



SaturnLab



清华 Tsinghua
iCenter

深度学习硬件

Hardware-Architectures for Deep Learning

教师: 陈 震

单位: 清华大学基础工业训练中心

深度学习软硬件布局

- DL硬件
- DL软件
- AI产业



深度学习框架

- **Tensor Flow:** *Google Deep Learning Library*
 - Supports general deep learning with symbolic diff.
 - Python on top of C++ (Easy + Fast)
 - GPU, cluster, and mobile implementations
- **pyTorch:** *Facebook AI research*
 - Tensor Library
 - File I/O Interface Library
- **Berkeley Caffe:** *GPU accelerated Computer Vision*
 - Focused on computer vision and GPU acceleration
 - C++ with Python support (Very Fast + somewhat easy)
 - Rich library of pre-trained models (Caffe Model Zoo)
- **Theano:** *U of Montreal*
 - General Symbolic Diff. Modeling Framework
 - Covers many recent research models
 - Python only (easy but not fast)
- **DMLC/MXNET:** *Amazon*
- **CNTK:** *Microsoft*
- **Baidu/PaddlePaddle**

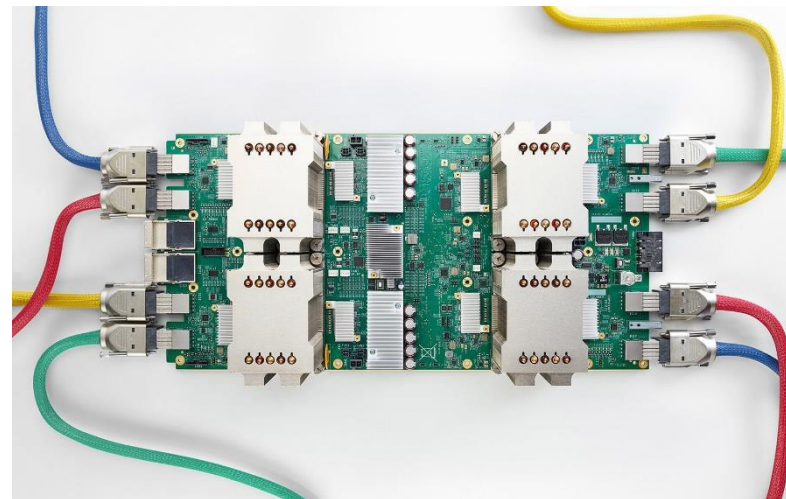
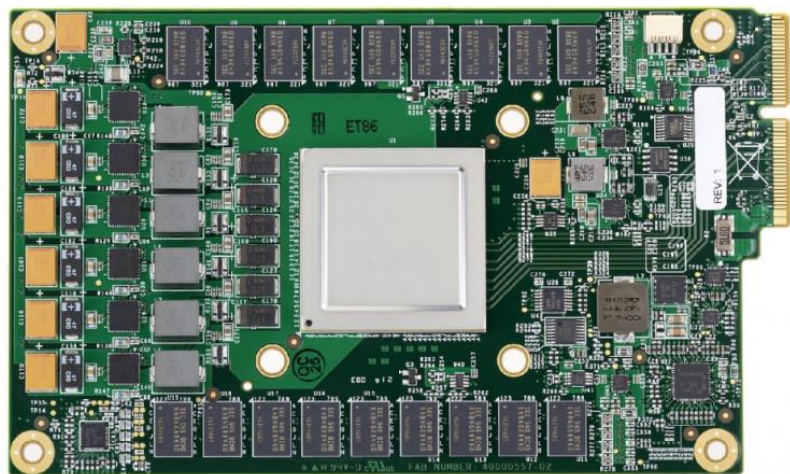


深度学习硬件-专用加速器

- Deep Learning Accelerator (DLA)
- 谷歌
 - 2016年发布TPU一代（用于推断）
 - 2017年发布TPU二代和Cloud TPU（用于训练和推断，2017年10月）
- 英伟达
 - 开源深度学习加速器XAVIER DLA（2017年5月）
 - <http://nvdla.org/>
- TFLOPS（万亿次浮点运算）

谷歌TPU

- 协处理模式工作 (coprocessor) PCIe-v3
- 张量处理器TPU (tensor processing unit)
- TPU一代
- TPU二代 Cloud TPU
- 运算速度: 180 teraflops 64GB
- 二维高速环形网络, 单精度浮点MXU (matrix Unit)



GPU编程

- NVIDIA GPU显卡
- 协处理模式工作 (Coprocessor)
- 开发卡: Titan XP (帕斯卡架构)
- 开发工具: CUDA
- 编程语言: C++
- 运算速度: 12 TFLOPs



