Last Mile On Democratizing AI

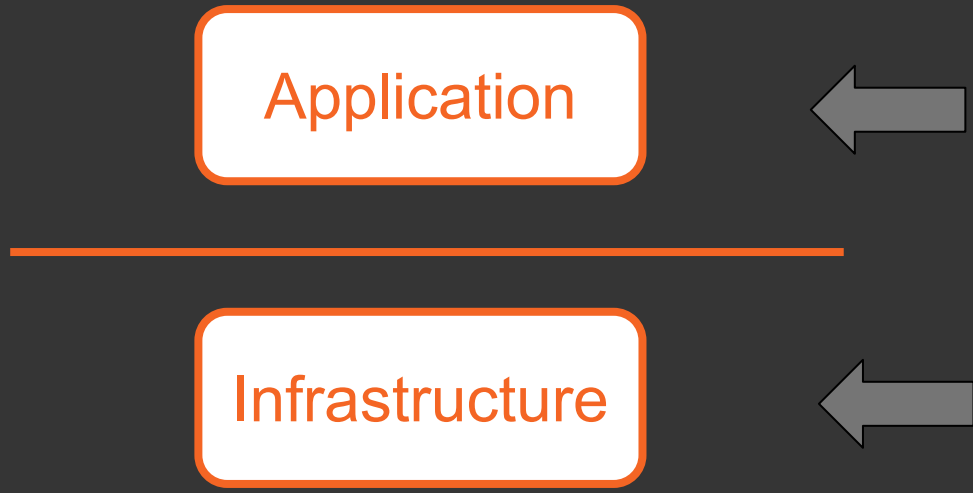
Howard Huang, Huawei
Jianfeng Ding, Intel

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



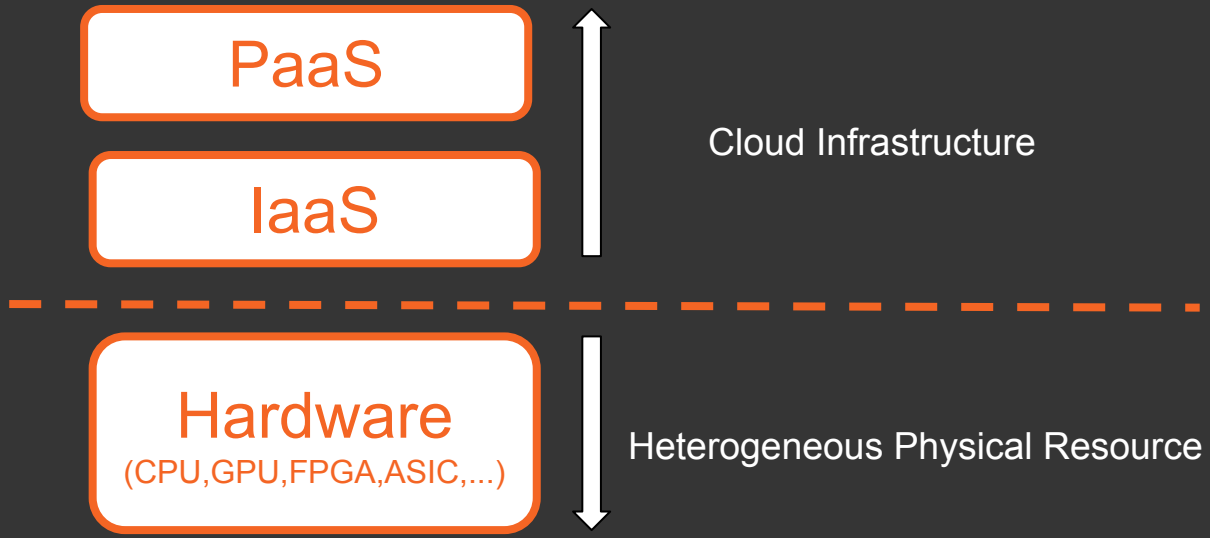
- Tensorflow, CNTK, Pytorch, Caffe, MXNET, … Basically everything you can find now about major AI related open source projects
- Same goes to majority of the research papers

???

Interesting comparison on Blockchain and AI

Most hyped technologies	Blockchain	AI
Level of understanding on Infrastructure	Good Sense Few work	Few Sense Good work

Define a Cloud Infrastructure For AI



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) Fascilates the e2e AI development

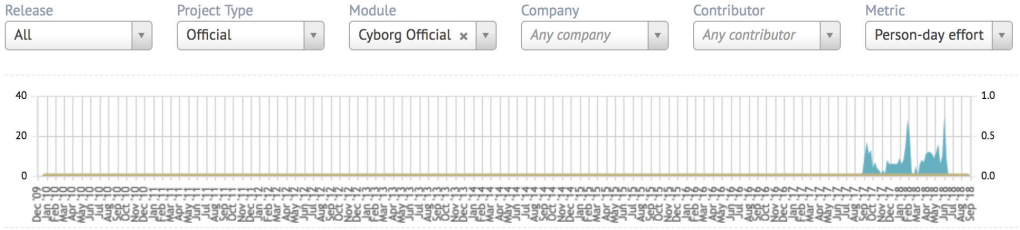
Outline

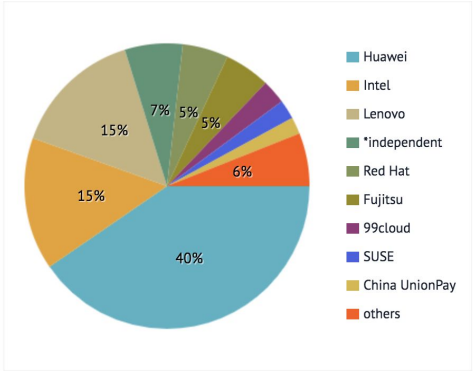
- ➔ Diffuse The Hype
- ➔ Introduce Cyborg Project
- ➔ Intel's Recent Effort in AI
- ➔ Look Into The Future

