# DianNao系列AI加速器
## ——2020年高性能评测与优化课程小组讨论

中国人民解放军国防科技大学
NATIONAL UNIVERSITY OF DEFENSE TECHNOLOGY

报告人：徐云鹏

组　员：李新龙 曹政 徐云鹏

指导：龚春叶、甘新标、杨博

# 简介

1.2007，Larochelle和Bengio增加网络层数可以获得比浅层模型更强大的建模能力。

2.2010，ISCA，Olivier Temam在“The Rebirth of Neural Networks”演讲中首次提出机器学习加速器的思想。

3.2012，ISCA，Olivier Temam提出了第一个AI加速器的设计

# DianNao系列AI加速器

一、需求分析

二、动机

三、技术方案

四、效果

五、分析

# 一、需求分析

AI　　→　　芯片

陈云霁、陈天石与Olivier Temam合作

项目名：DIANNAO

核心：设计一系列定制的AI芯片

公司：寒武纪

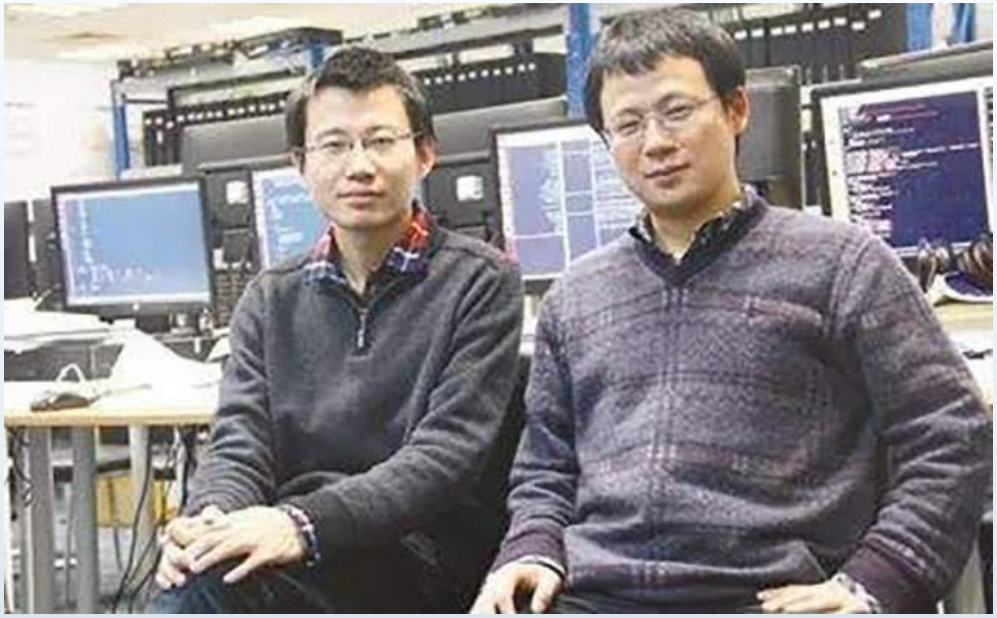
# 一、需求分析

- DIANNAO——第一个设计
- DADIANNAO——DianNao的多片版本
- SHIDIANNAO——与传感器直连
- PUDIANNAO——支持多种常规机器学习算法

# 一、需求分析

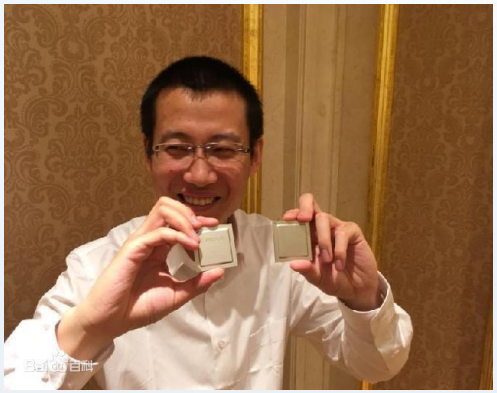

2017年，华为发布世界首款手机AI芯片麒麟970，核心模块NPU，正是来自于中科寒武纪的1A处理器。

# 简介



陈天石（左）与陈云霁（右）

# 简介



陈天石，男，1985年6月生，江西南昌人，中国科学院计算技术研究所 研究员，中科寒武纪科技CEO。

陈云霁，男，1983年生，江西南昌人，中国科学院计算技术研究所研究员，博士生导师。他带领智能处理器研究中心，研制了国际上首个深度学习专用处理器芯片。

# DianNao系列AI加速器

一、需求分析

二、动机

三、技术方案

四、效果

五、分析

# 二、动机

1.DianNao: A Small-Footprint High-Throughput Accelerator for Ubiquitous Machine-Learning是DianNao项目的第一篇，是一篇开创性论文，发表在了ASPLOS14上，并且获得了当年的最佳论文。

