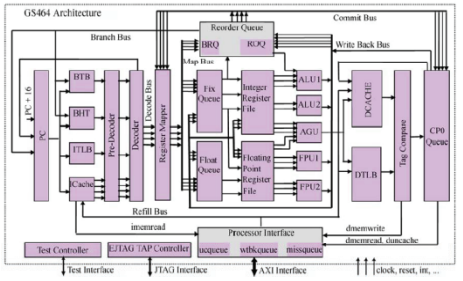
寒武纪神经网络计算机

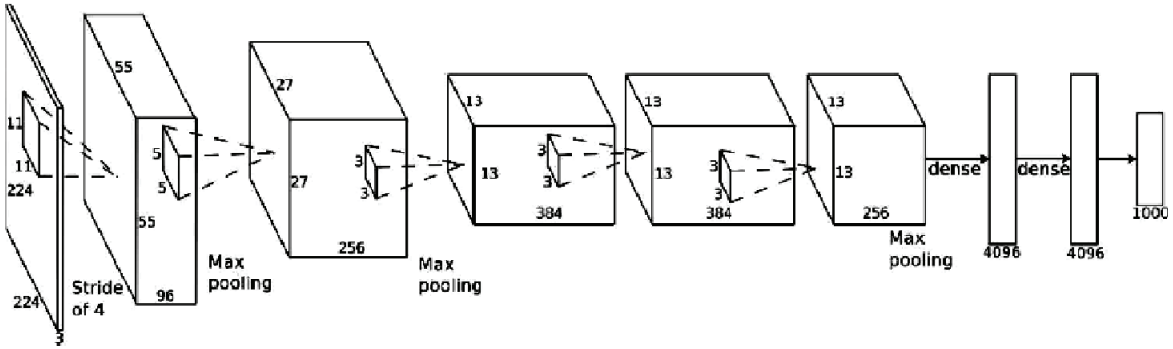
中国科学院计算技术研究所
计算机体系结构国家重点实验室
陈天石

体系结构与机器学习的交叉

- **体系结构的角度**
  - 内部结构复杂的系统
  - AI方法辅助设计



- **机器学习的角度**
  - 支撑机器学习的手段是计算
  - 计算系统的能力是潜在瓶颈



提纲

- 基于机器学习方法的处理器研发
  - 优化处理器结构参数
  - 优化处理器片上网络
  - 处理器功能验证
- 寒武纪神经网络计算机
  - 寒武纪1号（DianNao）
  - 寒武纪2号（DaDianNao）
  - 未来展望

优化处理器结构参数

- 处理器研发初期的**最关键步骤**
  - 选择最合适的处理器参数以最大化处理器的性能/效能
  - 待定参数很多：发射宽度、功能部件数量、各级缓存大小…
  - 设计约束不少：“功耗<50瓦”、“面积<10mm^2”…

Abbr.	Parameter	Value
WIDTH	Fetch/Issue/Commit Width	2,4,6,8
FUNIT	FPALU/FPMULT Units	2,4,6,8
IUNIT	IALU/IMULT Units	2,4,6,8
L1IC	L1-ICache	1,2,4,8,16,32KB
L1DC	L1-DCache	1,2,4,8,16,32KB
L2UC	L2-UCache	256,512,1024,2048,4096KB
ROB	ROB size	16-256 with a step of 16
LSQ	LSQ size	8-128 with a step of 8
GSHARE	GShare size	1,2,4,8,16,32K
BTB	BTB size	512,1024,2048,4096
Total	**10 parameters**	**70,778,880 Options**

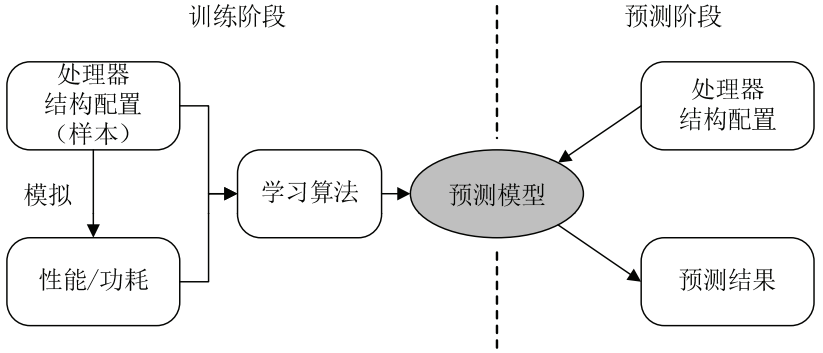
优化处理器结构参数

- 处理器可选参数组合极多
  - 处理器结构空间规模随参数数量指数增长
    - 数千万甚至上亿可能的参数组合
- 处理器模拟速度极为缓慢
  - 芯片还没制造出来，只以软件形式模拟执行
    - 模拟速度比真实处理器运算速度差了3~5 个数量级！
- 蛮力遍历整个参数空间？
  - 总耗时 = 模拟单个处理器参数组合的时间 x 可能的参数组合总数
  - 演化算法 + 代理模型

Abbr.	Parameter	Value
WIDTH	Fetch/Issue/Commit Width	2,4,6,8
FUNIT	FPALU/FPMULT Units	2,4,6,8
IUNIT	IALU/IMULT Units	2,4,6,8
L1IC	L1-ICache	1,2,4,8,16,32KB
L1DC	L1-DCache	1,2,4,8,16,32KB
L2UC	L2-UCache	256,512,1024,2048,4096KB
ROB	ROB size	16-256 with a step of 16
LSQ	LSQ size	8-128 with a step of 8
GSHARE	GShare size	1,2,4,8,16,32K
BTB	BTB size	512,1024,2048,4096
Total	**10 parameters**	**70,778,880 Options**

优化处理器结构参数

- 处理器性能的回归建模(IJCAI'11; TIST'13)
  - 采样：模拟少量处理器参数组合
    - 处理器结构优化的主要时间开销
  - 建模：通过样本训练回归模型
    - 半监督学习(co-training)、主动学习
  - 预测：通过模型预测处理器参数组合对应的绝对性能/功耗



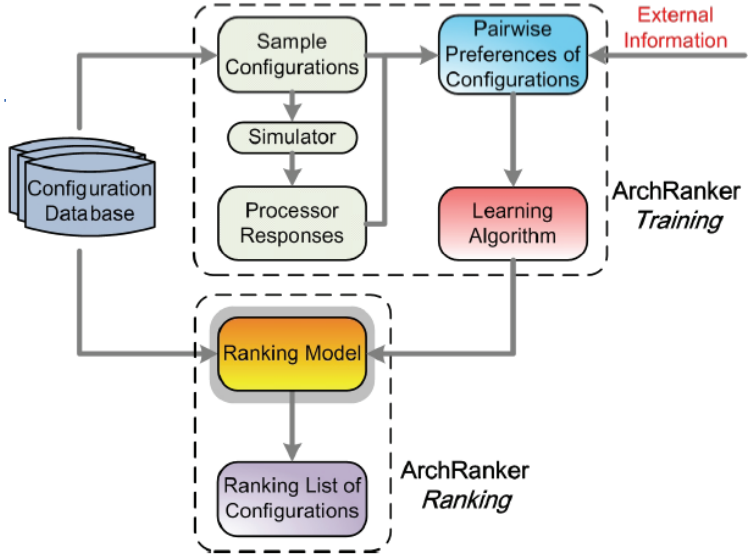
the labeled data set should decrease the most if the most confident unlabeled instance is labeled. Formally, for each unlabeled instance $\mathbf{x}_s$, the quality of $\mathbf{x}_s$ can be measured using a criterion as shown in Equation (1):

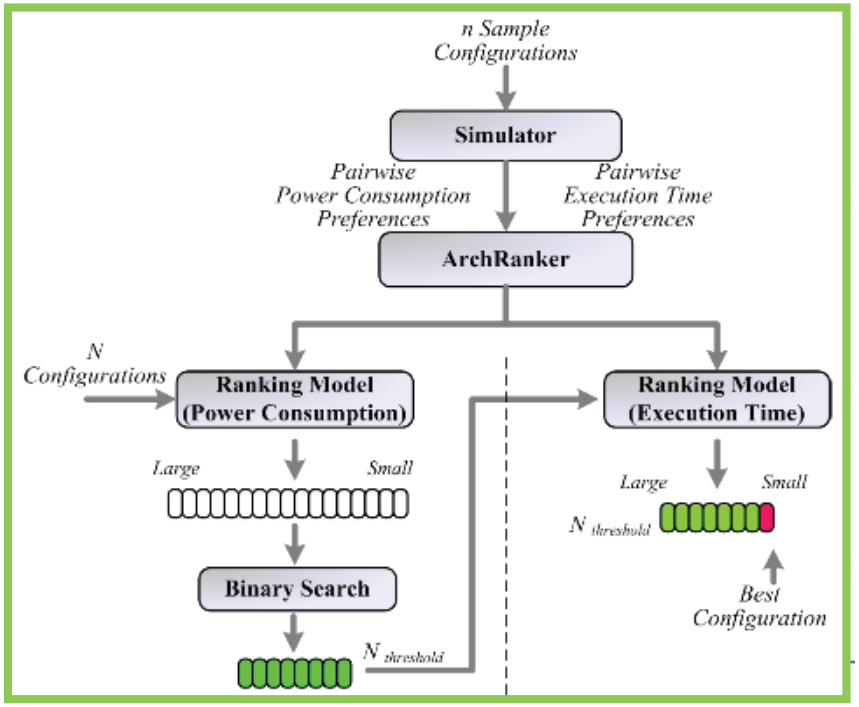
$$\Delta(\mathbf{x}_s) = \frac{1}{|L|}\sum_{\mathbf{x}\in L}\left(\left(y - h(\mathbf{x})\right)^2 - \left(y - h'(\mathbf{x})\right)^2\right), \quad (1)$$

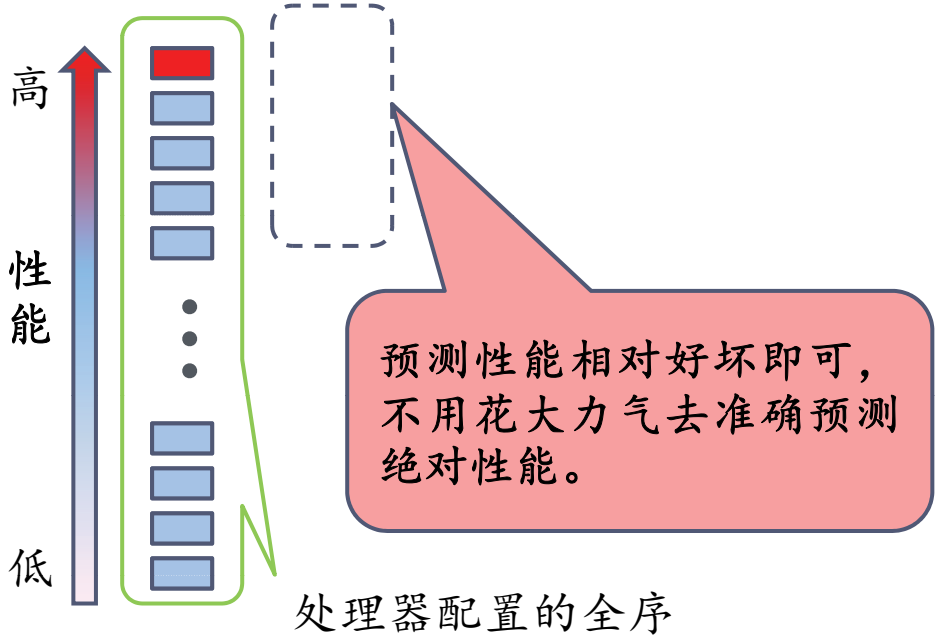
1. 两棵回归树互相为对方标记样本
2. 每次仅标记使最小均方误差降低最大的（未标记）样本