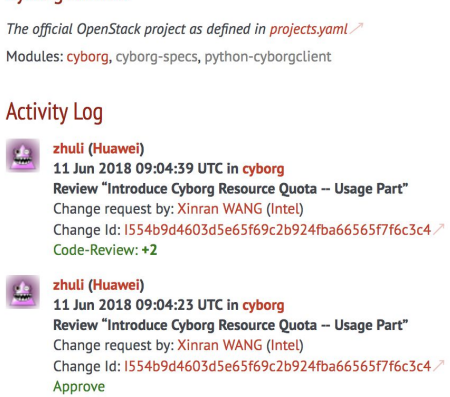
Cyborg is a general management framework for accelerators

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

Cyborg Project Overview

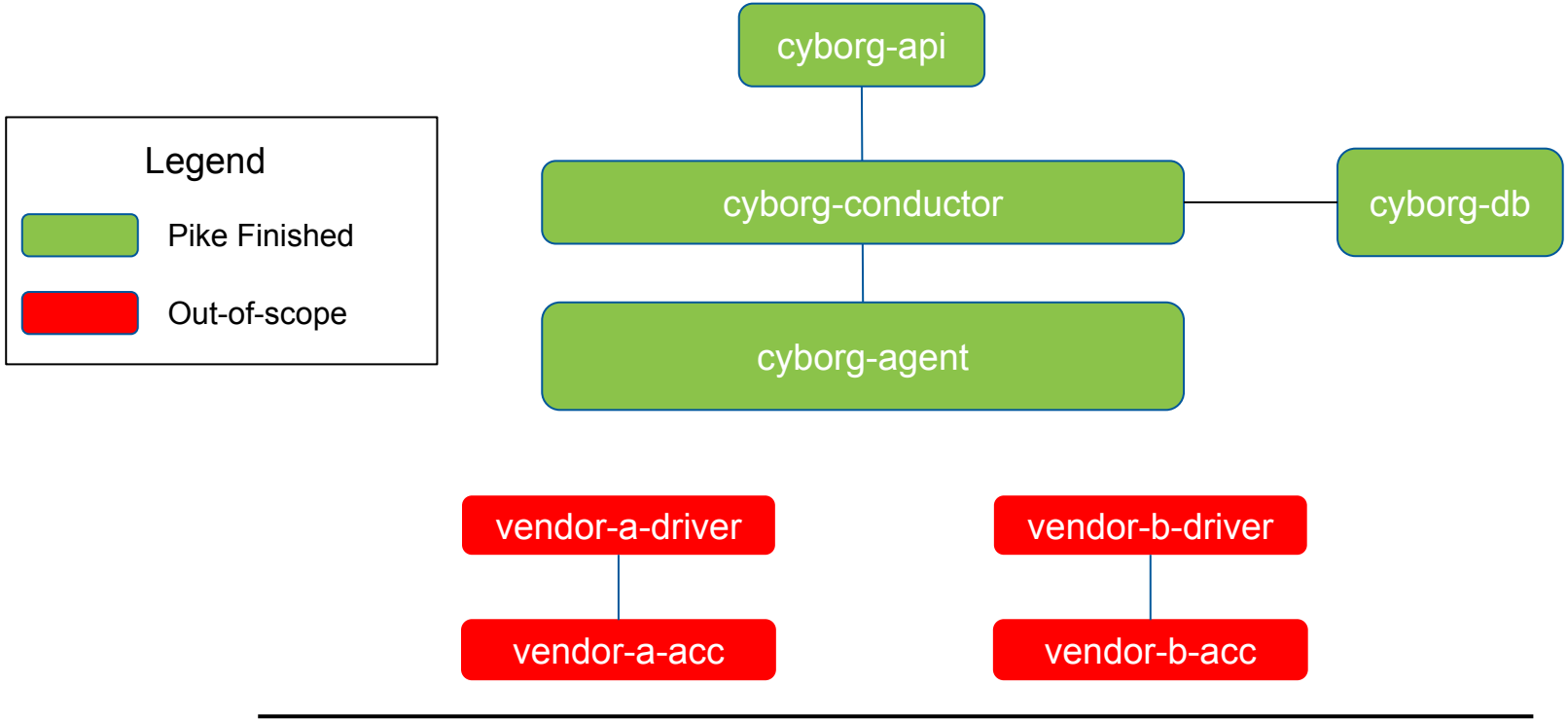




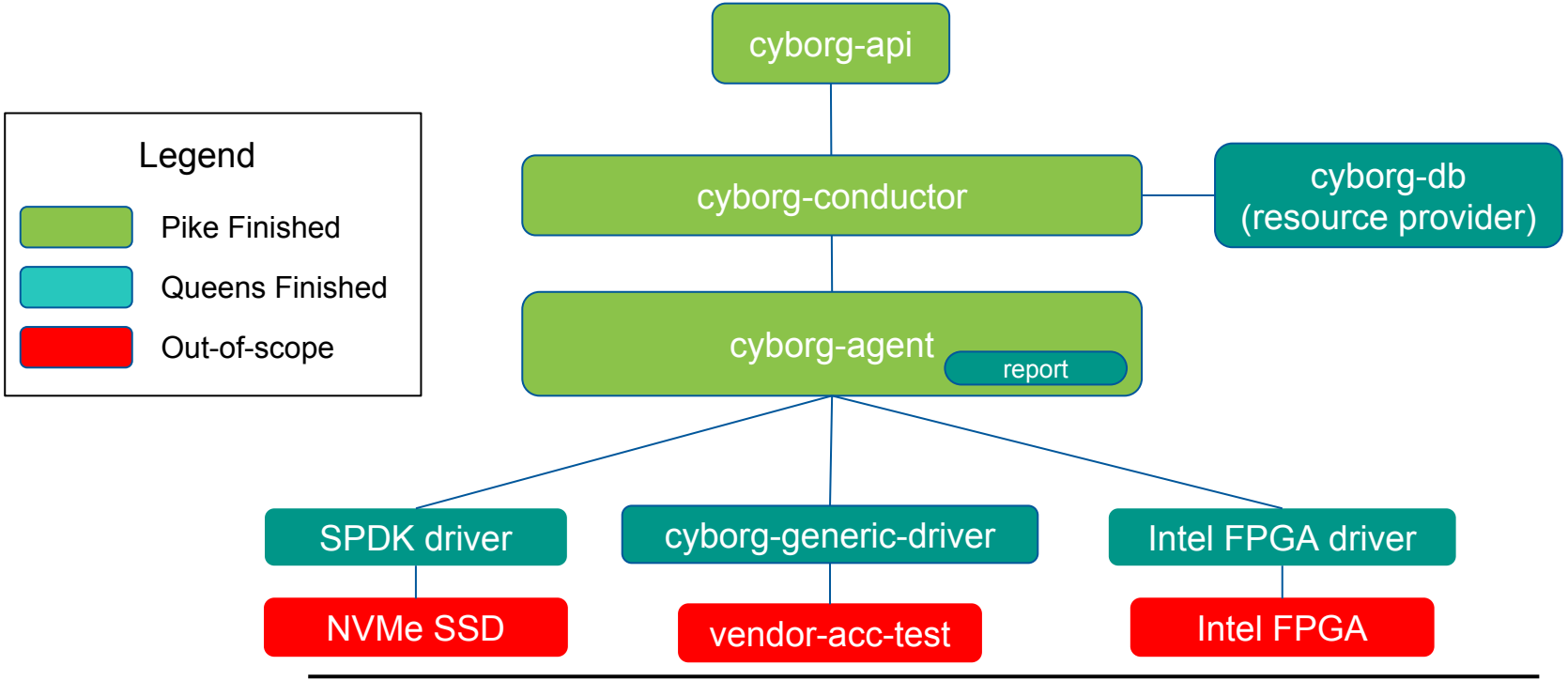


- Subteams: release, driver, doc
- Active Chinese Dev wechat group (48 members) from companies like Huawei, China Mobile, Intel, Lenovo, ZTE, Tencent, Nokia, Unionpay, 99Cloud, Xilinx, Inspur, iFlyTech, UC Berkeley, UIUC, CMU
- Lots of gifs ...