核心思想： 结合神经网络模型的数据局部性特点以及计算特性，进行存储体系以及专用硬件设计，从而获取更好的性能加速比以及计算功耗比。

# 二、动机

在这之前方案：

全硬件实现（full-hardware implementation）

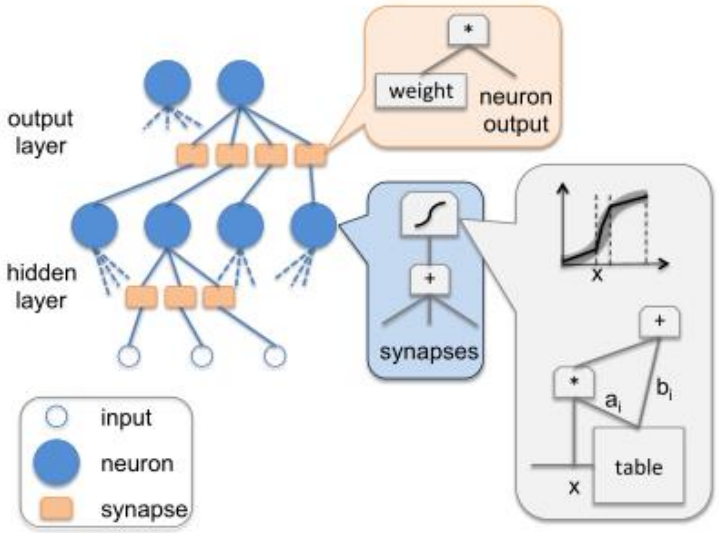



**Figure 9.** *Full hardware implementation of neural networks.*

将每个神经元都映射到具体的硬件计算单元上，模型权重参数则作为latch或是RAM块实现

# 二、动机

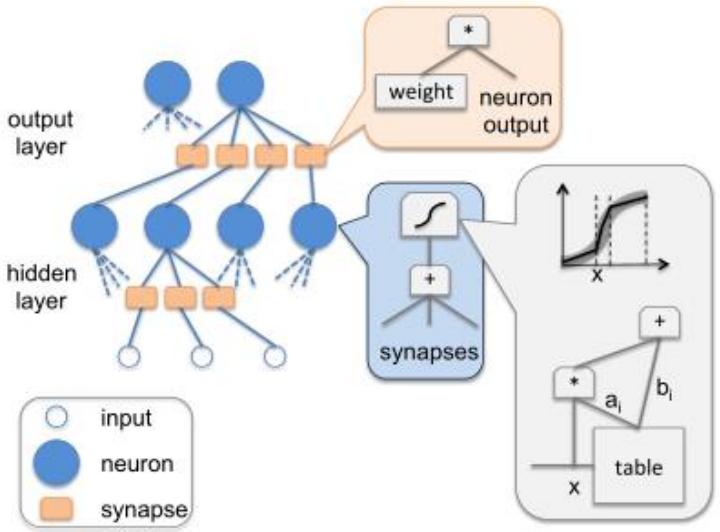



**Figure 9.** *Full hardware implementation of neural networks.*

优点：简洁，计算性能高，功耗低

缺点：扩展性太差

# 二、动机

针对不同输入神经元（input neurons）网络权重数的网络层（network layer）给出了全硬件实现方案在硬件关键路径延时/芯片面积/功耗上的变化趋势

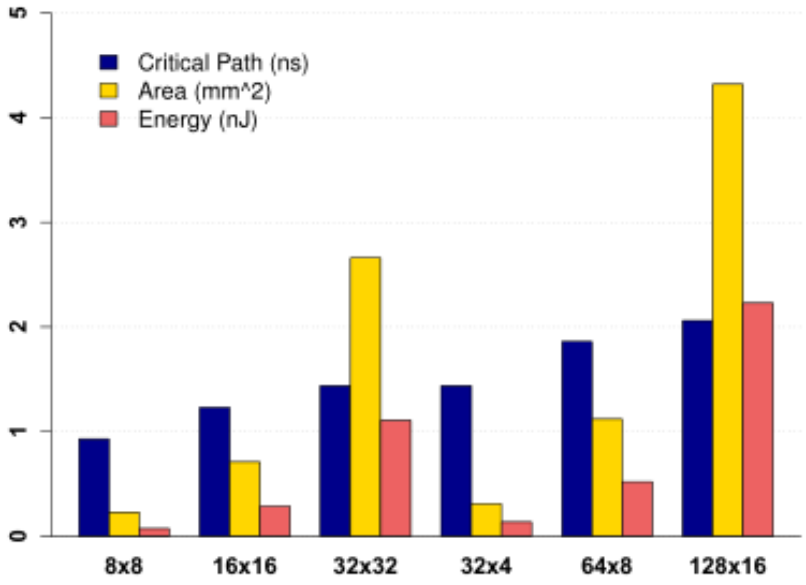



**Figure 10.** *Energy, critical path and area of full-hardware layers.*

# 二、动机

2.DaDianNao:
A Machine-Learning Supercomputer是DianNao项目的第二篇代表性论文，发表在Micro 2014，并且获得了当届的最佳论文。这篇论文针对主流神经网络模型尺寸较大的应用场景，提出了一种具备伸缩性，并通过这种伸缩性可以承载较大尺寸模型的加速器设计架构。

