Jiasen Huang

📞 +86-19921547781 · ✉ 13110720047@fudan.edu.cn · ⓞ [KKK10001](#) ·

ⓘ OBJECTIVE

- Senior CPU Architect

🎓 EDUCATION

- Ph.D in Microelectronics, Fudan University，2011.9~2016.6
- M.S. in Information & Communication Engineering, Shanghai University，2007.3~2011.3

💼 WORK EXPERIENCE

RVSilicon   **2023.12~2025.1**

Head of CPU Performance Modeling
Senior CPU Architect
Being in charge of performance modeling of the whole Middle-End and part of the Back-End including the Load Store Unit (LSU) and the Data Prefetcher for both our 8-decode baseline and 3-decode legacy branch.

- Proposed, implemented, and tuned the area-saving precise store-set. Performance gain of 5.06% was then achieved.
- Proposed a weighted temporally correlated load value predictor (LVP). Prediction accuracy of the proposed LVP proves to vary in accuracy from 57% to 99% with an average of 84%, that was about 7% better than the know best FCM predictor of order 3.
- Proposed an architecture of Lookahead Renaming, which snoops the available dispatching entries in advance at Rename Stage, and then dynamically schedule free entries in between the dispatch buffers. This feature detects possible stalls due to lacking free entries at Dispatch Stage, and exploits more dispatching capabilities.
- Porting the VLDP data prefetcher to gem5.
- Microarchitecture Analysis of the LSU in XuanTie C910 RISCV CPU Core.

ICubeCorp   **2023.7~2023.11**

R&D Head of Shanghai Branch
Head of GPGPU Arch. & Performance Modeling
Head of TensorCore

- Performance modeling and analysis of the in-house IC1001P GPU
  - Led porting the in-house MVP ISA to gem5-gcn to completion.
  - Setup framework of cycle-accurate (CA) simulator, and correlated the thread scheduling module (THDC) with RTL
  - Ported the simulator above to UVM, enabling co-sim in between the CA simulator and the reference model.
- Led the development and verification of TensorCore v0.2 to completion.

Baidu KUNLUNXIN   **2022.11~2023.6**

Head of RISCV CPU Simulator
Head of Cluster-Next (XCN) Simulator

- Setup DiffTest-based co-sim environment of the in-house K* processor and Spike. Smoke tests passed.
- Proposed and setup Trace-Driven simulation framework, driving gem5-gcn with the dump of the in-house functional simulator.
- Analyzed architectural difference of AMD-GCN and XCN
- Profiled instructions latency of XCN v1.0
- Correlated for the Thread Scheduler, Vector Register File (VRF), Instruction Buffer (iBuff), Operand Buffer (OpBuff)

T-Head Semiconductor   **2020.3~2022.8**

CPU Architect

- Development of the ISA of T-Head C906 CPU IP Core
  - Implemented C906 ISA, including BF16 conversion, Exception, CSR Checker, SC, SEG Comparison, FOF etc.
  - Fixed several bugs in C910 ISA
  - Refinement of C910 ISA implementation
- Gem5-based performance modeling and correlation for C910 memory sub-system
- Preliminary Studies of C908 architecture, including data prefetcher, cache replacement policy, and inclusive property across L1D/L1I and L2C.
- Implemented C910 ISA Extension, and setup in-line assembly based test framework. All 73 smoke cases passed with cross-checking among ISA, gem5, and RTL.
- Built the cycle-accurate (CA) simulator for C910 L2 Cache (L2C)
  - Proposed a Trace-Driven correlation method for tightly-coupled CIU and L2C architecture
  - Setup C++ based standalone CA framework
  - Correlated the CA simulator with C910 L2C
- Proposed four architectural improvement suggestions for C910 memory sub-system, and one of four suggestion was adopted in C908
- Produced C910 L2C architectural specification document
- Total 8 contributions to gem5 community

Hygon Information Technology Co Ltd   **2019.2~2019.12**

CPU Architect

- Deployment of AI framework on high-performance enterprise GPUs
  - Deployed Caffe/Tensorflow on NVIDIA GTX750/GTX1660 and AMD VEGA20.
  - Performance analysis of CNN
  - Participated in the short-term breakthrough project of HPL led by SUGON.
- Performance modeling and correlation of L3 Cache for the in-house Moksha_Plus/Dharma CPU IPs
  - Fixed several performance bugs in L3 Cache model
- Performance studies on L3C replacement policies based on dead-zone detection and prediction
- Development of one Perfect L3 Cache

HiSilicon Descartes Development Department   **2017.4~2018.8**

GPU Architect

- Development of de-compression algorithms including BC, ETC, ETC2, EAC, and ASTC
- Stripped the ARM Frame Buffer Compression (AFBC) decoder from Hivoyager CModel as a standalone module for UT
- Built the Unit Test (UT) framework, and supported co-sim with RTL
- Owned the CModel of Texel Sample Generation (TSG) and Z
- Owned the system test (ST) validation of the whole Texture Unit (TU) covering Addressing/Surface Format/Resource Format Groups
- Proposed two suggestions on performance enhancement for TSG
  - Take an offset of -0.5 before addressing to avoid redundant checking for those texels that would be definitely out-of-boundary.
  - Let the Non-power-of-two (NP2) Loop and Seamless-Cube Loop share cross-face checking logic.
- Development and upgrade of debugging tools
  - Introduced CMake for cross-platform compilation for Hivoyager DVM
  - Upgraded the HEX to BMP Image Converter Tool to support texture format containing over 8-bit texels per channel.
  - Built behavior model covering NP2 Split and Seamless-Cube Split data paths for self-checking for the bit-accurate CModel.
  - Visualized the whole GPU data flow with Graphviz