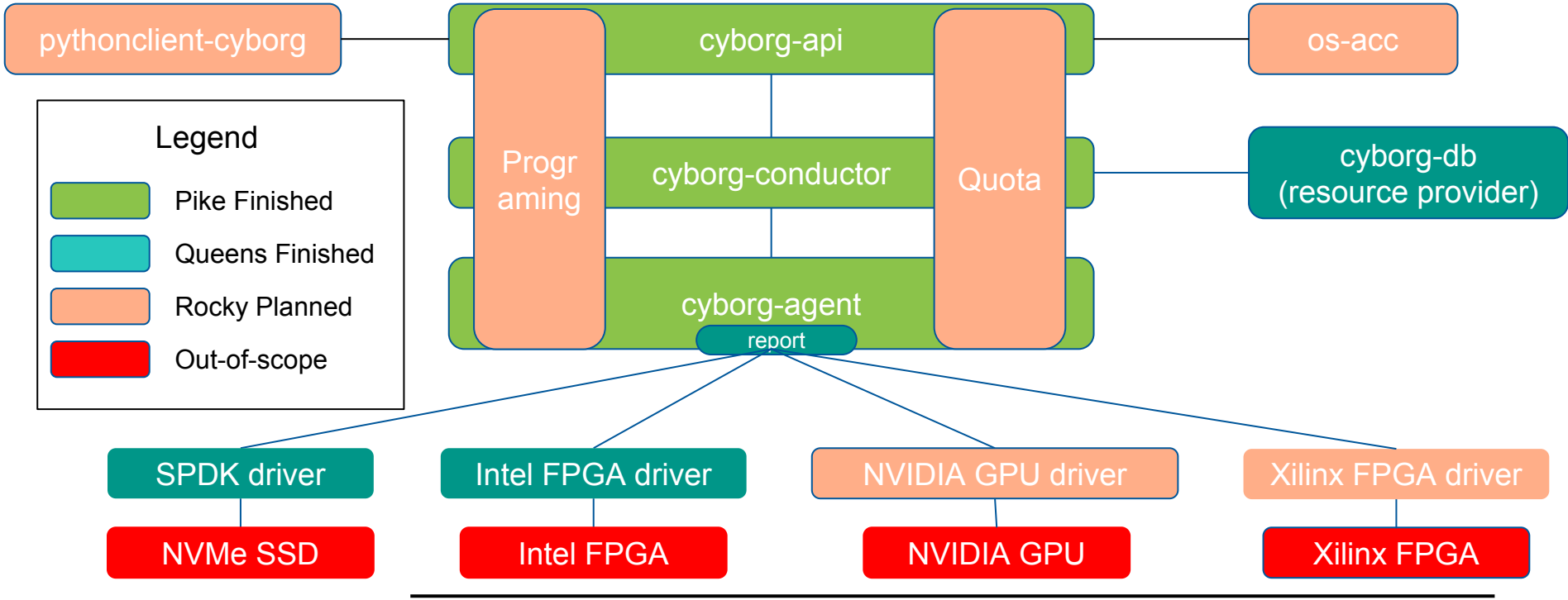
Cyborg Pike Release



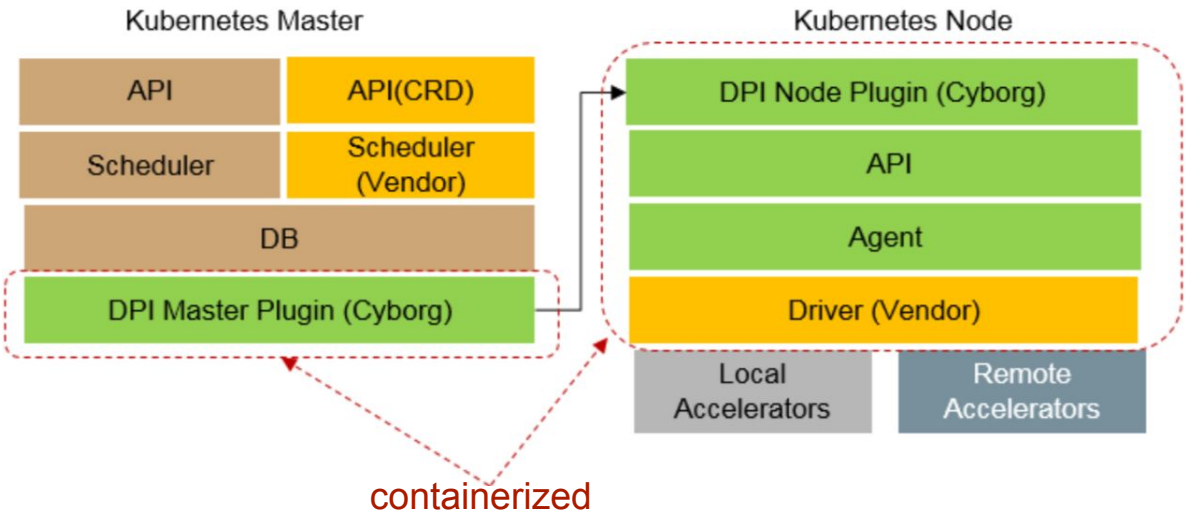
Cyborg Queens Release



Cyborg Rocky Release Planning (OpenStack)



Cyborg Rocky Release Planning (Kubernetes)



- Align Cyborg data model with DPI before 1.13 release
- Cyborg DPI Plugin reready 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 AI PORTFOLIO

SOLUTIONS

Data Scientists | Technical Services | Reference Solutions

PLATFORMS

Intel® AI DevCloud

Intel® Deep Learning System‡

intel Saffron™ REASONING

TOOLS

Intel® Deep Learning Studio‡

Intel® Deep Learning Deployment Toolkit†

Intel® Computer Vision SDK†

Intel® Movidius™ Software Development Kit (SDK)

FRAMEWORKS

TensorFlow | Caffe | mxnet | BigDL Spark | Caffe2 | PYTORCH | CNTK | PaddlePaddle

LIBRARIES

Intel® MKL/MKL-DNN, clDNN, DAAL, Intel Python Distribution, etc.

DIRECT OPTIMIZATION

Intel® nGraph™ Compiler*

CPU Transformer† | NNP Transformer‡ | Other

TECHNOLOGY

XEON | NERVANA | ARRIA 10 | ATOM | CORE i7 | Iris Graphics | Movidius

END-TO-END COMPUTE

SYSTEMS & COMPONENTS

Intel® Xeon® Scalable Processors

企业 IT

