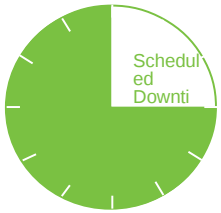
Accelerate your AI cloud: A virtualization perspective

Liang Yan (lyan)
SUSE Labs



Outline

1. Background
   AI Cloud
   Hardware Accelerator
2. GPU virtualization
3. FPGA virtualization
4. AI chips and their virtualization
5. Conclusion
6. Q&A

Background

AI & Cloud

Cloud is well known today.

Public Cloud:
- Goolge, AWS, Azure

Private Cloud:
- Vmware, Xenserver, openstack

AI:

AI/Machine Learning/ Deep Learning

AI Cloud

Gartner Report on Cloud market:

2017: 145 billions
2018: 175 billions
2019: 206 billions

Gartner Report on AI market:

Table 1. Forecast of Global AI-Derived Business Value (Billions of U.S. Dollars)

	2017	2018	2019	2020	2021	2022
Business Value	692	1,175	1,901	2,649	3,346	3,923
Growth (%)		70	62	39	26	17

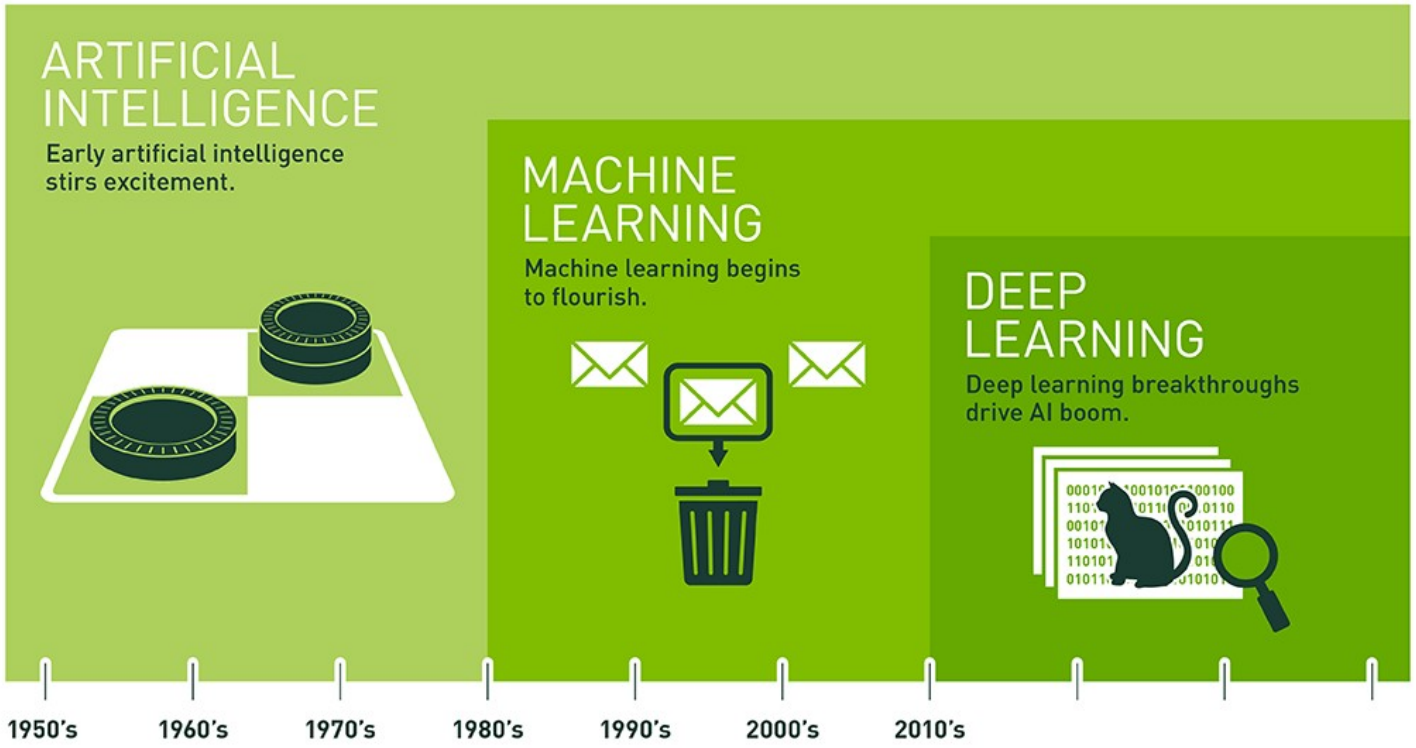
Source: Gartner (April 2018)