ACADEMIC RESEARCH

Research on key technologies for accelerating sparse matrix computations based on FPGA   2014.3~2016.6

Balanced K-Way Graph Partitioning Heuristic Algorithm for Parallel SpMxV on FPGAs [J1][C1][C2][C3]

Optimization and Acceleration of Photoacoustic Imaging Algorithms Based on Fully Differential Gradient Descent Strategy   2013.7~2014.3

"National Natural Science Foundation of China (Project No. 10974035)"
"Shanghai Chief Scientist Program (Project No. 10XD140060)" [C1][C2][C3][C4][C5]

"National Science and Technology Major Project: Design and Application Demonstration of DC-OFDM Ultra-Wideband Wireless Communication Chip (Project No. 2009ZX03007-002-02)"   2011.3~2013.5

A Novel Joint Estimation and Compensation Algorithm for Non-idealities of Analog Front-end in DC-OFDM System [C6]
A Novel Approach Based on Recursive Indexing for FFT Data Reordering [J2][C7]

PATENTS

- ® CN111143245B "A cache data processing method, circuit, processor, and chip"
- Ⓟ CN202411043592.5 "A Load-Set Based Area-Saving Speculative Load Circuit and Method"
- Ⓟ CN202410927066.9 "A Method for Processing CPU Instructions and Associated Device"

HONORS

- "HUAWEI Future Star 2017"
- "National Scholarship for the 2015 Academic Year"
- "First-Class Scholarship of Fudan University (2014)"
- "Second-Class Scholarship of Fudan University (2012)"
- "Second-Class Scholarship of Fudan University (2011)"
- "Outstanding Graduate of Shanghai (2011)"
- "Third Prize, 2010 Intel Cup National Undergraduate Embedded System Design Invitational (China)"
- "Second Prize, 2010 TI Cup National Undergraduate Analog System Design Invitational (China)"
- "Outstanding Student of Shanghai University in 2010"
- "Outstanding Student of Shanghai University in 2009"
- "Special Grade Scholarship of Shanghai University (2009)"
- "Special Grade Scholarship of Shanghai University (2008)"
- "Outstanding Student Award of Shanghai University (2008)"
- "Excellent Participant in Shanghai University's 2007 Military Training"

PAPERS

- Conference Proceedings Papers
  - [C1] Jiasen Huang, Yuanyuan Wang, Wei Li, Ning Li, and Junyan Ren, "Greedy Approach Based Heuristics for Partitioning the Measurement Matrices in Photoacoustic Imaging," IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), San Diego, California, USA, Nov. 2015
  - [C2] Jiasen Huang, Weina Lu, and Junyan Ren, "Greedy Approach Based Heuristics for Partitioning SpMxV on FPGAs," International Conference on Field-programmable Logic and Applications (FPL), London, UK, Sept. 2015, pp.1-2.
  - [C3] Jiasen Huang, Junyan Ren, Wenbo Yin, and Lingli Wang, "No Zeros Padded Sparse MatrixVector Multiplication on FPGAs," International Conference on Field-programmable Technology (FPT), Shanghai, China, Dec. 2014, pp. 290-291.
  - [C4] Jiasen Huang, Junyan Ren, Jun Xu, and Yuanyuan Wang, "General Expression Based Inner Loop Unrolling Scheme for TV-GD Algorithm Adopted in Photoacoustic Imaging," IEEE Biomedical Circuits and Systems Conference (BioCAS), Lausanne, Switzerland, Oct. 2014, pp.129-132.
  - [C5] Jiasen Huang, Chen Zhang, Junyan Ren, and Yuanyuan Wang, "General Expression Based Loop Unrolling Scheme for Real-Time Implementation of PAI," IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), Seattle, WA, Nov. 2014, pp. 1-3.
  - [C6] Jiasen Huang, Hao Chen, Junyan Ren, and Fan Ye, "A Novel Joint Estimation and Compensation Algorithm for Non-idealities of Analog Front-end in DC-OFDM System," IEEE International Conference on ASIC (ASICON), Shenzhen, China, Oct. 2013, pp.1-4.
  - [C7] Jiasen Huang, Shuai Chen, Hao Chen, and Fan Ye, "A Novel Approach Based on Recursive Indexing for FFT Data Reordering," International Conference on Information & Communications Technologies (IETICT), Beijing, China, Apr. 2013, pp.394-398.
- Journal Articles
  - [J1] Jiasen Huang, Junyan Ren, and Wei Li, "Greedy Approach Based Heuristics for Partitioning Sparse Matrices," IEICE Transactions on Information and Systems, vol. E98-D, no.10, Oct. 2015, pp.1847-1851.

- ○ [J2] Jiasen Huang, Shuai Chen, Xiaolong Wang, Fan Ye, and Junyan Ren, "FFT Reordering Algorithm Design Based on Mapping Recursion Strategy", Computer Engineering, vol. 39, no. 9, pp. 285-288, Sept. 2013.

⚒ SKILLS

- ★★★★ uArch.
- ★★★★ Performance Analysis, Profiling, Correlation
- ★★★★ C++, gem5
- ★★★ Shell, Python
- ★★★ RTL
- ★★★ PCB
- ★★★ 89C51/89C52 series MCU