四路 HammerDB OLTP 数据库

E7-4870
使用 4 年的系统

5 倍

E7-8890 v4
上一代

1.5 倍

intel XEON PLATINUM inside
8180

混合云基础设施

虚拟化应用

E5-2690
使用 4 年的系统

4.2 倍

E5-2699 v4
上一代

1.5 倍

intel XEON PLATINUM inside
8180

通信服务提供商

DPDK L3 转发

E5-2658
使用 4 年的系统

2.7 倍

E5-2658 v4
上一代

1.7 倍

intel XEON GOLD inside
6152

技术计算

LINPACK 英特尔® 分发版

E5-2690
使用 4 年的系统

8.2 倍

E5-2699 v4
上一代

2.2 倍

intel XEON PLATINUM inside
8180

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[1]

Ultra low power speech & audio inference

Integrated graphics

Built-in deep learning inference

Data Center

Edge *Small scale clusters to a few on-premise server & workstations*

Device *User-touch end-devices typically with lower power requirements*

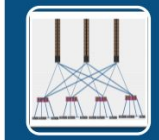
Intel® Omni-Path Architecture

CPU/Fabric Integration

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

Optimized Host Implementation

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



Enhanced Fabric Architecture

- 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

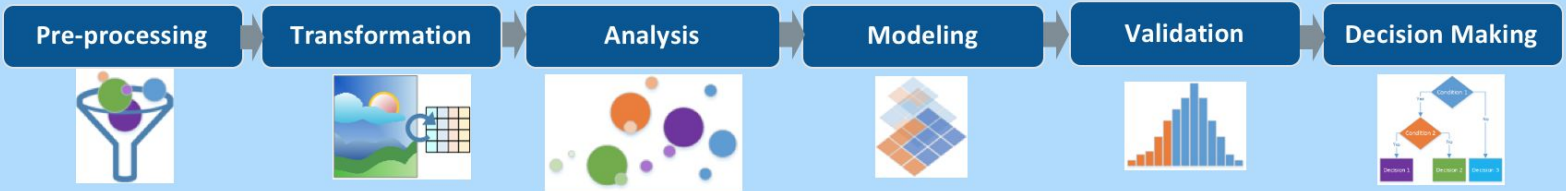
Advancing Python* Performance Closer to Native Speeds

