OpenStack Berlin Summit, Nov. 2018

What's Behind 8K Video Streaming on Cloud Edge?

Jinghua Gao, Zhi Chang (Staff Researcher, Lenovo Research)
Shaohe Feng (Senior Cloud Engineer, Intel)
2018-11-15

Agenda

1. 8K Video Background

New Buzzword

LG

Samsung

First commercially ready 8K TV

Sharp

TCL

Aug 2012

ITU approved NHK's 8K SHV standard

Jan 2013

Sharp announced first 8K TV in CES

Aug 2016

NHK kicked off the first 8K satellite broadcasts

Mar 2018

Alibaba released 8K video cloud solution.

Sep 2018

IFA CONSUMER ELECTRONICS UNLIMITED

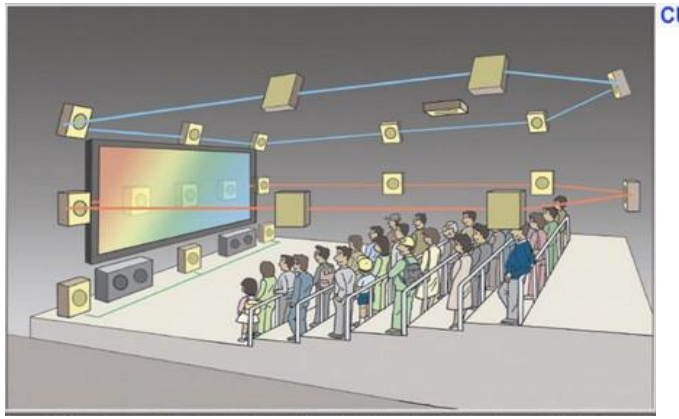
Samsung, LG, and Sony all displayed 8K TVs in IFA consumer technology show.

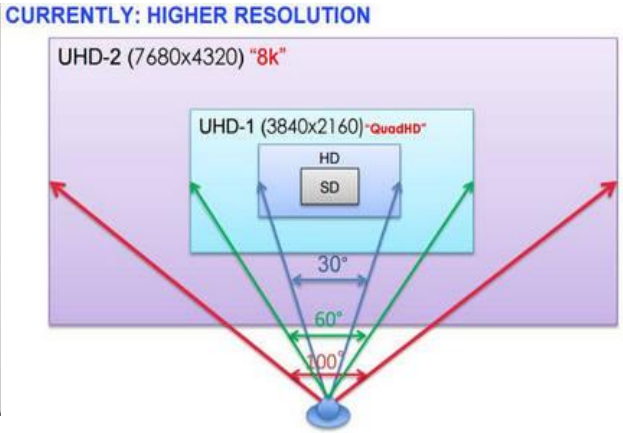
July & Aug 2020

Tokyo Olympics will shot and broadcast in 8K by NHK

8K Video Overview

	HD	4K	8K
Resolution	1920*1080	3840*2160	7680*4320
Audio Channel	5.1	5.1	22.2
Viewing Angle	30	60	100
Main Coding Format	MPEG, H.264, H.265, VP9		
Network bps	20Mbps	40-50Mbps	120-150Mbps





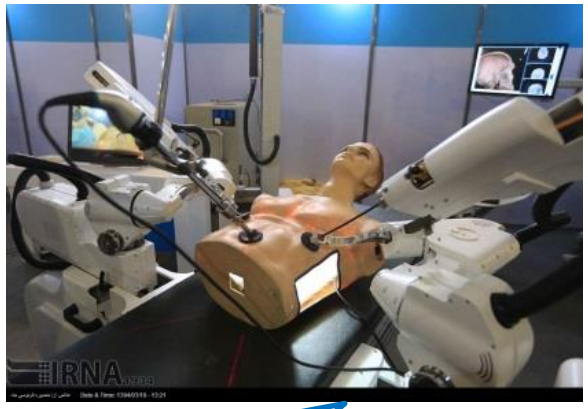
Usage Scenarios

- 8K offers **stronger sense of presence and realness**.
  - Commercial TVs
  - Immersive video applications, such as **panoramic video** and **virtual reality** (AR/VR)
  - Real-life applications, such as **remote healthcare** and **high-precision monitoring**

Panoramic Video

AR/VR

Remote Healthcare



High-precision monitoring

Internal structures of blood vessels,
Boundaries between cancer and normal tissues.
Tiny sutures

Challenges

More parallel computing in codec algorithm (8K: H.265, VP9）

Cameras are separated and far away from the datacenter

Need to manage devices and topology
of different separated edge nodes

Codec
Encoding
Decoding
Transcoding

Storage
…

Cameras

Analytics

Transmission

Need to manage different accelerators
to empowering the DL/ML models and video codec algorithms.