优化处理器结构参数

- 处理器性能的排序建模(ISCA'14)
  - 预测处理器配置的相对好坏
    - 排序学习
  - 处理器参数选择加速3-10倍







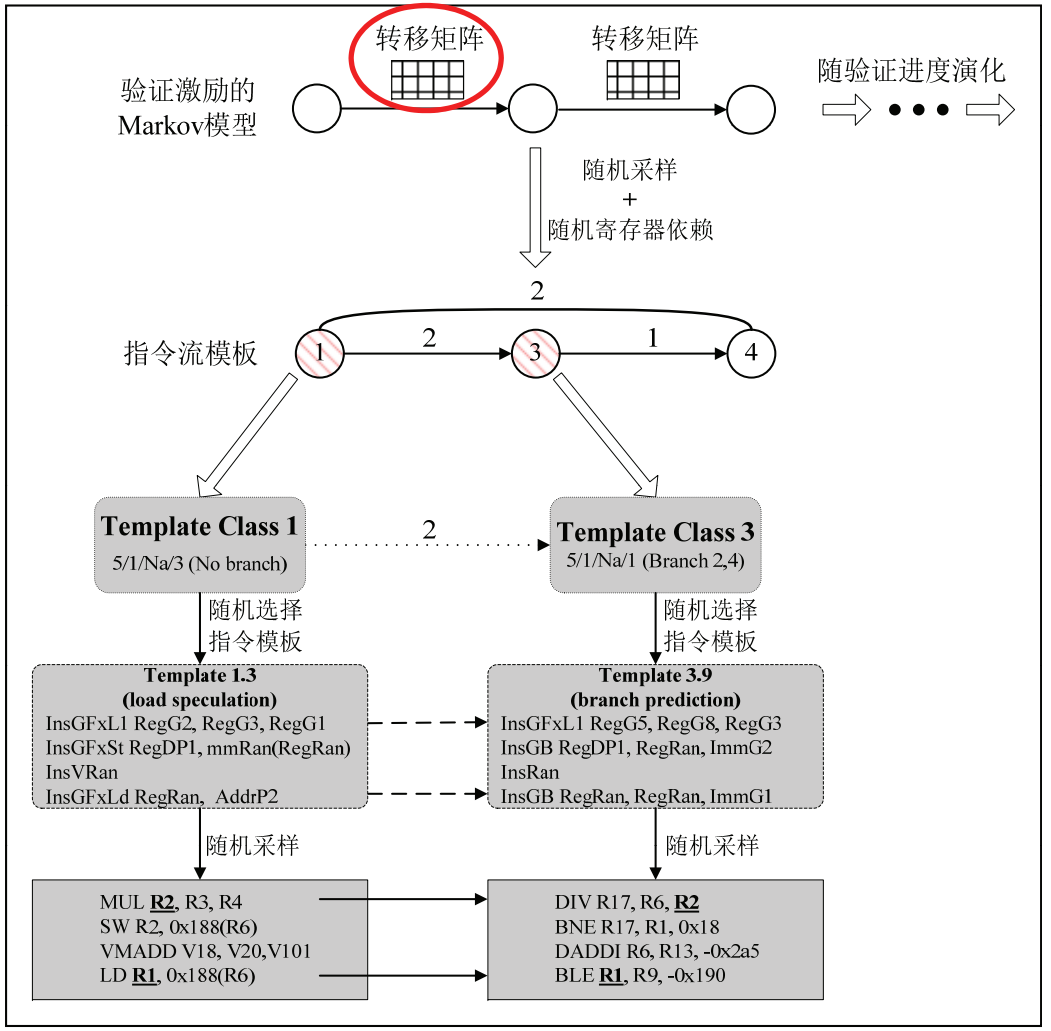
处理器功能验证

- 多核处理器设计70%的时间耗于功能验证
- 一旦流片错误难以修复
  - 数千万重流片费用，上亿召回代价
- 处理器状态空间爆炸
  - 状态数随晶体管数指数增长
- 处理器仿真速度极慢
- 大量功能点需要长验证激励才能触发
  - 百万量级覆盖点
  - 手写验证激励杯水车薪
  - 全随机生成：盲目、维数灾难

Intel花4.75亿美元
召回奔腾

处理器功能验证

- 验证激励的Markov模型
  - 转移矩阵刻画指令上下文关系
  - 每节点对应一小段未决指令流
- 转移矩阵随验证进度演化
  - 推动验证激励向未覆盖点倾斜
  - 避免重复覆盖
- Markov模型采样得到指令流
  - 随机加入节点间寄存器依赖
  - 多节点连结得到长验证激励
- 模块级验证资源分配技术
  - 由历次改版信息建立资源模型
  - 由模型指导验证资源分配



提纲

- 基于机器学习方法的处理器研发
  - 优化处理器结构参数
  - 优化处理器片上网络
  - 处理器功能验证
- 寒武纪神经网络计算机
  - 寒武纪1号（DianNao）
  - 寒武纪2号（DaDianNao）
  - 未来展望

背景：神经网络的复兴

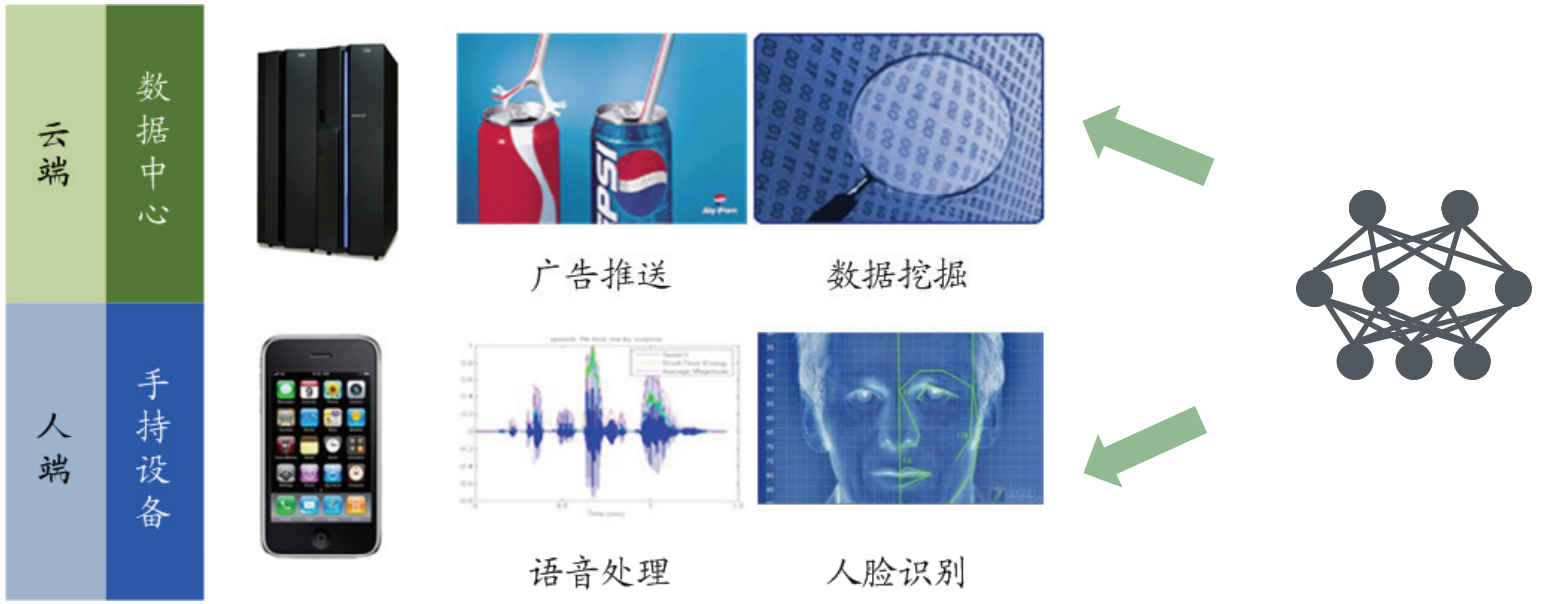
Deep Learning:

*" **With massive amounts of computational power**, machines can now recognize objects and translate speech in real time."*

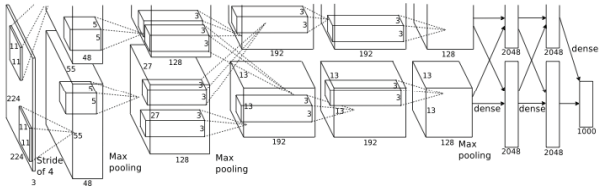
---《MIT技术评论》将深度学习技术评为

2013年十大突破性技术之首

但也暗示深度学习的最主要瓶颈是**计算能力**

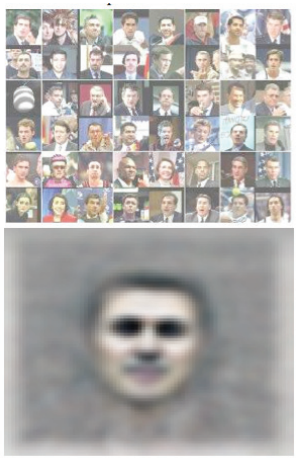


趋势：神经网络规模快速增长



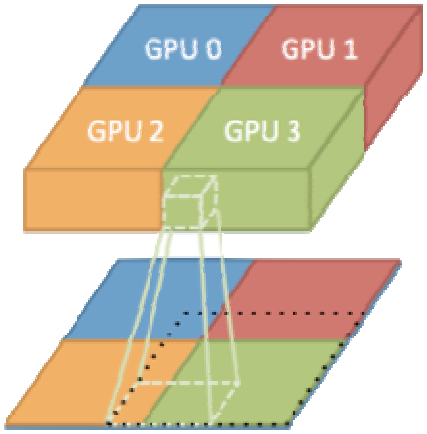
6千万突触

Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). Imagenet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems.



10亿突触

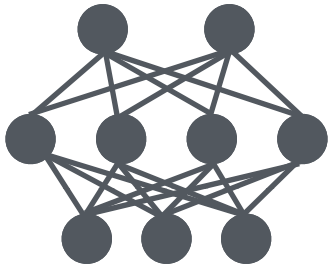
Le, Q. V., Ranzato, M. A., Monga, R., Devin, M., Chen, K., Corrado, G. S., ... Ng, A. Y. (2012). Building High-level Features Using Large Scale Unsupervised Learning. In International Conference on Machine Learning.



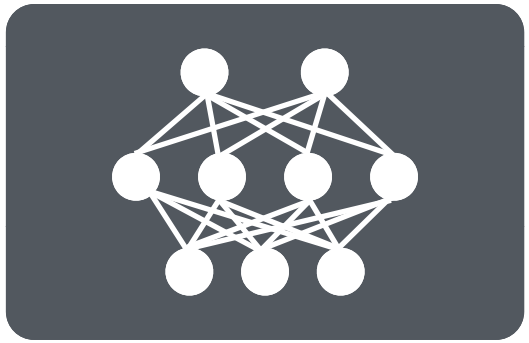
110亿突触

Coates, A., Huval, B., Wang, T., Wu, D. J., & Ng, A. Y. (2013). Deep learning with cots hpc systems. In International Conference on Machine Learning.

