A Brief Introduction to The DianNao Project

Before the DianNao Project

In 2010, during a keynote at ISCA, Prof. Temam outlined that a remarkable convergence of trends in technology, applications and machine-learning, were pointing to machine-learning accelerators as a very attractive scalability path for micro-architectures.

At ISCA in 2012, Prof. Temam proposed a first machine-learning accelerator design, showing that it was possible to achieve high performance with a small area and power footprint on a large set of neural network based applications. The main limitation of that accelerator was the memory bandwidth.

The DianNao Project

The goal of the DianNao research project was to develop accelerator architectures for machine-learning. The project was an academic collaboration between Prof. Yunji Chen (ICT) and Prof. Olivier Temam (Inria), within a joint ICT/Inria lab.

The academic collaboration between Prof. Temam and Prof. Chen started with the second accelerator, called DianNao (the first member of DianNao family). This accelerator extended the ISCA 2012 accelerator with local memories in order to capture the locality properties of deep neural networks and overcome memory bandwidth limitations. This design was published at ASPLOS in 2014 and got the best paper award.

The second accelerator of the DianNao family was a multi-chip version of DianNao and had two main goals: show that machine-learning accelerators have excellent scalability properties thanks to the partitioning properties of neural network layers, and to aggregate enough memory capacity to store the whole machine-learning model on-chip in order to overcome memory bandwidth limitations. This design, called DaDianNao, was published at MICRO in 2014 and got the best paper award.

As another way to overcome memory bandwidth limitations in embedded applications, we have also shown that such machine-learning accelerators can be directly connected to a sensor, bypassing memory. We applied this to a vision sensor, leading to the design called ShiDianNao (the third member of DianNao family), and published at ISCA in 2015.

Finally, we also demonstrated that the application scope of such accelerators can be extended to multiple machine-learning algorithms because they share common primitives; the corresponding design, called PuDianNao (the fourth and final member of DianNao family) was published at ASPLOS 2015.

After the DianNao Project

Prof. Chen and his ICT team designed an Instruction Set Architecture (ISA) for a broad range of neural network accelerators, called Cambricon. This ISA design was published at ISCA 2016, and got the highest score in peer review.

中法联合DianNao项目简介

DianNao项目之前

2010年，Temam教授在ISCA的主题报告上提到，机器学习硬件加速器是处理器微结构领域极有吸引力的一个发展方向，是处理器技术、应用和机器学习发展的大势所趋。在2012年的ISCA上，Temam教授提出了第一个机器学习加速器设计，表明在以神经网络为基础的一大类应用上是可以以很小的面积和功耗获得高性能的。但此工作的主要局限性在于其内存带宽。

DianNao项目

DianNao学术项目的目标是面向机器学习研究加速器架构。本项目是中科院计算所的陈云霁教授和法国Inria的Olivier Temam间的一个学术合作项目，双方为此设立了联合实验室。

Temam教授和陈教授的合作始于第二个加速器，名为DianNao（这也是DianNao家族的第一个成员）。DianNao在ISCA-2012加速器的基础上增加了局部存储，使其可以捕捉深度神经网路的数据局部性并由此克服内存带宽的限制。DianNao加速器的设计发表于ASPLOS-2014，获得了该会议的最佳论文奖。

DianNao家族的第二个加速器是DianNao的多片版本，有两个主要的设计目标：一是揭示神经网络层的可分特性使得加速器可具备极好的可扩展性，二是聚集足够多的片上存储来将整个机器学习模型都放在片上，从而克服内存带宽的限制。这个被称为DaDianNao的设计发表在MICRO-2014上，获得了该会议的最佳论文奖。

作为克服嵌入式应用中内存带宽限制的另一种方法，我们揭示可以通过加速器和传感器的直连来绕过内存。我们将此思想应用于视觉传感器，从而提出了DianNao家族的第三个加速器ShiDianNao，发表于2015年的ISCA上。

最后，我们也揭示这类加速器的应用领域可以被拓展至多种机器学习算法，因为这些算法多具有类似的运算操作。相应的加速器设计称为PuDianNao（DianNao家族的第四个以及最后一个成员），发表于ASPLOS-2015。

DianNao项目之后

陈云霁教授和他的中科院计算所团队为一大类神经网络加速器设计了一套名为Cambricon的指令集。该指令集发表于ISCA-2016，在该会议的同行评议中获得了最高分。