AI Cloud

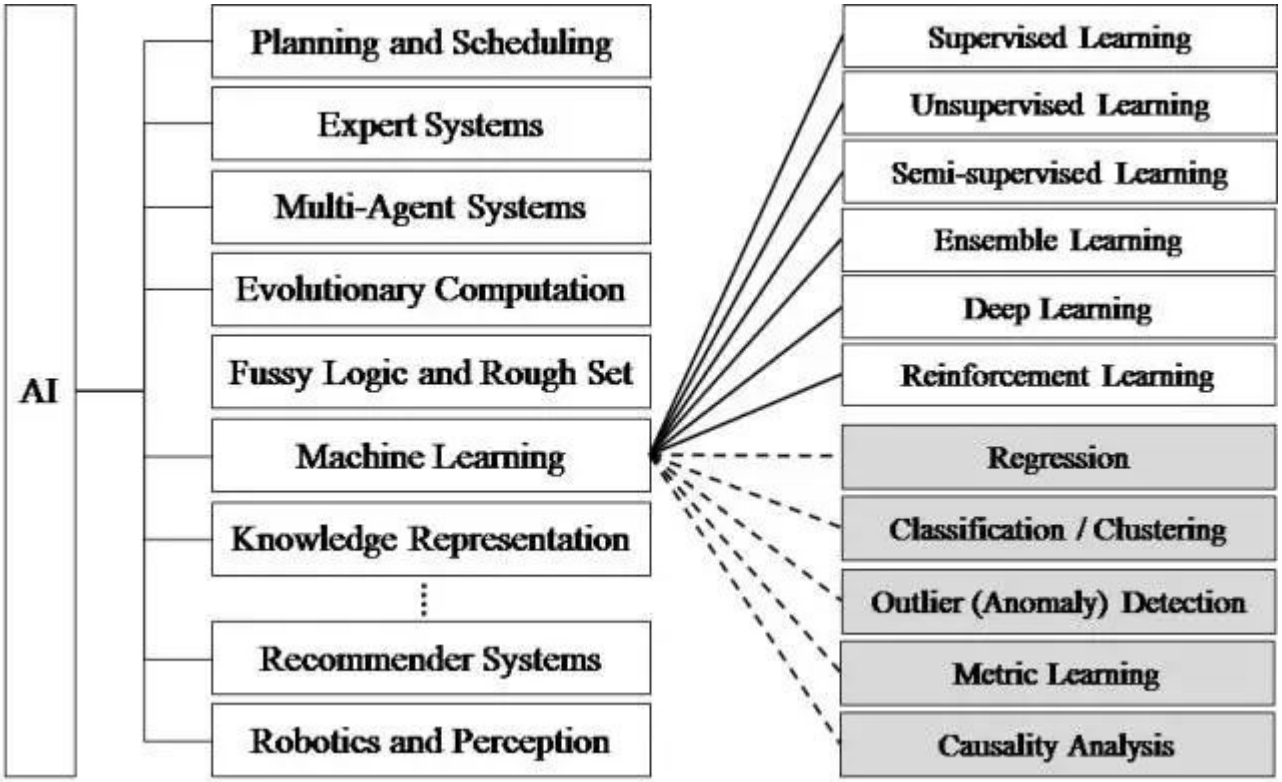
AI User Cases:

Auto drive:
Medical area:
Finance:
Electronic Commerce: Delivery Transport, Recommend Sale
Language Translate:
Face/Image Recognition

AI & Cloud


ARTIFICIAL INTELLIGENCE
Early artificial intelligence stirs excitement.
MACHINE LEARNING
Machine learning begins to flourish.
DEEP LEARNING
Deep learning breakthroughs drive AI boom.
1950's
1960's
1970's
1980's
1990's
2000's
2010's

