

交叉信息核心技术研究院

**前沿构架与智能芯片研究中心**

启明920芯片测试报告

|  |  |
| --- | --- |
| **负责人:** | 马恺声 |
| **组 员:** | 谭展宏 伍毅夫 |
| 张延年 张武科 |
| 张军超 王成勋 |
| 王涛 钟坛 |
| 陈声发 陈红洋 |
| Jan-Henrik Lambrechts |
| **时 间:** | 2020年09月04日 |

**目录**

[一、 设计简介 1](#_Toc28010924)

[1、 研究背景 1](#_Toc28010925)

[2、 启明920芯片设计介绍 3](#_Toc28010926)

[二、 芯片物理测试 7](#_Toc28010927)

[1、 芯片基本信息 7](#_Toc28010928)

[三、 芯片测试板与测试系统介绍 7](#_Toc28010929)

[四、 芯片基本模块测试 8](#_Toc28010930)

[1、 芯片寄存器测试 8](#_Toc28010931)

[2、 Memory测试 8](#_Toc28010932)

[3、 芯片PLL测试 8](#_Toc28010933)

[4、 复位测试 9](#_Toc28010934)

[5、 功能测试 9](#_Toc28010935)

[6、 异常测试 9](#_Toc28010936)

[五、 基于测试数据的性能评估 10](#_Toc28010937)

[1、 测试数据描述 10](#_Toc28010938)

[2、 性能评估 10](#_Toc28010939)

[3、 整体性能效果 12](#_Toc28010940)

[4、 设计比较 13](#_Toc28010941)

[六、 总结与展望 13](#_Toc28010942)

[1、 启明920工作总结 13](#_Toc28010943)

[2、 未来展望与计划 14](#_Toc28010944)

[附录—启明920调试记录 16](#_Toc28010945)

# 设计简介

## 研究背景

深度神经网络（Deep Neural Networks, DNNs）是目前许多现代人工智能应用的基础。目前无论学术界还是产业界，都致力于研究满足各种应用需求的神经网络模型。如图1所示，伴随着DNN模型快速发展的是越来越庞大的参数量以及计算量。为了能够将庞大的模型部署到我们实际的产品应用中，不得不对原始模型进行优化，其中一种重要手段就是模型压缩。