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

github.com/01org/mkl-dnn

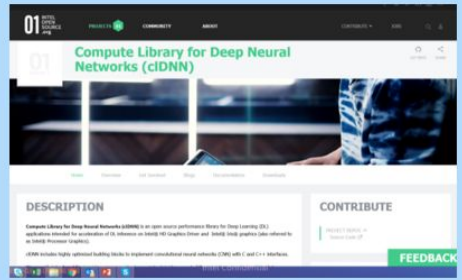
Intel's Open-Source Math Kernel Library for Deep Neural Networks

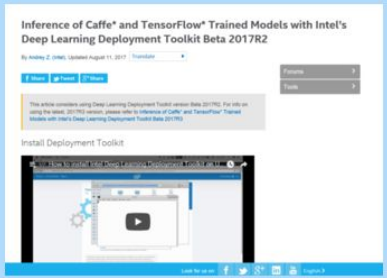
For developers of deep learning frameworks featuring optimized performance on

Intel® clDNN - *Intel GPU DL acceleration middleware*

Compute Library for Deep Neural Networks on Intel Integrated Graphics

https://01.org/cldnn



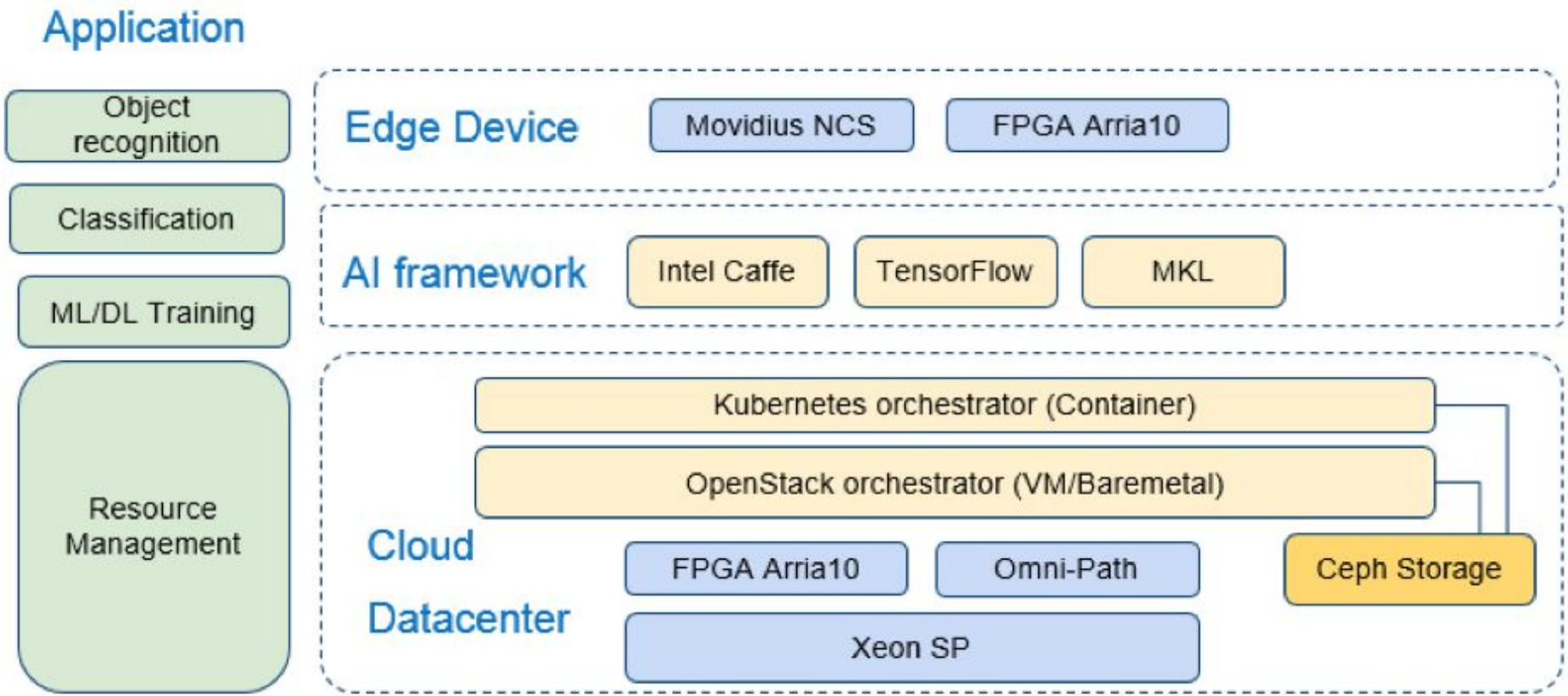


Deep Learning Frameworks

Many Popular DL Frameworks are now optimized for CPU

More under optimization: Caffe2 PYTORCH Microsoft CNTK PaddlePaddle and more...