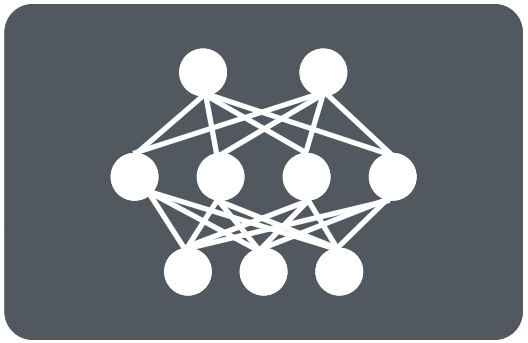
传统的办法



算法

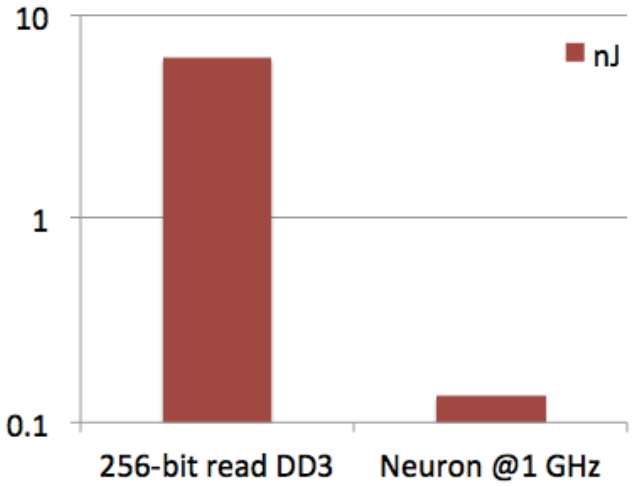


硬件运算单元和算法神经元一一对应



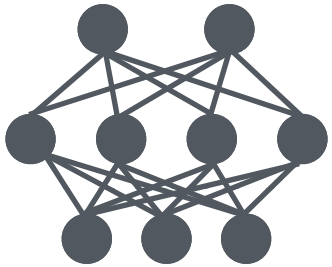
片外内存

把数据从内存搬运到硬件运算单元，甚至比运算本身更耗能量



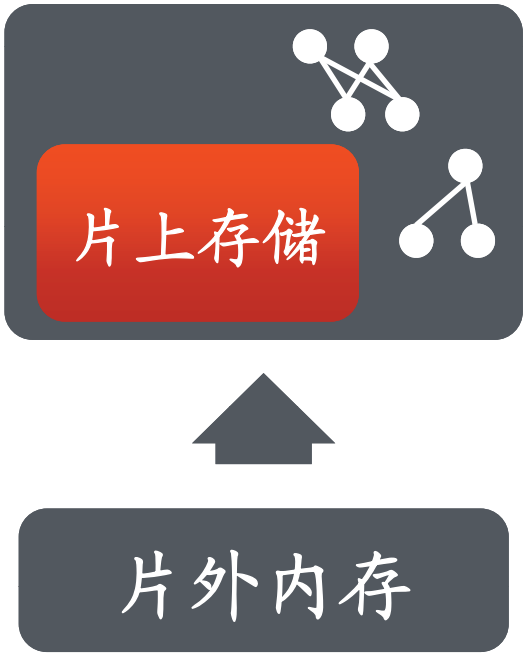
硬件运算单元数量稍微一多，访存带宽就供应不上数据

我们的策略



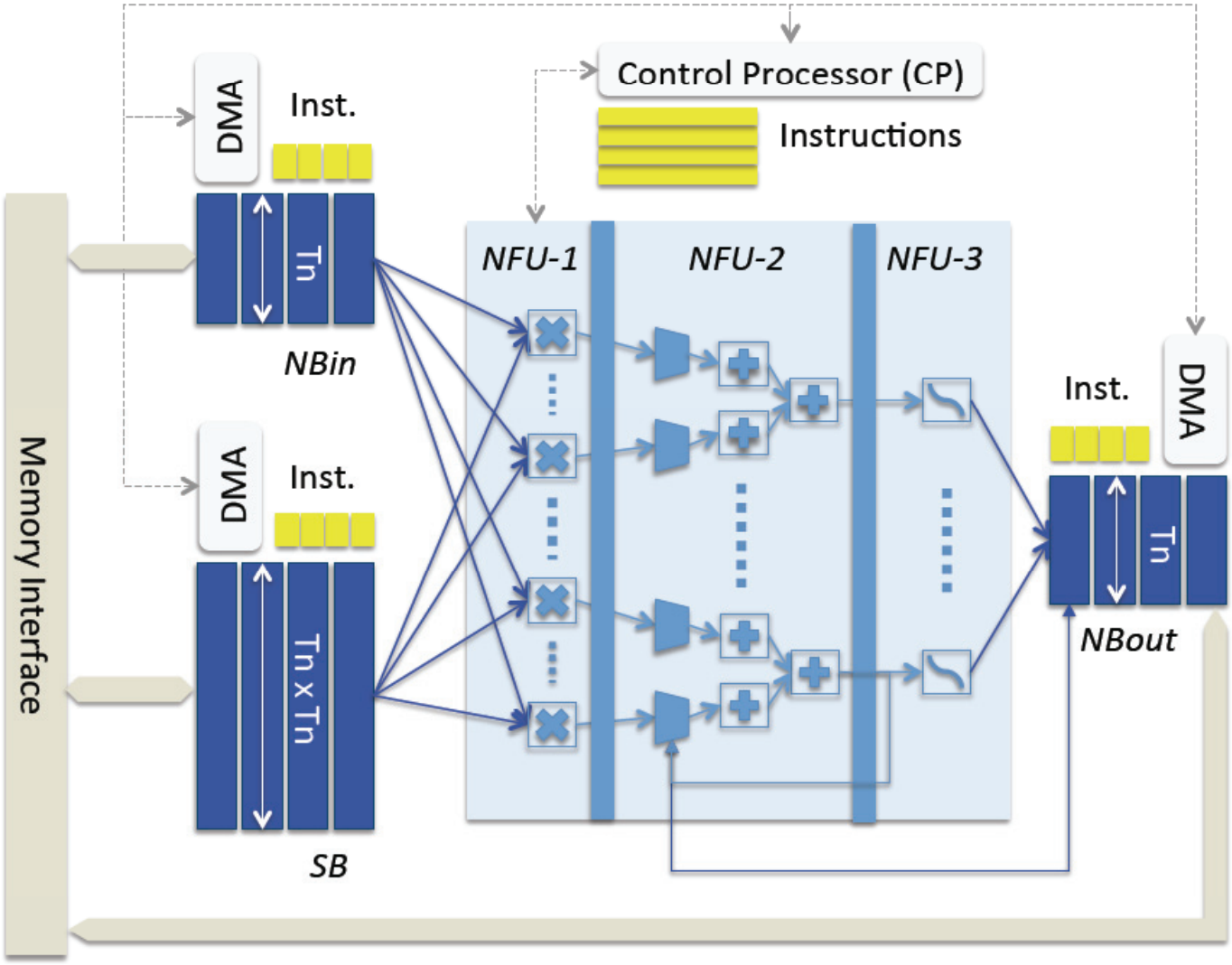
算法

对硬件运算单元
分时复用