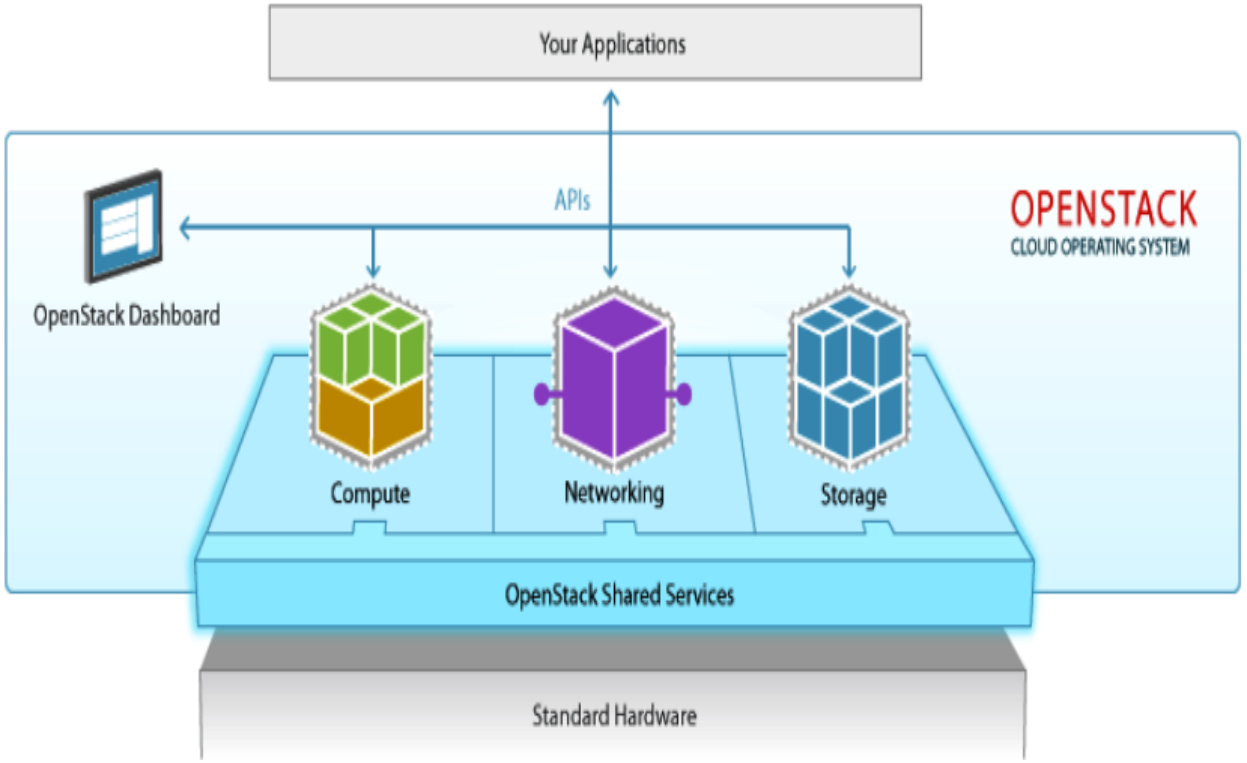
Need edge computing technologies.

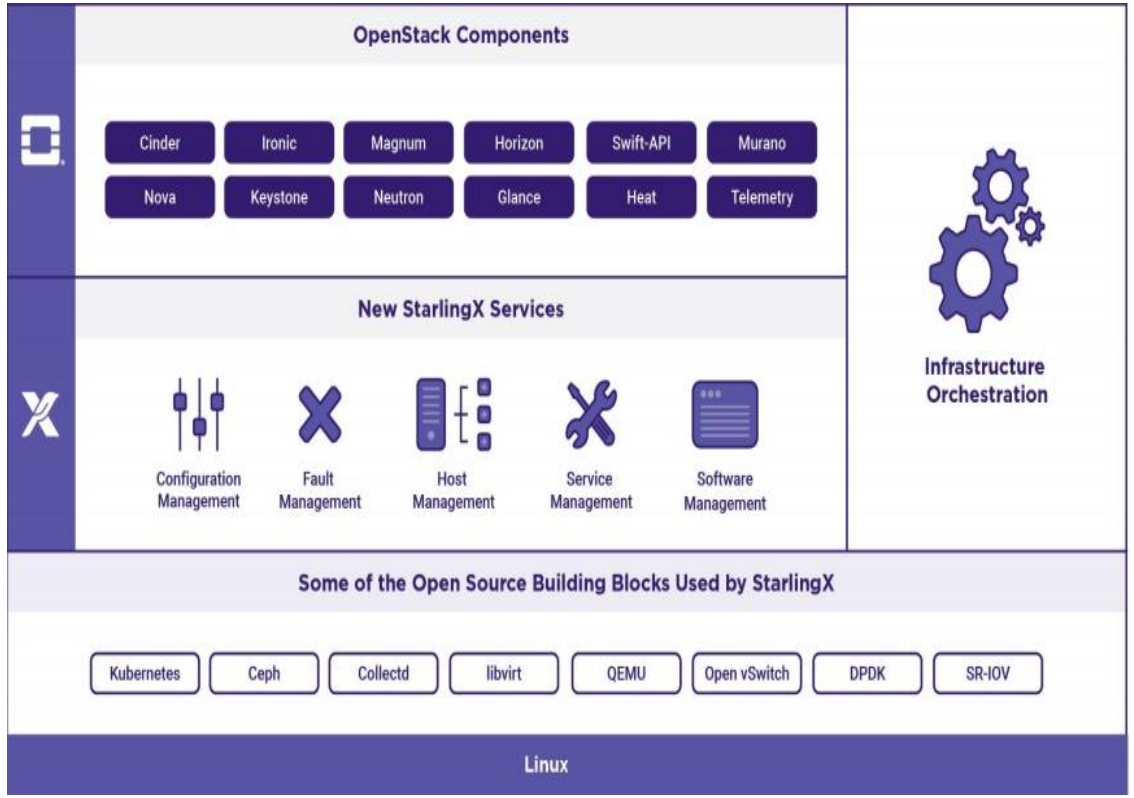
Add complexity to data processing of DL/ML model