AI & Cloud



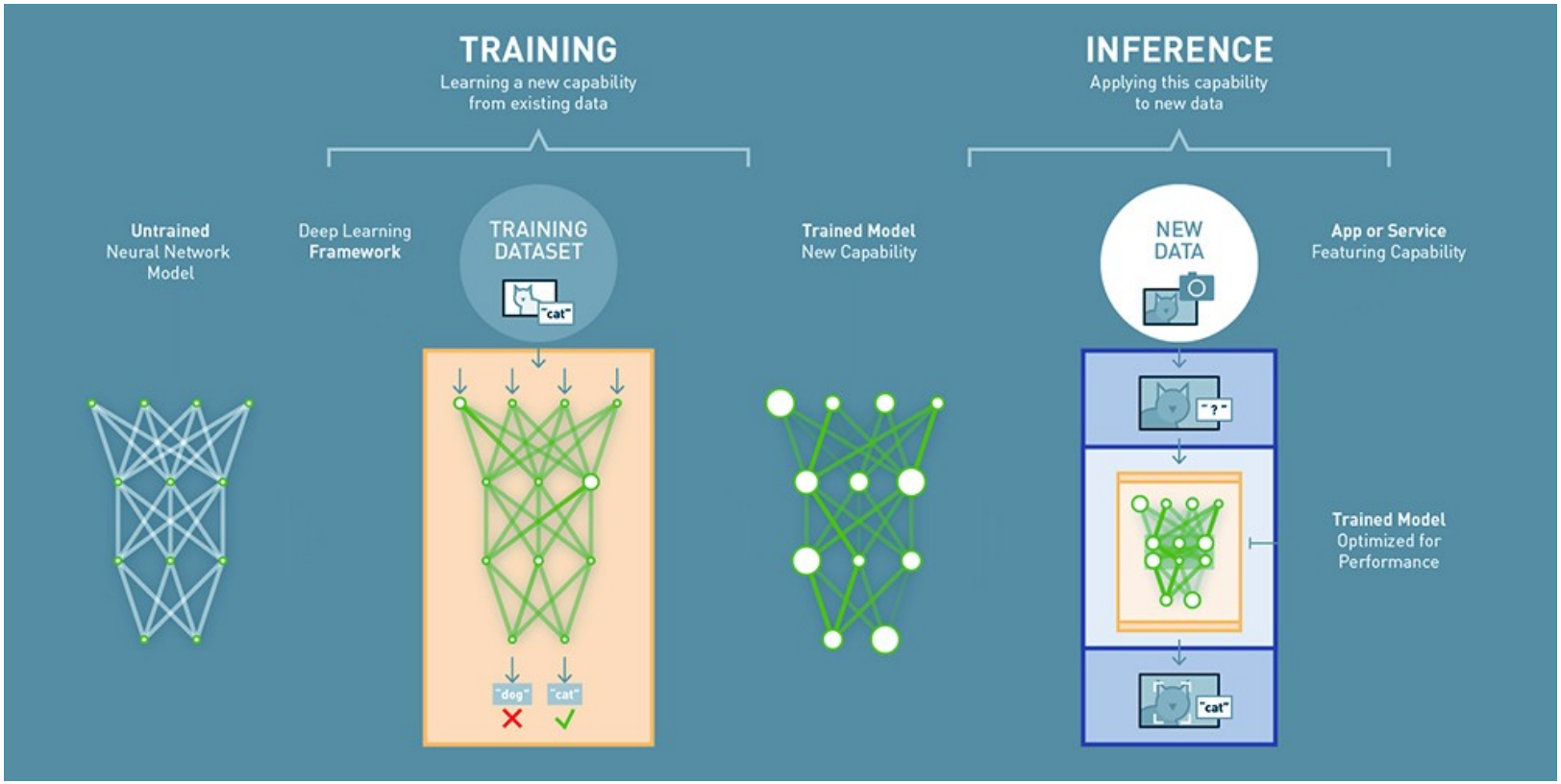
https://www.zhihu.com/question/57770020

Deep Learning

Deep Learning is a branch of machine learning in which the models (typically neural networks) are graphed like “deep” structures with multiple layers. Deep learning is used to learn features & patterns that best represent data.

It achieves an algorithm not designed by human, the more data and layer it has, the more accurate result it will get.

Artificial neural networks such as：
DNN=>CNN=>RNN Model



https://blog.exxactcorp.com/discover-difference-deep-learning-training-inference/

Deep Learning

A lot of Framework today, too many actually!

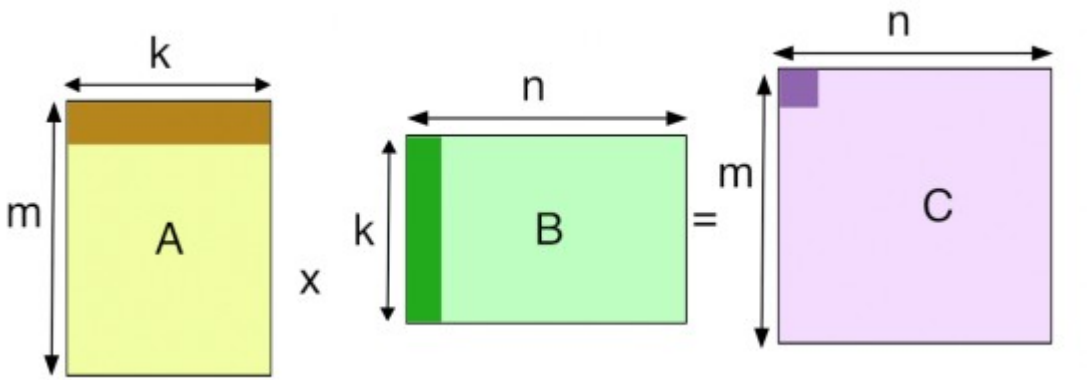
- TensorFlow
- PyTorch
- MxNet
- Keras
- Theano