# DianNao系列AI加速器

一、需求分析

二、动机

三、技术方案

四、效果

五、分析

# 三、技术方案

## 1.DIANNAO:

针对全硬件实现方案的不足，文章提出了基于时分复用原则的加速器设计结构

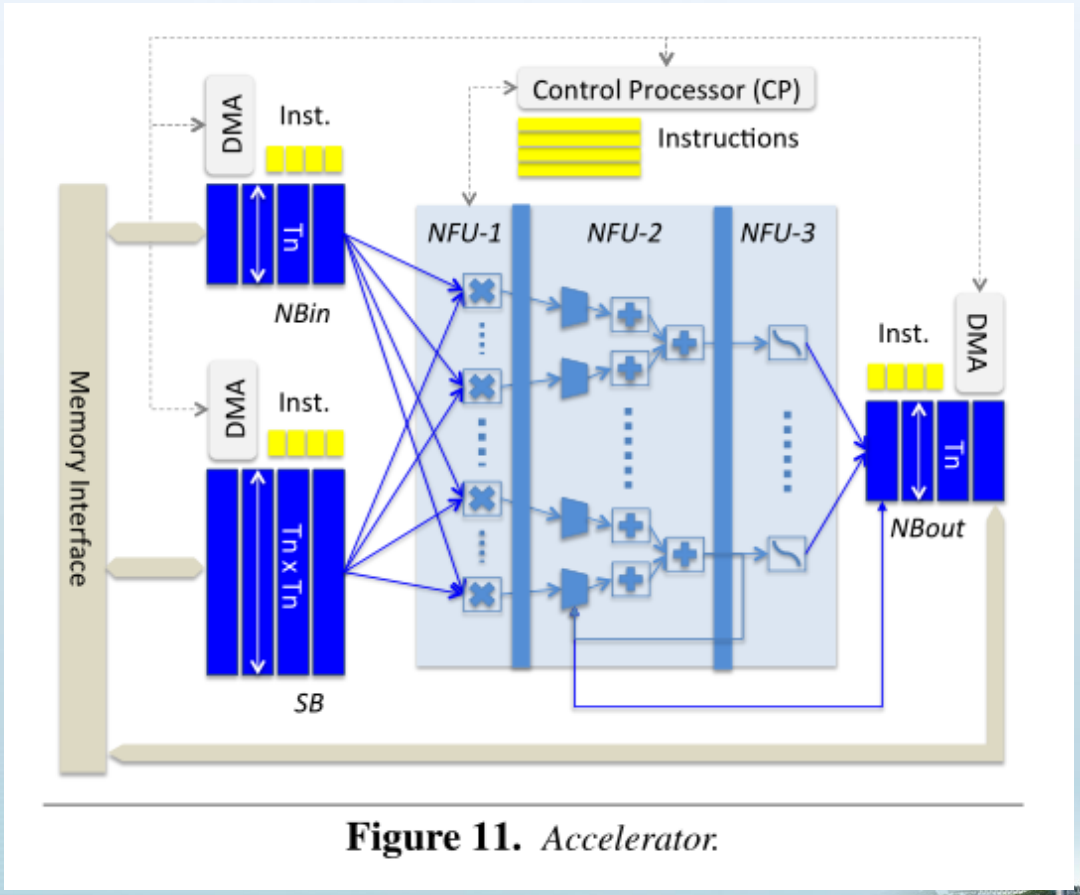



**Figure 11.** *Accelerator.*

# 三、技术方案

## 1.DIANNAO:

加速器芯片里包含三块片上存储：

- 存储输入神经元的NBin、
- 存储输出神经元（output neuron）的NBout
- 存储神经网络模型权重参数的SB。

# 三、技术方案

## 1.DIANNAO:

另一核心部件——NFU(Neural Functional Unit)，由三级流水线组成，完成神经网络的核心计算逻辑。

NFU提基础计算building block——乘法、加法操作以及非线性函数变换

# 三、技术方案

## 1.DIANNAO:

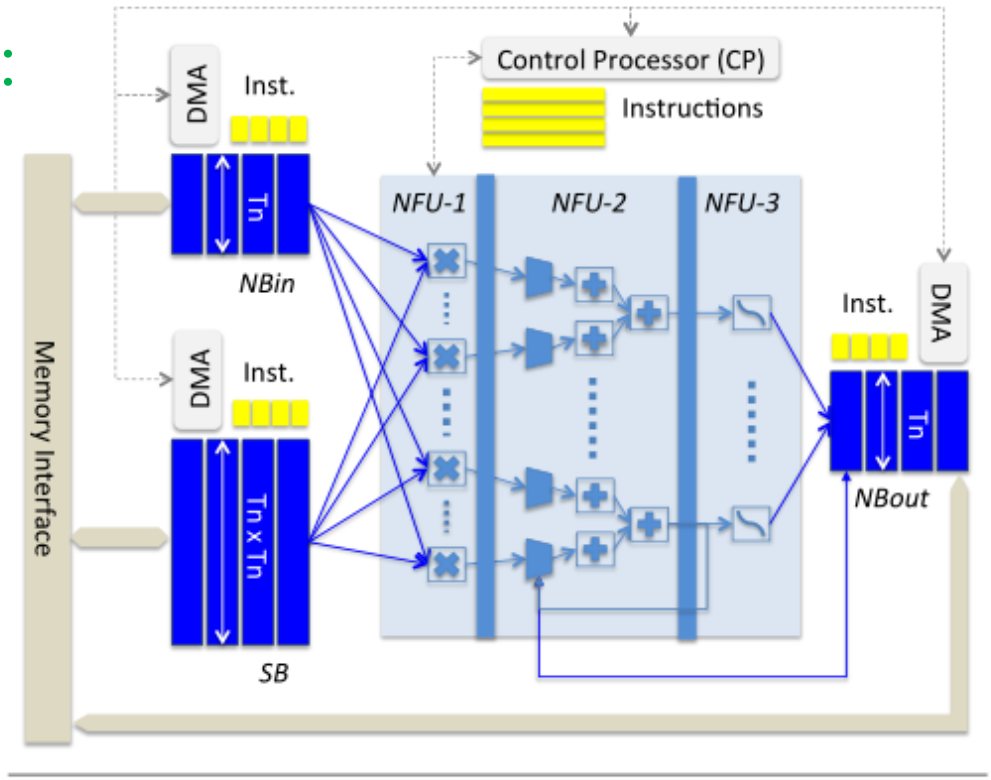



**Figure 11.** *Accelerator.*

时分复用的思想：

- 模型参数 → SB
- 每层神经层的输入数据 → NBin
- layer计算结果 → NBout

# 三、技术方案

## 1.DIANNAO:

第一个设计方案中一些重要细节。

# 三、技术方案

1.以小的模型精度损失（16位定点代替32位浮点），在芯片面积和功耗上都取得了明显的收益

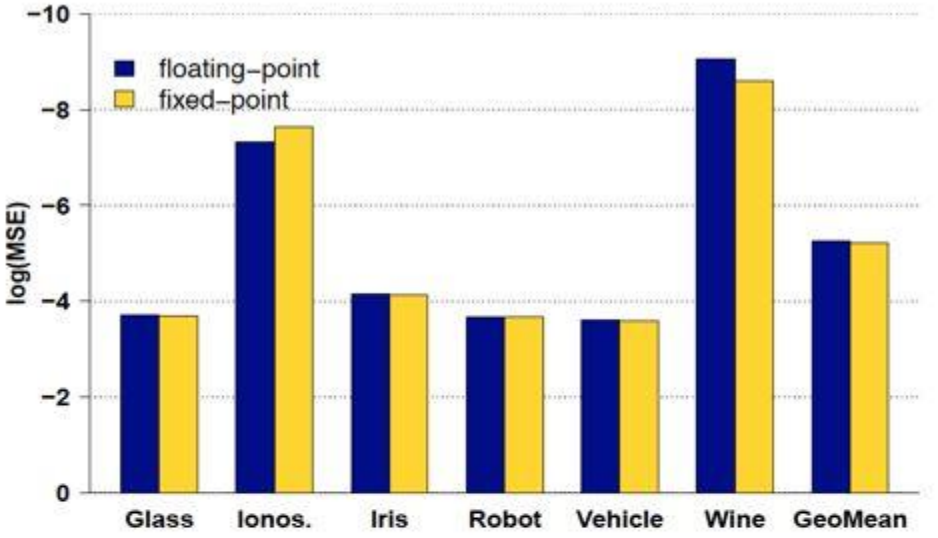



**Figure 12.** *32-bit floating-point vs. 16-bit fixed-point accuracy for UCI data sets (metric: log(*Mean Squared Error*)).*

| Type | Error Rate |
|---|---|
| 32-bit floating-point | 0.0311 |
| 16-bit fixed-point | 0.0337 |

**Table 1.** *32-bit floating-point vs. 16-bit fixed-point accuracy for MNIST (metric: error rate).*

| Type | Area ($\mu m^2$) | Power ($\mu W$) |
|---|---|---|
| 16-bit truncated fixed-point multiplier | 1309.32 | 576.90 |
| 32-bit floating-point multiplier | 7997.76 | 4229.60 |