Extreme high bandwidth and low latency（8K: 150Mbps）

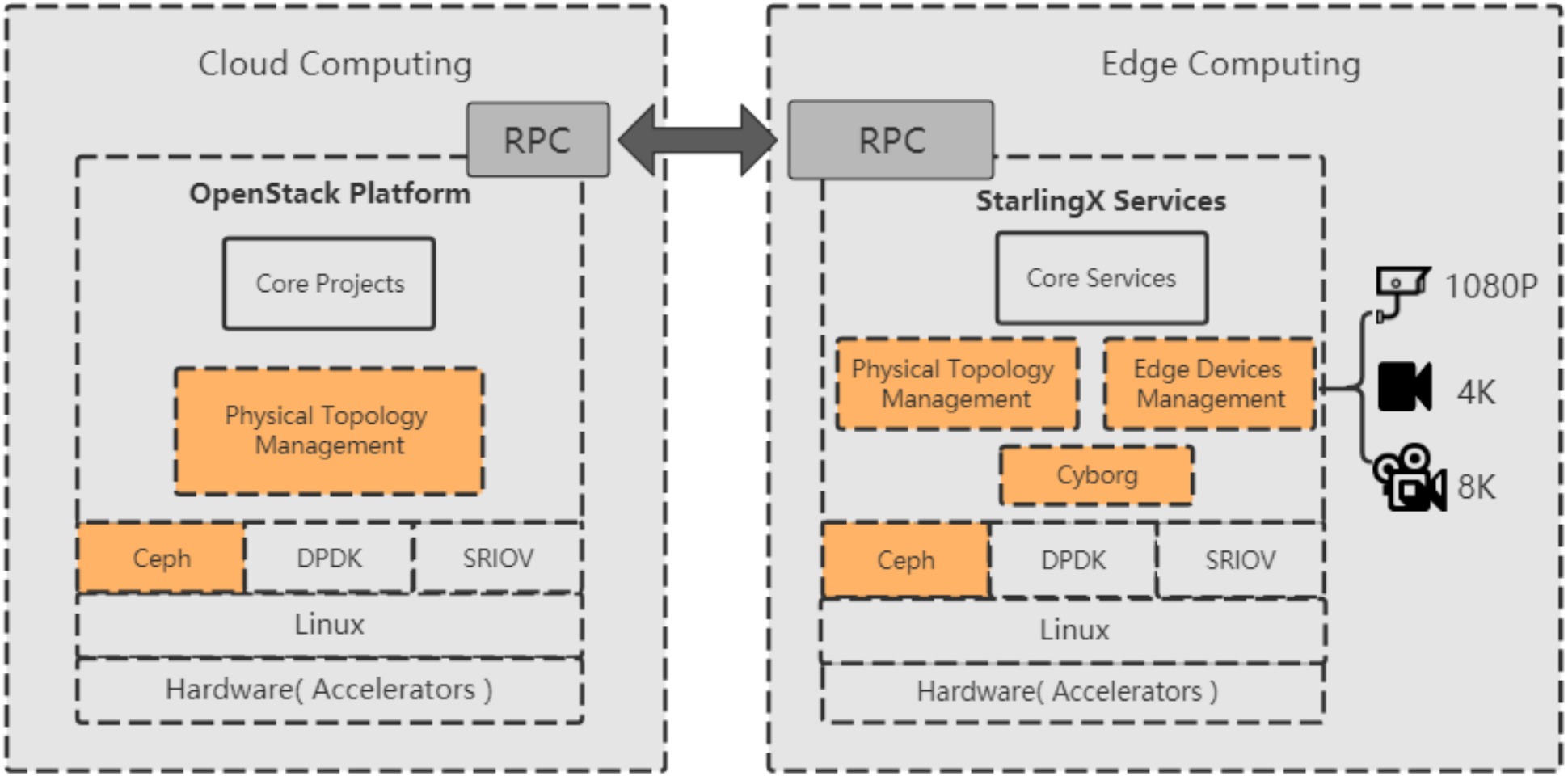
2. Edge Optimized Architecture

Cloud & Edge



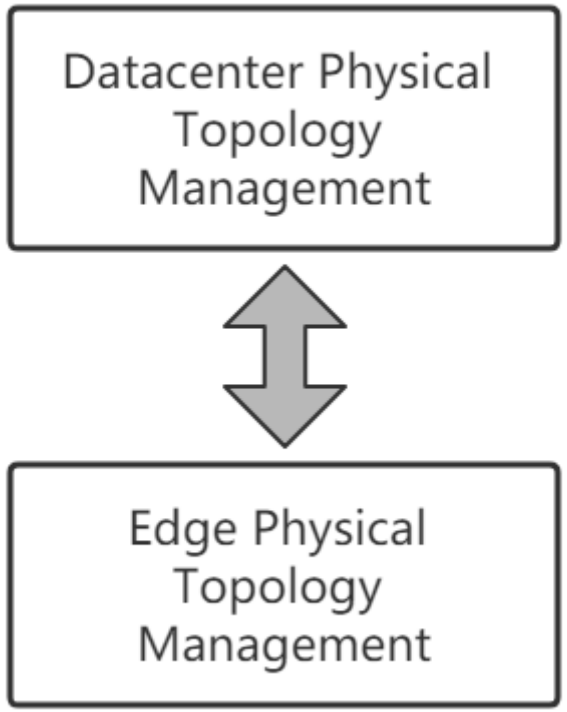


Edge Computing Solution Architecture



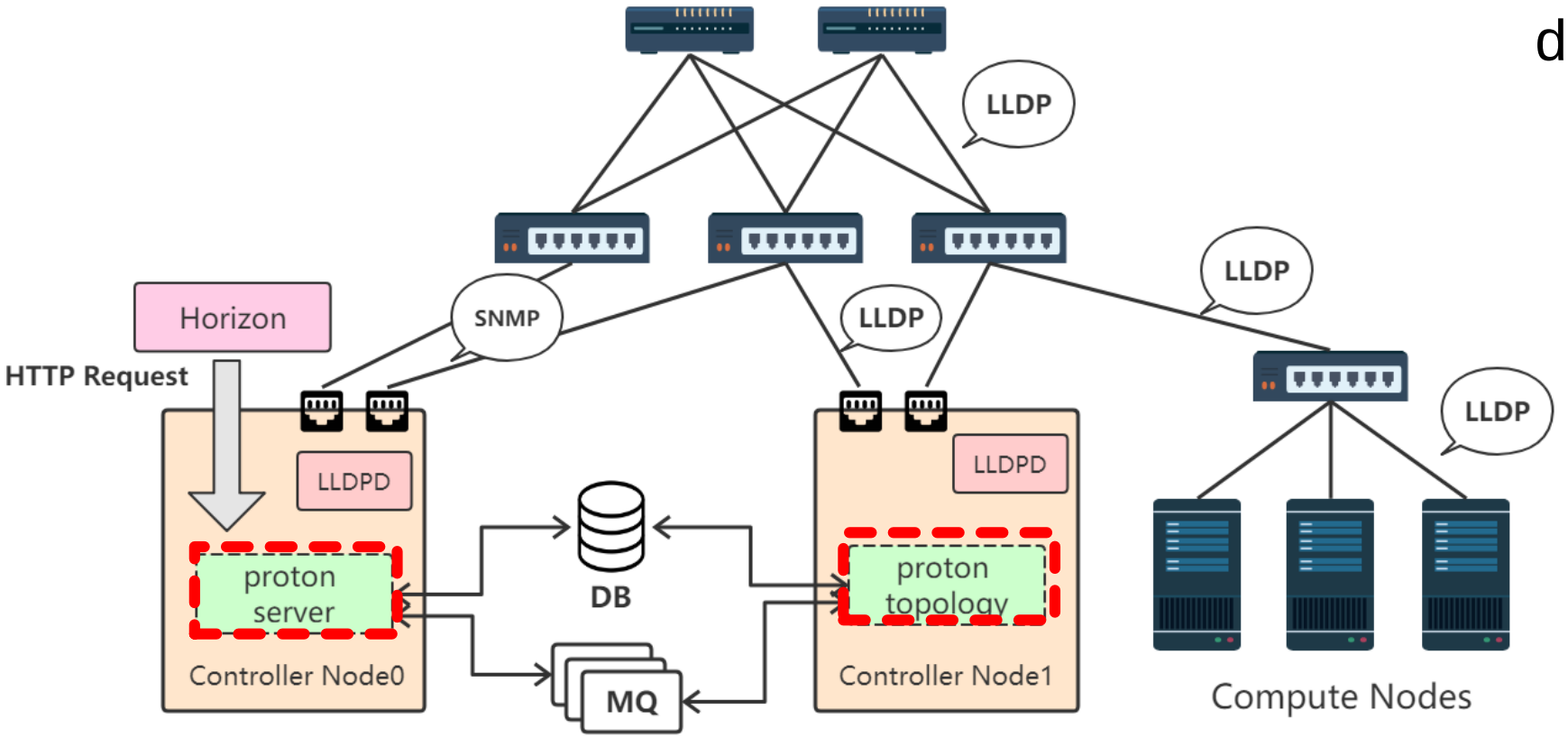
3. Key Edge Technologies