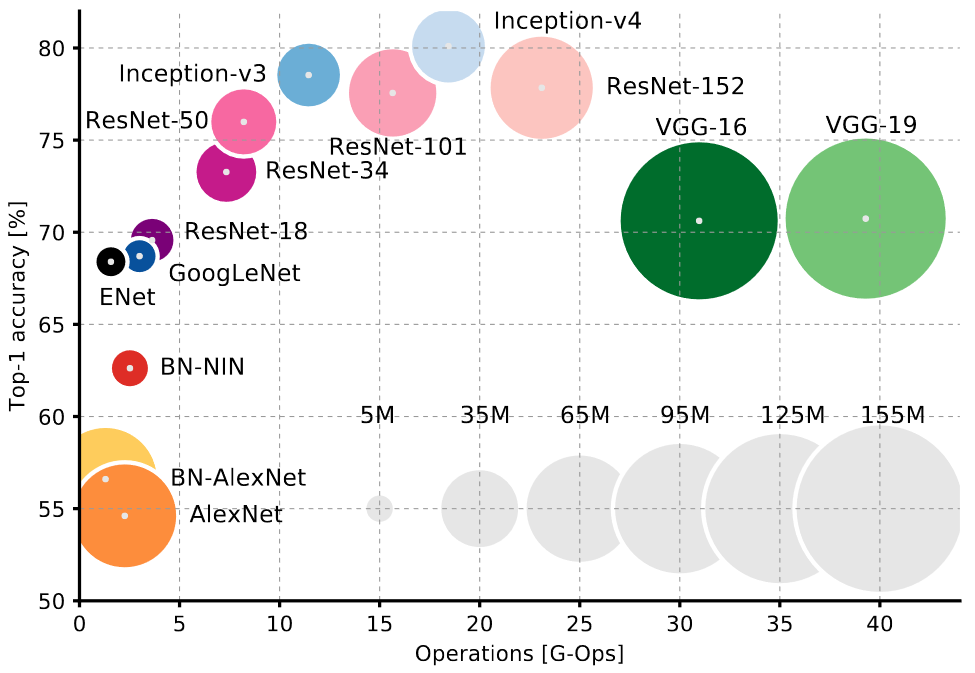


图1 DNN模型精度和计算量及参数规模的发展

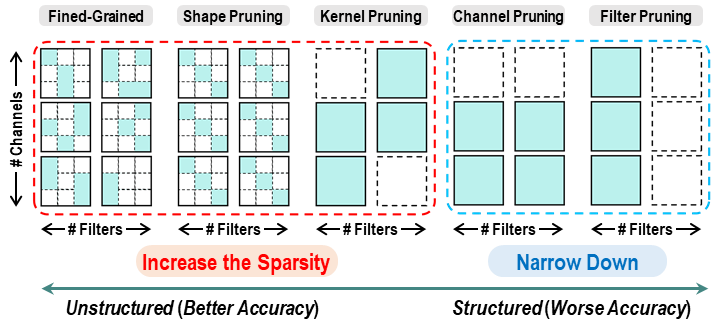


图2 DNN模型的权重剪枝方式

对DNN模型权重的剪枝以及量化是模型压缩的主要方式。权重剪枝的主流手段如图2所示，包括细粒度剪枝（非结构化）、形状剪枝、卷积核剪枝、（输入）通道剪枝以及滤波器剪枝。越在左边的剪枝方式精度越友好，但是硬件对稀疏的利用效益越小；越在右边的剪枝方式精度会越容易下降，但是硬件对稀疏的利用效益越大。

为了寻求更好的硬件友好与模型精度的折中，我们核心院和清华大学合作提出了高效的图案剪枝方式，该思想如下图所示。图案剪枝是对3×3卷积进行优化，在不同的卷积核中采用相同的稀疏比，但允许其位置各不相同，并进而通过算法筛选选取少量种图案。这样，我们可以通过很少的比特来编码不同的图案（如4比特），并且在存储中只需要存储每个卷积核中“非零”参数，这样可以降低模型的存储和带宽需求，同时不会引入太多的索引存储。此外，由于不同卷积核中稀疏比相同，这样有助于平衡不同卷积核之间的计算时间，从而提升计算资源利用率。该剪枝方法我们在ResNet以及VGG等网络，于CIFAR-10和ImageNet数据集下进行测试，在把3×3卷积基本只剩一个“非零”参数的压缩策略下，精度损失小于1%。

因此，在剪枝方面需要有一个统一的硬件架构为不同的剪枝模式提供支持。在现有的硬件中，大多数只能对通道剪枝和滤波器剪枝有所支持，这是因为这是完全结构化的剪枝策略，在硬件上对应的实现是配置卷积层的通道宽度即可，相当于是计算更窄的一个卷积层，所以我们一般把这两种剪枝成为“Narrow Down”即“变窄”的处理，这种剪枝方式的精度敏感性很高，很难把模型进行高倍率的压缩。而对于我们的图案剪枝，包括前人所采用的形状剪枝、卷积核剪枝等，这是会引入相对不规则的数据稀疏性。对于图案剪枝和形状剪枝，引入的是卷积核内部的每个参数层次稀疏；而对于卷积核剪枝，引入的是卷积核自身的稀疏。为了去利用这些稀疏数据，以实现在计算时间上的减少，需要专用的硬件予以支持，从而发挥稀疏所带来缩短计算延时的机会，突破硬件自身固有的算力，实现由稀疏带来的算力增益。