**Table 2.** *Characteristics of multipliers.*

# 三、技术方案

2.片上SRAM存储划分为NBin/NBout/SB这三个分离的模块

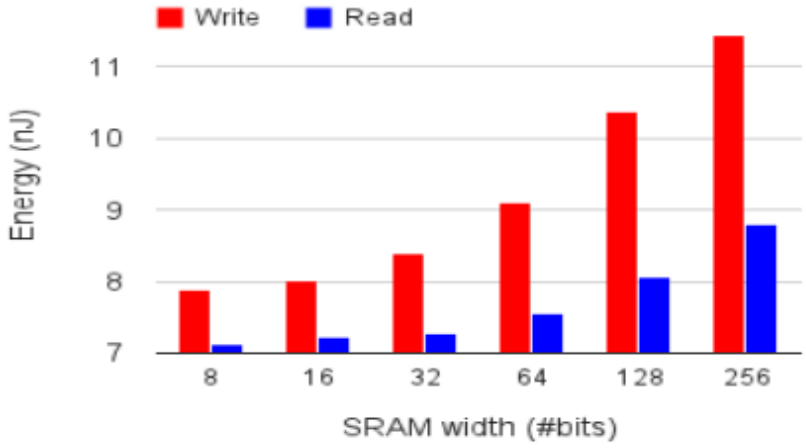



**Figure 13.** *Read energy vs. SRAM width.*

# 三、技术方案

3.对输入神经元数据以及SB数据局部性的挖掘。

- 输入数据的加载与计算过程给重叠起来
- 神经元进行计算
- DMA启动下一组输入神经元/SB参数的加载
- SRAM存储需要支持双端口访问

4.对输出神经元数据以及SB数据局部性的挖掘。

- 引入专用寄存器
- 一定程度上减少存储的性能开销

# 三、技术方案

## 2.DADIANNAO:

### 设计思想:

1.用eDRAM代替SRAM/DRAM，在存储密度/访存延迟/功耗之间获得了大模型所需的更适宜的trade-off。

2.在体系结构设计中以模型参数为中心。

3.神经网络模型具备良好的模型可分特性。

# 三、技术方案

## 2.DADIANNAO: 设计方案

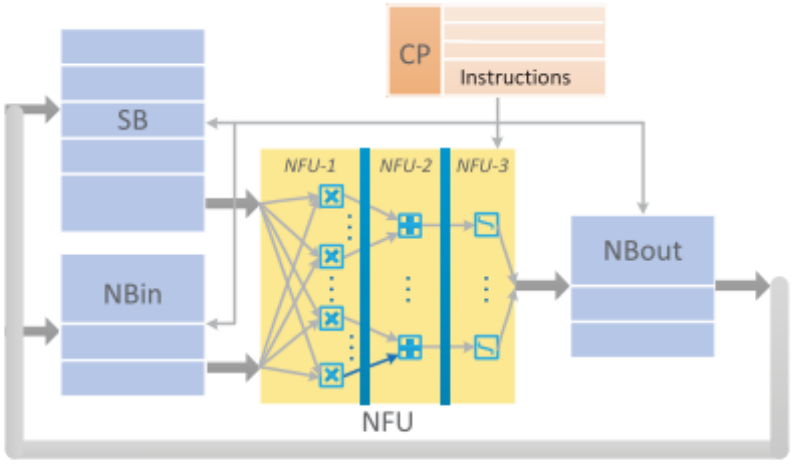



Figure 3: *Block diagram of the DianNao accelerator [5].*

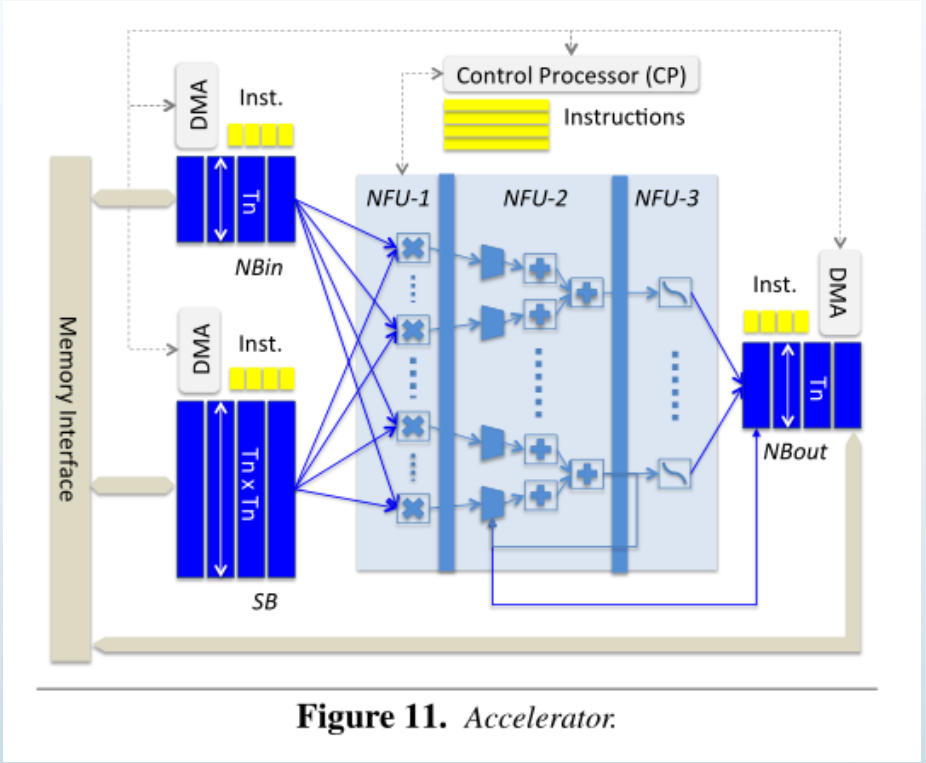



**Figure 11.** *Accelerator.*

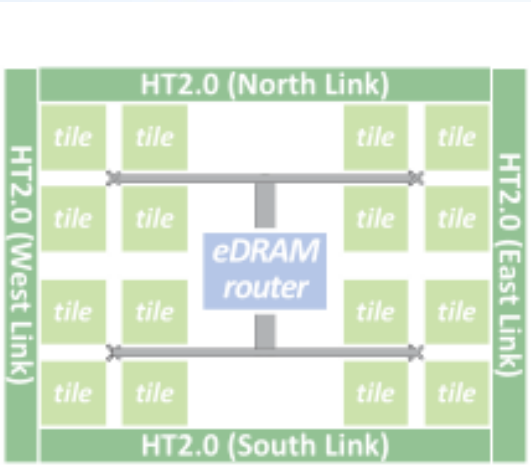
DaDianNao的逻辑结构与DianNao非常相似

主要区别：

- NBin，NBout，SB的组织方式，
- 与NFU的交互方式

# 三、技术方案

单个DaDianNao芯片



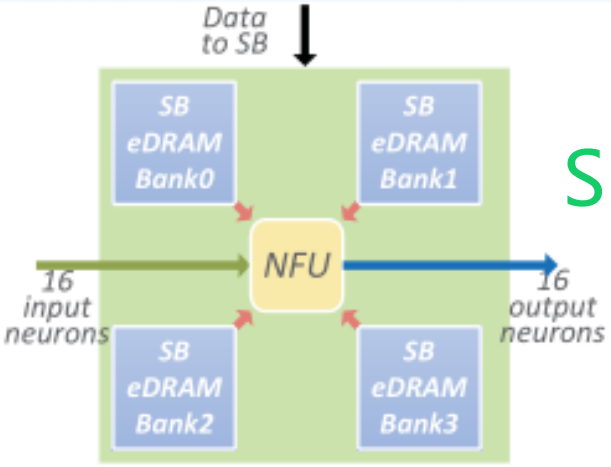




Figure 5: *Tile-based organization of a node (left) and tile architecture (right). A node contains 16 tiles, two central eDRAM banks and fat tree interconnect; a tile has an NFU, four eDRAM banks and input/output interfaces to/from the central eDRAM banks.*

