Last Mile On Democratizing AI

Howard Huang, Huawei
Jianfeng Ding, Intel
Outline

➔ Diffuse The Hype
➔ Introduce Cyborg Project
➔ Intel’s Recent Effort in AI
➔ Look Into The Future
Everyone is talking about democratizing AI
But it can't be truly done without an open cloud infrastructure

- Tensorflow, CNTK, Pytorch, Caffe, MXNET, … Basically everything you can find now about major AI related open source projects
- Same goes to majority of the research papers
Interesting comparison on Blockchain and AI

<table>
<thead>
<tr>
<th>Most hyped technologies</th>
<th>Blockchain</th>
<th>AI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level of understanding on Infrastructure</td>
<td>Good Sense</td>
<td>Few Sense</td>
</tr>
<tr>
<td></td>
<td>Few work</td>
<td>Good work</td>
</tr>
</tbody>
</table>
Define a **Cloud Infrastructure For AI**

- **PaaS**
- **IaaS**
- **Hardware** (CPU, GPU, FPGA, ASIC, ...)

Cloud Infrastructure

Heterogeneous Physical Resource
Asking the right question

Can we have an AI cloud infrastructure software which
(1) Provides nice abstraction and management of the heterogeneous resources
(2) Is open source and driven by an open community
(3) Facilitates the e2e AI development
Outline

➔ Diffuse The Hype

➔ Introduce Cyborg Project

➔ Intel’s Recent Effort in AI

➔ Look Into The Future
Cyborg is a general management framework for accelerators

Proud OpenStack Official Project since 2017.09
(https://github.com/openstack/cyborg)
Cyborg Project Overview

- Subteams: release, driver, doc
- Active Chinese Dev wechat group (48 members) from companies like Huawei, China Mobile, Intel, Lenovo, ZTE, Tencent, Nokia, Unionpay, 99Cloud, Xilinx, Inspur, iFlyTech, UC Berkeley, UIUC, CMU
- Lots of gifs ...
Cyborg Pike Release

Legend
- Green: Pike Finished
- Red: Out-of-scope

Diagram:
- cyborg-api
- cyborg-conductor
- cyborg-agent
- cyborg-db
- vendor-a-driver
- vendor-a-acc
- vendor-b-driver
- vendor-b-acc
Cyborg Queens Release

Legend
- Pike Finished
- Queens Finished
- Out-of-scope

- cyborg-api
- cyborg-conductor
- cyborg-agent
- cyborg-db (resource provider)
- SPDK driver
  - NVMe SSD
- cyborg-generic-driver
  - vendor-acc-test
- Intel FPGA driver
  - Intel FPGA
Cyborg Rocky Release Planning (Kubernetes)

- Align Cyborg data model with DPI before 1.13 release
- Cyborg DPI Plugin reledy when DPI GA
- Consider the possibility of a CRD Acc controller
- Could be utilized by Kubeflow
Cyborg is forming
Outline

➔ Diffuse The Hype
➔ Introduce Cyborg Project
➔ Intel's Recent Effort in AI
➔ Look Into The Future
**Intel AI Compute Continuum**

**Cloud/Data Center**
Large scale data centers such as public cloud or comms service providers, gov’t and academia, large enterprise IT

**Edge**
Small scale data centers, small business IT infrastructure, to few on-premise server racks and workstations

**Devices**
User-touch endpoint devices with lower power requirements such as laptops, tablets, smart home devices, drones
Intel® Xeon® Scalable Processors

企业 IT
四路 HammerDB OLTP 数据库

E7-4870 使用 4 年的系统
E7-8890 v4 上一代
5 倍
8180

混合云基础设施
虚拟化应用

E5-2690 使用 4 年的系统
E5-2699 v4 上一代
4.2 倍
8180

通信服务提供商
DPDK L3 转发

E5-2658 使用 4 年的系统
E5-2658 v4 上一代
2.7 倍
6152

技术计算
LINPACK 英特尔® 分发版

E5-2690 使用 4 年的系统
E5-2699 v4 上一代
8.2 倍
8180

2.2 倍
Intel® Deep Learning Inference Accelerators

Intel® FPGA
Custom deep learning inference

Intel® Movidius™ VPU
Low power computer vision & inference

Intel® Mobileye EyeQ
Autonomous driving inference platform

Intel® GNA IP
Ultra low power speech & audio inference

Integrated graphics
Built-in deep learning inference

Data Center
Small scale clusters to a few on-premise server & workstations

Edge
User-touch end-devices typically with lower power requirements
Intel® Omni-Path Architecture

- Improved cost, power, and density
- Increased node bandwidth
- Reduced communication latency

- High MPI message rate
- Low latency scalable architecture
- Complementary storage traffic support

- Very low end-to-end latency
- Efficient transient error detection & correction
- Improved quality-of-service delivery
- Support extreme scalability, millions of nodes
Intel® Next Generation High Performance Storage

Intel® Optane™ Technology

ONCE-IN-A-GENERATION INNOVATION

This is Intel® Optane™ technology. After 25 years, the first new major breakthrough in storage & memory is here.
Intel® Optimized AI Libraries

Intel distribution for python

Advancing Python* Performance Closer to Native Speeds

software.intel.com/intel-distribution-for-python

Intel® Data Analytics Acceleration Library (Intel® DAAL)

High Performance Machine Learning and Data Analytics Library

Building blocks for all data analytics stages, including data preparation, data mining & machine learning
Intel® Optimized AI Libraries (continue)

Intel® MKL-dnN

Intel’s Open-Source Math Kernel Library for Deep Neural Networks

For developers of deep learning frameworks featuring optimized performance on

Intel® cLDNN - Intel GPU DL acceleration middleware

Compute Library for Deep Neural Networks on Intel Integrated Graphics

https://01.org/cldnn
Deep Learning Frameworks

Many Popular DL Frameworks are now optimized for CPU

Frameworks optimized by Intel

More under optimization: Caffe2, PyTorch, CNTK, PaddlePaddle, and more...
Outline

➔ Diffuse The Hype

➔ Introduce Cyborg Project

➔ Intel’s Recent Effort in AI

➔ Look Into The Future
Future #1: AI Native Open Infrastructure

Infrastructure As Code - Programmable Framework

Infrastructure As Model - Learnable Framework
Future #1: AI Native Open Infrastructure

Where is the scheduler?

Software 1.0: Write a complex fixed scheduler
Future #1: AI Native Open Infrastructure

Where is the scheduler?

Software 2.0: Write a model that learns the most appropriate scheduling functionality
Future #2: Truly Disruptive AI Technologies

Causal Model
Neural Network (Application)

Evolution Strategy
Hyperparameter Tuning (Application)

Brain Inspired Circuit
Neuromorphic Computing (Infrastructure)
Future #2: Truly Disruptive AI Technologies
Future #3: AI Diaspora

- Full Implementation (API+Sched+DB+Agent+Driver)
- Agent + Driver
- Only Driver
Now we can build a cradle for the westworld
Then how do we effectively manage the hosts?

Come checkout the talk for K8S Policy WG at 309A (16:40)
Thank You
Backup