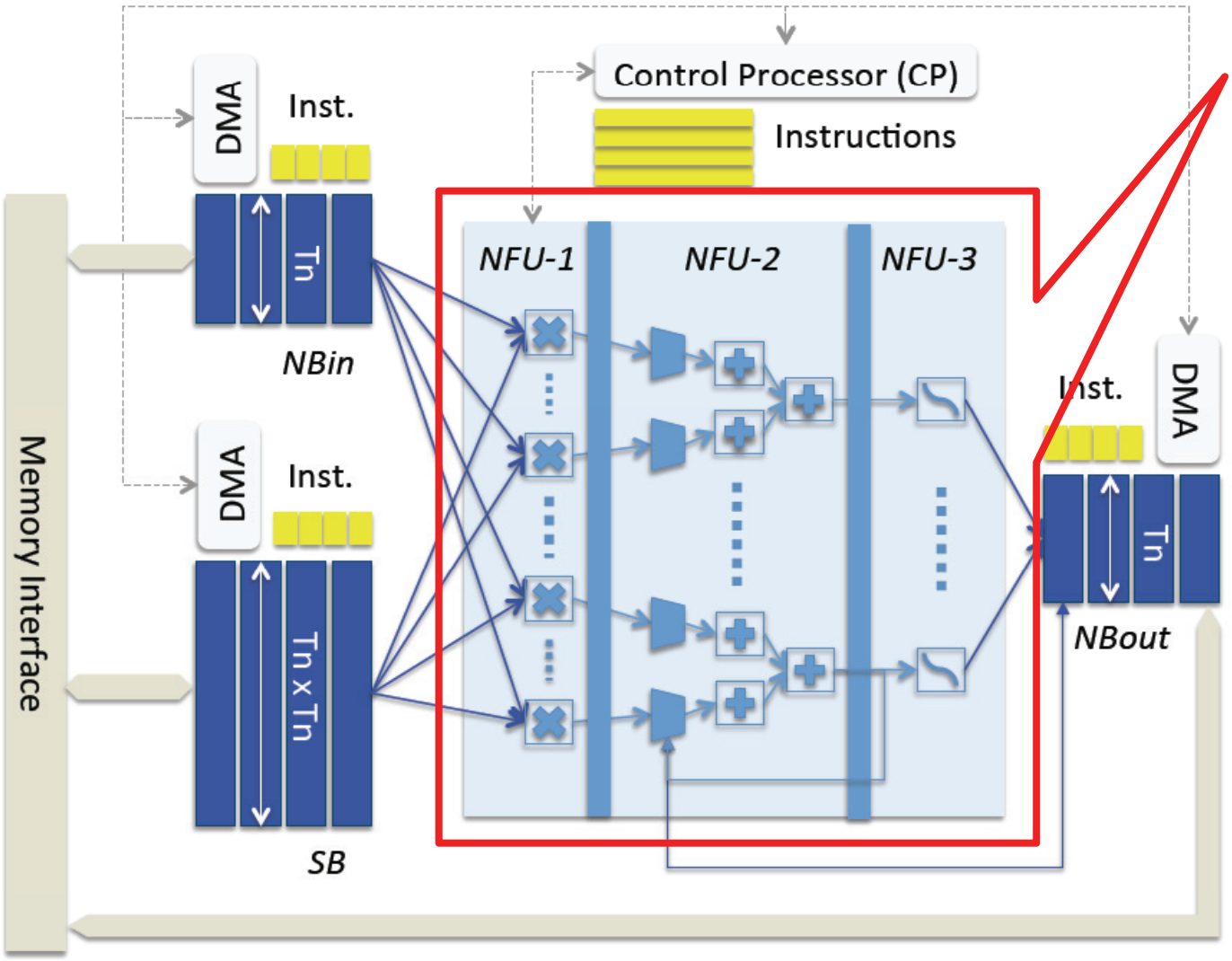


- 小尺度但支持大规模神经网络
- 速度：把访存带宽用起来，尽可能提高性能
- 能耗：通过优化片上存储层次**尽量减少访存次数**

寒武纪1号神经网络处理器架构

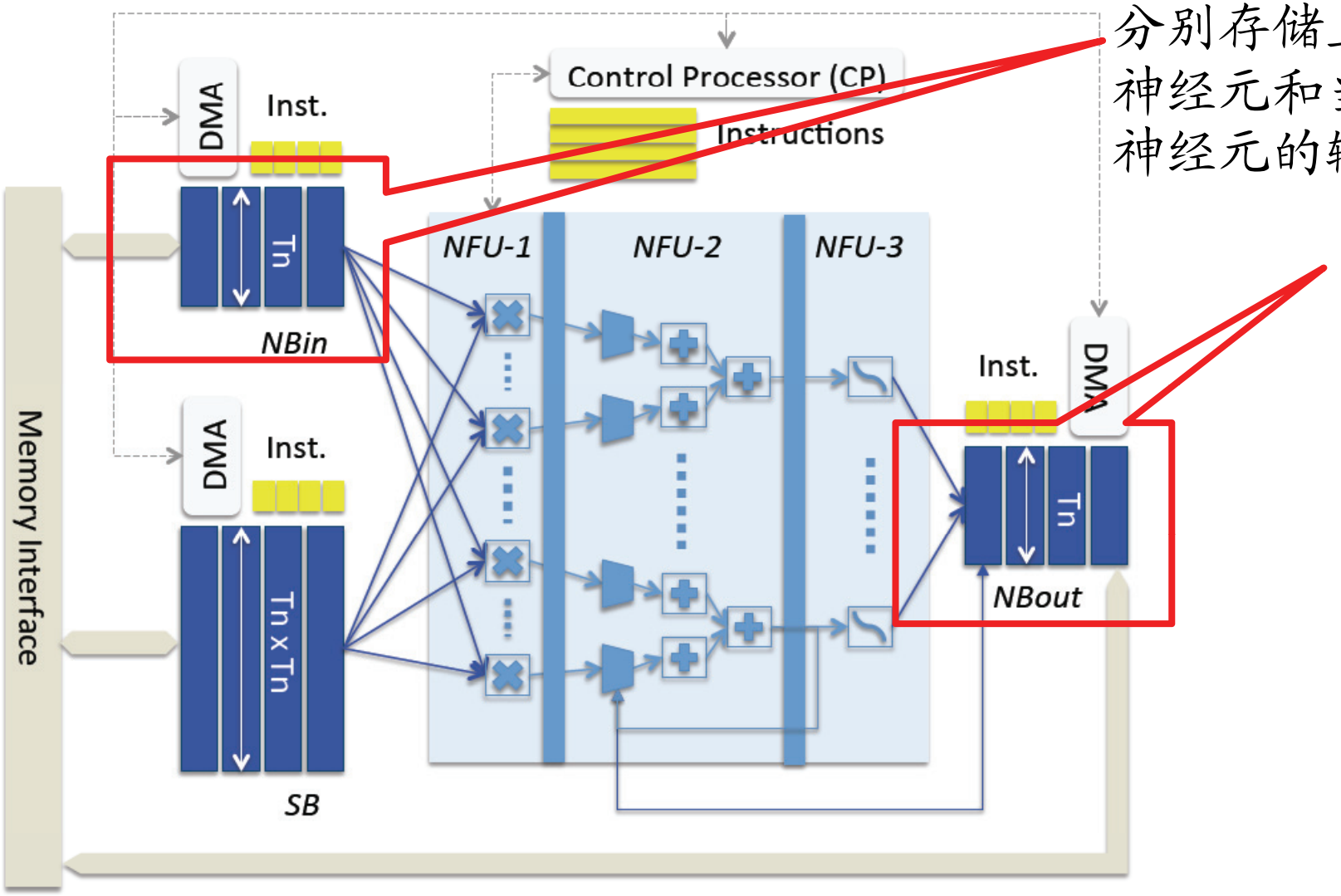


寒武纪1号神经网络处理器架构

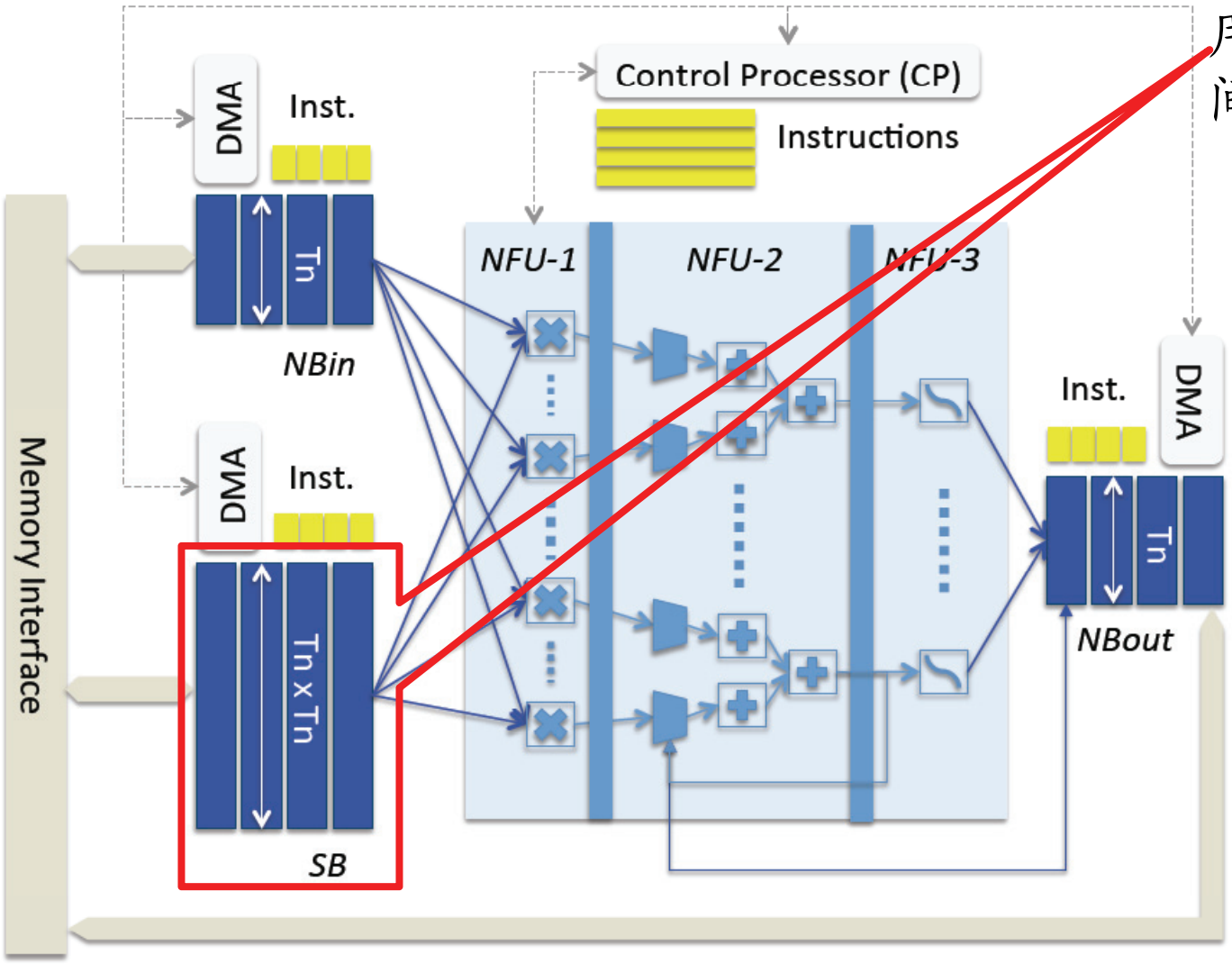


16个神经元硬件运算单元

寒武纪1号神经网络处理器架构

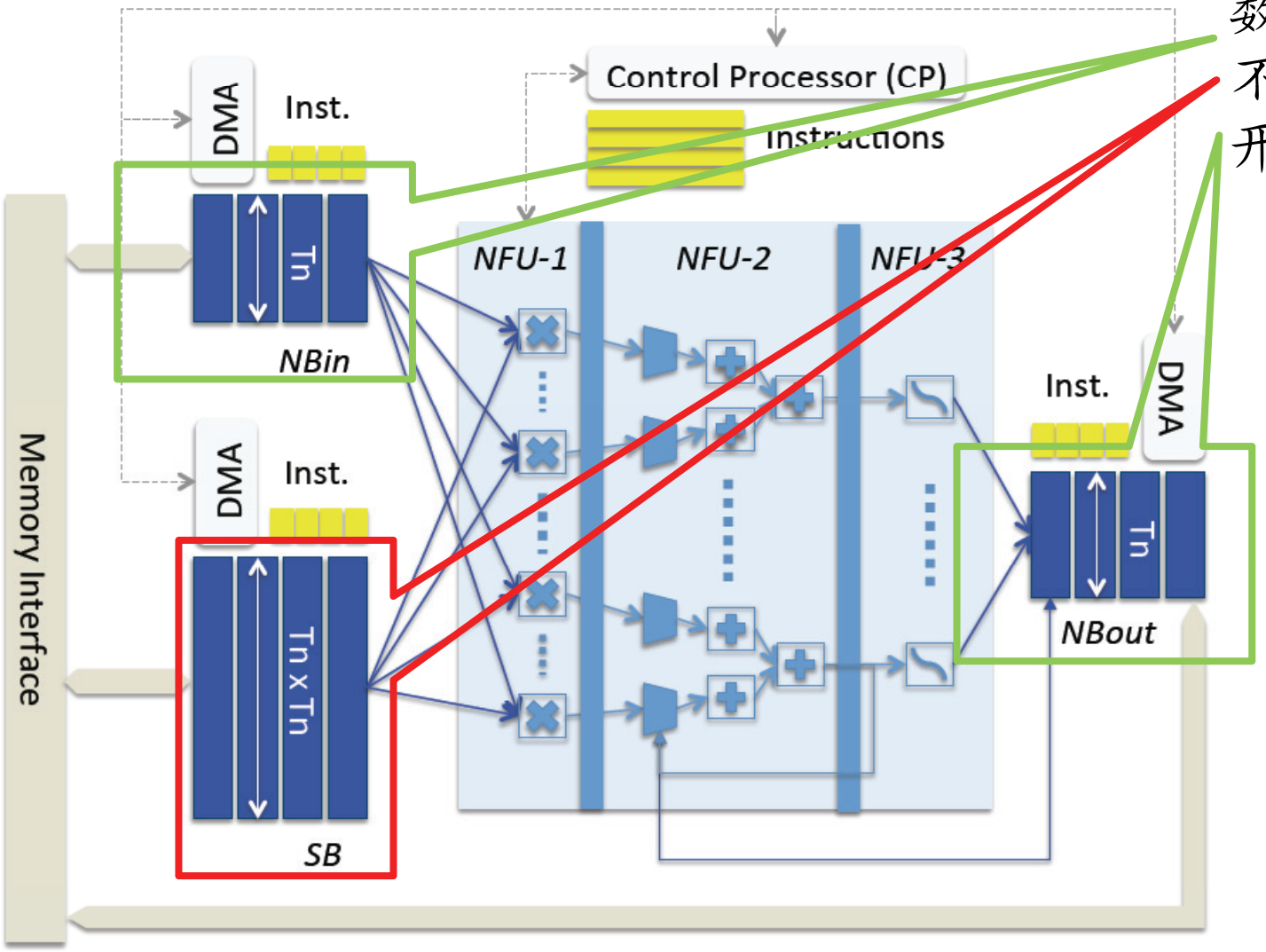


寒武纪1号神经网络处理器架构



用于存储神经元间的突触权重