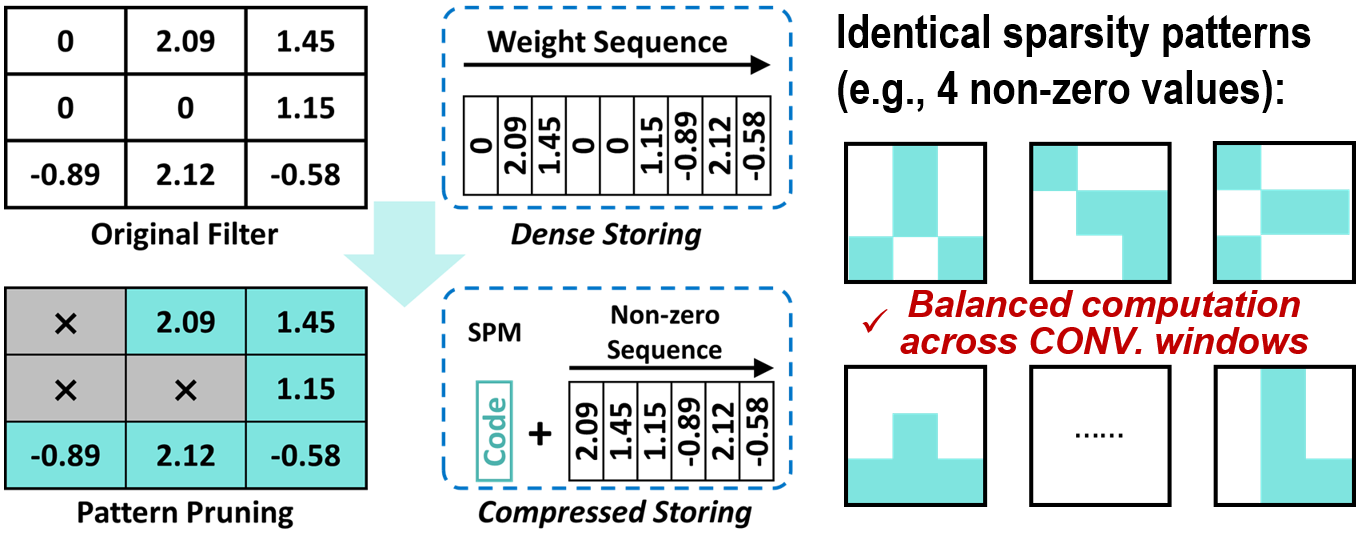


图3 DNN模型图案剪枝示意

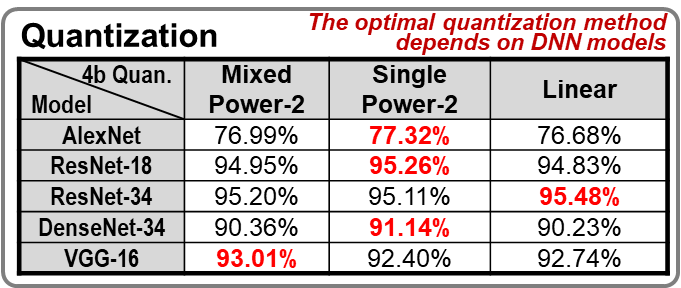


图4 不同模型的量化效果差异

而在量化方面，我们发现不同模型有其合适的权重量化方式。如图4所示，我们在CIFAR10上测试了5个经典模型，量化策略采用了两种非线性量化以及一种线性量化模式。因此为了高效地适配不同场景、不同模型的量化需求，硬件有必要提供相应的动态配置的解决方案。

基于以上的研究背景，“启明920”应运而生，旨在用统一的硬件架构，对不同的剪枝模式以及不同的权重量化策略，予以高效的支持。通过对不同剪枝模式、量化的可配置支持，可适配在边缘不同场景的需求：例如在对压缩需求小、希望更高精度的场景，可以采用细粒度压缩或者不压缩的处理，这时候“启明920”会检测在输入数据（包括激活与权重）中随机分布的零值，尽可能跳过无效计算的周期，提升计算效率。在我们的测试中，仅利用模型参数和激活输入中固有零值，“启明920”也可由77.5%的性能提升；而对于压缩需求大的应用场景，我们最高可以实现764%的性能提升，即总等效算力为硬件固有算力的近9倍。

## 启明920芯片设计介绍

“启明920”芯片为“启明910”的迭代版本，其峰值性能为前代版本的近10倍，并且有更高的灵活性以及更高的面积效率。“启明920”在UMC CMOS 55nm的工艺下完成流片，面积为3.5×3.9mm，片上集成SRAM合计约243KB，并部署有256个MAC单元，整个处理流程均以流水线形式实现，峰值频率可达300MHz，硬件固有的峰值算力为153.6GOPS，而通过利用稀疏计算的效益，等效的峰值算力可达到1327.2GOPS。在电气设计上，计算核心和SRAM的标准电压为1.0V，IO标准电压为3.3V，IO数量合计216个，其中160个为信号管脚，其余56个为电源管脚。

“启明920”对深度学习模型计算具有较高的灵活性，在卷积核配置上支持3×3和1×1两种主流尺寸，步长支持1和2两种，特征图的尺寸可以任意指定。基于这些计算配置，可以完成VGG，ResNet，MobileNet等主流分类模型，以及Yolo，SSD等检测模型的运算。