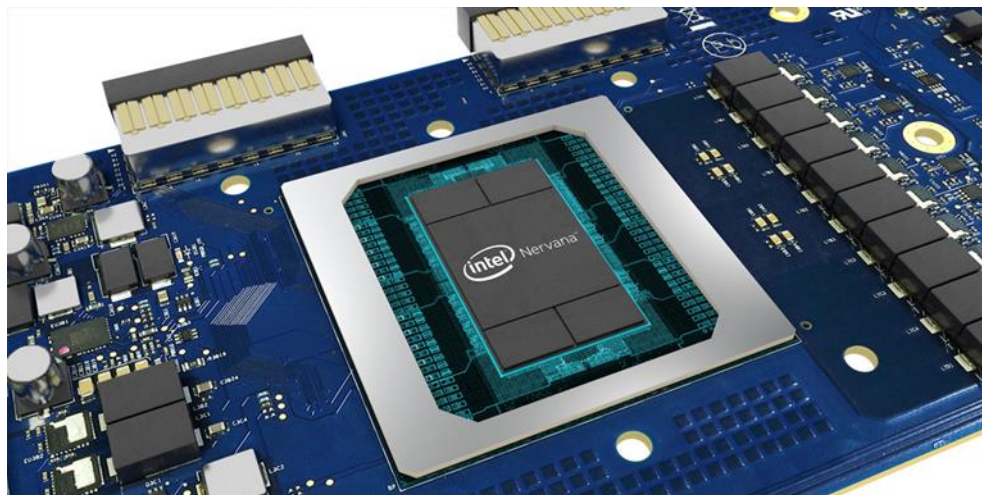
Intel公司

- Intel公司
 - 至强融核（Xeon Phi处理器） Knights Mill
 - Deep Learning Inference Accelerator (DLIA)
- 协处理模式工作
- 指令集
 - “四倍融合乘加指令”（QFMA: Quad Fused Multiply Add）
 - “四倍虚拟神经网络指令”（QVNNI: Quad Virtual Neural Network Instruction）。
- QFMA把Knights Mill的单精度性能提高一倍，QVNNI指令则可以进一步降低精度，同时满足深度学习框架的精度需求



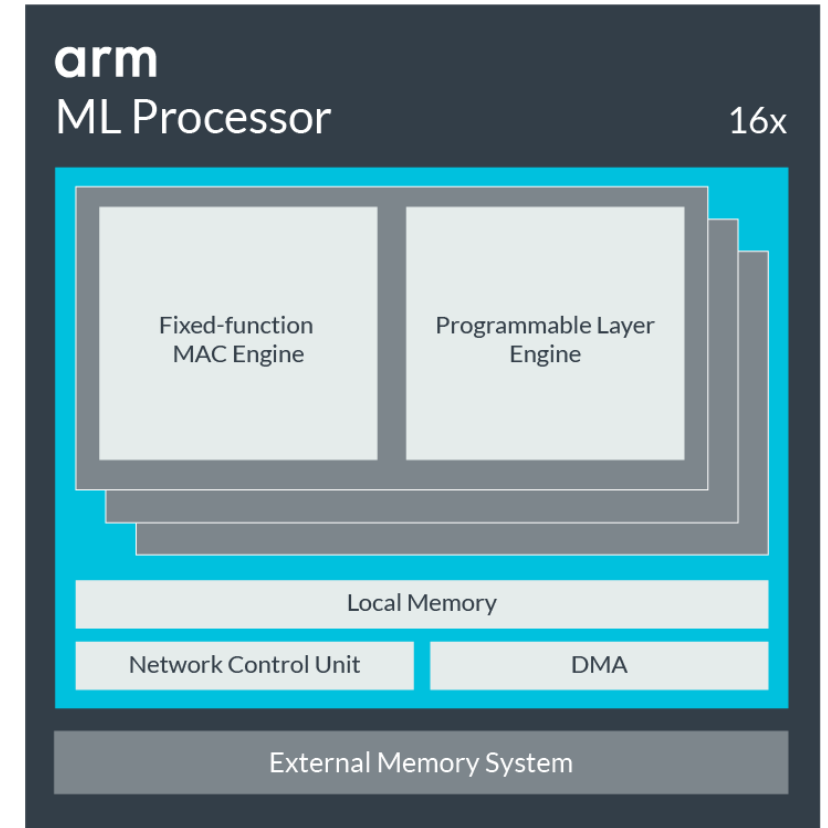
Intel公司

- Intel收购的Nervana System
- NNP (Neural Network Processors) (2017年10月)
- 协处理模式工作



安谋ARM

- Project Trillium, Arm's Machine Learning (ML) platform
- Arm ML processor, Arm OD processor 和 Arm NN SDK
- 性能4.6 TOPs, 功耗3 TOPs/ W
- <https://developer.arm.com/products/processors/machine-learning>



FPGA编程

- PYNQ: Xilinx APSoCs
- ARM+Zynq-7000
- 开发工具: Xilinx vivado 2017.3
- 程序: Python/C++/HLS
- 应用:
 - Binary Neural Network
 - <https://github.com/Xilinx/BNN-PYNQ/>
 - CNN Example
 - <https://github.com/awai54st/PYNQ-Classification>



谢谢指正！

zhenchen@tsinghua.edu.cn