论文列表

Cache Replacement policy

- Perceptron learning for reuse prediction
- Back to the Future: Leveraging Belady's Algorithm for Improved Cache Replacement
- SHiP + + : Enhancing Signature–Based Hit Predictor for Improved Cache Performance
- Multiperspective Reuse Prediction
- Maximizing Cache Performance Under Uncertainty
- Exploring Predictive Replacement Policies for Instruction Cache and Branch Target Buffer
- Applying Deep Learning to the Cache Replacement Problem
- Designing a Cost–Effective Cache Replacement Policy using Machine Learning
- P–OPT: Practical Optimal Cache Replacement for Graph Analytics
- An Imitation Learning Approach for Cache Replacement
- Ripple: Profile–Guided Instruction Cache Replacement for Data Center Applications

分支预测

- Bit–Level Perceptron Prediction for Indirect Branch Prediction
- An Alternative TAGE–like Conditional Branch Predictor
- Auto–Predication of Critical Branches
- Towards the Adoption of Local Branch Predictors in Out–Of–Order Superscalar Processors

- The IBM z15 High Frequency Mainframe Branch Predictor
- BranchNet: A Convolutional Neural Network to Predict Hard-to-Predict Branches
- COBRA: A Framework for Evaluating Compositions of Hardware Branch Predictors

数据预取

- Bingo Spatial Data Prefetcher
- Perceptron-Based Prefetch Filtering
- Stream-based Memory Access Specialization for General Purpose Processors
- DSPatch: Dual Spatial Pattern Prefetcher
- Temporal Prefetching Without the Off-Chip Metadata
- RnR: A Software-Assisted Record-and-Replay Hardware Prefetcher
- Bouquet of Instruction Pointers: Instruction Pointer Classifier-based Spatial Hardware Prefetching
- Efficient Meta-Data Management for Irregular Data Prefetching
- Prodigy: Improving the Memory Latency of Data-Indirect Irregular Workloads Using Hardware-Software Co-Design
- Compiler-Assisted Data Streaming for Regular Code Structures
- Boosting Store Buffer Efficiency with Store-Prefetch Bursts

投机执行攻击（幽灵/熔断）

- Conditional Speculation: An Effective Approach to Safeguard Out-of-Order Execution Against Spectre Attacks.
- Efficient Invisible Speculative Execution through Selective Delay and Value Prediction
- CleanupSpec: An "Undo" Approach to Safe Speculation
- NDA: Preventing Speculative Execution Attacks at Their Source
- Speculation Invariance (InvarSpec): Faster Safe Execution through Program Analysis
- Speculative Data-Oblivious Execution: Mobilizing Safe Prediction For Safe and Efficient Speculative Execution
- MuonTrap: Preventing Cross-Domain Spectre-Like Attacks by Capturing Speculative State
- Hardware-Software Contracts for Secure Speculation
- Context-Sensitive Fencing: Securing Speculative Execution via Microcode Customization
- New Models for Understanding and Reasoning about Speculative Execution Attacks

Processing in Memory / Near data Processing

- Active-Routing: Compute on the Way for Near-Data Processing
- Duality Caches for Data Parallel Acceleration
- CoNDA: Enabling Efficient Near-Data Accelerator Communication by Optimizing Data Movement
- TensorDIMM: A Practical Near-Memory Processing Architecture for Embeddings and Tensor Operations in Deep Learning
- FReaC Cache: Folded-Logic Reconfigurable Computing in the Last Level Cache
- FAFNIR: Accelerating Sparse Gathering by Using Efficient Near-Memory Intelligent Reduction
- SynCron: Efficient Synchronization Support for Near-Data-Processing Architectures
- ABC-DIMM: Alleviating the Bottleneck of Communication in DIMM-Based Near-Memory Processing with Inter-DIMM Broadcast
- BOSS: Bandwidth-Optimized Search Accelerator for Storage-Class Memory
- PIM-VR: Erasing Motion Anomalies In Highly-Interactive Virtual Reality World with Customized Memory Cube.
- NAND-Net: Minimizing Computational Complexity of In-Memory Processing for Binary Neural Networks.
- FloatPIM: In-Memory Acceleration of Deep Neural Network Training with High Precision
- GraphQ: Scalable PIM-Based Graph Processing
- eAP: A Scalable and Efficient In-Memory Accelerator for Automata Processing
- ComputeDRAM: In-Memory Compute Using Off-the-Shelf DRAMs
- CASCADE: Connecting RRAMs to Extend Analog Dataflow in an End-to-End In-Memory Processing Paradigm
- Enabling Highly Efficient Capsule Networks Processing Through A PIM-Based Architecture Design
- ELP2IM: Efficient and Low Power Bitwise Operation Processing in DRAM
- Impala: Algorithm/Architecture Co-Design for In-Memory Multi-Stride Pattern Matching
- Look-Up Table based Energy Efficient Processing in Cache Support for Neural Network Acceleration
- Enabling Highly Efficient Capsule Networks Processing Through A PIM-Based Architecture Design
- GradPIM: A Practical Processing-in-DRAM Architecture for Gradient Descent
- SpaceA: Sparse Matrix Vector Multiplication on Processing-in-Memory Accelerator