“启明920”的整体架构如图5所示，整体为软硬件协同的设计，其中虚线部分为算法优化及数据预处理，实线部分为硬件模块。硬件模块可分为激活缓冲与调度模块（ICS），64组处理引擎（PE），64组与PE相适配的权重缓冲区，以及其他处理模块（包括池化模块、后处理与量化模块、输出缓冲区）。在权重缓冲区中，上层算法与软件会将经过卷积核剪枝和（或）图案剪枝后的权重按照一定规则导入存储中，同时ICS模块也会导入由上层软件提供的卷积核剪枝的信息，提供输入通道数据多播的掩码。在PE模块中，压缩后的卷积核会经过解码后与相应激活进行稀疏检测，计算核心会挑选出有效的A-W操作对。从而实现跳过零数据的运算周期，提升整体的计算效率。

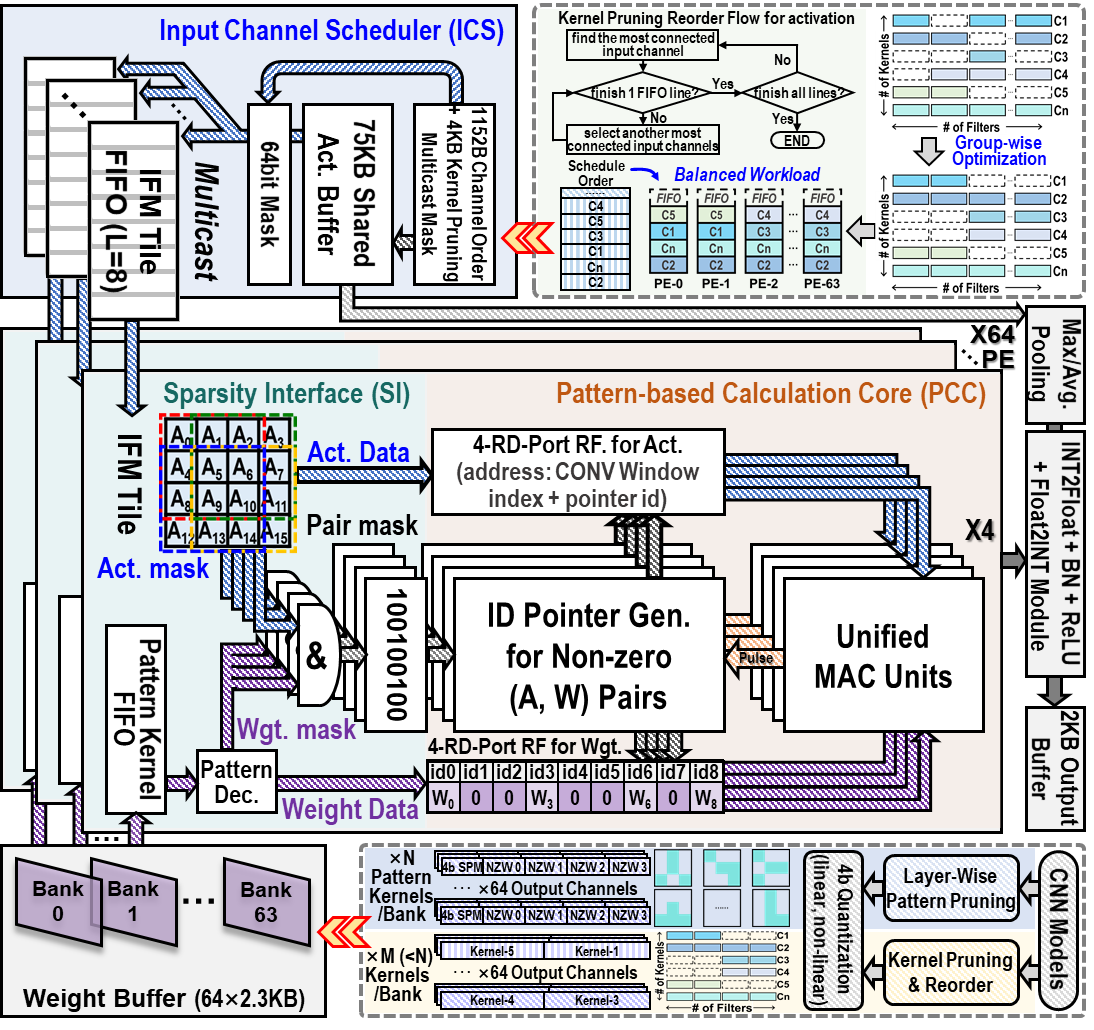