Interaction Between Datacenter and Edge



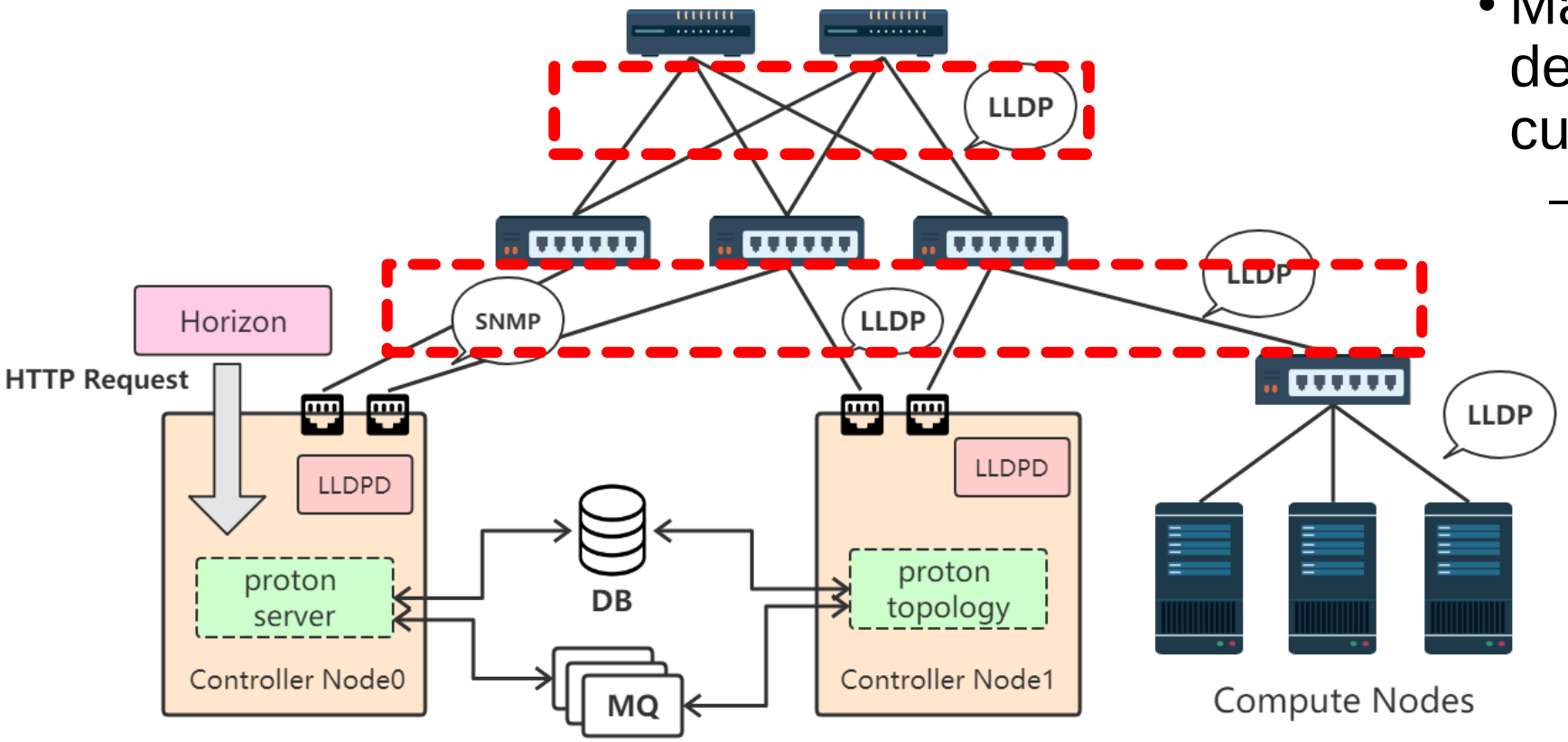
- Data Center
  - Monitor every edge cluster
  - Store edge cluster's info
- Edge Cluster
  - Register into Data Center
  - Report its status and topology info periodically

Physical Topology Management



