Yixian Shen

+31649779351 | y.shen@uva.nl | LinkedIn | GitHub | Amsterdam, Netherlands

Education

University of Amsterdam — Amsterdam, Netherlands
Ph.D. in Computer Science; — *Dec 2019 – Feb 2024*

- **Main coursework:** Hardware and System Security, Efficient Deep Learning, Morden High-Performance Computing, Virtualization Technologies for Cloud Computing, Software-hardware Co-designs.

Sun Yat-sen University — Guangzhou, China
M.Sc in computer science; ***GPA: 4.1/5, Rank: 2/72*** — *Aug 2017 – July 2019*

- **Main coursework:** Modern Artificial Intelligence Technology, Artificial Intelligence in Software Engineering, System Analysis and Design (Software Engineering), Digital Image Processing.

Communication University of Zhejiang — Hangzhou, China
B.Sc in Electronic and Information Engineering; ***GPA: 4.06/5, Rank: 1/307*** — *Sep 2013 – Jun 2017*

- **Main coursework:** Algorithms and Data Structures, Embedded System Analysis and Design, Data Communication and Computer Networks, Techniques and Applications of Database.

Research Experience

Parallel Computing Systems Group — Amsterdam, Netherlands
Supervised by Anuj Pathania and Prof. Andy Pimentel — *Jun 2020 – Present*

- Deep Neural Networks (DNNs) are increasingly used in embedded applications at the edge, powered by Heterogeneous Multi-Processors System-on-Chips (HMPSoCs) equipped with Neural Processing Units (NPUs). These NPUs outperform CPUs and GPUs in power consumption and performance for DNN inference. However, unlike CPUs and GPUs that can perform full precision inference, NPUs often only support quantized inference, which can lead to accuracy loss due to quantization. To address this, the PiQi framework has been introduced. It enables layer-wise switching between CPU, GPU, and NPU for DNN inference on HMPSoCs, allowing for partially quantized inference with minimal overhead. PiQi incorporates a multi-objective Genetic Algorithm (GA) to optimize power-performance under an accuracy constraint through selective multi-layer quantization during inference. Additionally, it employs a specialized neural network to predict accuracy when assigning DNN layers to the appropriate cores. PiQi's effectiveness is demonstrated on the RK399Pro HMPSoC, showing a significant improvement in the hyper-volume of power-performance Pareto-frontier for Yolov3 and MobileNetv1 under an accuracy constraint, compared to the state-of-the-art.
- 3D-stacked processor-memory systems offer significant boosts in computing performance. However, they grapple with intensified thermal challenges, stemming from elevated power density and limited heat dissipation avenues. Presently, prevalent power management strategies, including Dynamic Voltage and Frequency Scaling (DVFS) for cores and Low Power Mode (LPM) for memory banks, function in deeply interconnected domains. Their singular application often results in a less-than-ideal thermal equilibrium. Addressing this shortfall, our research augments the cutting-edge CoMeT interval thermal simulator by integrating an LPM feature. We introduce a sophisticated learning-based thermal management approach, harmonizing both DVFS and LPM in a seamless strategy. This orchestrated approach promises more efficient thermal regulation, potentially leading to enhanced system longevity and consistent peak performance without compromising thermal safety benchmarks.
- The heat produced during computation is a major bottleneck for the performance of multi-/many-core processors. This issue is even more pronounced in 3D-stacked systems compared to their planar 2D counterparts. While power budgeting techniques for 2D processors exist, they often fall short in addressing the vertical thermal coupling unique to 3D-stacked architectures and necessitate more refined RC thermal models. Addressing this gap, our study unveils pioneering linear algebra-based, time-invariant transformations specifically designed for power budgeting in 3D-stacked configurations. We introduce "3D-TTP", an innovative transient-temperature-aware technique. Empirical tests using the advanced CoMeT simulator for interval thermal simulations affirm that our 3D-TTP approach effectively eliminates thermal violations.

