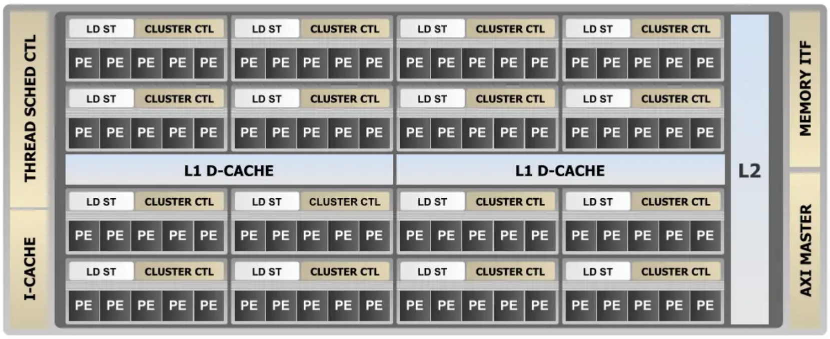
# ASPLOS 2021

## 加速器架构

### DiAG: a dataflow-inspired architecture for general-purpose processors

- 基于数据流的 RISCV 众核处理器架构
- 通用并行处理器，512PE，类 GPU 架构
- FPGA systemVerilog 验证，比乱序处理器更快，能效比更高。物理版图。



## 加速器单元

### Gamma: Leveraging Gustavson’s Algorithm to Accelerate Sparse Matrix Multiplication

- 稀疏矩阵乘法加速单元
- Gustavson’s algorithm
- 数据流

## 算法+硬件

### Training for multi-resolution inference using reusable quantization terms

- 一种多精度 DNN 训练方法
- 多精度乘加器及其脉动矩阵实现

## 编译器

### A Compiler Infrastructure for Accelerator Generators

- HLS 编译器，将高级程序直接编译成电路，性能比商用 HLS 更好

### VeGen: a vectorizer generator for SIMD and beyond

- 通用程序编译中自动矢量化，性能高于 LLVM vectorizer

### Analytical characterization and design space exploration for optimization of CNNs

- 为多核 CPU 上的 CNN 查找循环优化，提升访存性能

### Mind mappings: enabling efficient algorithm-accelerator mapping space search

- 为 algorithm-accelerator mapping space search 提出了一种基于梯度的搜索方法

### Neural architecture search as program transformation exploration

- DNN 网络结构搜索，既提升精度，又提升执行效率，超过 TVM

## 传统硬件上的加速

### MERCI: Efficient Embedding Reduction on Commodity Hardware via Sub-query Memoization

- 改进 Embedding 优化内存访问

## other

### Statistical robustness of Markov chain Monte Carlo accelerators

- Markov Chain Monte Carlo (MCMC) 加速器的评估

### NeuroEngine: a hardware-based event-driven simulation system for advanced brain-inspired computing

- 类脑计算仿真器

### Defensive approximation: securing CNNs using approximate computing

- CNN 抗攻击

# HPCA 2021

## 加速器架构

### Heterogeneous Dataflow Accelerators for Multi-DNN Workloads

- DNN 不同层的数据流特征存在差异，可重构 DNN 加速器 (RDAs)能进行数据流格式的配置，但是硬件代价很高
- 提出异构数据流加速器 (HDA)，并用软硬件协同的方式探索，提出一种 HDA 架构 Maelstrom，性能等方面优于其他加速器