Intel® Open AI Cloud Reference Architecture



Outline

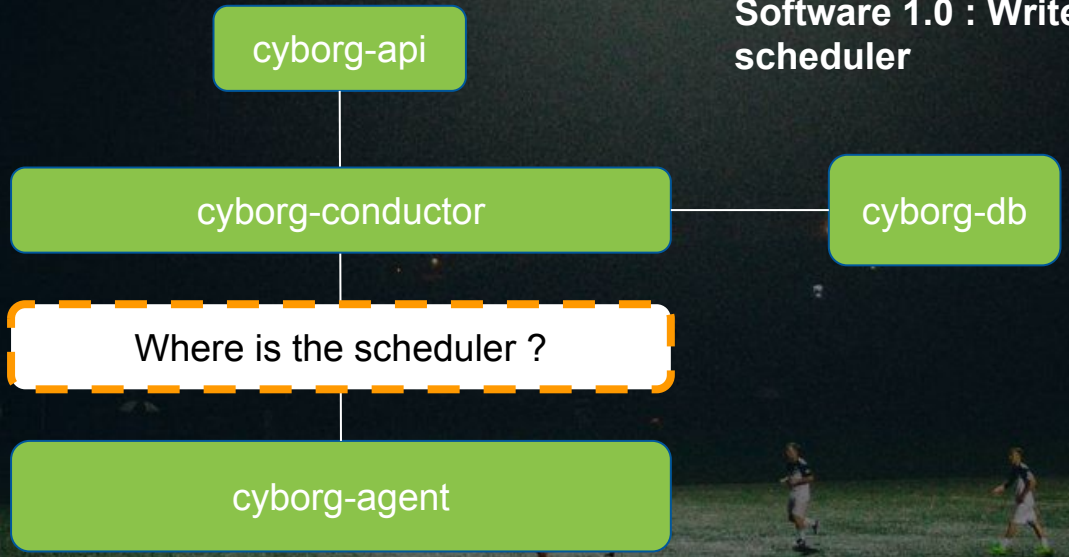
- ➔ Diffuse The Hype
- ➔ Introduce Cyborg Project
- ➔ Intel's Recent Effort in AI
- ➔ Look Into The Future

Future #1 : AI Native Open Infrastructure

Infrastructure As Code - Programmable Framework

Infrastructure As Model - Learnable Framework

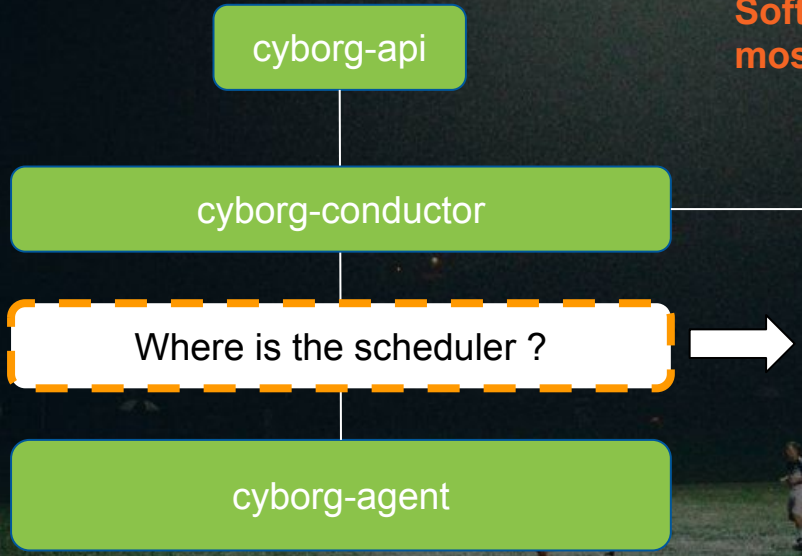
Future #1 : AI Native Open Infrastructure

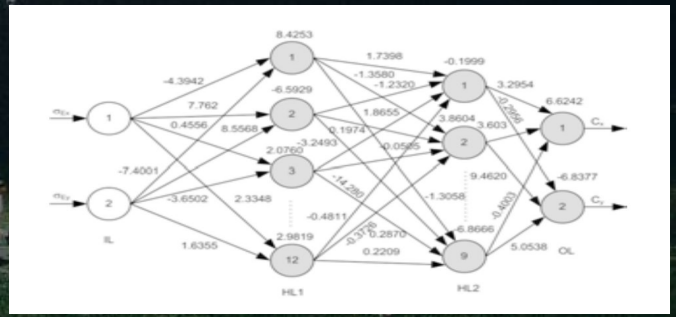


Software 1.0 : Write a complex fixed scheduler

Future #1 : AI Native Open Infrastructure

Software 2.0 : Write a model that learns the most appropriate scheduling functionality





Future #2 : Truly Disruptive AI Technologies

Causal Model

Neural Network
(**Application**)