- Hardware Architecture and Software Stack for PIM Based on Commercial DRAM Technology
- An Analog Preconditioner for Solving Linear Systems
- Fulcrum: a Simplified Control and Access Mechanism toward Flexible and Practical in-situ Accelerators
- Memristive Data Ranking
- CATCAM: Constant-time Alteration Ternary CAM with Scalable In-Memory Architecture
- DUAL: Acceleration of Clustering Algorithms using Digital-Based Processing In-Memory
- Newton: A DRAM-Maker's Accelerator-in-Memory (AiM) Architecture for Machine Learning
- MOUSE: Inference In Non-volatile Memory for Energy Harvesting Applications

DSA-Sparse Matrix Vector Multiplication

- ExTensor: An Accelerator for Sparse Tensor Algebra
- Efficient SpMV Operation for Large and Highly Sparse Matrices Using Scalable Multi-Way Merge Parallelization
- SMASH: Co-designing Software Compression and Hardware-Accelerated Indexing for Efficient Sparse Matrix Operations
- SIGMA: A Sparse and Irregular GEMM Accelerator with Flexible Interconnects for DNN Training
- SpArch: Efficient Architecture for Sparse Matrix Multiplication
- Tensaurus: A Versatile Accelerator for Mixed Sparse-Dense Tensor Computations
- A Hybrid Systolic-Dataflow Architecture for Inductive Matrix Algorithms
- MatRaptor: A Sparse-Sparse Matrix Multiplication Accelerator Based on Row-Wise Product
- SPAGHETTI: Streaming Accelerators for Highly Sparse GEMM on FPGAs
- VIA: A Smart Scratchpad for Vector Units With Application to Sparse Matrix Computations
- SpaceA: Sparse Matrix Vector Multiplication on Processing-in-Memory Accelerator
- Dual-Side Sparse Tensor Core

DSA-NPU

- HyPar: Towards Hybrid Parallelism for Deep Learning Accelerator Array
- Shortcut Mining: Exploiting Cross-Layer Shortcut Reuse in DCNN Accelerators
- Sparse ReRAM Engine: Joint exploration of activation and weight sparsity on compressed neural network
- TIE: Energy-efficient tensor train-based inference engine for deep neural network
- Laconic Deep Learning Inference Acceleration

- DeepAttest: An End-to-End Attestation Framework for Deep Neural Networks
- Cambricon-F: Machine Learning Computers with Fractal von Neumann Architecture
- Wire-Aware Architecture and Dataflow for CNN Accelerators
- ShapeShifter: Enabling Fine-Grain Data Width Adaptation in Deep Learning
- Simba: Scaling Deep-Learning Inference with Multi-Chip-Module-Based Architecture
- ZCOMP: Reducing DNN Cross-Layer Memory Footprint Using Vector Extensions
- Boosting the Performance of CNN Accelerators with Dynamic Fine-Grained Channel Gating
- SparTen: A Sparse Tensor Accelerator for Convolutional Neural Networks
- eCNN: a Block-Based and Highly-Parallel CNN Accelerator for Edge Inference
- SuperNPU: An Extremely Fast Neural Processing Unit Using Superconducting Logic Devices
- Procrustes: A Dataflow and Accelerator for Sparse Deep Neural Network Training
- DUET: Boosting Deep Neural Network Efficiency on Dual-Module Architecture
- TFE: Energy-Efficient Transferred Filter-Based Engine to Compress and Accelerate Convolutional Neural Networks
- TensorDash: Exploiting Sparsity to Accelerate Deep Neural Network Training
- SAVE: Sparsity-Aware Vector Engine for Accelerating DNN Training and Inference on CPUs
- ConfuciuX: Autonomous Hardware Resource Assignment for DNN Accelerators using Reinforcement Learning
- Mix and Match: A Novel FPGA-Centric Deep Neural Network Quantization Framework
- η-LSTM: Co-Designing Highly-Efficient Large LSTM Training via Exploiting Memory-Saving and Architectural Design Opportunities
- FuseKNA: Fused Kernel Convolution based Accelerator for Deep Neural Networks
- Layerweaver: Maximizing Resource Utilization of Neural Processing Units via Layer-Wise Scheduling
- CSCNN: Algorithm-hardware Co-design for CNN Accelerators using Centrosymmetric Filters
- Ascend: a Scalable and Unified Architecture for Ubiquitous Deep Neural Network Computing

Hardware Acceleration

- Trainbox: An extreme-scale neural network training server architecture by systematically balancing operations
- Shredder: GPU-accelerated incremental storage and computation
- FIDR: A scalable storage system for fine-grain inline data reduction with efficient memory handling
- CIDR: A cost-effective in-line data reduction system for terabit-per-second scale SSD arrays

- SPIN: Seamless operating system integration of peer–to–peer DMA between SSDs and GPUs
- Solros: A Data–Centric Operating System Architecture for Heterogeneous Computing
- Hippogriff: Efficiently moving data in heterogeneous computing systems
- Morpheus: Creating Application Objects Efficiently for Heterogeneous Computing
- DCS–ctrl: A fast and flexible device–Control mechanism for device–Centric server architecture
- DCS: A fast and scalable device–centric server architecture