深度学习硬件

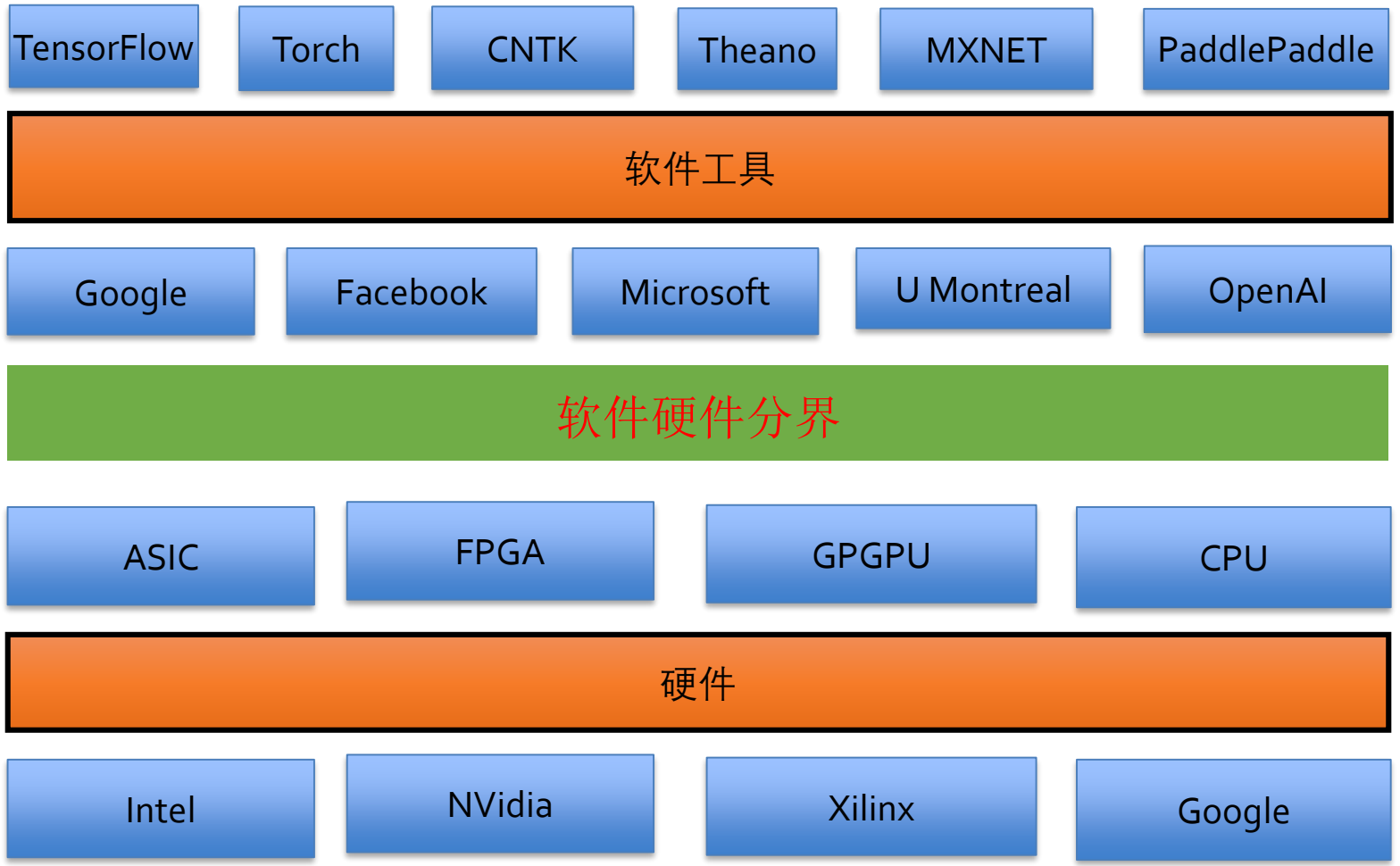
Hardware-Architectures for Deep Learning

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)

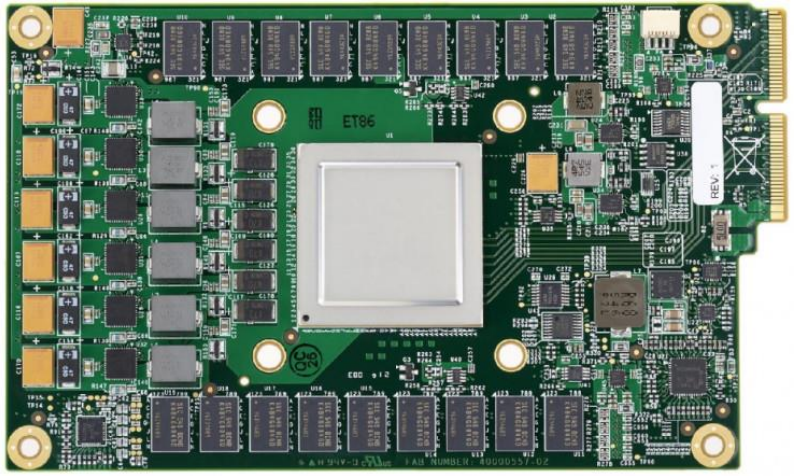
- 谷歌
  - 2016年发布TPU一代（用于推断）
  - 2017年发布TPU二代和Cloud TPU（用于训练和推断，2017年10月）