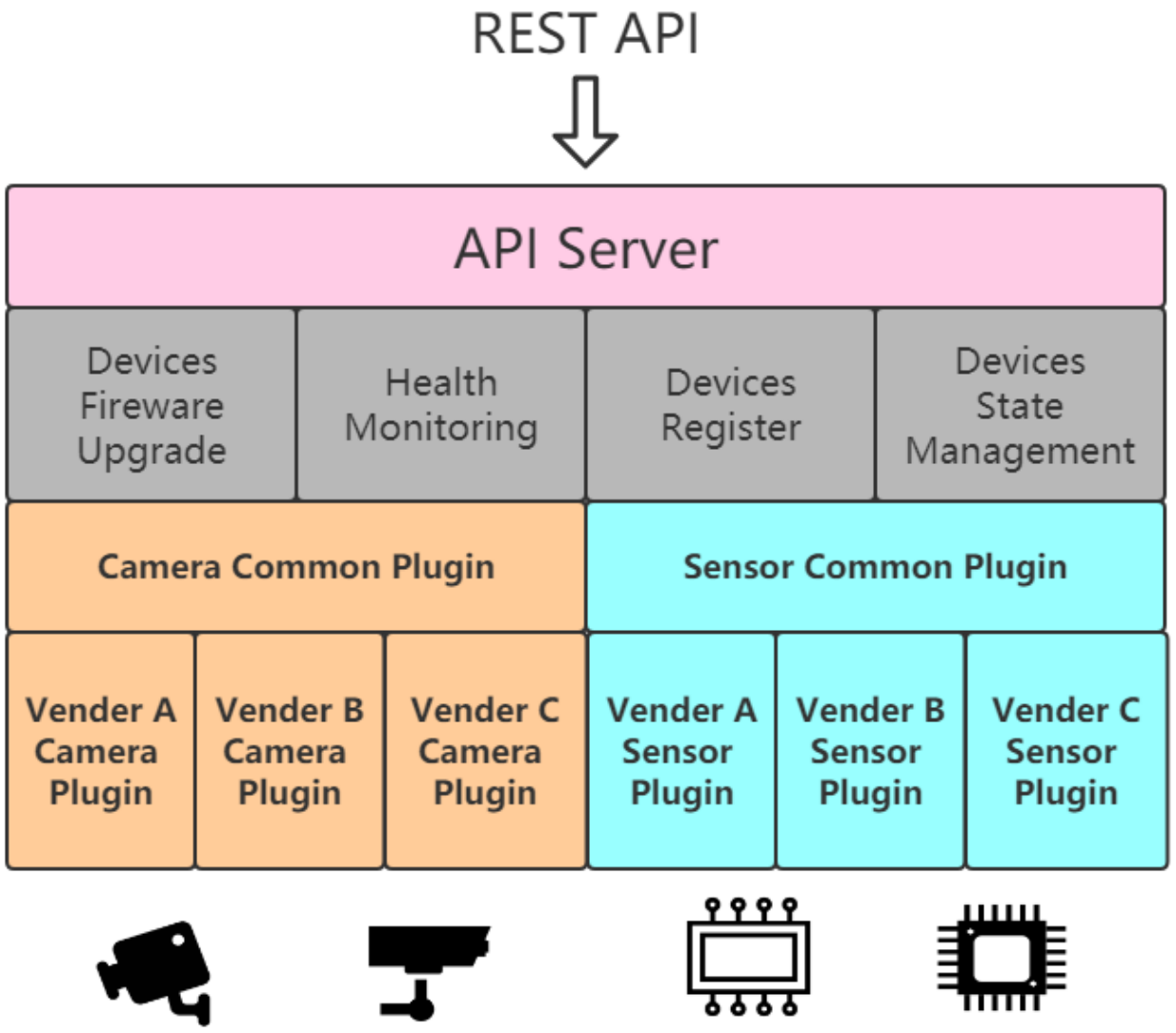
- Display and automatic discover network topology.
  - After enable LLDP protocol, Proton can find all network switches and servers automatically and the relationships between them.
  - If network topology is changed, Proton can detect the change within few minutes.