### GCNAX: A Flexible and Energy-efficient Accelerator for Graph Convolutional Neural Networks

- 图卷积神经网络数据流分析，提出一种灵活的数据流

- 提出 GCNAX 加速器

### Ascend: a Scalable and Unified Architecture for Ubiquitous Deep Neural Network Computing

- 华为昇腾910 AI 芯片

### VIA: A Smart Scratchpad for Vector Units with Application to Sparse Matrix Computations

- CPU 向量计算单元微架构改进，更好地处理稀疏矩阵

## 加速器单元

### SPAGHETTI: Streaming Accelerators for Highly Sparse GEMM on FPGAs

- 稀疏矩阵乘法加速器，数据流优化
- FPGA, Chisel

### FuseKNA: Fused Kernel Convolution based Accelerator for Deep Neural Networks

- 利用稀疏性进行核融合，CNN 加速器

### GradPIM: A Practical Processing-in-DRAM Architecture for Gradient Descent

- 存内计算加速参数更新
- 新型架构

### SpaceA: Sparse Matrix Vector Multiplication on Processing-in-Memory Accelerator

- 存内计算加速稀疏矩阵乘法

## 算法+硬件

### Mix and Match: A Novel FPGA-Centric Deep Neural Network Quantization Framework

- 提出新型的量化方案，并在 FPGA 上实现加速

### CSCNN: Algorithm-hardware Co-design for CNN Accelerators using Centrosymmetric Filters

- 提出了一种 CNN 压缩算法和硬件实现

## 编译器

### A Computational Stack for Cross-Domain Acceleration

- 跨领域计算的编译器

### 传统硬件上的加速

#### SpAtten: Efficient Sparse Attention Architecture with Cascade Token and Head Pruning

- 注意力机制训练推理优化，在 CPU, GPU 上最高实现数百倍加速和数千倍能效

#### Tensor Casting: Co-Designing Algorithm-Architecture for Personalized Recommendation Training

- 推荐系统训练加速，分析负载特征，实现加速方案

### other

#### Revisiting HyperDimensional Learning for FPGA and Low-Power Architectures

- HyperDimensional Learning

#### Lazy Batching: An SLA-aware Batching System for Cloud Machine Learning Inference

- 服务器 AI 推理任务调度

#### Ultra-Elastic CGRAs for Irregular Loop Specialization

- 可重构加速器, 粗粒度可重构阵列 (CGRA)
- 一般加速器集成到 Soc 后无法用于不规则负载
- 提出超弹性 CGRA，通过调整 PE 单元的电压、频率，提高性能和能效

#### Layerweaver: Maximizing Resource Utilization of Neural Processing Units via Layer-Wise Scheduling

- AI 推理调度

#### NeuroMeter: An Integrated Power, Area, and Timing Modeling Framework for Machine Learning Accelerators

- 谷歌设计 TPU 用的一种 IC 验证框架

## ISCA 2021

### 加速器架构

#### Ten Lessons From Three Generations Shaped Google’s TPUv4i : Industrial Product

- 谷歌 TPUv4i

#### Sparsity-Aware and Re-configurable NPU Architecture for Samsung Flagship Mobile SoC

- 三星的 NPU

### Energy Efficiency Boost in the AI-Infused POWER10 Processor

- IBM公司 POWER10 AI 加速

### RaPiD: AI Accelerator for Ultra-low Precision Training and Inference

- 一款低精度 AI 加速芯片

## 加速器单元

### FORMS: Fine-grained Polarized ReRAM-based In-situ Computation for Mixed-signal DNN Accelerator

- 忆阻器加速 DNN, 存内计算

### NASGuard: A Novel Accelerator Architecture for Robust Neural Architecture Search (NAS) Networks

- 利用网络结构搜索得出的模型具有多分枝的特性
- 提出一种加速 NAS 网络的架构

### Albireo: Energy-Efficient Acceleration of Convolutional Neural Networks via Silicon Photonics

- 光子器件加速 DNN

### Dual-side Sparse Tensor Core

- 既利用权重稀疏性，又利用激活稀疏性的张量计算单元设计

### GoSPA: An Energy-efficient High-performance Globally Optimized SParse Convolutional Neural Network Accelerator

- 一种稀疏 CNN 加速器

## 算法+硬件

### η-LSTM: Co-Designing Highly-Efficient Large LSTM Training via Exploiting Memory-Saving and Architectural Design Opportunities

- 软硬件加速 LSTM

### ELSA: Hardware-Software Co-design for Efficient, Lightweight Self-Attention Mechanism in Neural Networks

- 软硬件加速自注意力

- 专用硬件

### Cambricon-Q: A Hybrid Architecture for Efficient Training

- 对神经网络量化训练算法进行硬件加速

### NASA: Accelerating Neural Network Design with a NAS Processor

- 对 one-shot based NAS 算法进行加速的加速器

## 编译器

### CoSA: Scheduling by Constrained Optimization for Spatial Accelerators

- DNN 加速器运行时需要大量调度和运行时参数
- 将调度决策表示为一个约束优化问题，使用数学优化技术确定性地解决该问题

### TENET: A Framework for Modeling Tensor Dataflow Based on Relation-centric Notation

- 张量计算编译加速

### SARA: Scaling a Reconfigurable Dataflow Accelerator

- 半导体行业“暗硅”问题推动可重构数据流加速器 (RDA) 的兴起
- 面向 RDA 的编译器

### HASCO: Towards Agile HArdware and Software CO-design for Tensor Computation

- 张量计算编译器
- 软件、硬件设计空间探索

## 传统硬件上的加速

### REDUCT: Keep it Close, Keep it Cool! — Scaling DNN Inference on Multi-core CPUs with Near-Cache Compute

- CPU 上的 DNN 加速

## other

### RingCNN: Exploiting Algebraically-Sparse Ring Tensors for Energy-Efficient CNN-Based Computational Imaging

- 环代数 CNN 加速器

### Communication Algorithm-Architecture Co-Design for Distributed Deep Learning

- 大规模分布式训练

**Enabling Compute-Communication Overlap in Distributed Deep Learning Training Platforms**

- 分布式训练，集群

**NN-Baton: DNN Workload Orchestration and Chiplet Granularity Exploration for Multichip Accelerators**

- AI 芯片布局优化