●●●

Framework =》 IR =》 Backend API

Backend support:
CUDA/OpenCL
GPU, FPGA,TPU,

Deep Learning / Neural Network
GEMM, General Matrix to Matrix Multiplication




```
for(i = 0; i < n; i++) {
    for(j = 0; j < n; j++) {
        C[i][j] = 0;
        for(k = 0; k < n; k++)  {
            C[i][j] += A[i][k] * B[k][j];
        } }}
```

AI Accelerator

Thinks of about millions of layers and parameters

- GPGPU
  - Nvidia
  - AMD
  - Intel
- FPGA: Field Programmable Gate Array
  - Software Algorithm Oriented
  - Xilinux, Altera(acquired by intel in 2015 as $16.7 B)
    Both support OpenCL and HSL(High Level Synthesis)
    2011, Altera introduced OpenCL for FPGA, CNN algorithms
- AI chip(ASIC-Application Specific Integrated Circuits)
  A ASIC specific for DL algorithms

AI Accelerator

	Training	Inference
Cloud	GPU: Nvidia FPGA: Intel Xlinux ASIC: Google TPU	GPU: Nvidia AMD FPGA: Intel Xlinux MS AWS Google Tencent Ali ASIC: Google TPU, Groq WaveComputing DPU, Bitmain Cambricon DianNao, Horizon Robotics BPU,
End/Mobile		GPU: Nvidia ARM FPGA: Xlinux Deephi ASIC: QCOM Huawei Apple Intel Nervana,

AI Cloud

GPU/FPGA Virtualization in Cloud, providing machine learning service

GPU:

Google Colaboratory
Paperspace Gradient
FloydHub Workspace
Lambda GPU Cloud
AWS Deep Learning AMIs
GCP Deep Learning VM Images

FPGA:

AWS+Xlinux
Ali Cloud + Intel Altera
MS + Intel Altera

GPU Virtualization

Full GPU Virtualization

vGPU Investments Upstream

- NVIDIA (vComputeService)
- Intel (GVT-G)
- AMD(GIM)

Intel has no VRAM

AMD has IOMMU support

SRIOV 97%
MDEV 80~90%

Full GPU Virtualization

Run native graphics driver in VM
Achieve good performance and moderate multiplexing capability

- Split
  - Time Slices
  - framebuffer memory
- Isolate
  - Give a neat access between VM and Host Physical Device
  - IOMMU/Mdev and VFIO
  - DMA
  - Interrupt
- Schedule
  - Efficient and Robust
  - Pretty fix for AMD,
  - More flexible for NVIDIA, RR, BOND

Full GPU Virtualization

Nvidia
Tesla Series: Volta Pascal Maxwell M6 M10 M60 P4 P6 P40 P100 V100
Tensor processor, cuDNN
Inference production: Jetson Nano, Xavier, TX2
http://www.nvidia.com/object/grid-certified-servers.html

AMD
Radeon Instinct MI25, MI50
ROCm
https://lists.freedesktop.org/archives/amd-gfx/2016-December/004075.html

Intel

- Haswell(3VMs) Broadwell(7VMs) Skylake, Kaby Lake
- Dedicated GPU next year
- OpenVINO(Open Visual Inferencing and Neural Network Optimization)
  https://github.com/intel/gvt-linux/wiki