Physical Topology Management



- Manage physical network devices and display their current status.
  - Get detailed info about switches, such as software version, system name, interfaces status

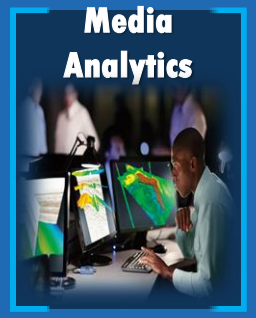
Edge Devices Management

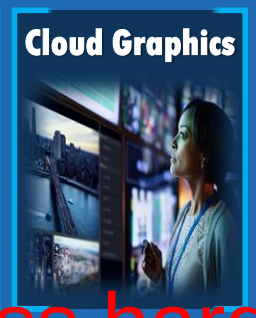


- Vender-based plugin
- Device lifecycle management
- Device monitoring

Intel is Unleashing Innovation in Visual Cloud

Visual Cloud Workloads





DECODE

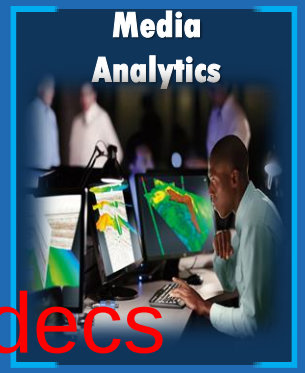
RENDER

INFERENCE

ENCODE

Four core building blocks – Five Key Workloads – Multiple Use Cases

Visual Cloud Pipeline



DECODE → INFERENCING → ENCODE

Building block selection & sequence are definitional

Access hardware accelerated codecs and programmable graphics on the latest Intel® processors.

Intel is Unleashing Innovation in Visual Cloud

Software accelerated workloads, faster time to market

Encode / Decode

- Intel Media Server Studio
- Collaboration Suite for WebRTC

Inference

- OpenVINO™ Toolkit

render

- Remote Display SDK
- OSPRay
- Embree
- OpenSWR

Scalable Software Supporting Industry Frameworks

Broad platform portfolio – scalability for power, density, quality

- Intel® Xeon® Processor
- Intel® Xeon-D Processor
- Intel® Visual Cloud Accelerator
- Intel® FPGA
- Intel® Optane Persistent Memory
- Intel® Gen Graphics
- Intel® 8th Gen Core™ Processor Radeon graphics
- Intel Movidius

Transformed Network

Workload convergence

Intel Architecture

Cloud-ready networks

VNFs

NFVi

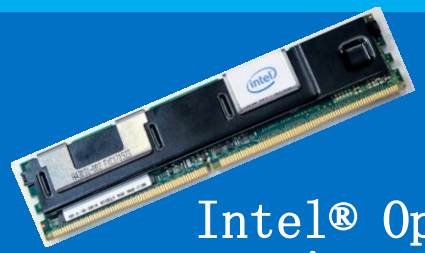
MANO

5G Infrastructure

Power 8k video processing with hardware

Xeon® Scalable Processor with Advanced Vector Extensions (AVX512)

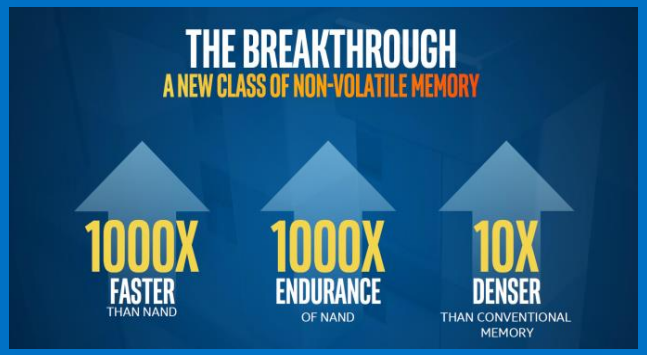
- Mesh Architecture delivers low latency and high bandwidth among cores, memory, and I/O controllers
- 1.72x video stitch
- 1.9x HEVC video encoding
- 1.5x transcoding
- 2.2x AI deep learning