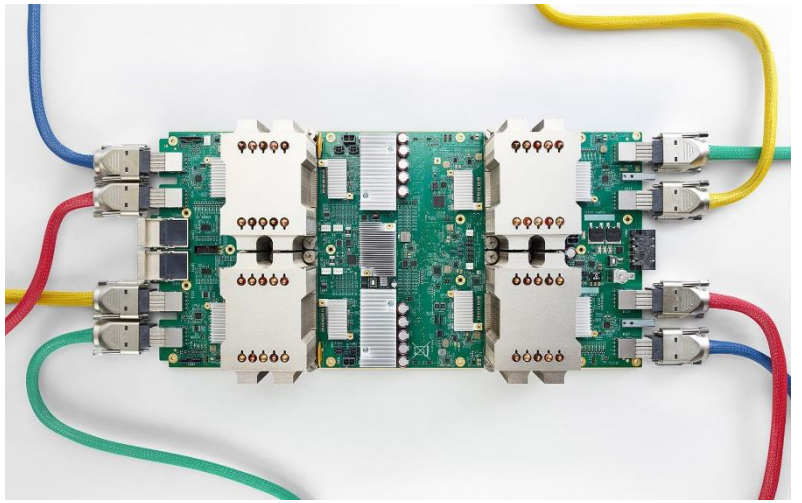
- 英伟达
  - 开源深度学习加速器XAVIER DLA（2017年5月）
  - http://nvdla.org/

- 运算能力单位:
  - TFLOPS tera floating point operation 1万亿次
  - PFLOPS peta floating point operation 1千万亿次
  - EFLOPS exa floating point operation 1百亿亿次

谷歌TPU

- 张量处理器TPU (tensor processing unit)
- 协处理模式工作（coprocessor）PCIe-v3
- 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）

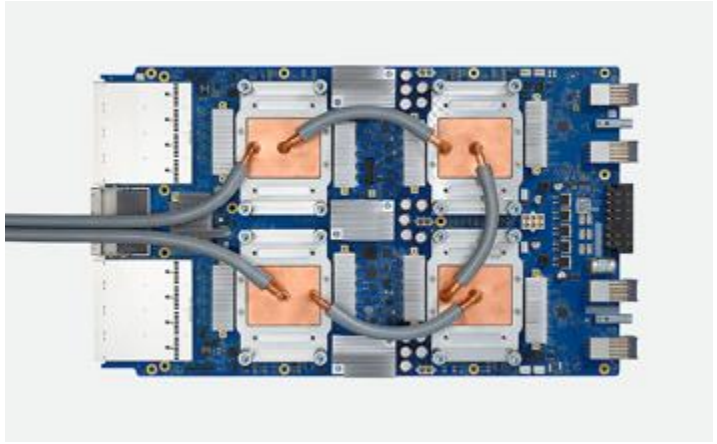




Cloud TPU v2
每秒 180 万亿次浮点运算
64 GB 高带宽内存 (HBM)

谷歌TPU

- 张量处理单元 (TPU)
- 为机器学习设计的 ASIC，应用于包括 Google 翻译、Google 相册、Google 搜索、Google 助理和 Gmail。