SUSE

- Intel KVMGT technical ready
- Nvidia vGPU technical ready
- AMD MxGPU on going

- GPU passthrough stage for Cloud
- GPU virtualization for CAASP

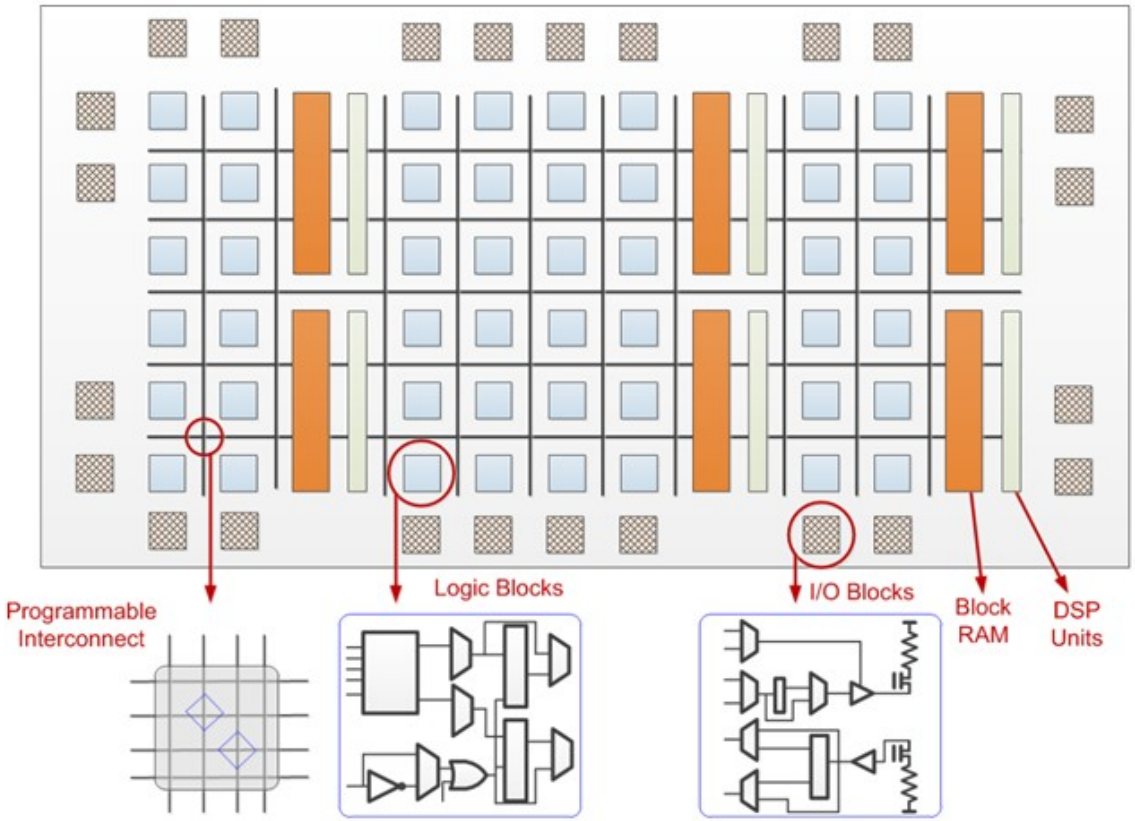
Upstream

- Live Migration
- Scalability/Schedule

FPGA Virtualization

FPGA: Field Programmable Gate Array

xilinux announced the first fpga(xc2000) in 1985.
It was used as prototype development by HDL(Verilog/VHDL).
Now start to be used as complicate situation, like DL.



https://medium.com/@ckyrkou/what-are-fpgas-c9121ac2a7ae

FPGA Virtualization

FPGA virtualization is a technique that provides an abstraction to the FPGA hardware.

Three main directions:

Overlays:

a programmable architecture on top of an FPGA

Dynamic modules:

PR(Partial Reconfiguration)

Intel Altera Stratix 10

Xilinx UltraScale

FPGA resource pool: MS catapult project

Multiple Nodes, Cloud Oriented

FPGA as a servce

Microsoft: Catapult Project V3
Bing Network => NN
FPGA resources pool
5670 servers with FPGAs over 15 countries