图5 启明920芯片整体架构

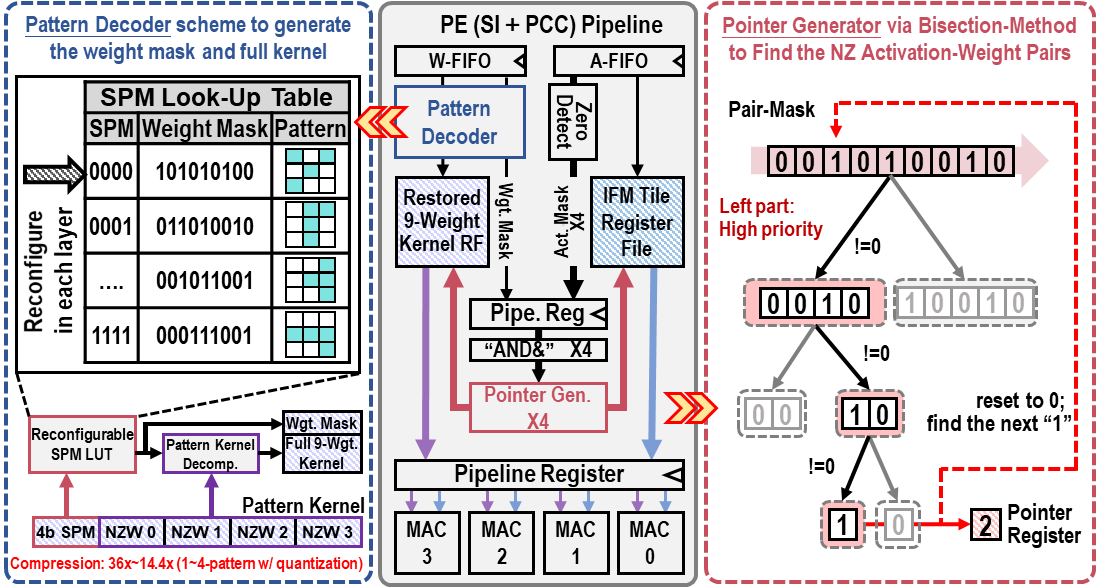


图6 面向图案剪枝的架构设计

图6为所示，中间部分为整体的计算流水线，其中在图案解码器为左边蓝框所示，即对捆绑到权重参数前的索引进行解码，将若干个非零参数还原回完整的3×3卷积核，再进行后续计算。而在指针生成器在右侧所示，采用二分发找到有效A-W数据对的位置。

在量化方面，由于在现有工作中，权重的定点量化、2的幂指数量化和混合的2的幂指数量化都有相关的研究，但缺乏一个统一的硬件架构同时支持这几种量化方式。这是有必要的，因为不同网络或不同的应用场景中，采取不同的量化效果会不一样。此外，在混合2的幂指数上我们做了优化，引入两个调整因子，在每一层都会采用不同的调整因子去使得参数最优。

