系统级文章：

DIANA\_An\_End-to-End\_Hybrid\_DIgital\_and\_ANAlog\_Neural\_Network\_SoC\_for\_the\_Edge

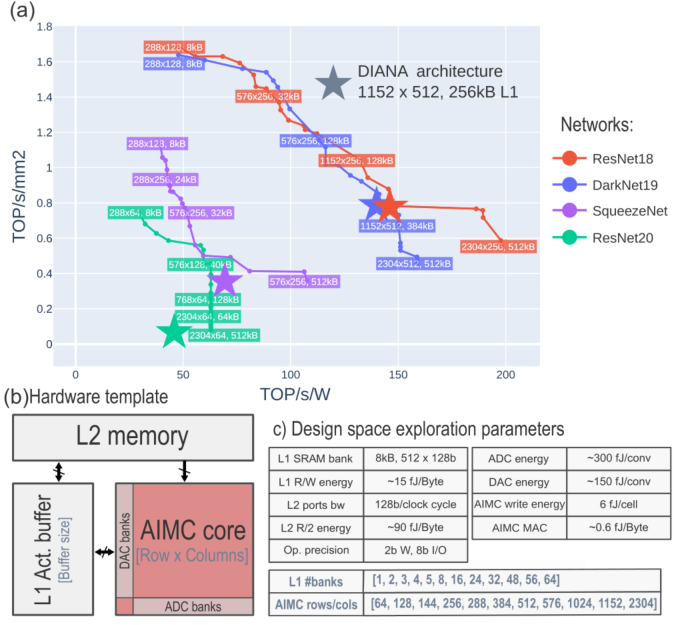
图示

描述已自动生成

DIANA体系结构，主要包含CPU，数字核(数字PE阵列)以及模拟核(AIMC)。

该系统通过两个通信网络连接各SoC组件：用于数据通信的TCDM总线以及控制指令的APB总线。

片上空间探索：



作者通过自己设计的硬件模板对AIMC的硬件空间设计进行了探索。图中说明了该硬件模板卸载不同网络进行计算时的能效。其中ResNet18以及DarkNet19能效均得到很好的效果（这两个网络均包含大量的non-pointwise 卷积），而SqueezeNet以及ResNet20的优化效果则不是很理想（这些网络有着大量的1x1的卷积）。

图示

描述已自动生成

**控制线路：**该架构的工作步骤如图，分为5步。1.RISC-V内核初始化L2 2.L2通过DMA将激活和权重从L2->L1 3.内核启动（在单个内核中运行单个负载或者多个负载）4.在计算启动时，DMA持续工作。5.输出数据通过I/O送出芯片。

**存储层次：**DIANA包含三个层次的存储器1.全局L2 SRAM（512KB），存储RISC-V二进制代码和系统配置数据 2. 加速器专用L1 SRAM（256KB），一个存储体占用16字节，存储特征图，作为加速器之间的通信数据缓冲区 3. 本地Reg

AIMC计算核

图示, 示意图

描述已自动生成

AIMC核针对大规模MVM操作的节能执行进行了优化处理。

设计这些模块有两个挑战（1）如何在每个周期提供阵列所需的大规模向量 （2）最大化重用操作数。

AIMC核的工作流程如下： MCU从L1提取输入特征图的向量，并送入激活缓冲区，出发AIMC核计算。计算完成后输出缓冲收集结果顺序将6b\*512的结果送入SIMD。其流水线如图（c）所示。

MCU模块工作流程：

图示

描述已自动生成

其工作流程如右图，在做一个3x3的卷积时，第一拍取出特征图中第一个3x3的数据块，进行第一轮乘法运算，由于卷积核向右滑动，所需输入的特征图切换为2，但是此时并没有加载IX4这一列的数据，因此在在第三拍完成数据替换。

AIMC宏：

每个单元由两个SRAM构成，分别存放weight的符号位以及值。对于每个MVM操作，激活所有1152个字线。

输出buffer以及SIMD单元：

SIMD使用六级流水线处理输出buffer的数据。

图示, 示意图

描述已自动生成

图示

中度可信度描述已自动生成

数字核：

内部主要由一个16x16的PE阵列构成，具有2、4、8位精度可扩展的MAC操作，根据选定的精度，并行度可以增加到32x16核64x16。这一块和脉动阵列差不多。

实验：

图表, 直方图

描述已自动生成

使用模拟核进行计算在TOP/s/W性能以及TOP/s效率方面均优于数字核的95倍。

工作负载分配方面，文章分析了CIFAR10的ResNet20网络和ImageNet的ResNet18网络具有不同形状的层。

表格

描述已自动生成

系统级论文：

A Programmable Neural-Network Inference Accelerator Based on Scalable In-Memory Computing