SB是分布式的设计

# 三、技术方案

NFU的内部结构：

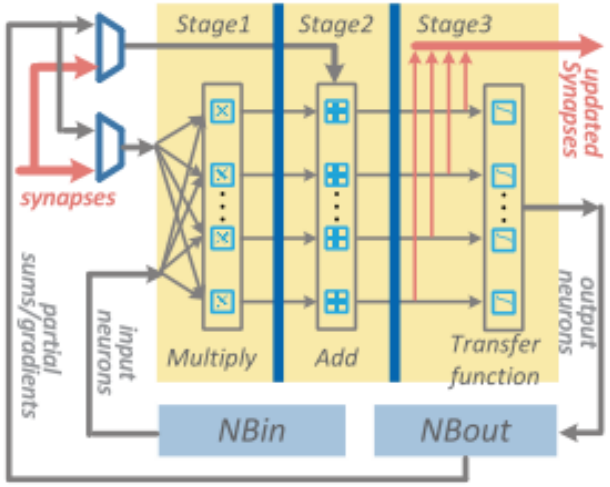



Figure 6: *The different (parallel) operators of an NFU: multipliers, adders, max, transfer function.*

# 三、技术方案

NFU的流水线工作模式：

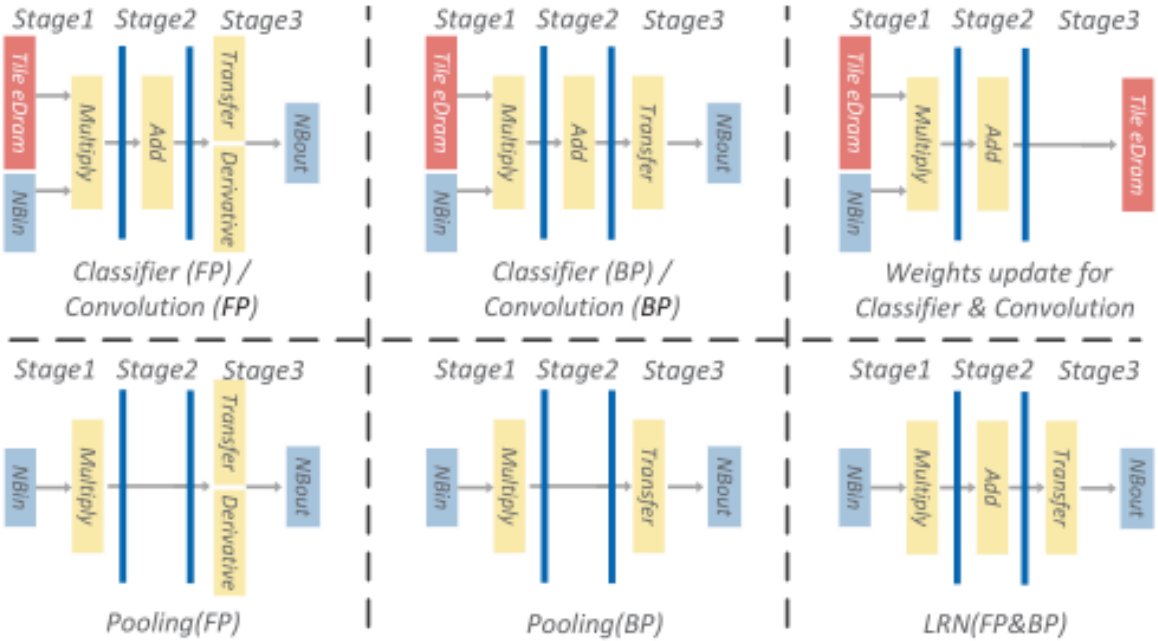



Figure 7: *Different pipeline configurations for CONV, LRN, POOL and CLASS layers.*

单片上的SB存储仍然有限，为了支持大模型，就需要由多个DaDianNao芯片构成的多片系统。

# DianNao系列AI加速器

一、需求分析

二、动机

三、技术方案

四、效果

五、分析

# 四、效果

## 1.DIANNAO:

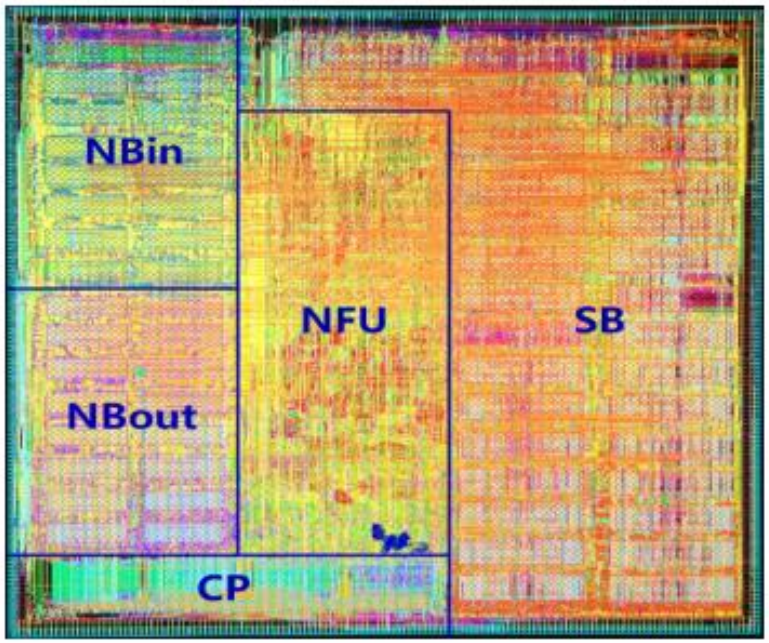



**Figure 15.** *Layout (65nm).*

整体布局

# 四、效果

## 1.DIANNAO:

| Component or Block | Area in $\mu m^2$ | (%) | Power in $mW$ | (%) | Critical path in $ns$ |
|---|---|---|---|---|---|
| ACCELERATOR | 3,023,077 | | 485 | | 1.02 |
| Combinational | 608,842 | (20.14%) | 89 | (18.41%) | |
| Memory | 1,158,000 | (38.31%) | 177 | (36.59%) | |
| Registers | 375,882 | (12.43%) | 86 | (17.84%) | |
| Clock network | 68,721 | (2.27%) | 132 | (27.16%) | |
| Filler cell | 811,632 | (26.85%) | | | |
| SB | 1,153,814 | (38.17%) | 105 | (22.65%) | |
| NBin | 427,992 | (14.16%) | 91 | (19.76%) | |
| NBout | 433,906 | (14.35%) | 92 | (19.97%) | |
| NFU | 846,563 | (28.00%) | 132 | (27.22%) | |
| CP | 141,809 | (5.69%) | 31 | (6.39%) | |
| AXIMUX | 9,767 | (0.32%) | 8 | (2.65%) | |
| Other | 9,226 | (0.31%) | 26 | (5.36%) | |

**Table 6.** *Characteristics of accelerator and breakdown by component type (first 5 lines), and functional block (last 7 lines).*