Intel® Optane™ DC persistent memory

- Higher performance
  - Bandwidth, IOPS , latency
- Higher endurance

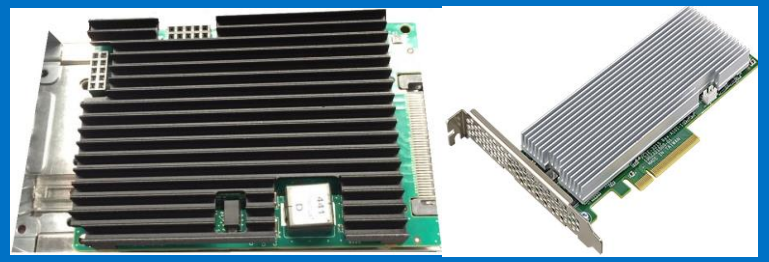
Intel® Optane™ SSD



Power 8k video processing with hardware

high-performance demands of high-throughput systems

- 10 TFLOPS of floating-point performance
- Up to 28.3 Gbps Transceiver support
- Over 2.3 Tbps bandwidth for parallel memory interfaces.
- Inference, HPC



QAT(QuickAssist) provides hardware acceleration for compute-intensive workloads

- 100 Gbs Cryptography and Public key encyption workload acceleration.
- 100 Gbs data compression
- 100 kops RSA
- 2k Decrypt

Power 8k video processing with hardware

Cascade Glacier (SmartNic) accelerates network performance

- 2x25GbE Today, 100GbE In Future
- Full Open vSwitch Acceleration On SmartNIC
- Programmability with Ease of Deployment
- Offload networking functions (NFs) from host processors

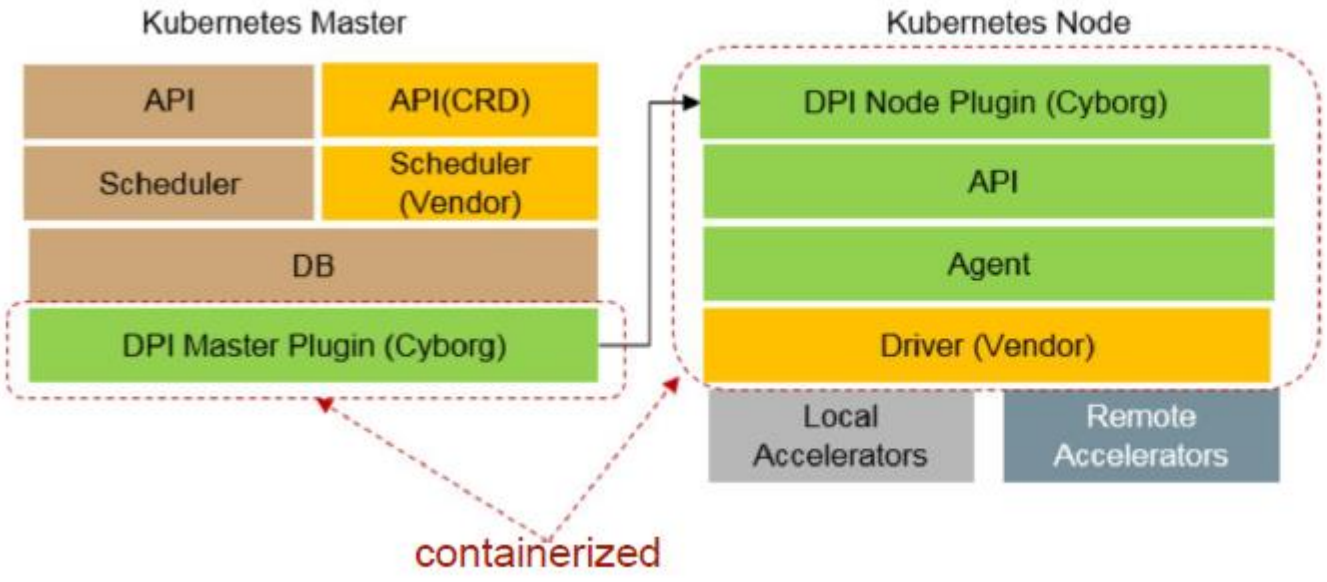
Accelerator Upstream Status

- FPGA in Cyborg https://review.openstack.org/#/c/531129/
- QAT Crypto in Ceph https://github.com/ceph/ceph/pull/15168
- QAT Compress in Ceph https://github.com/ceph/ceph/pull/19714
- Persistent Memory in Nova https://review.openstack.org/#/c/601596/
- Persistent Memory for Read Cache in Ceph https://github.com/ceph/ceph/pull/22573
- Persistent Memory for Write Cache in Ceph https://github.com/ceph/ceph/pull/20375
- AVX512 VNNI in libvirt https://www.spinics.net/linux/fedora/libvir/msg139936.html
- QAT/GPU/FPGA in Kubernets https://github.com/intel/intel-device-plugins-for-kubernetes

Accelerator Upsteam Status

- FPGA plugin in Kubernets
  - Support two mode: af and region
  - Multi pod can share FPGA by SR-IOV
- GPU plugin in Kubernets
  - Multi pod can share GPU by i915 driver
- QAT plugin in Kubernets
  - Data Plane Development Kit (DPDK) drivers must be loaded and configured
  - Intel QuickAssist Technology software for Linux must be installed and configured
  - Multi pod can share QAT by SR-IOV

