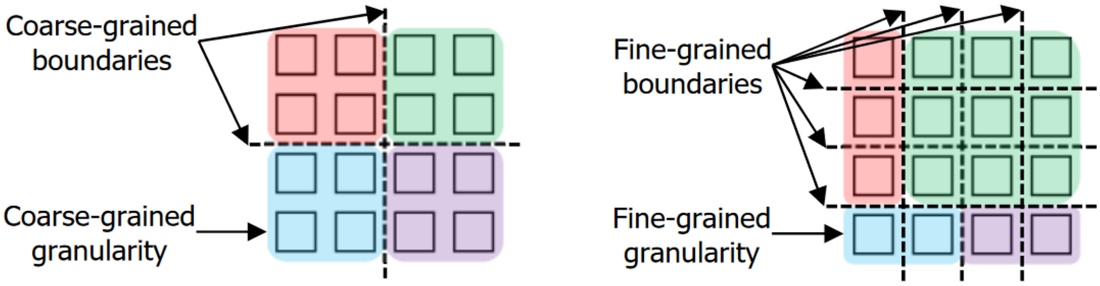
Dataflow Mirroring: Architectural Support for Highly Efficient Fine-Grained Spatial Multitasking on Systolic-Array NPUs

DAC-21



(a) Coarse-grained [5] (b) Fine-grained (Ours)

[5] S. Ghodrati et al., "Planaria: Dynamic Architecture Fission for Spatial Multi-Tenant Acceleration of Deep Neural Networks," in *MICRO*, 2020.

a software runtime，动态的硬件资源分配

[3] Y. Choi and M. Rhu, "PREMA: A Predictive Multi-task Scheduling Algorithm For Preemptible NPUs," in *HPCA*, 2020.

Benchmark

[9] P. Mattson et al., "MLPerf: An Industry Standard Benchmark Suite for Machine Learning Performance," *IEEE Micro*, 40(2), 2020.

[10] V. J. Reddi et al., "MLPerf Inference Benchmark," in *ISCA*, 2020.

模拟片下 DMA 访存时间

[7] S. Li et al., "DRAMsim3: A Cycle-Accurate, Thermal-Capable DRAM Simulator," *IEEE CAL*, 2020.

周期精确的模拟器

[11] A. Samajdar et al., "A Systematic Methodology for Characterizing Scalability of DNN Accelerators using SCALE-Sim," in *ISPASS*, 2020.

模拟能耗、面积

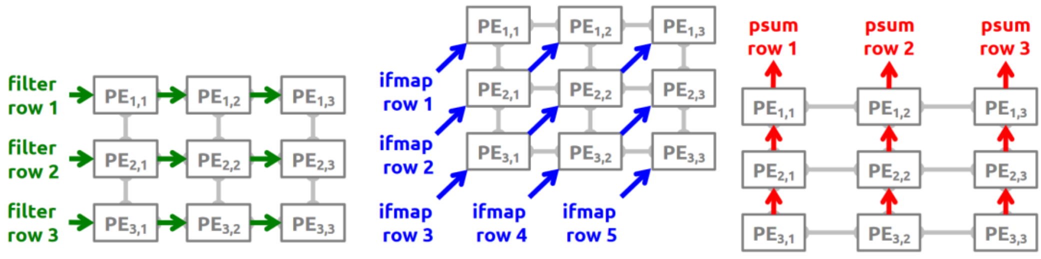
[17] Y. N. Wu et al., "Accelergy: An Architecture-Level Energy Estimation Methodology for Accelerator Designs," in *ICCAD*, 2019.

负载：STP, ANTT

[4] S. Eyerman and L. Eeckhout, "System-Level Performance Metrics for Multiprogram Workloads," *IEEE Micro*, 2008.