- Dynamic Voltage and Frequency Scaling (DVFS) has long been the go-to technique for thermal management in multi/many-core systems. By adjusting a core's frequency and voltage, DVFS can curtail power consumption, leading to a subsequent reduction in core temperature. Yet, its adoption often comes at the cost of diminished application performance. In our latest research, we introduce a novel, DVFS-agnostic approach to thermal management tailored for S-NUCA many-core systems. Instead of altering frequency and voltage, our method emphasizes the synchronized rotation or migration of threads across the S-NUCA many-core landscape. This ensures that no individual core ever exceeds its designated thermal limits. The result? A marked reduction in overheating risks and a path to a more consistent, optimized system performance.
- Cache interference in the Last Level Cache (LLC) hinders the predictability of multicore systems, which can be detrimental for hard real-time systems where timing is crucial. In response, we embarked on an in-depth exploration of two potential mitigative strategies: explicit modeling of cache interference and the implementation of partitioned cache. Investigated within both global and partitioned scheduling contexts, our focus was on optimizing cache behavior to ensure consistent task execution times. Through explicit modeling, we aimed to predict and counteract interference-induced disruptions. Meanwhile, with partitioned cache, we sought to isolate tasks, providing each with dedicated cache segments to minimize cross-task interference, thereby aiming for a more deterministic and enhanced schedulability.

High-Performance Computing Group Guangzhou, China

Supervised by Prof. Yunfei Du *Aug 2017 – Jun 2019*

- The increasing volume of data from high-performance computing presents challenges in data management and analysis. Current query technologies face issues with slow index construction and redundant data retrieval. This study introduces a high-performance query framework for scientific data, featuring a two-tier index data structure for parallel index construction and real-time indexing. A two-tier parallel query mechanism, dynamic union read strategy, and adaptive scheduling strategy optimize data retrieval efficiency.
- Deep learning (DL) excels in complex applications like computer vision and natural language processing, but larger models and datasets lead to extended training times, hindering progress. Modern high-performance hardware, such as GPUs, has been adopted in DL frameworks like Caffe, Torch, and TensorFlow, but their computational efficiency remains low. We introduce SingleCaffe, a DL framework designed to fully utilize high-performance hardware and improve training efficiency. SingleCaffe employs multiple threads within a single node and uses data parallelism across threads. It designates one thread as a parameter server and the others as worker threads, distributing both data and workloads. Additionally, the framework carefully manages memory allocation to reduce overhead.
- Large-scale, loosely-coupled applications face implementation challenges on high-performance computing platforms, and their deployment and maintenance can lead to wasted resources. To address this issue, we propose Teno, a high-throughput computing job execution framework designed for Tianhe-2 without modifying its existing Slurm configuration. Teno employs hierarchical scheduling through Slurm, optimizing the traditional Master-Worker model to accelerate high-throughput operations and enhance cluster resource utilization. The framework also incorporates effective fault-tolerance mechanisms such as fault recovery and error retry.

Research Output

- **PiQi: Partially Quantized DNN Inference on HMPSoCs**. Ehsan Aghapour, **Yixian Shen**, Dolly Sapra, Andy Pimentel, Anuj Pathania. 2024 Design Automation Conference(DAC 2024) in submission.
- **Thermal Management for 3D-Stacked Systems via Unified Core-Memory Power Regulation. Yixian Shen**, Leo Schreuders, Anuj Pathania, Andy Pimentel. 2023 International Conference on Hardware/Software Codesign and System Synthesis (CODES+ISSS 2023) and ACM Transactions on Embedded Computing Systems (TECS 2023),https://dl.acm.org/doi/full/10.1145/3608040.
- **3D-TTP: Efficient Transient Temperature-Aware Power Budgeting for 3D-Stacked Processor-Memory Systems.** Sobhan Nicknam*, **Yixian Shen***, Anuj Pathania, Andy Pimentel. 2023 IEEE Computer Society Annual Symposium on VLSI (ISVLSI 2023),https://ieeexplore.ieee.org/abstract/document/10238664.
- **Thermal Management for S-NUCA Many-Cores via Synchronous Thread Rotations. Yixian Shen**, Sobhan Nicknam, Anuj Pathania, Andy Pimentel. 2023 Design, Automation & Test in Europe Conference & Exhibition (DATE 2023),https://ieeexplore.ieee.org/document/10136895.
- **TCPS: A Task and Cache-aware Partitioned Scheduler for Hard Real-time Multi-core Systems. Yixian Shen**, Jun Xiao, Andy Pimentel. Proceedings of the 23rd ACM SIGPLAN/SIGBED International Conference on Languages, Compilers, and Tools for Embedded Systems (LCTES 2022), https://dl.acm.org/doi/abs/10.1145/3519941.3535067.