分别按组件类型和功能块划分的面积和功耗及其占比

# 四、效果

## 1.DIANNAO:

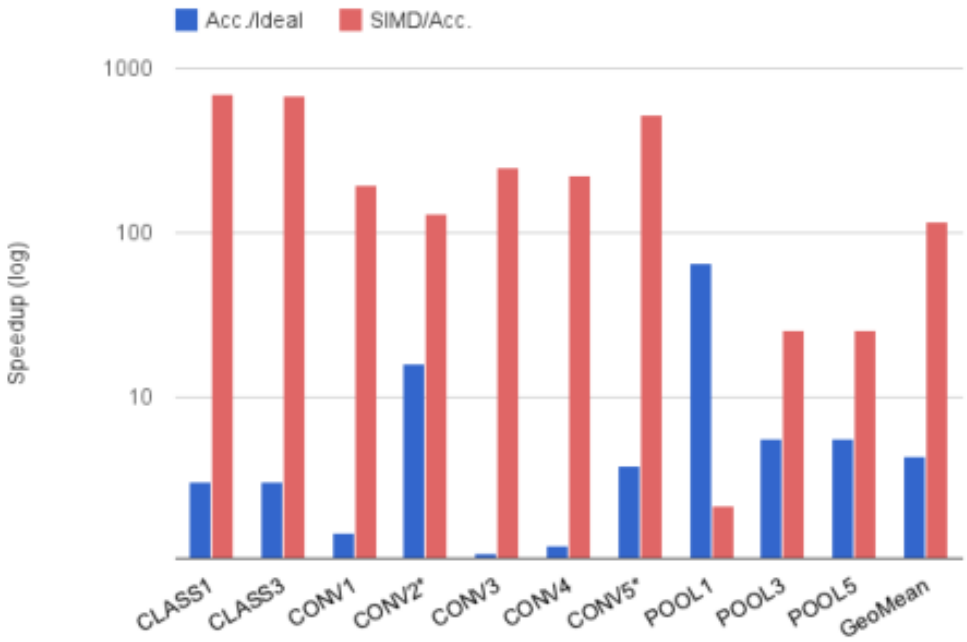



**Figure 16.** *Speedup of accelerator over SIMD, and of ideal accelerator over accelerator.*

相对于SIMA芯片的加速比

# 四、效果

## 1.DIANNAO:

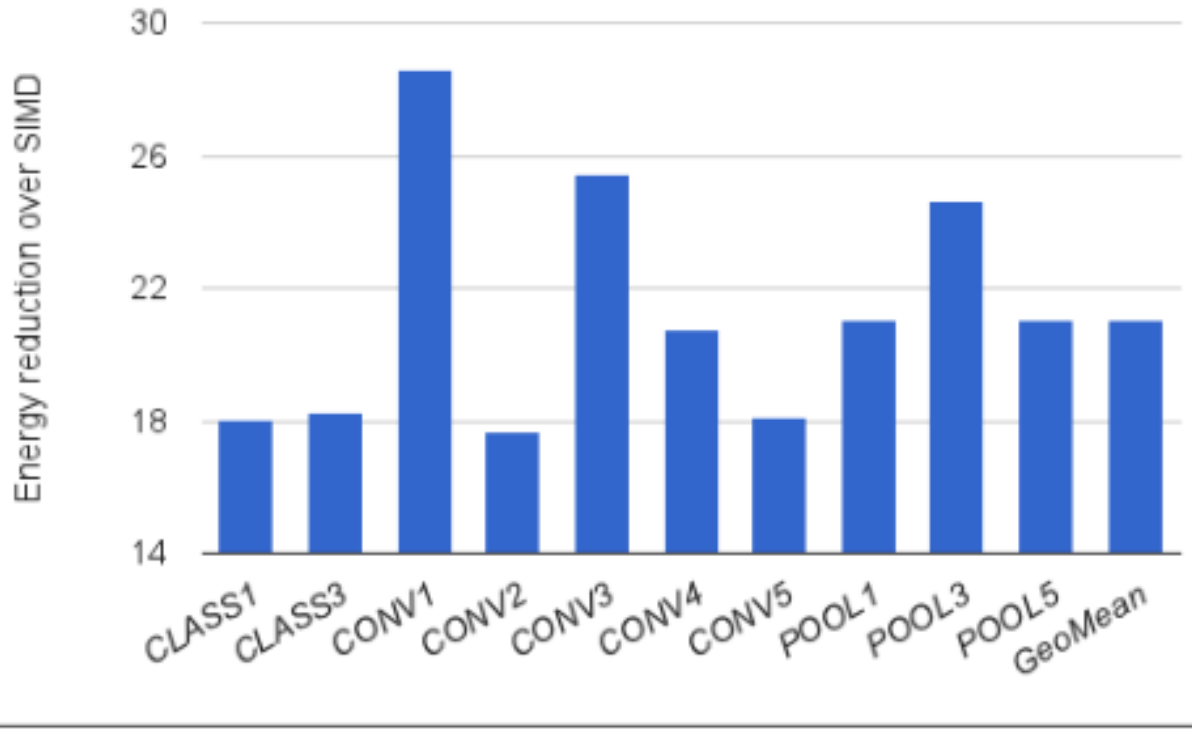



**Figure 17.** *Energy reduction of accelerator over SIMD.*

功耗减少的情况

# 四、效果

## 1.DIANNAO:

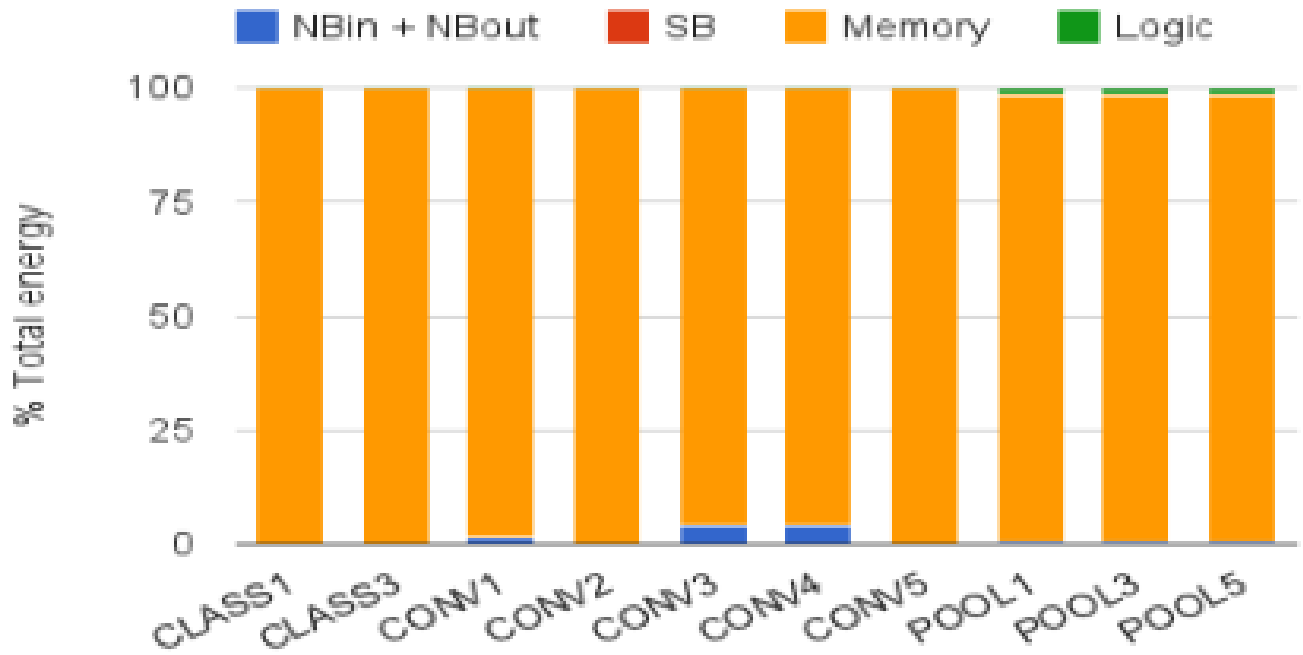



**Figure 18.** *Breakdown of accelerator energy.*

DianNao加速器的能耗分布

# 四、效果

## 1.DIANNAO:

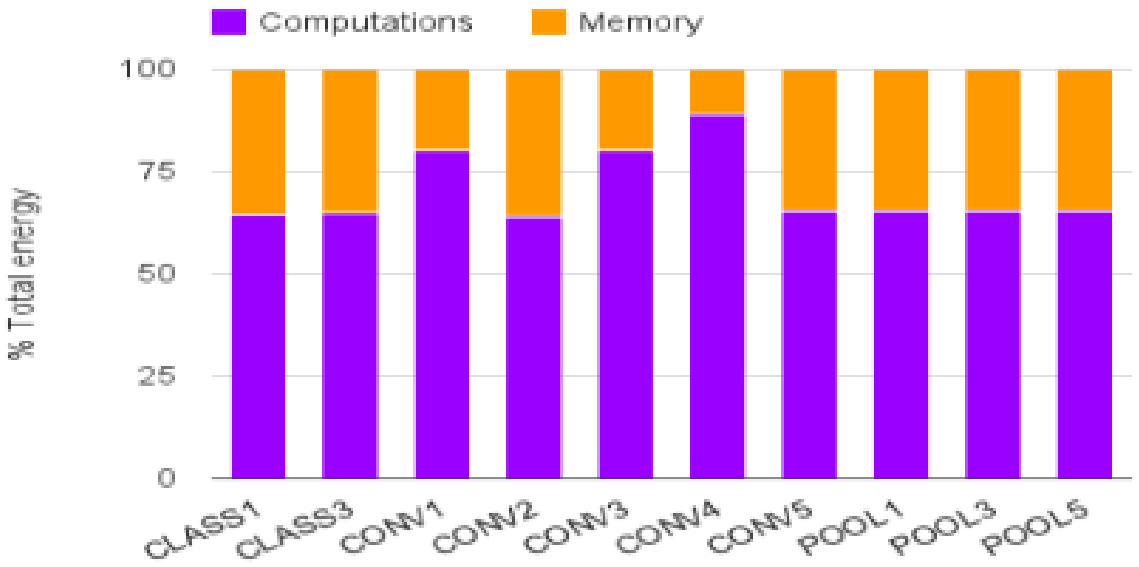



**Figure 19.** *Breakdown of SIMD energy.*

SIMA的能耗分布

# 四、效果

## 2.DADIANNAO:

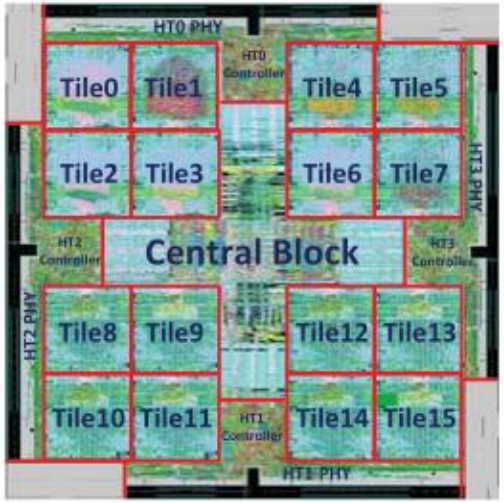



Figure 9: *Snapshot of the node layout.*

| Component/Block | Area ($\mu m^2$) | (%) | Power ($W$) | (%) |
|---|---|---|---|---|
| WHOLE CHIP | 67,732,900 | | 15.97 | |