寒武纪1号神经网络处理器架构



数据重用模式不同，因此分开存储

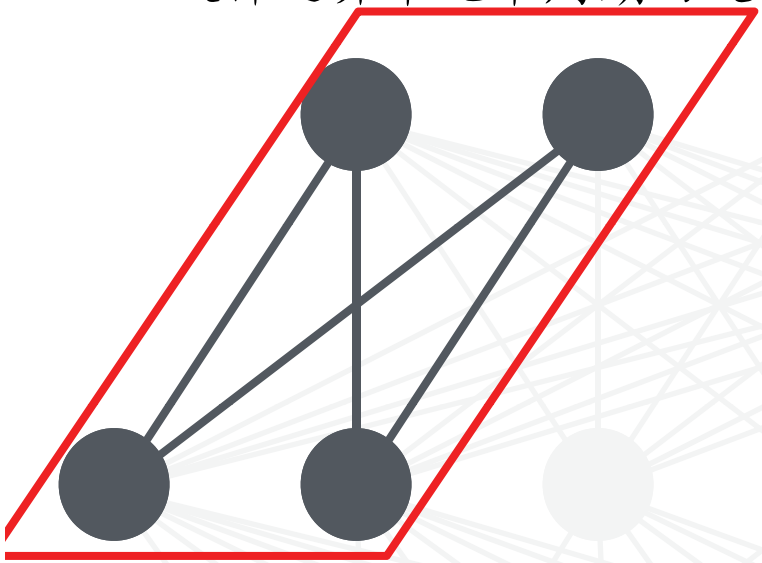
硬件运算单元的分时复用

甲层神经元（输出已知）

乙层神经元（输出待计算）

硬件运算单元的分时复用

硬件运算单元单周期的处理能力



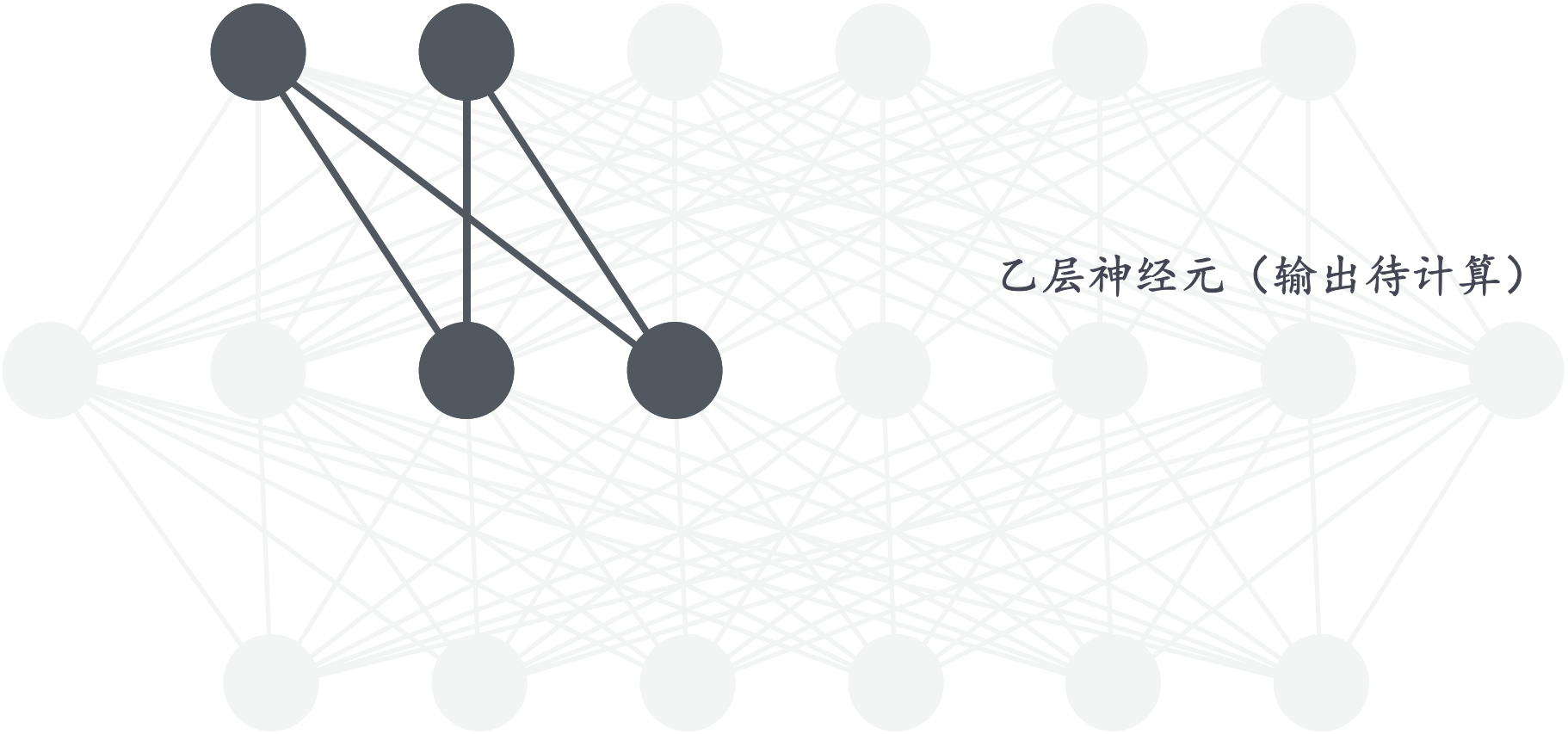
甲层神经元（输出已知）

乙层神经元（输出待计算）

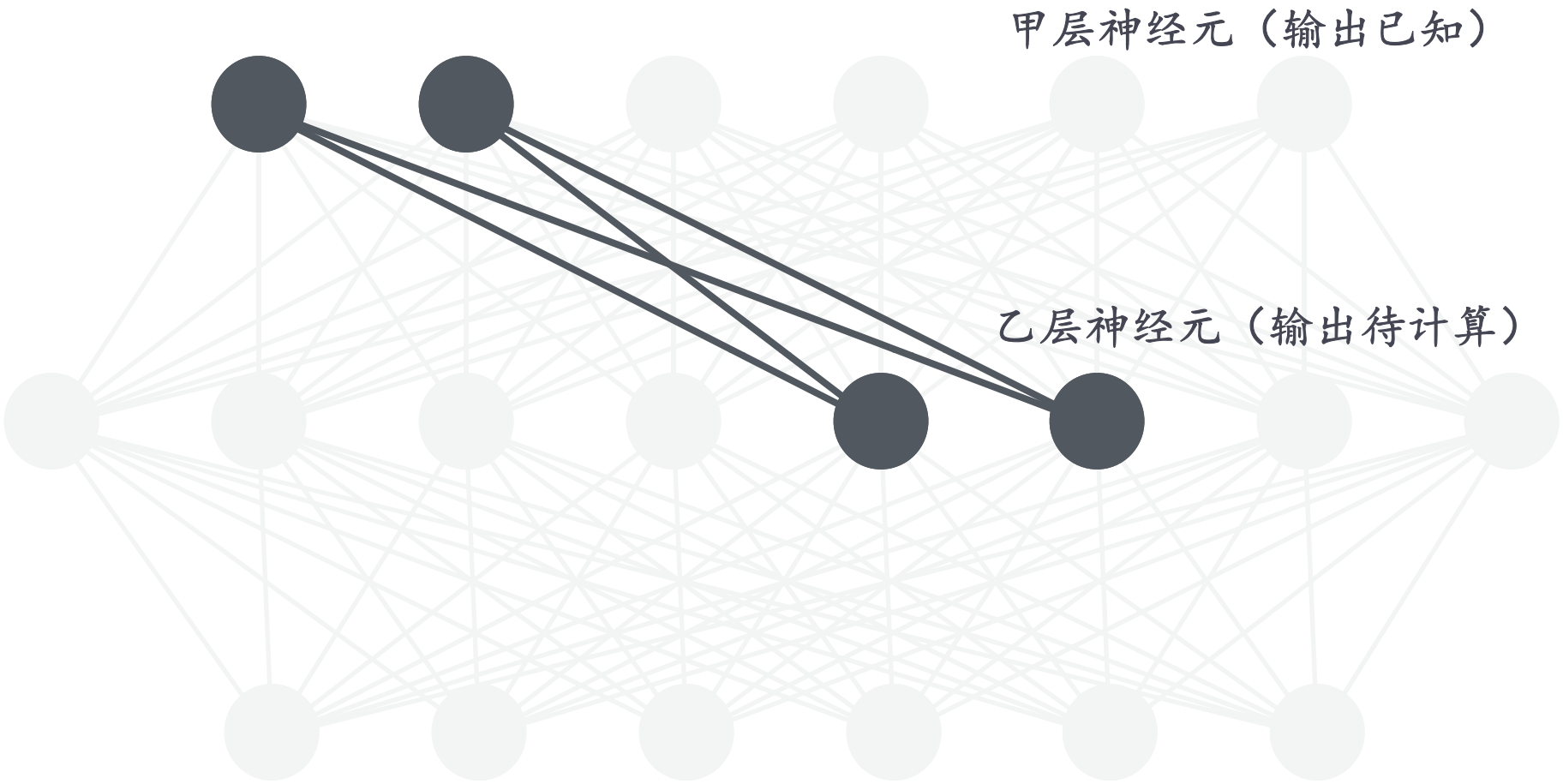
硬件运算单元的分时复用

甲层神经元（输出已知）

乙层神经元（输出待计算）

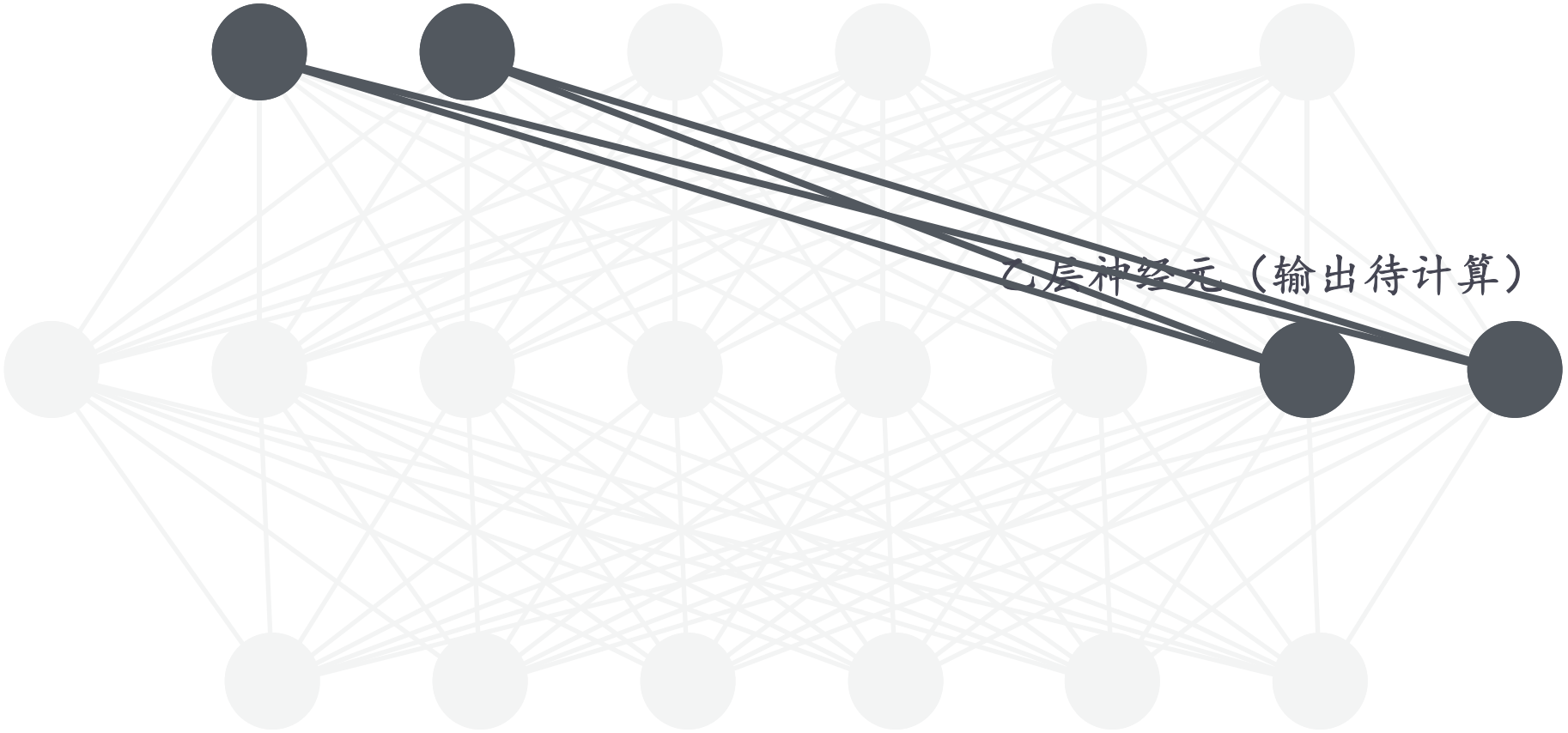