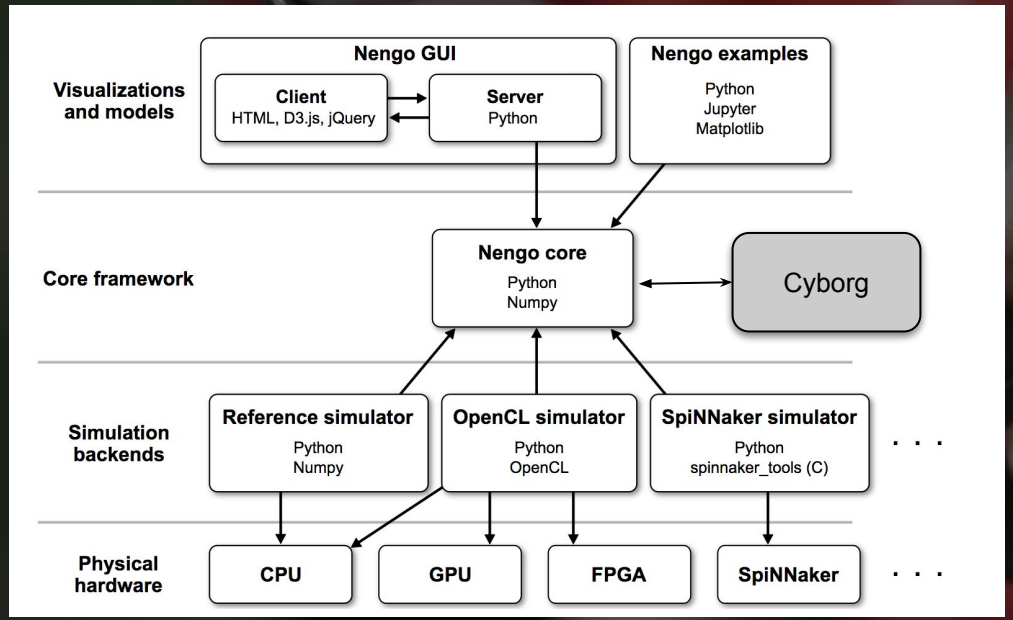
Evolution Strategy

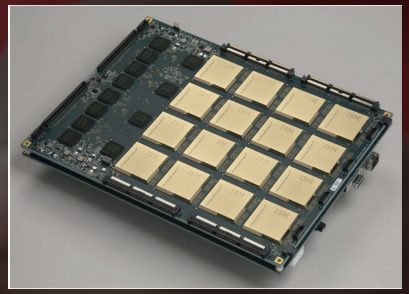
Hyperparameter Tuning
(**Application**)

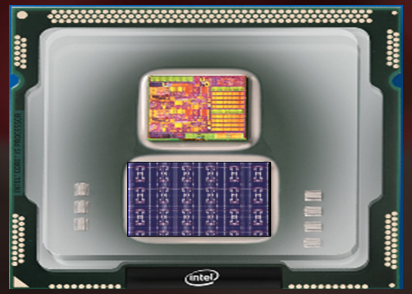
Brain Inspired Circuit

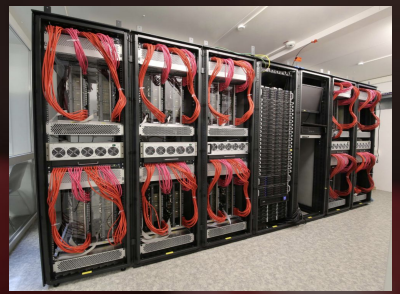
Neuromorphic Computing
(**Infrastructure**)

Future #2 : Truly Disruptive AI Technologies









Future #3 : AI Diaspora

Full Implementation
(API+Sched+DB+Agent+Driver)

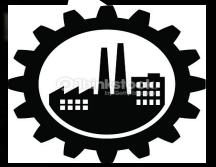
Cyborg
(Public Cloud)

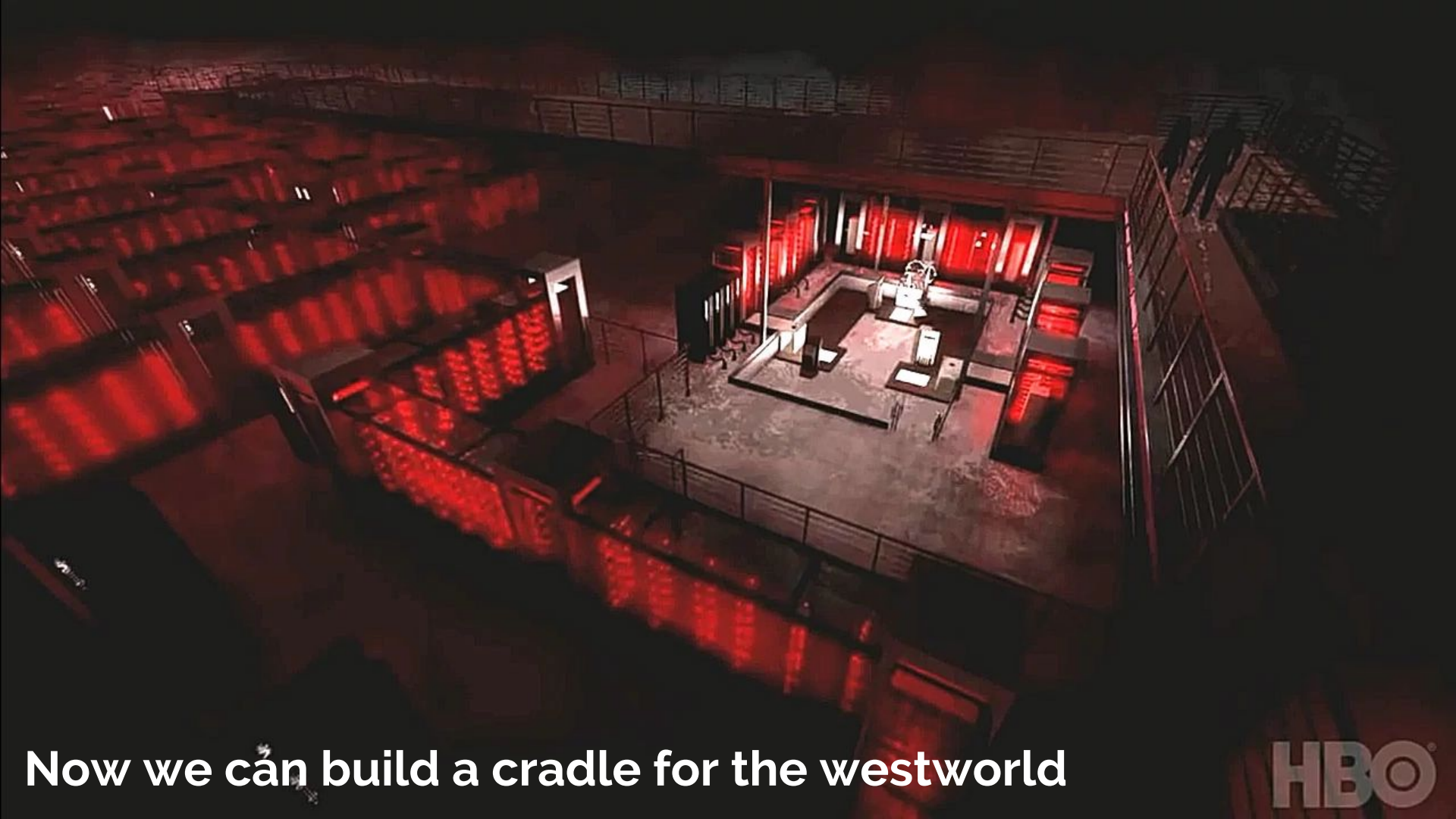
Cyborg
(Private Cloud)