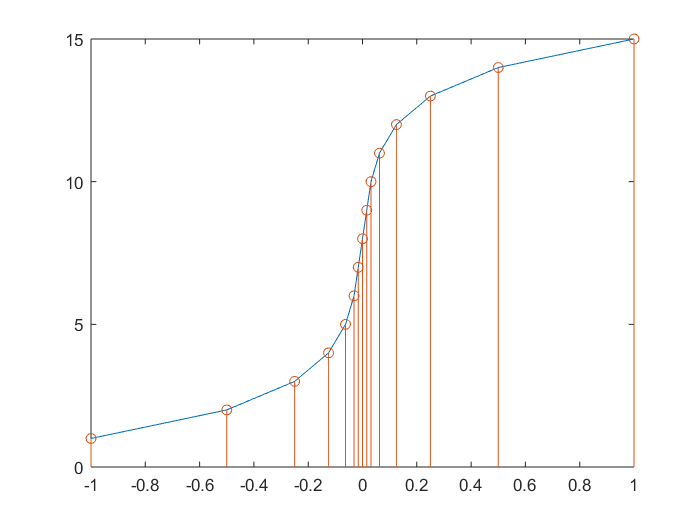
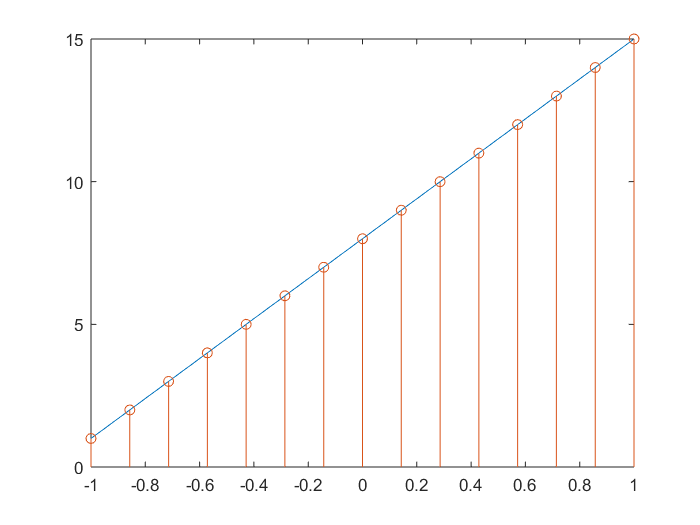
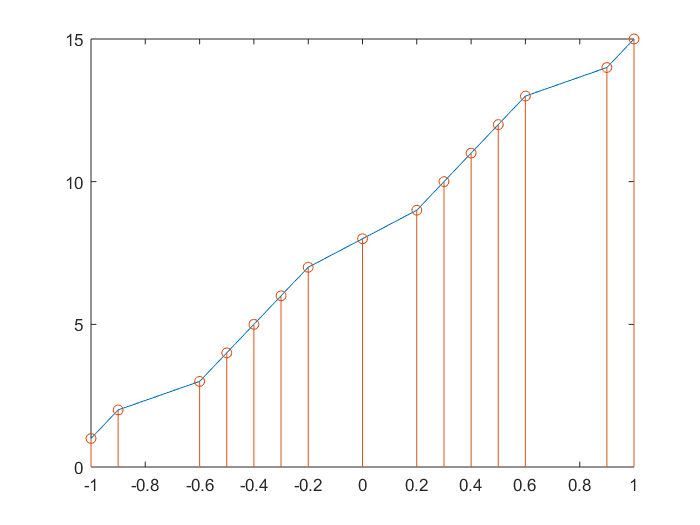
 

图7 线性量化、2的幂指数量化、混合2的幂指数量化的数值分布

图7为三种权重量化方式的数值分布，线性量化的数值分布是均匀的，指数量化呈现典型的小数部分密集、大数部分稀疏的分布，而混合幂指数量化为两者之间。在不同的网络、任务中，有模型自身较好的一种权重参数数值分布形式。

图8为面向三种量化的统一MAC单元架构，以及相应的计算方式。我们将这三种量化方式拆解为移位计算以及多个小型加法器，从而搭建出一个面向三种不同权重量化的等效乘法器，从而避免采用复杂的译码模块以及传统乘法器。相对传统设计，我们这种统一的MAC单元设计节省了约30.7%的功耗以及3.5%的面积开销。