| Central Block | 7,898,081 | (11.66%) | 1.80 | (11.27%) |
| Tiles | 30,161,968 | (44.53%) | 6.15 | ( 38.53%) |
| HTs | 17,620,440 | (26.02%) | 8.01 | ( 50.14%) |
| Wires | 6,078,608 | (8.97%) | 0.01 | (0.06%) |
| Other | 5,973,803 | (8.82%) | | |
| Combinational | 3,979,345 | (5.88%) | 6.06 | (37.97%) |
| Memory | 32207390 | (47.55%) | 6.12 | (38.30%) |
| Registers | 3,348,677 | (4.94%) | 3.07 | (19.25%) |
| Clock network | 586323 | (0.87%) | 0.71 | (4.48%) |
| Filler cell | 27,611,165 | (40.76%) | | |

Table VI: *Node layout characteristics.*

布局、多芯片系统的性能和节能效果

# 四、效果

## 2.DADIANNAO:

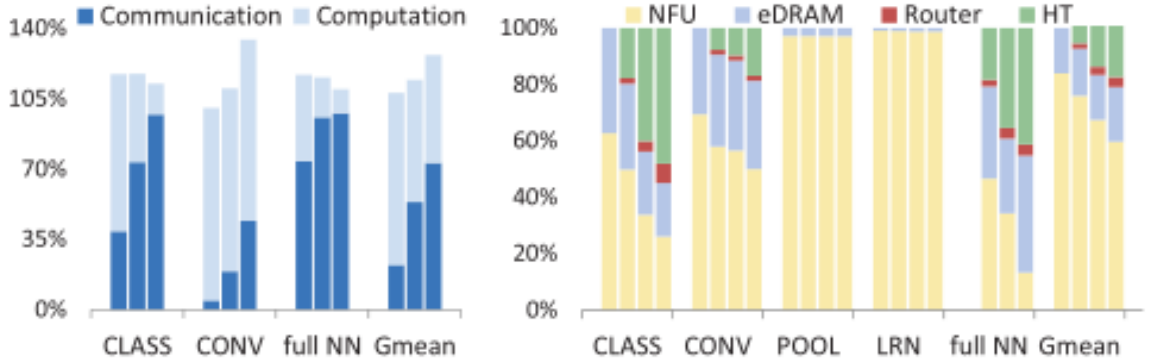



Figure 11: *Time breakdown (left) for 4, 16 and 64 nodes, (right) breakdown for 1, 4, 16, 64 nodes; CLASS, CONV, POOL, LRN stand for the geometric means of all layers of the corresponding type, Gmean for the global geometric mean.*