AWS F1

Huawei Cloud

Ali Cloud

SUSE

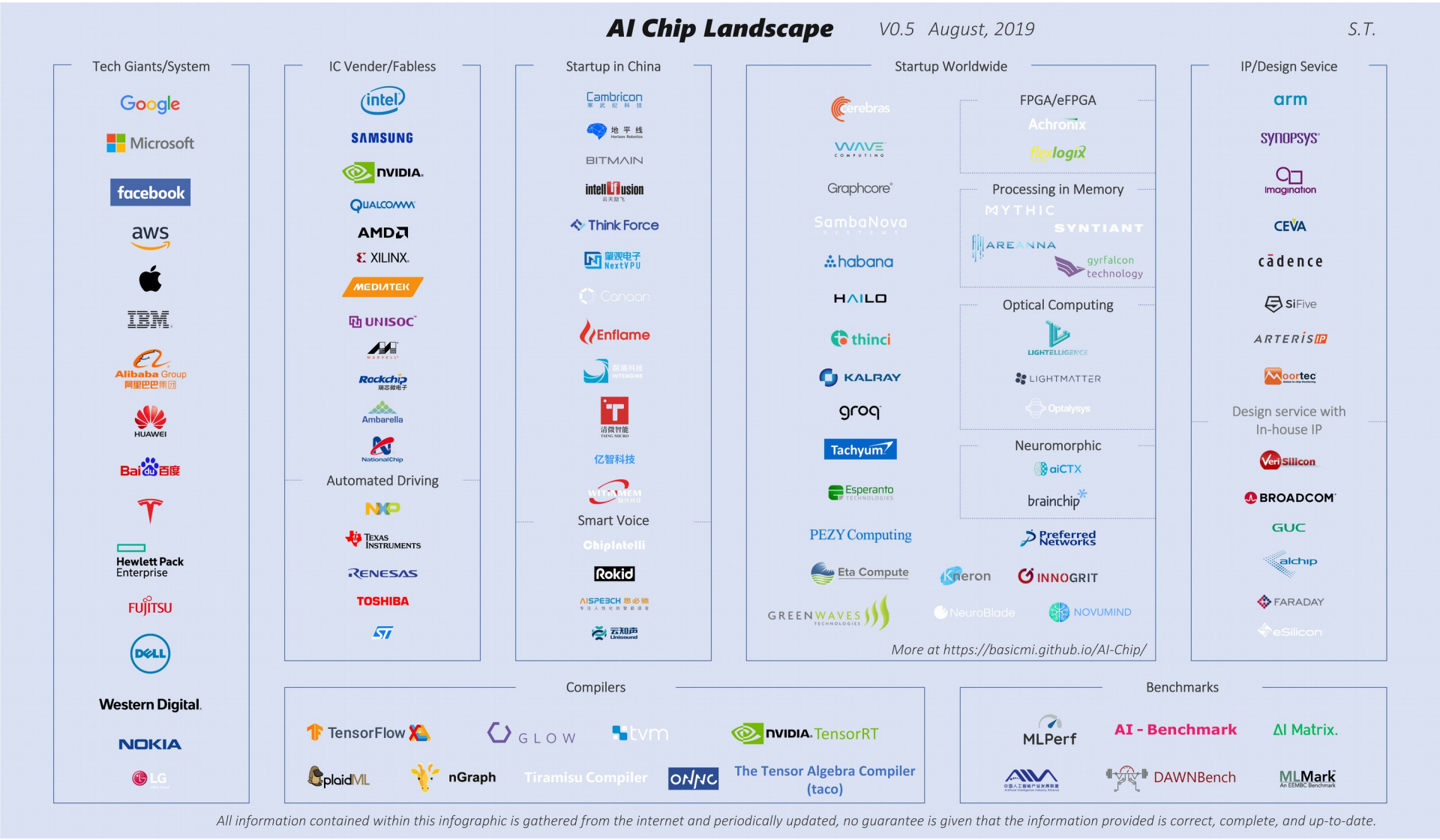
- Not yet
- Start work with Intel FPGA

Upstream

- Linux kernel FPGA subsystem
- Xilinx contributes Alveo FPGA Accelerator Drivers
- Vendor keep releasing new hardware

AI Chips

AI Chip Company



All information contained within this infographic is gathered from the internet and periodically updated, no guarantee is given that the information provided is correct, complete, and up-to-date.

AI Chips Today

Designless-Fabless: Software Define Chip

1. Mostly for Deep Learning/ Neural Network
   - MatrixMultply

2. Mostly for inference

3. Mostly for specific usage, no general platform like CUDA

AI Chips

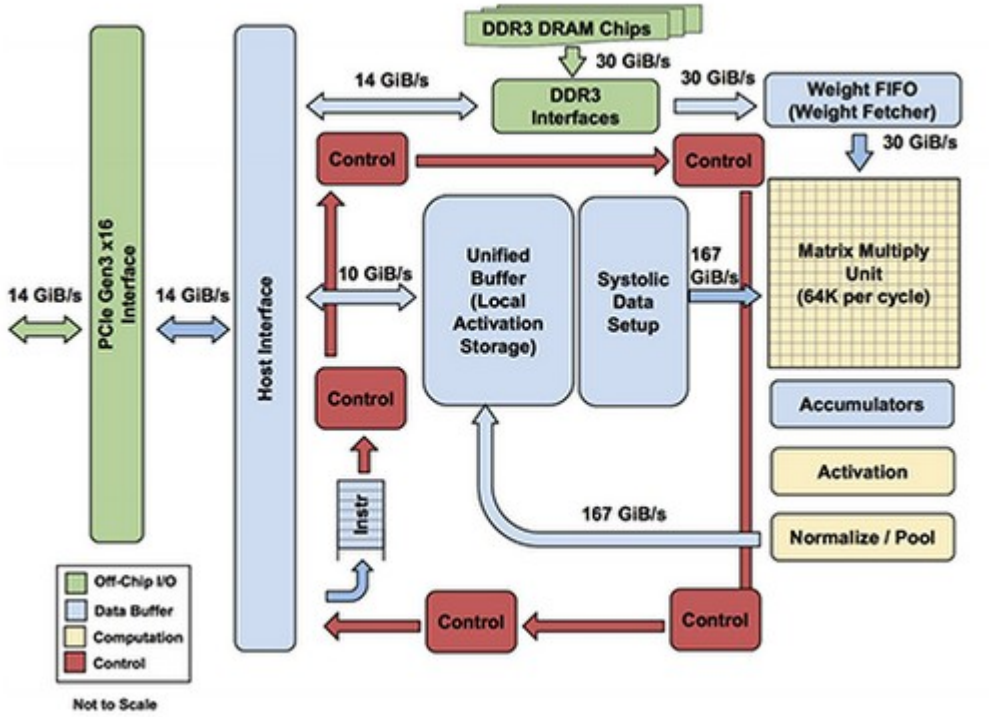
TPU(CISC)

Currently V3, which is 8 times than V2
You could try it on Google Colab(CPU, GPU(K80), TPU(v2))

Could support train stage since V2, TPU is around 3 times faster than GPU

https://colab.research.google.com/notebooks/intro.ipynb#

TPU architecture



Hardware implementation for CNN

Matrix Multiplier Unit (MXU):

