fall @ UVA, TA for CS 4444: Introduction to Parallel Computing (undergraduate course)

- Designed programming assignments in CUDA.

24 Fall @ UVA, TA for CS 6354: Computer Architecture (graduate course)

- Designed a pipelined RISC-V simulator in Python as programming assignment.

AWARD

[**Travel Grant**] ISPASS'25 ISCA'25

[**Fellowship@UVA**] 2023

[**Scholarship@JLU**] 2019 2020 2021 2022

SERVICE

[**SIG**] Co-organizer of Systems-Interest-Group meetings at UVA since Summer 2024.

[**Artifact Evaluation Committee**] IISWC'25

SKILLS

Languages	C/C++ · Assembly · Python · Go
Frameworks	CUDA · Pytorch · LLVM · TDX · QEMU/KVM · NVIDIA Linux Open GPU Kernel Module
Software	LINUX · LaTeX · GNU compiler (gcc, *etc.*) · GPGPU-Sim/Accel-Sim · Varius-TC · Z3 Solver