硬件运算单元的分时复用

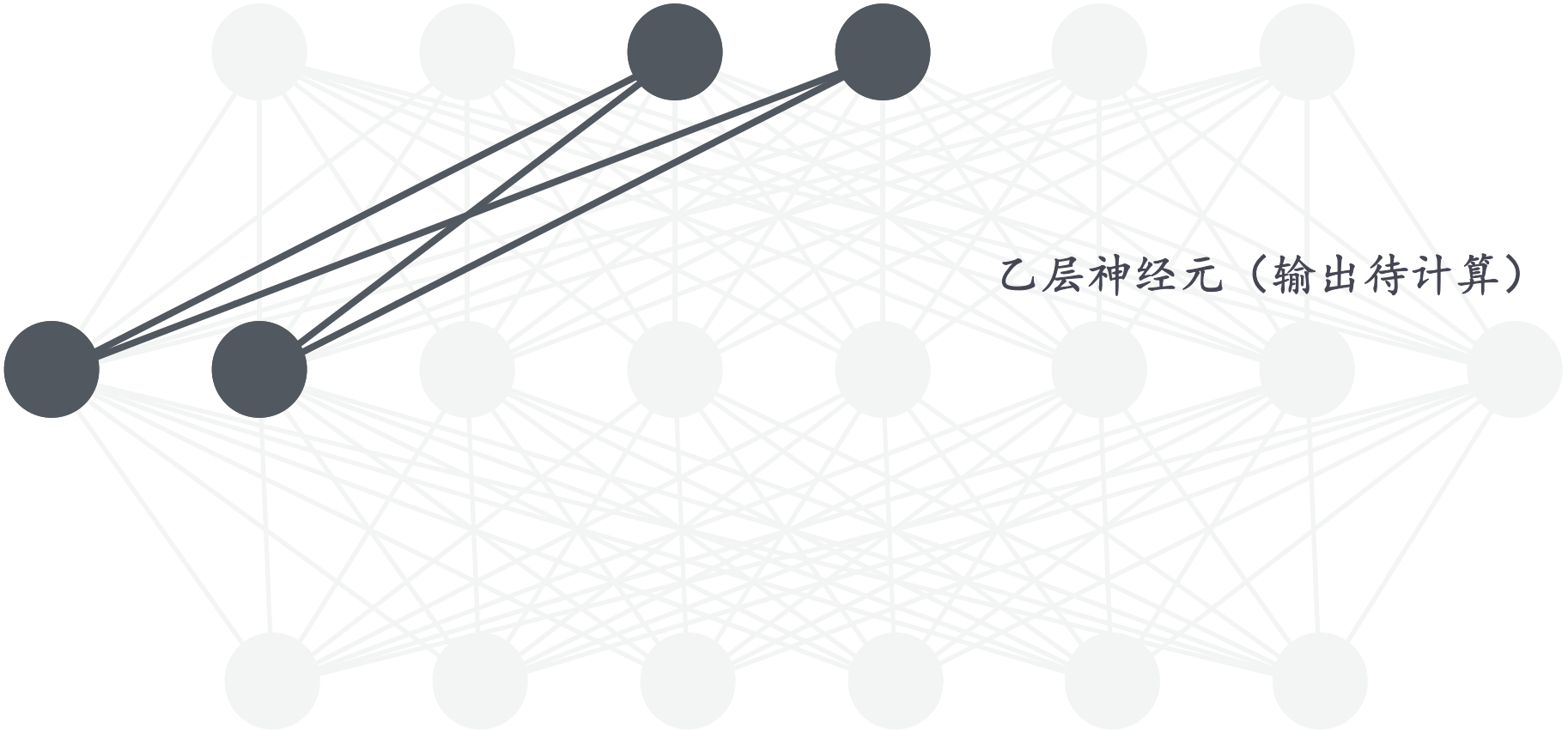


硬件运算单元的分时复用

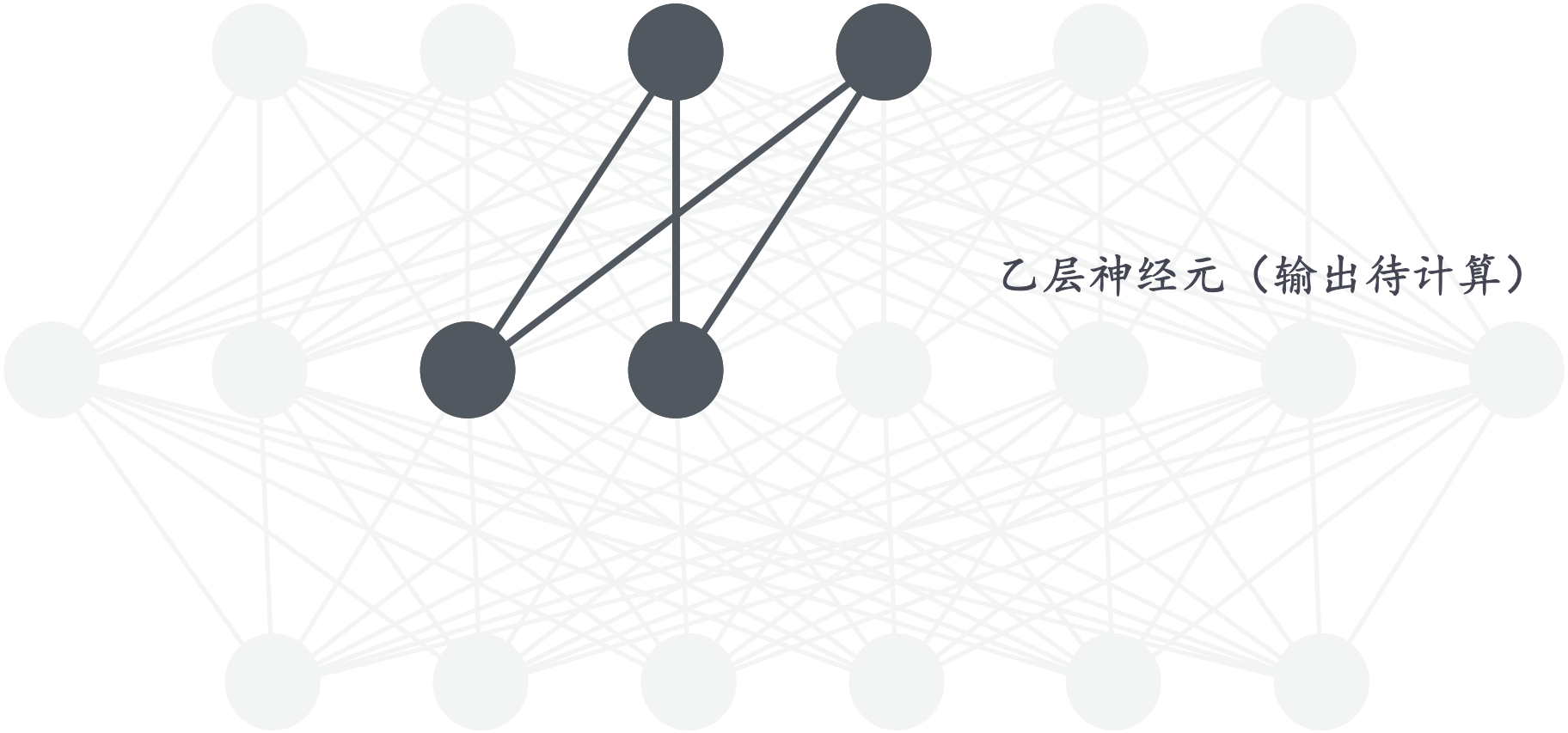
这块神经元的输出数据不再能重用



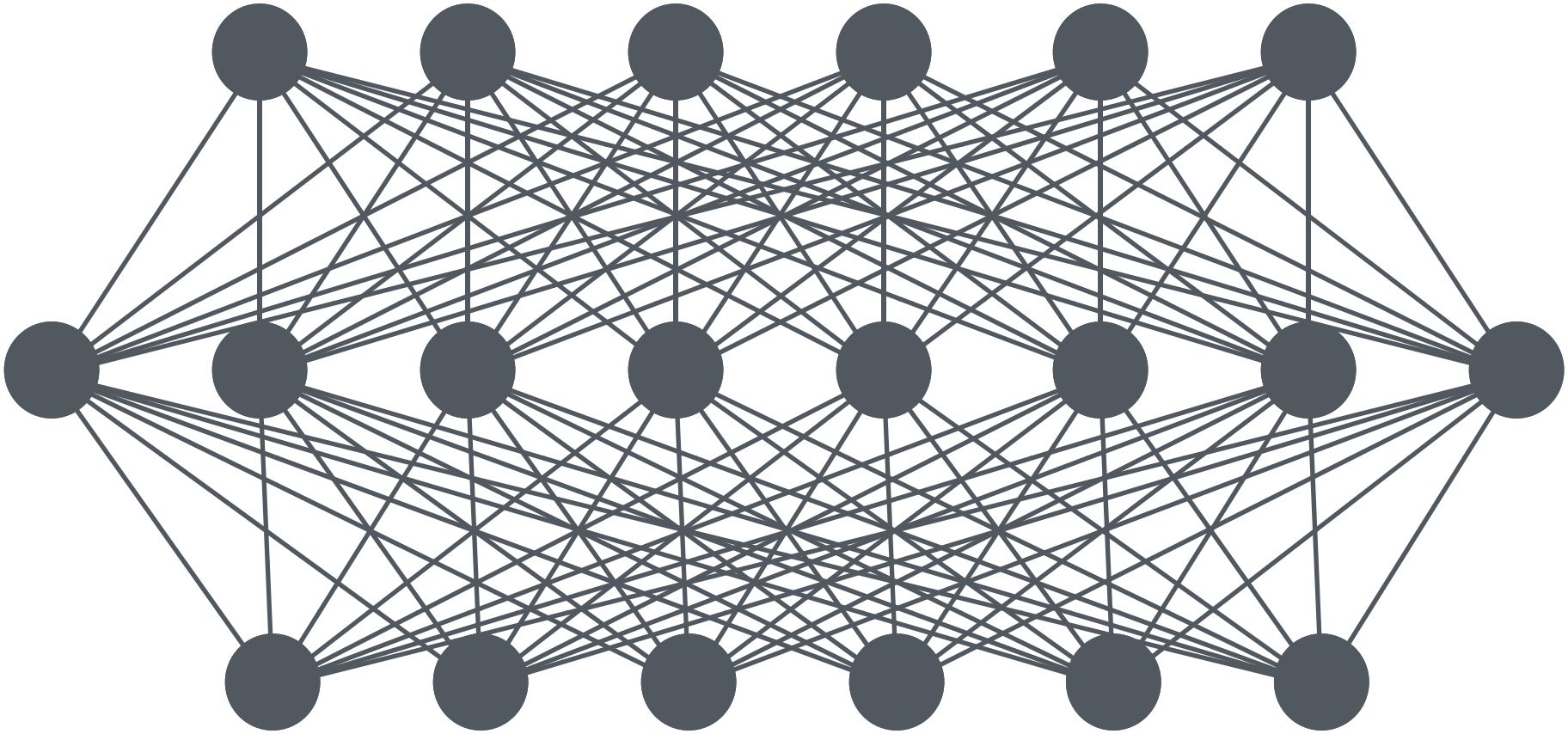
硬件运算单元的分时复用



硬件运算单元的分时复用

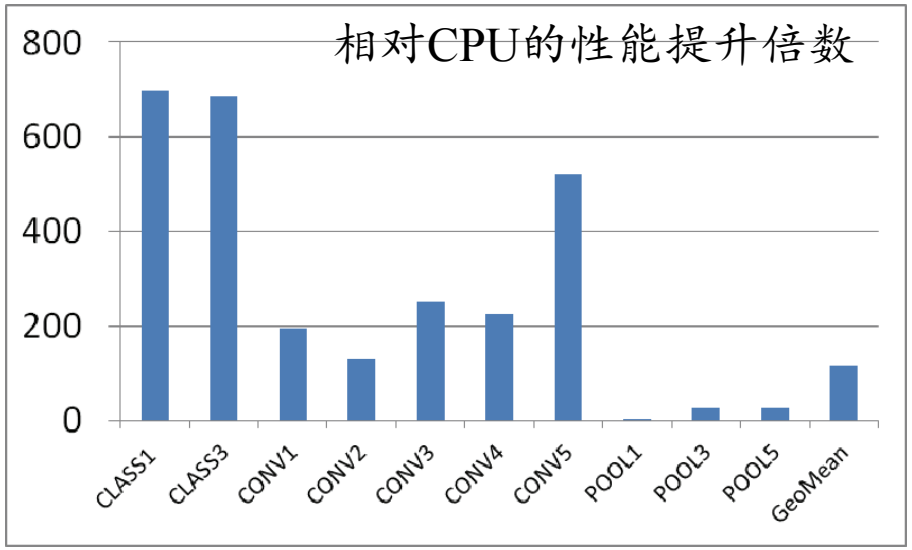


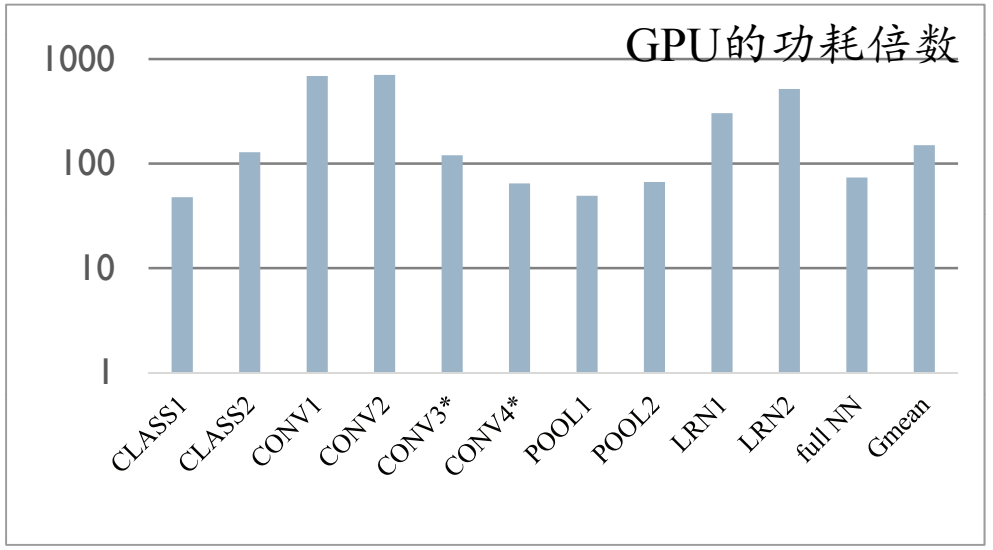
硬件运算单元的分时复用



寒武纪1号神经网络处理器