65,536 8-bit multiply-and-add units for matrix operations

Unified Buffer (UB):

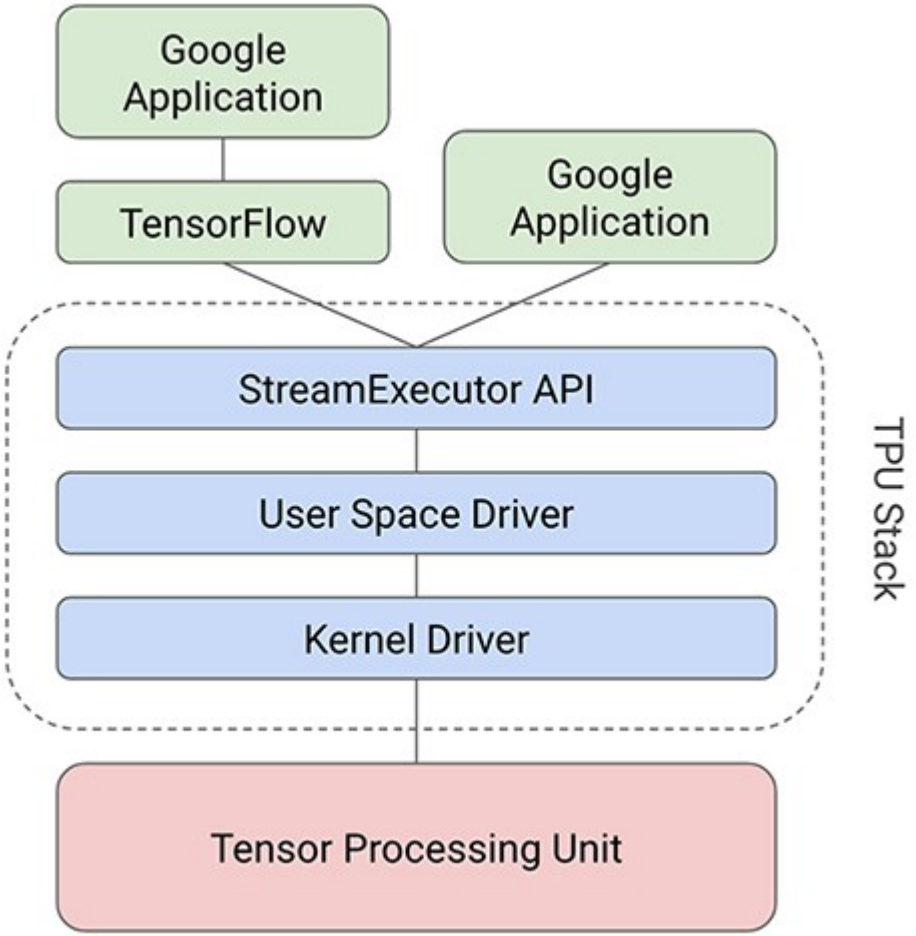
24MB of SRAM that work as registers

Activation Unit (AU):

Hardwired activation functions

https://cloud.google.com/blog/products/gcp/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu