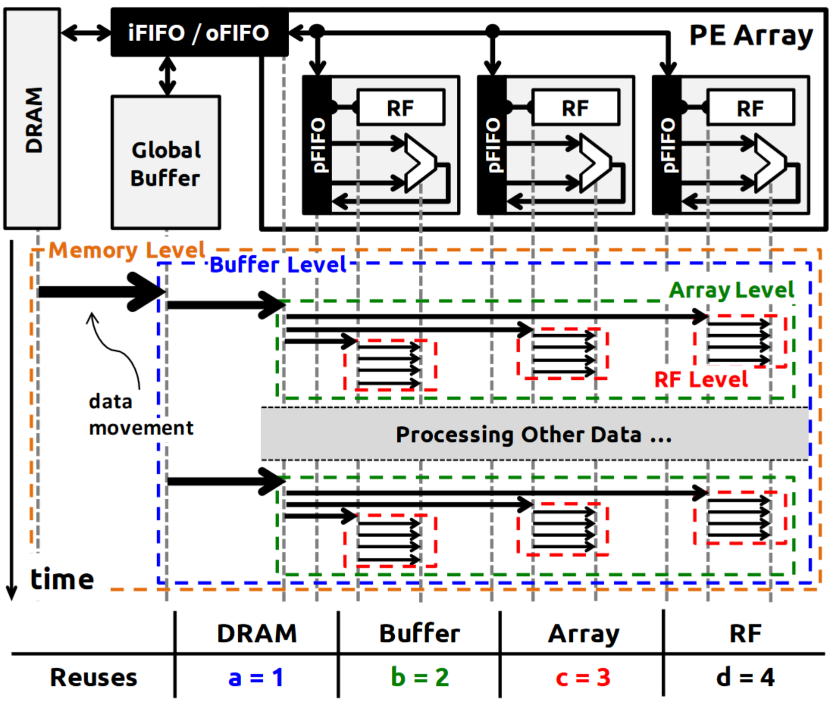
Eyeriss: A Spatial Architecture for Energy-Efficient Dataflow for Convolutional Neural Networks

ISCA-16



(a) (b) (c)

Dataflow	Data Handling
WS	Maximize *convolutional reuse* and *filter reuse* of weights in the RF.
SOC-MOP OS	Maximize *psum accumulation* in RF. *Convolutional reuse* in array.
MOC-MOP OS	Maximize *psum accumulation* in RF. *Convolutional reuse* and *ifmap reuse* in array.
MOC-SOP OS	Maximize *psum accumulation* in RF. *Ifmap reuse* in array.
NLR	*Psum accumulation* and *ifmap reuse* in array.



Efficient Processing of Deep Neural Networks: A Tutorial and Survey

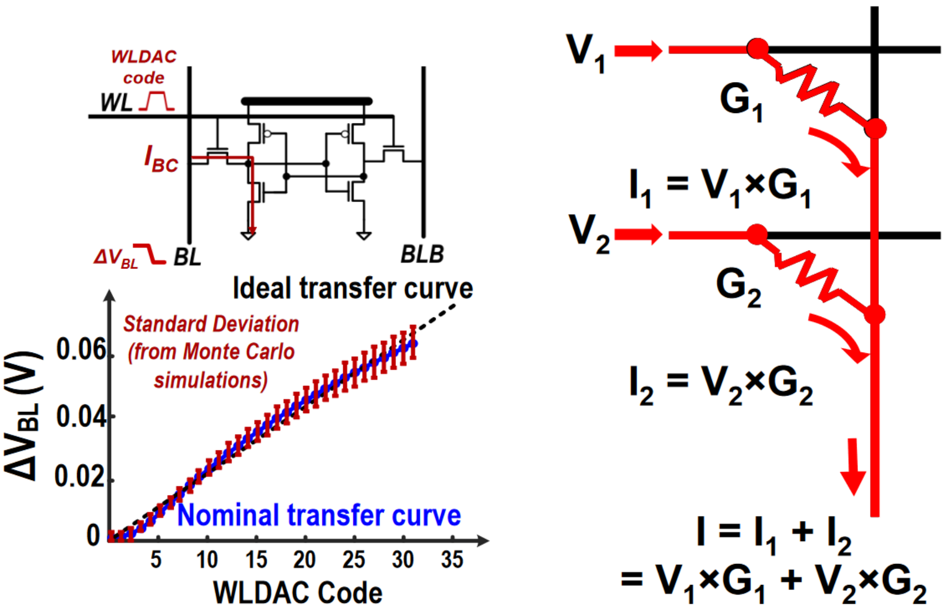
MICRO-17

1. *DNN 基础*

2. 众多用于 DNNs 的硬件平台和优化技术（不损失精度）

- For *temporal* architectures such as CPUs and GPUs, we will discuss how *computational transforms* on the kernel can reduce the number of multiplications to *increase throughput*.
- For *spatial* architectures used in accelerators, we will discuss how *dataflows* can increase data reuse from low cost memories in the memory hierarchy to *reduce energy consumption*.

3. 近数据处理(NDP)解决数据搬移的功耗



(a) Multiplication performed by bit-cell (Figure from [102])

(b) G_i is conductance of resistive memory (Figure from [104])

4. 通过降低精度提升吞吐量和能效的联合算法和硬件平台

   - *Reduce precision of operations and operands*. This includes going from floating point to fixed point, reducing the bitwidth, non-linear quantization and weight sharing.
   - *Reduce number of operations and model size*. This includes techniques such as compression, pruning and compact network architectures.

5. 对比不同硬件效果必须考虑的关键矩阵负载

   - DNN model 需要用到的矩阵特性
   - DNN hardware 需要用到的矩阵特性