inter-chip的工作模式下数据通信量

# 四、效果

## 2.DADIANNAO:

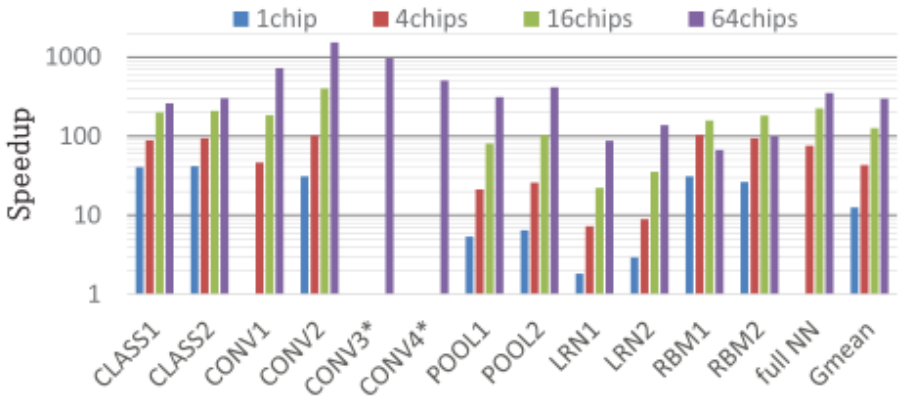



Figure 12: *Speedup w.r.t. the GPU baseline (training).*

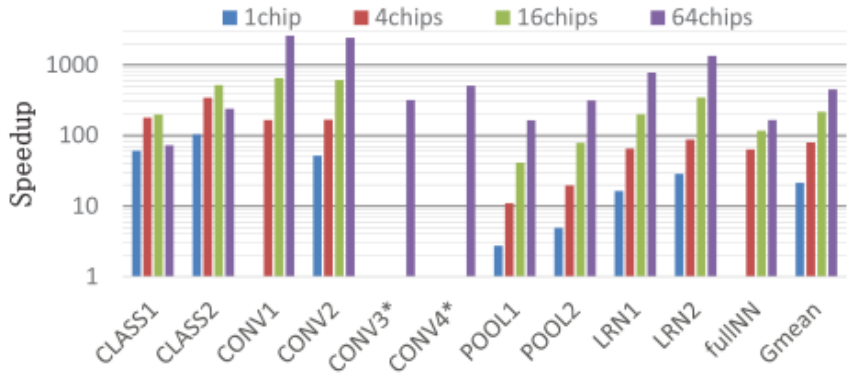



Figure 10: *Speedup w.r.t. the GPU baseline (inference). Note that CONV1 and the full NN need a 4-node system, while CONV3* and CONV4* even need a 36-node system.*

以GPU为基线，training环节和inference环节的加速比。

# 四、效果

## 2.DADIANNAO:

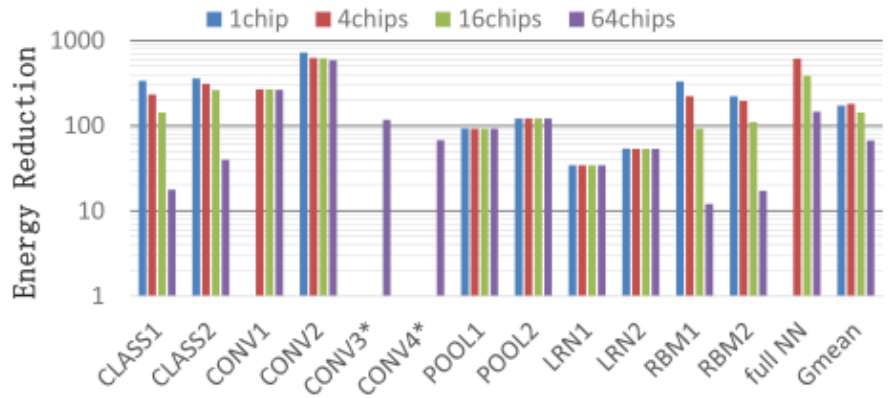



Figure 14: *Energy reduction w.r.t. the GPU baseline (training).*

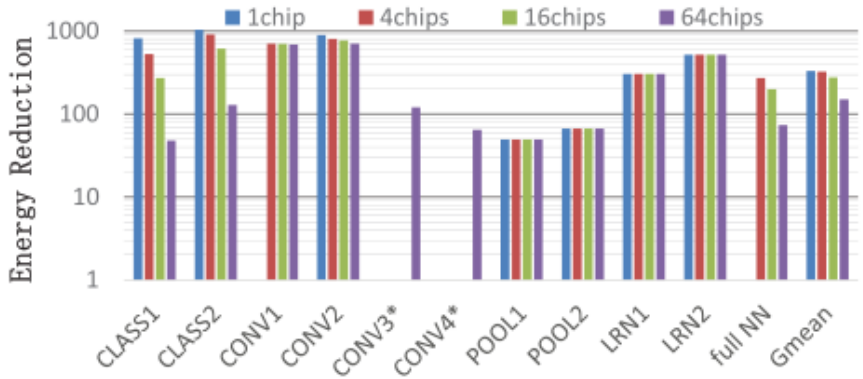



Figure 13: *Energy reduction w.r.t. the GPU baseline (inference).*

以GPU为基线，training环节和inference环节的功耗减少情况。

# DianNao系列AI加速器

一、需求分析

二、动机

三、技术方案

四、效果

五、分析

# 五、分析

## 1.DIANNAO:

DianNao在实验评估上，作为第一个里程碑式的工作，虽然很多细节有待琢磨，基线（baseline）选取上与后续的几篇论文相比，有些保守，在这篇论文里只选取了CPU作为基线，并未将GPU作为基线。

但是，这篇论文是DianNao项目的开山之作，为后续的工作打下了基础。提到下一步要对NFU进行修改和算法的改进，进一步减少主存储器传输延迟，研究可伸缩性，并提高实现工艺，这些在DaDianNao的方案中都有所体现。

# 五、分析

## 2.DADIANNAO:

对体系结构进行了改进，具备了可以承载较大尺寸模型的能力。与针对嵌入式设备应用场景提出的ShiDianNao（第三篇论文）不同，DaDianNao针对的应用场景是服务器端的高性能计算，所以在计算能耗比上虽然相比于基线（GPU/CPU)会有提升，但其设计核心还是专注于高性能地支持大尺寸模型，所以在硬件资源的使用上也远比ShiDianNao要更为大方一些。在GPU/CPU和同期提出的加速器中，DaDianNao算法具有良好的加速性能和节省空间的功能，但它们在很大程度上仍受带宽限制。下一步，主要沿着多个方向改进体系结构：提高NFU的时钟频率，多维环面互连以改善大型CLASS layer的可伸缩性，以每个节点简单的VLIW核和相关工具链的形式研究更灵活的控制。

# 谢谢，敬请批评指正！