Siyuan Chai Contact: siyuanc3@illinois.edu https://schai.me

Education

University of Illinois, Urbana Champaign, IL
Computer Science Ph.D. Candidate — Start Aug. 2021
Advisor: Prof. Tianyin Xu
Research Area: Operating Systems, Memory Systems, SW/HW Codesign

Northwestern University, Evanston, IL
M.S. Computer Science, B.S. Electrical Engineering — Graduated June 2021
GPA: 4.0/4.0 (Summa Cum Laude)

Publications

1. [**OSDI 2025**] **Siyuan Chai**, Jiyuan Zhang, Jongyul Kim, Alan Wang, Jovan Stojkovic, Weiwei Jia, Dimitrios Skarlatos, Josep Torrellas, and Tianyin Xu. "EMT: An Operating System Framework for New Memory Translation Architectures." In *Proceedings of the 19th USENIX Symposium on Operating Systems Design and Implementation.*

2. [**ASPLOS 2025**] Yan Sun, Jongyul Kim, Douglas Yu, Jiyuan Zhang, **Siyuan Chai**, Michael Jaemin Kim, Hwayong Nam, Jaehyun Park, Eojin Na, Yifan Yuan, Ren Wang, Jung Ho Ahn, Tianyin Xu, Nam Sung Kim. "M5: Mastering Page Migration and Memory Management for CXL-based Tiered Memory Systems." In *Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems.*

3. [**ASPLOS 2024**] Jiyuan Zhang, Weiwei Jia, **Siyuan Chai**, Peizhe Liu, Jongyul Kim, and Tianyin Xu. "Direct Memory Translation for Virtualized Cloud" In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems.*

4. [**ASPLOS 2022**] Brian Suchy, Souradip Ghosh, Drew Kersnar, **Siyuan Chai**, Zhen Huang, Aaron Nelson, Michael Cuevas, Alex Bernat, Gaurav Chaudhary, Nikos Hardavellas, Simone Campanoni, and Peter Dinda. "CARAT CAKE: replacing paging via compiler/kernel cooperation". In *Proceedings of the 27th ACM International Conference on Architectural Support for Programming Languages and Operating Systems.*

5. [**Radiology**] Ramsey M Wehbe, Jiayue Sheng, Shinjan Dutta, **Siyuan Chai**, Amil Dravid, Semih Barutcu, Yunan Wu, Donald R. Cantrell, Nicholas Xiao, Hatice Savas, Rishi Agrawal, Nishant Parekh, Aggelos K. Katsaggelos. "DeepCOVID-XR: An Artificial Intelligence Algorithm to Detect COVID-19 on Chest Radiographs Trained and Tested on a Large U.S. Clinical Data Set." *Radiological Society of North America.*

Work Experience

Meta, AI System Co-design, Software Engineering Intern — May to Aug. 2024
Chunked Prefill — Mentor: Dr. Jaewon Lee

- Prototyped and implemented chunked prefill, a technique mitigates prefill-decode interference in LLM serving by splitting prefill request into smaller chunks
- Compared to Meta's production baseline, it offers up to 1.7x better p99 inter-token latency and 1.3x higher serving capacity (max throughput under tail latency constraints).

Google, Google Cloud, Software Engineering Intern — May to Aug. 2022
Machine Model Population Pipeline — Mentor: Alex Tran

- Designed a distributed pipeline to collect data of all Google's server machines (4M+) to model their physical topology. It implements batch reads from Bigtable and capacitor or makes RPC calls with rate limitation
- Validated mac address of machines with as-maintained models across three data sources. Results will be stored in Spanner

Tencent, Network Group, Research Intern — June to Aug. 2021
Service Driven Network Verification Tool — Mentor: Dr. Congcong Miao

- Contributed to design a network verification tool for routing configurations (e,g. BGP, OSPF); it supports quantitative query and covers all data plane with global formal modeling and local simulation

Research Experience

UIUC Xlab, Prof. Tianyin Xu — Aug. 2021 to Present

EMT: An OS Framework for New Memory Translation Architectures

- Designed a pragmatic framework for Linux to empower different hardware schemes of memory translation such as radix tree and hash page table.
- Framework supports diverse memory translation architectures, enables hardware-specific optimizations, and has negligible overhead over hardwired implementations.
- Extensively modified address translation portion of QEMU, SST, and DynamoRIO to simulate performance of different architectures

Direct Memory Translation for Virtualized Clouds

- Proposed Direct Memory Translation (DMT), a practical hardware-software extension for x86-based address translation; it minimizes address translation overhead by directly fetching PTEs
- Speeded up page walks by 1.61x and overall application execution by 1.21x in virtualized environment

NU Parallelism Group, Prof. Peter Dinda — June 2020 to May 2021

CARAT CAKE: Replacing Paging via Compiler/Kernel Cooperation

- Designed and implemented an allocation level address space which aims to replace virtual memory and paging with protection checks inserted at compile time and allocations tracked in runtime
- Implemented a competitive paging address space with support for red black tree and splay tree data structures to track VA-PA mapping, transparent huge pages, and PCID; performance measured with performance monitoring counter

Image & Video Processing Lab, Prof. Aggelos Katsaggelos June 2019 to July 2021

DeepCOVID-XR

- Designed and implemented a CNN model to flag out positive COVID cases based on patients' chest X-ray images
- Outperformed experienced radiologists with an accuracy of 85% compared to 76 - 82% and AUC of 0.935 compared to 0.819 - 0.856

Projects

CPU-GPU Simulator for Collaborative Workloads Modeling

- Designed and prototyped a CPU-GPU memory subsystem simulator for workloads running on CPU-GPU unified virtual memory.
- Integrated gem5 and UVMSmart to model performance of CPU, GPU and on-demand page migration between them

C-style Language Compiler

- Created, from scratch, a compiler to translate C-style language to x86_64 assembly
- Implemented features including graph-coloring register allocation, liveness analysis, instruction selection with tiling, control flow graph, and memory access checking

Middle End Analysis for a C-based API

- Coded a LLVM pass to reduce calls to a custom C-based API by implementing analysis like reaching-definition, constant propagation and folding, alias analysis, function inlining, and dead code elimination.

Skills

Programming languages:
C/C++, Assembly, Python, Java, Go, JavaScript, MATLAB

Artificial Intelligence:
CUDA, PyTorch, Tensorflow, Keras, Image Processing, Computer Vision

System-level Development:
Linux Kernel Development, QEMU, Docker, GDB, Make, Linker, LLVM, OpenMP

Hardware:

Raspberry Pi, Arduino, VHDL, Verilog

Web Development:

HTML, CSS, Flask, Django, React