- **Cache Interference-aware Task Partitioning for Non-preemptive Real-time Multi-core Systems.** Jun Xiao, **Yixian Shen**, Andy Pimentel. ACM Transactions on Embedded Computing Systems (TECS 2021), https://dl.acm.org/doi/full/10.1145/3487581.

- **Bi-Cluster: A High-Performance Data Query Framework for Large-Scale Scientific Data. Yixian Shen**, Peng cheng, Yunfei Du, Yutong Lu. 2019 IEEE 21st International Conference on High Performance Computing and Communications (HPCC 2019),https://ieeexplore.ieee.org/document/8855425.

- **SingleCaffe: An Efficient Framework for Deep Learning on a Single Node.** Chenxu Wang, **Yixian Shen**, Yunfei Du, Yutong Lu. IEEE Access 2018,https://ieeexplore.ieee.org/document/8528440.

- **Teno: An Efficient High-Throughput Computing Job Execution Framework on Tianhe-2.** Wei Yu, **Yixian Shen**, Lin Li, Yunfei Du, Zhiguang Chen, Yutong Lu. In Proceedings of the 20th IEEE International Conference on High Performance Computing and Communications (HPCC 2018),https://ieeexplore.ieee.org/document/8622823.

Professional Service

- **ICCAD 2023:** External Reviewer, The International Conference on Computer-Aided Design.

- **CGO 2024:** Artifact Evaluation Committee Member, IEEE/ACM International Symposium on Code Generation and Optimization.

Internship

- I have been doing an internship at **Imec** since Oct. 2023. My research topic is Chiplet-Based Co-Design of Heterogeneous Systems with Thermal-Aware Placement and 2.5/3D Integration.

Awards & Achievements

National/Province-level Awards

- **Outstanding Graduate Awards of Zhejiang Province:** in 2017
- **China National Scholarship:** in 2015

University-level Awards

- **The Second Prize Scholarship at Sun Yat-sen University:** totally three times in 2017,2018, 2019.
- **The First Prize Scholarship at Communication University of Zhejiang:** totally seven times in 2013, 2014, 2015, 2017.
- **Excellent Student Awards at Communication University of Zhejiang:** in 2017.

Competition Awards

- **Undergraduate Physics Competition of Zhejiang Province:** The First Prize Award, totally three times in 2014, 2015, 2016.
- **Undergraduate Calculus Competition of Zhejiang Province:** The Second Prize Award, in 2016.
- **Undergraduate Electronic Design Competition of Zhejiang Province:** The Second Prize Award, in 2014.

Other Awards

- **New-seedings Talent Training Project of Zhejiang Province Awards:** 10000 RMB for the Development of a Real-time Positioning and Tracking System for IoT, in 2016.
- **Alibaba Group Scholarship:** (10000 RMB) in 2016.
- **Dahua Group Scholarship:** (5000RMB) in 2014.

Skills

- **Programming:** Scala, Python, C++, Java and intermediate in Javascript, SQL,VHDL
- **Big Data Processing:** rich experiences in using Hadoop, Kafka, Spark, Hive, Spark Streaming for big data and streaming data processing and analytics
- **Devops and Cloud Computing:**familiar with DevOps ecosystem like Docker, Travis, Swagger and Postman
- **Machine Learning:** familiar with commonly used machine learning algorithms and models
- **Deep Learning:** Proficient in popular deep learning frameworks including PyTorch, TensorFlow, and Caffe.
- **Languages:** native Chinese and English