图示

描述已自动生成

文章提出的体系结构如图所示，主要由CIMU，On-Chip Network构成。

其中CIMU内部由CIMA，buffer，control 和 SIMD四个模块组成。

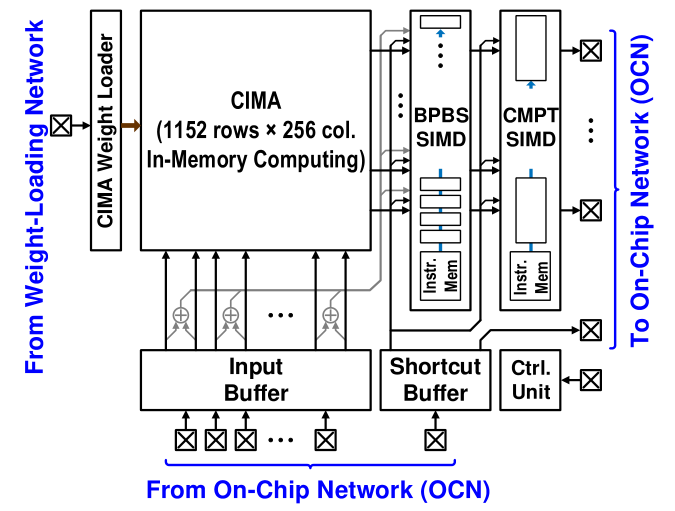
此外，还有一个专用的WL网络，用于将权重从基于SRAM的权重缓冲区移动到CIMU，该缓冲区最终将在芯片上，但未在图中画出。

图示

描述已自动生成

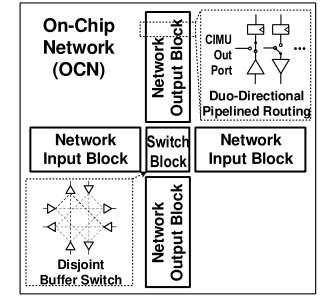
图中展示了NN层级的调度（作者采用了8个CIMU执行计算），一次将层权重的子集映射到并行CIMU（首先是第1-3层，然后是第4-5层），然后进行流水线执行，以此在不浪费时间的情况下完成各层级的运算。

CIMU核结构：



与第一篇文章类似，该CIMU核内部由CIMA（ADC占据了20%的面积和29%的能耗），输入buffer，shortcut buffer以及近存SIMD组成。其中CIMA进行权值与激活的乘加计算，输入buffer向CIMA传输激活值，shortcut buffer与输入buffer并行，当输入buffer向CIMA提供数据时，shortcut buffer 绕过CIMA直接向SIMD提供数据，以实现NN计算的进一步可配置性。在CIMA之后，SIMD接收数字化列计算结果并支持进一步的数字操作，BPBS SIMD执行加法和移位等运算，CMPT SIMD执行激活、池化等操作。

OCN模块：OCN由输入块，输出块以及切换块构成，主要实现各个CIMU间的互联，通过switch block，左下的CIMU可以与这个tile内部的其它所有CIMU互联。



图示, 示意图

描述已自动生成

图（a）展示了一个步长1卷积核尺寸3x3的卷积层计算，卷积核倍展平并且权重沿着列维度被映射到CIMA中，输入/输出对应于输入/输出激活。

图（b）说明了shortcut buffer的作用，其主要从OCN网络中接受激活数据并直接将其传递给SIMD进行合并计算，主要用于解决梯度消失问题（残差链接）

图（c）说明了dense层和其他无重用（例如，1×1卷积）层的映射。输入缓冲器接收来自OCN的高带宽激活，并将数据直接排序到CIMA。

图（d）说明了递归/记忆增强层的映射。OCN被配置为提供输出到输入路由。然后，输入缓冲器将新的状态输入x t与之前的隐藏状态yt−1连接起来。以LSTM为例，四个LSTM门之前的权重矩阵在CIMA列之间交错，从而可以在相邻的NMC SIMD数据路径内执行MVM之后的元素顺序和跨元素操作，以获得输出数据位置。

来自各CIMU的结果经过OCN的整合馈送到后续NN层。