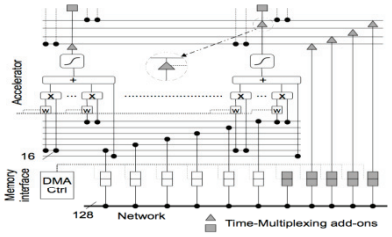
- 支持任意规模DNN、CNN、MLP、SOM等多种神经网络算法
- 0.98GHz，452 GOPS，3mm^2，0.485W @ 65nm
- 10000（甲层）x 10000（乙层）神经网络运算耗时约0.2毫秒

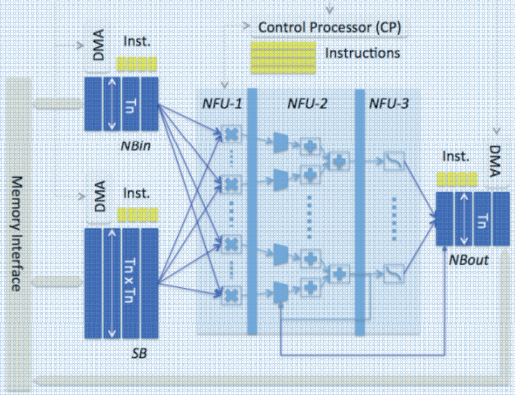


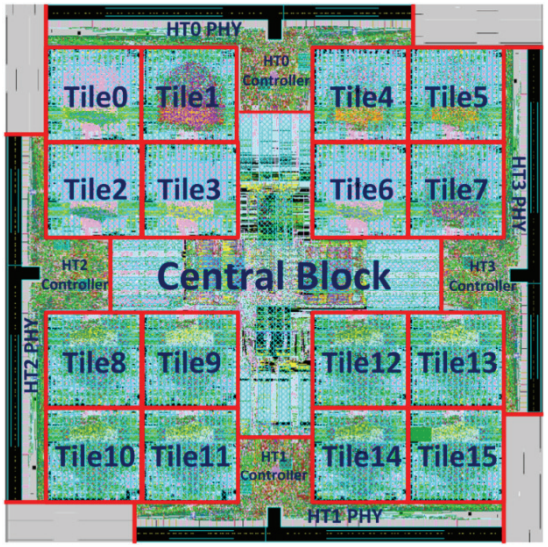


	性能	功耗	效能比	面积
CPU(Xeon E5-4620,2012年)	117x	0.09x	1300x	~0.1x
GPU(K20M,2012年)	1.1x	0.002x	550x	~0.01x