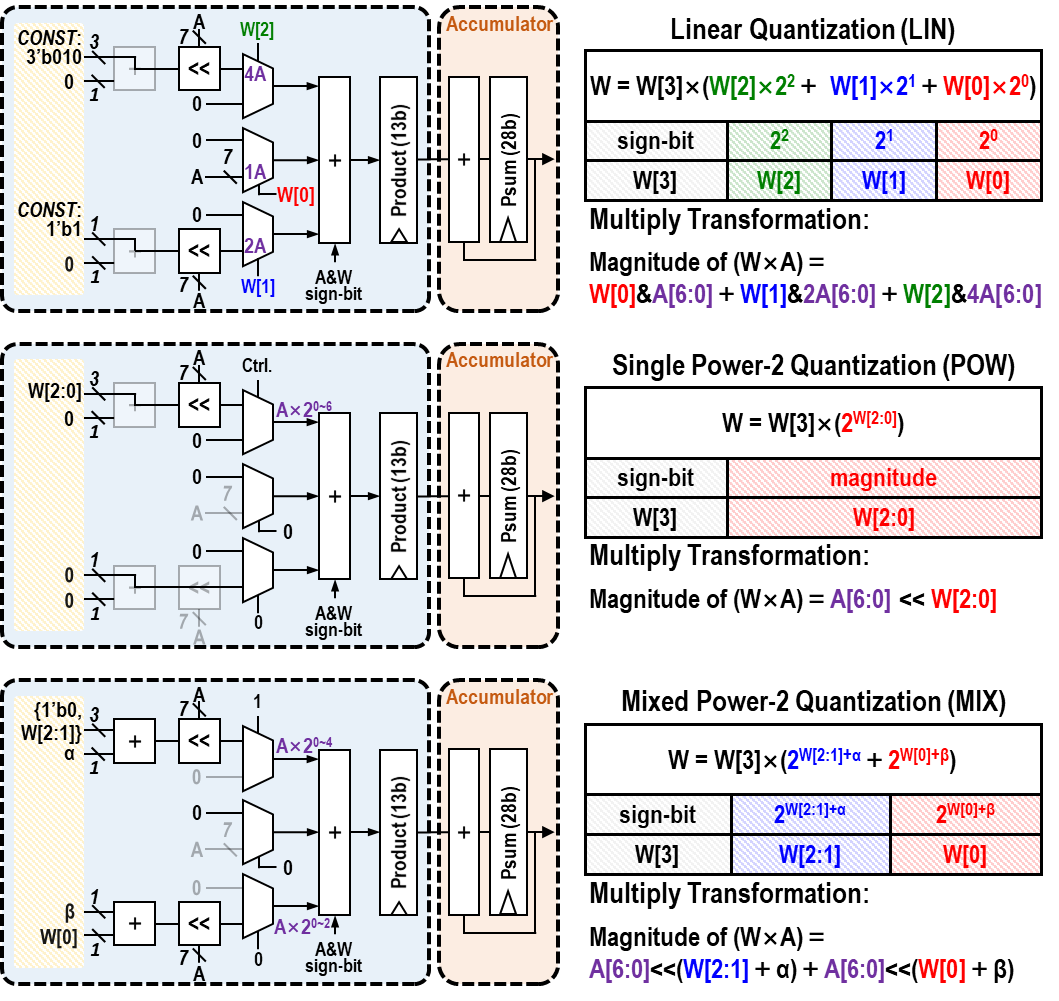


图8 支持三种量化的统一MAC单元架构



图9 混合量化计算处理

另外，在整体运算上，推断运算采用混合精度的方式进行（如图9所示）。我们把原始的运算拆解为卷积的乘加运算以及对结果的放缩与偏移。在计算密集型的卷积阶段，使用定点计算模式，这样可以大大降低计算的开销；在放缩与偏移计算中，我们对批归一化与权重的量化缩放因子做了统一处理，最终得到放缩与偏移的线性运算中的参数k和b，量化精度采用16比特的半精度浮点，这样以确保最终的硬件计算的精度和开销最优。

# 芯片基本信息与测试系统

## 芯片基本信息

“启明920”芯片架构名称为MUSE V2，整个芯片采用UMC 55nm SP工艺进行设计生产，裸片面积3.5mm x 3.9mm，数字逻辑约四百万门。IO管脚216个，其中信号管脚160个，电源管脚56个，分别为数字IO电源3.3V、模拟IO电源3.3V、数字Core电源1.0V和模拟Core电源1.0V。由于IO频率约为100MHz，片外数据位宽为128比特，从而提供12.5Gbps的片外带宽。片上时钟由PLL倍频产生，可稳定运行在300MHz频率。

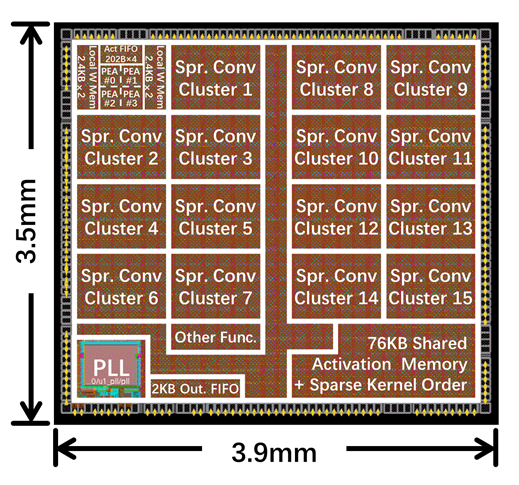
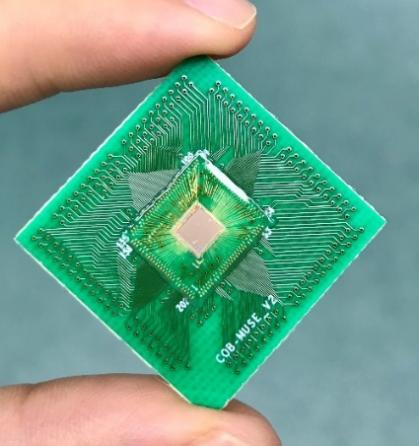
 