Cloud TPU v3 测试版
每秒 420 万亿次浮点运算
128 GB HBM

Cloud TPU v2 Pod Alpha 版
每秒 11.5 千万亿次浮点运算
4 TB HBM
二维环形网状网络

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

Nvidia GPU卡参数

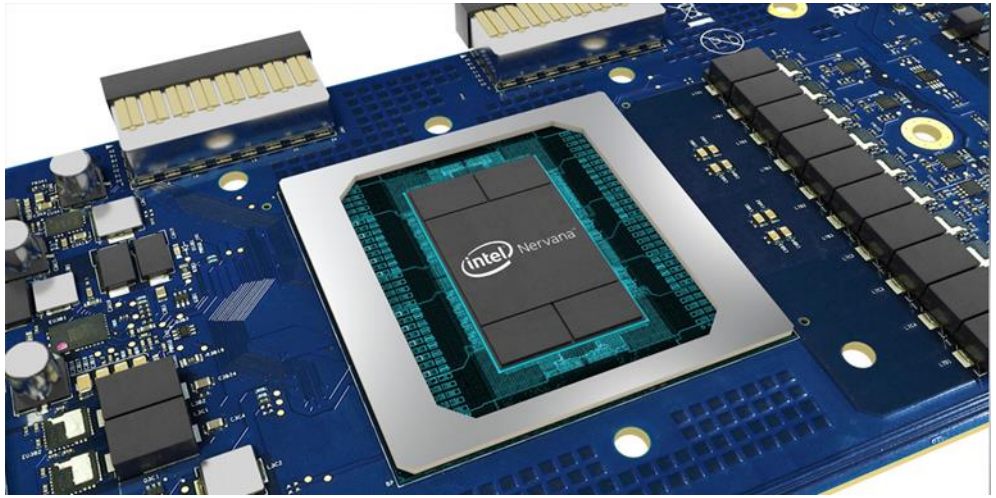
产品名称	GTX 1080Ti	RTX 2080Ti	TITAN Xp	TITAN V	TITAN RTX
GPU架构	Pascal	**Turing**	Pascal	Volta	**Turing**
GPU芯片数量	（GP102）x 1	**（TU102）x 1**	（GP102）x 1	（GV100）x 1	**（TU100）x 1**
CUDA核心数	3584	**4352**	3840	5120	**4608**
单精度计算峰值	11 Tflops	**14.2 Tflops**	12 Tflops	15 Tflops	**16.3 Tflops**
双精度计算峰值	N/A	**N/A**	N/A	N/A	**N/A**
内存容量	11GB	**11GB**	12GB	12GB	**24GB**
内存带宽	320GB/s	**616GB/s**	480GB/s	652.8GB/s	**672GB/s**
总功耗	250W	**260W**	250W	250W	**280W**
散热方式	主动散热	**主动散热**	主动散热	主动散热	**主动散热**
Display端口	1	**1**	1	4	**4**

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

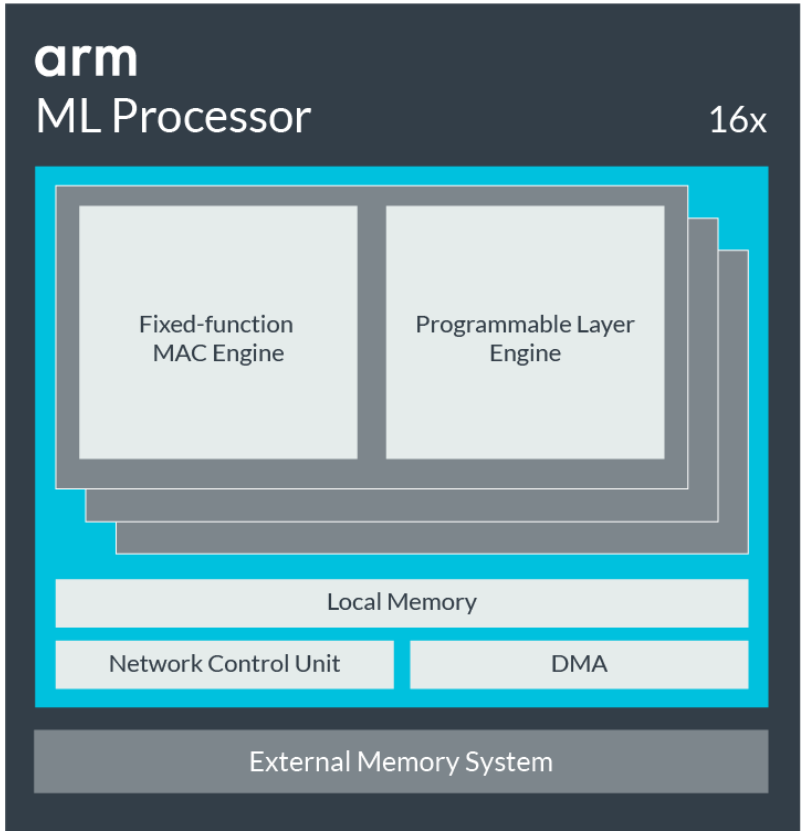
- Intel收购的Nervana System
- NNP（Neural Network Processors）（2017年10月）
- 协处理模式工作



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

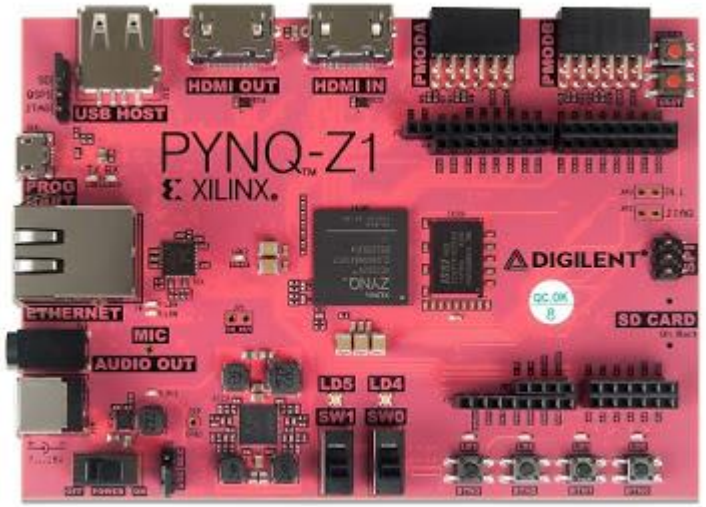
- https://developer.arm.com/products/processors/machine-learning



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