Agent + Driver

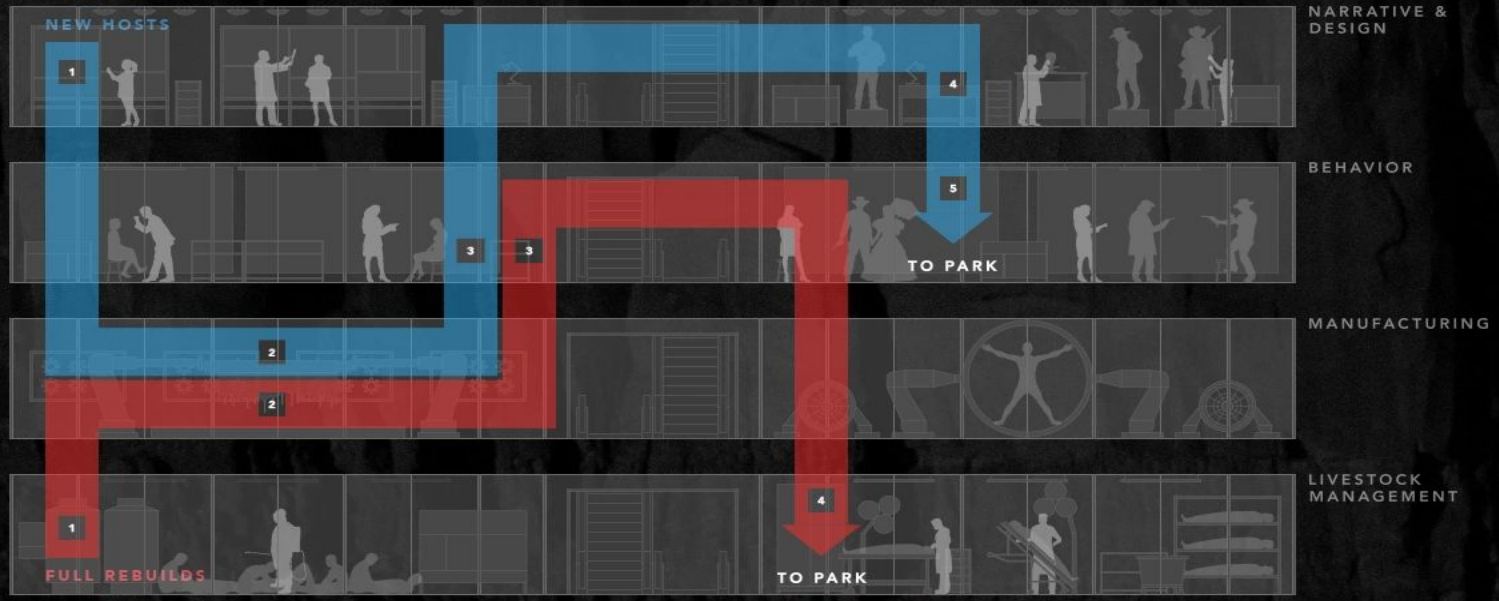
Cyborg
(Edge Cloud)

Only Driver




Now we cản build a cradle for the westworld
HBO

Then how do we effectively manage the hosts ?
Come checkout the talk for K8S Policy WG at 309A (16:40)



NEW HOSTS

1. DEVELOP THE CHARACTERS AND WORK WITH DESIGN TO CONCEPTUALIZE PHYSICAL LOOK.
2. PRINT HOST BODIES AND ASSEMBLE ACCORDING TO DESIGN SPECS.
3. ONCE BASIC CONTROL UNIT BRAINS ARE INSTALLED, CALIBRATE BASIC MOTOR FUNCTIONS AND COGNITION.
4. REVIEW HOSTS FOR QUALITY CONTROL, APPLY COSMETIC TOUCHES AND COSTUME. FINE TUNE BACKSTORIES AND STORYLINES.
5. UPLOAD FINAL PERSONALITIES, CORNERSTONES AND DRIVES. CALIBRATE CHARACTER-SPECIFIC MOTIONS AND NUANCED COGNITION.

FULL REBUILDS

1. REMOVE CONTROL UNIT FROM DAMAGED HOST REMAINS.
2. PRINT HOST BODIES AND ASSEMBLE ACCORDING TO ESTABLISHED DESIGN SPECS.
3. CALIBRATE BASIC COGNITION AND UPLOAD CURRENT CHARACTER CONFIGURATIONS.
4. REVIEW HOSTS FOR QUALITY CONTROL, APPLY FINAL COSMETIC TOUCHES AND COSTUME ACCORDING TO ESTABLISHED CHARACTER.

Thank You

Backup