图10 芯片布局图 图11 芯片物图

## 芯片测试板与测试系统介绍

芯片测试系统有三部分组成，包括软件系统、FPGA测试板卡以及启明920测试板。

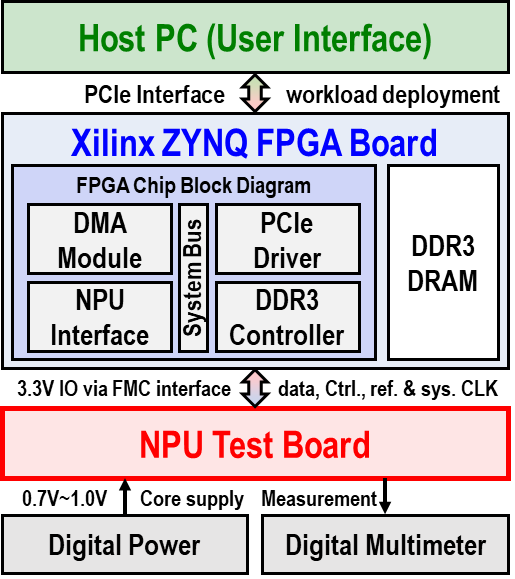
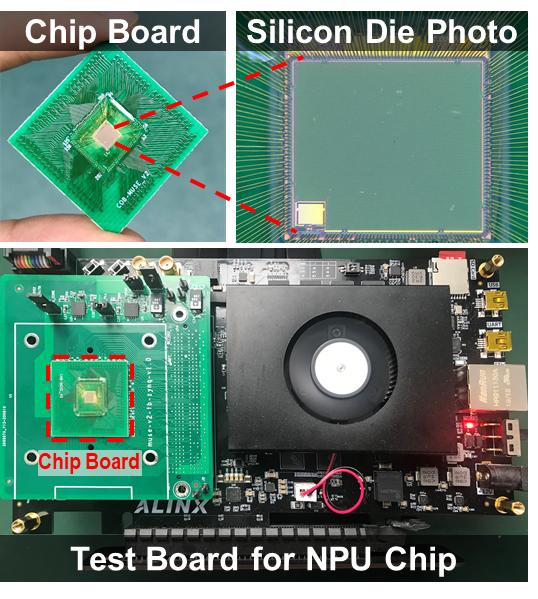
 

图12 芯片测试系统框图 图13 芯片测试硬件环境

# 芯片性能评估

## 整体性能效果

|  |  |
| --- | --- |
| **工艺** | UMC 55nm SP |
| **电压** | 0.75V – 1.00V |
| **面积** | 3.5mm x 3.9mm |
| **SRAM** | 243KB |
| **静态功耗** | 28mW (@Core: 1.0V; IO: 3.3V) |
| **频率** | 20 - 400MHz |
| **峰值性能** | 1407.15 GOPS  (VGG-16 Conv-8, 1-Pattern, 75% Kernel Pruning, 95% Activation Sparsity @380MHz, Core: 1.0V) |
| **功耗** | 24.71~312.84 mW  (min: VGG-16 Conv-11, @20MHz, Core: 0.75V  max: VGG-16 Conv-8, @360MHz, Core: 1.0V) |
| **最优能效** | 7.24TOPS/W  (VGG-16 Conv11, 1-Pattern, 75% Kernel Pruning, 95% Activation sparsity  @260MHz, Core: 0.87V) |

## 设计比较

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **设计工作** | **工艺** | **时钟频率** | **峰值性能** | **能耗** | **最优能效** |
| 启明920 | UMC 55nm SP | 20~400MHz | 1407.15 GOPS | 24.71~312.84mW | 7.24TOPS/W |
| 启明910[1] | UMC 55nm SP | 10-400MHz | 234.6 GOPS | 90mW | 3.16TOPS/W |
| Eyeriss [2] | TSMC 65nm LP | 200MHz | 84 GOPS | 278mW | 302.2GOPS/W |
| VLSI’16 [3] | TSMC 40nm LP | 12-204MHz | 102 GOPS | 76mW | 1.75TOPS/W |
| VLSIC’17 [4] | TSMC 65nm LP | 10-200MHz | 368.4 GOPs | 290mW | 1.27TOPS/W |
| QUEST [5] | 40nm CMOS | 75-330MHz | 1960 GOPS | 3.3W | 590GOPS/W |

[1]启明910峰值性能、功耗、能率比在400MHz频率下测试得到，

[2] Y.H. Chen, et al, ISSCC 2016 ， [3] B.Moons et al, VLSI 2016,

[4] S.Yin et al, VLSIC 2017 ， [5] K. Ueyoshi et al, ISSCC 2018