TPU Software Stack

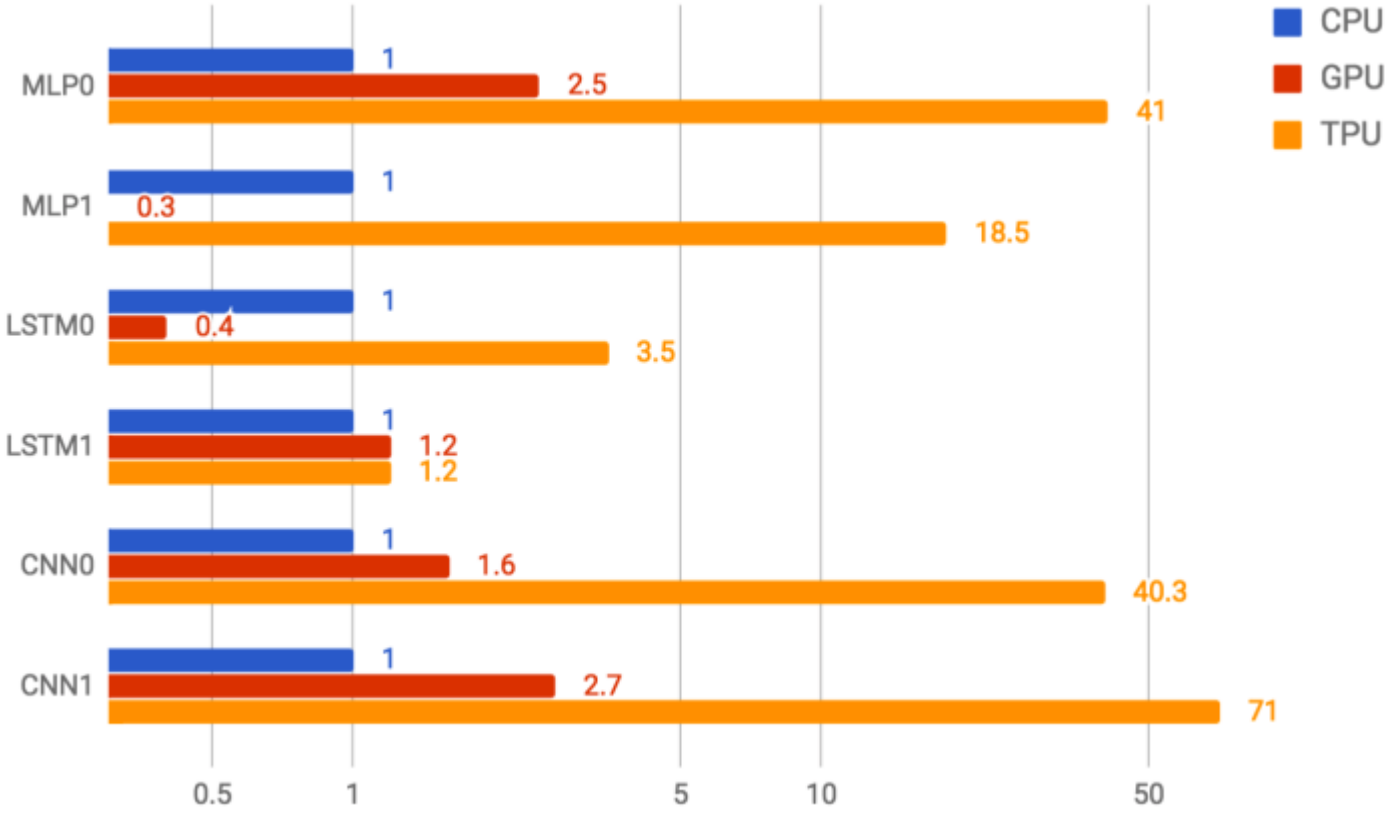


https://cloud.google.com/blog/products/gcp/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu

TPU performance

CPU, GPU and TPU performance on six reference workloads

Multilayer Perceptrons (MLPs) Convolutional Neural Networks (CNNs) Long Short Term Memory(LSTM)



https://cloud.google.com/blog/products/gcp/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu

AI Chips Future

Others:

Linux Hardware Accelerator Subsystem

ARM+ASIC

- Need a general platform

PCIE 4.0

Software ecosystem

Conclusion

Heterogeneous Architecture

Hardware Acceleration:

- Work load is so heavy that computing load is even higher than data movement and hardware latency.
- Hardware implementation for software algorithm

New IO virtualization trends: customize

- Device is focus on virtualization implementation(resource split, kernel bypass)
- Work is re-assigned to backend driver.
- Data plane between virtualization stack is much simple today.

SUSE(MY) Focus

1. Keep close with Hardware Vendors? New Feature, Enablement

2. Focus on Frameworks backend, fit in our production

3. Assistant work
   - Split Compute Task automatically
   - IO virtualization techniques

Question?

Thank you.

REFERENCE

VGPU ON KVM

https://blog.exxactcorp.com/discover-difference-deep-learning-training-inference/,

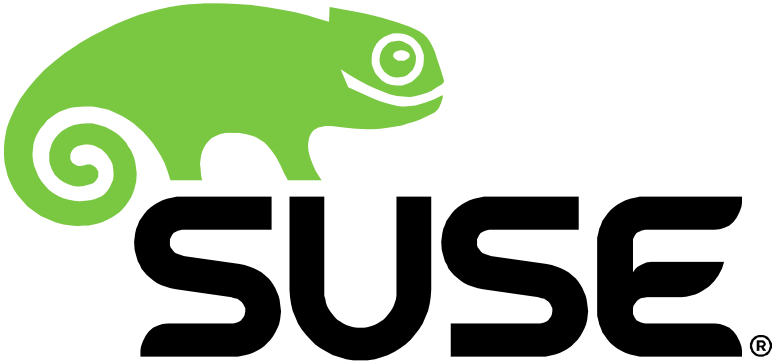
https://cloud.google.com/blog/products/gcp/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu

https://petewarden.com/2015/04/20/why-gemm-is-at-the-heart-of-deep-learning/

https://zhuanlan.zhihu.com/p/35489035

https://basicmi.github.io/AI-Chip/

http://cjc.ict.ac.cn/online/bfpub/zhm-201813141316.pdf



Corporate Headquarters
Maxfeldstrasse 5
90409 Nuremberg
Germany

+49 911 740 53 0 (Worldwide)
www.suse.com

Join us on:
www.opensuse.org

257-000020-001