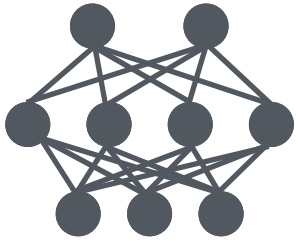
寒武纪处理器系列

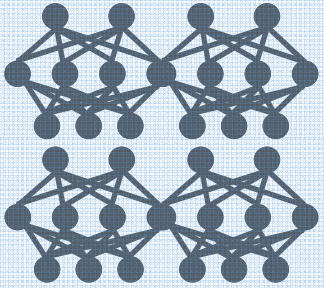




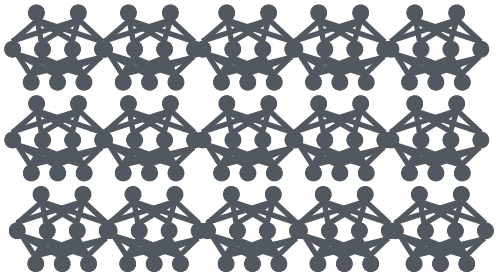


嵌入式设备 | 数据中心 | 超级计算机



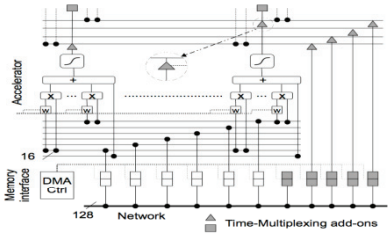


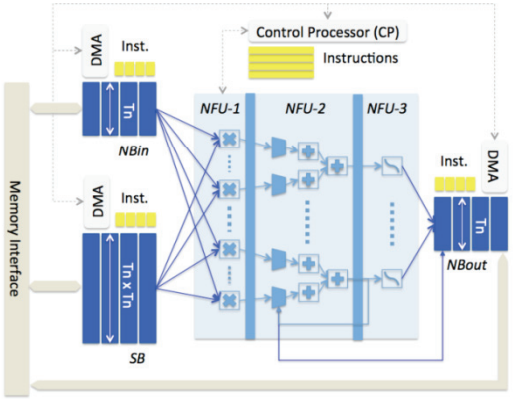
寒武纪1号

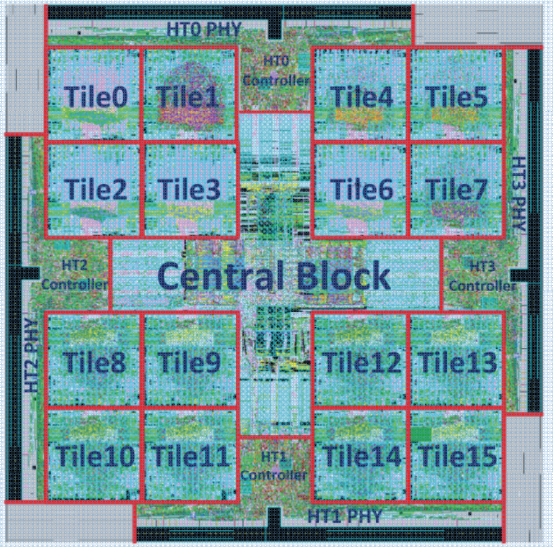


寒武纪2号

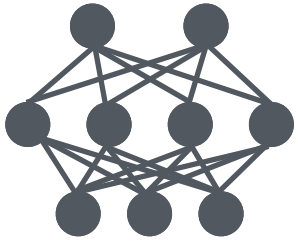
寒武纪处理器系列

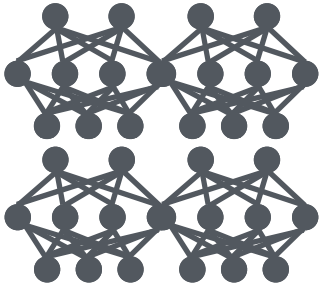




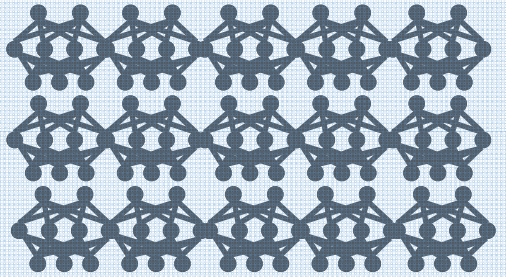


嵌入式设备 → 数据中心 → 超级计算机





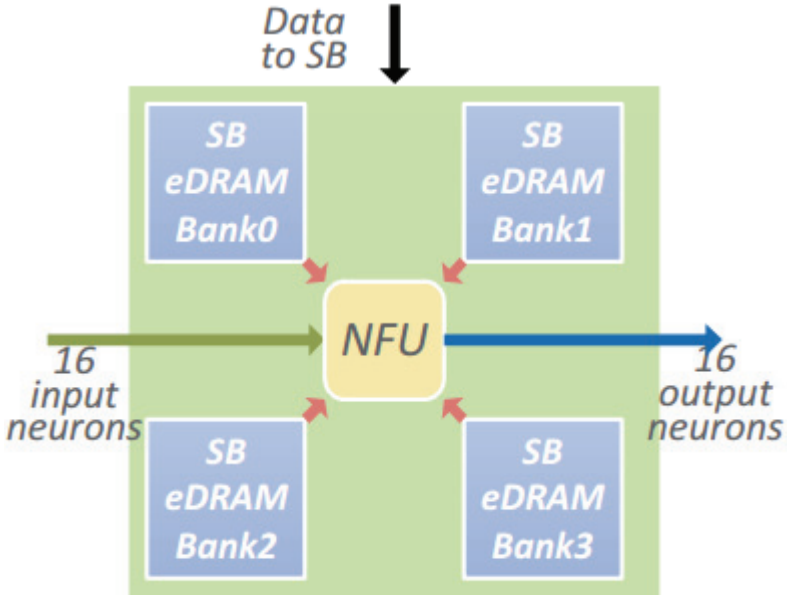
寒武纪1号



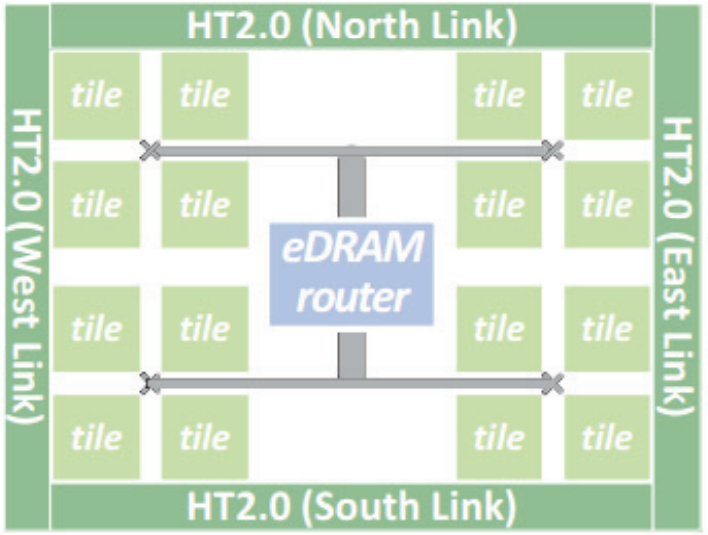
寒武纪2号

寒武纪2号神经网络处理器

- 运算所需数据全部存储在片上
  - 每个芯片存储一部分数据
    - 单芯片片上存储（eDRAM）达36MB
  - 芯片间高速互连共享数据
  - 完全摆脱了内存带宽的瓶颈
  - 单芯片运算速度达5.58 TeraOps/s
  - 可编程，灵活支持多种算法

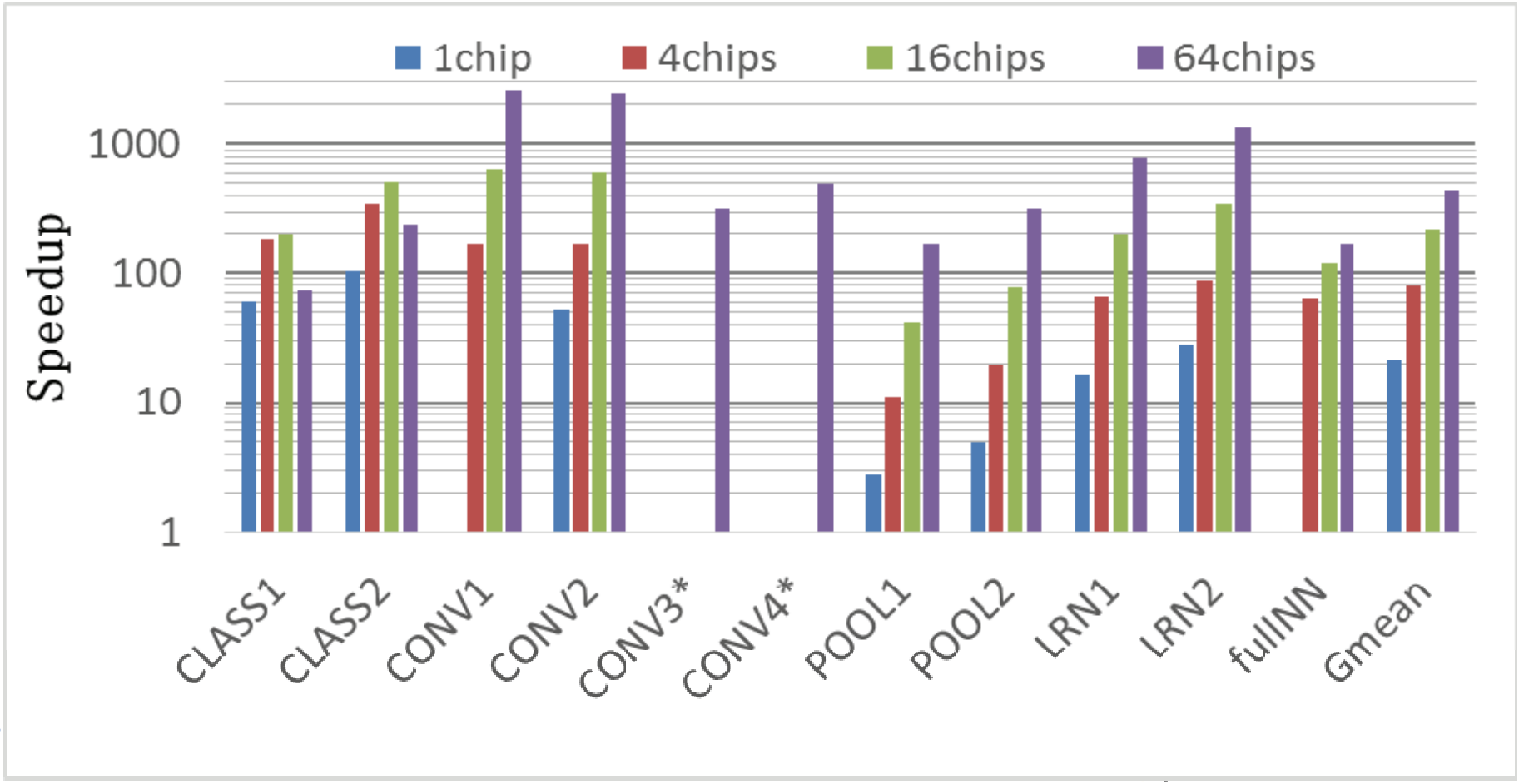


Parameters	Settings	Parameters	Settings
Frequency	606MHz	tile eDRAM latency	~3 cycles
# of tiles	16	central eDRAM size	4MB
# of 16-bit multipliers/tile	256+32	central eDRAM latency	~10 cycles
# of 16-bit adders/tile	256+32	Link bandwidth	6.4x4GB/s
tile eDRAM size/tile	2MB	Link latency	80ns



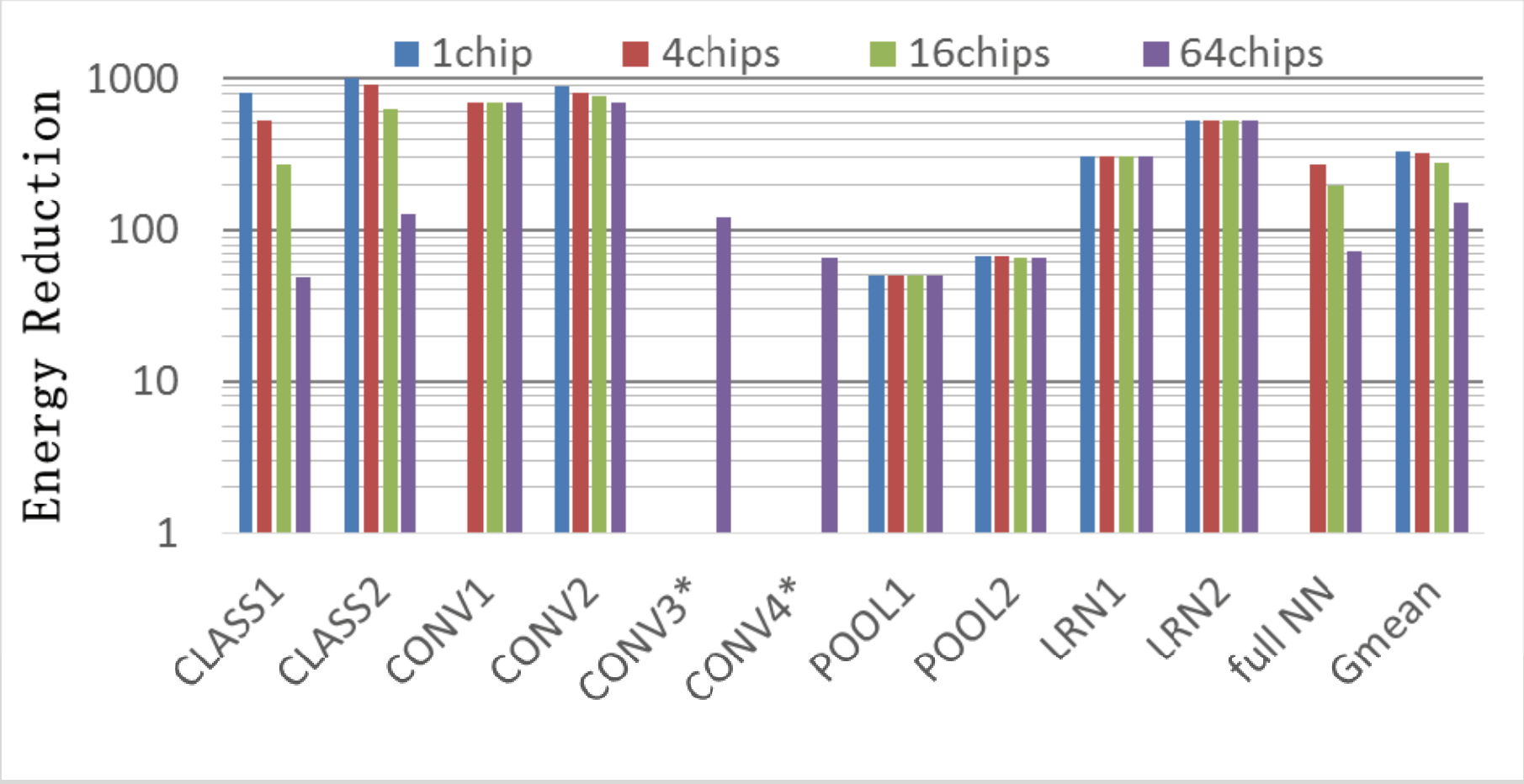
寒武纪2号性能

- 单芯片相对当前主流GPU(Nvidia K20)性能提升21倍
- 64结点系统相对GPU性能提升450倍



寒武纪2号能耗

- 单芯片相对当前主流GPU(Nvidia K20)能耗降低330倍
- 64结点系统相对GPU能耗降低150倍



未来展望

- 寒武纪（DianNao）系列神经网络/机器学习处理器
  - 面向从手机到超级计算机等不同计算平台
  - 体系结构→原型系统→量产芯片
- 机器学习超级计算机
  - 以寒武纪系列处理器为基础
  - 满足大规模机器学习商用的需求
- 面向认知、推理的大规模智能系统
  - 算法和应用原型设计
  - 模拟验证系统
  - 硬件实现

谢谢大家！