Cyborg at the edge



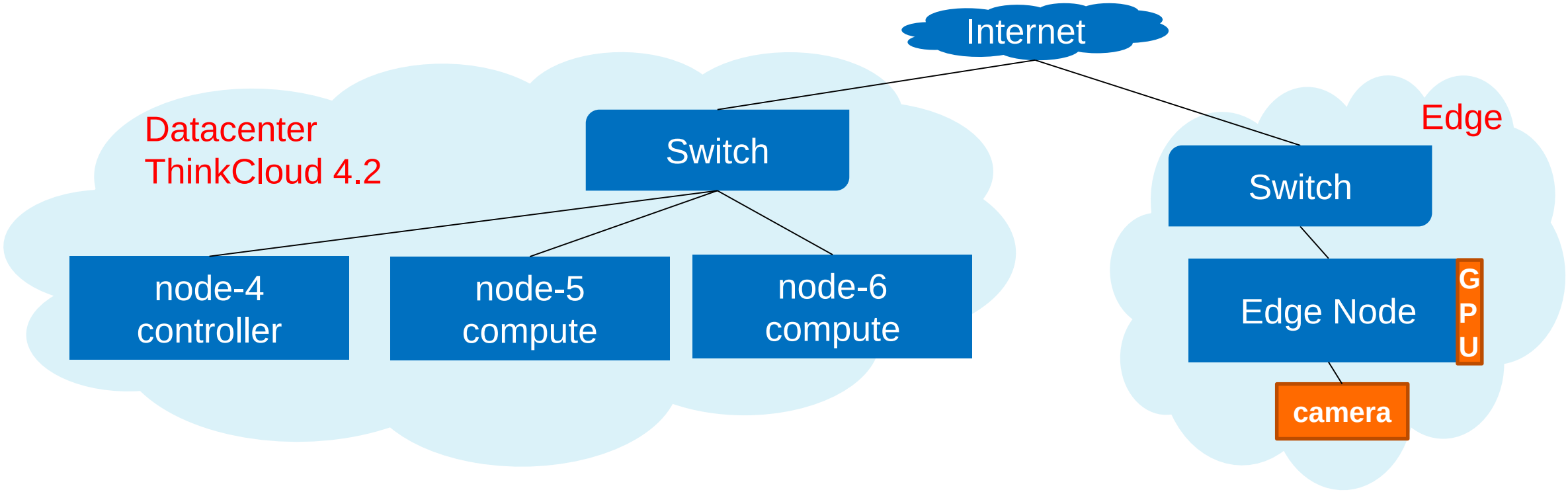
- Align Cyborg data model with DPI before 1.13 release
- Cyborg DPI Plugin releady when DPI GA
- Consider the possibility of a CRD Acc controller
- Could be utilized by Kubeflow

4. Demo

1. Management of Edge Clusters with Multiple Cameras
2. Object Detection

Management of Edge Clusters with Multiple Cameras

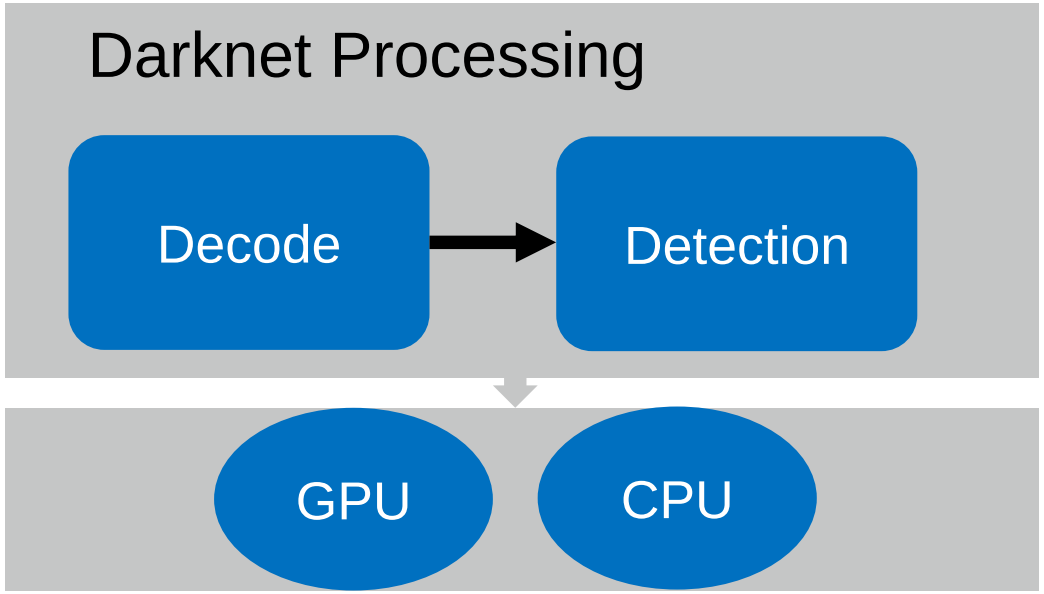
- DC： Lenovo ThinkCloud 4.2 Version
  - 3 nodes, 1 controller node and 2 compute nodes.
- Edge：
  - 1 node with 1 NVIDIA Titian XP, video camera and other devices



Object Detection

- Software:
  - CUDA 9.2
  - OpenCV 3.2.0
  - FFmpeg 3.4.4
  - **Object detection model: darknet + yolov3-spp**
  - Training data: [coco.data]

- Comparison:
  - 1) Processing
    - Decode: GPU, Detection: GPU
    - Decode: CPU, Detection: GPU
    - Decode: CPU, Detection: CPU
  - 2) 8K video vs. 1080p



5. Summary

Summary

- Conclusion
  - Edge optimized architecture based on StarlingX.
  - Manage devices & network topology at the edge.
  - Manage different accelerators using cyborg.
  - Video Demo: 8K video provide much more details in analysis. & For the real-time analysis, 8K video should be accelerated in every processing procedure.

- Future Work
  - Cyborg containerized implementation.
  - Using user-space network stack to accelerate edge networking.

Q&A

- Jinghua Gao
  - Email: gaojh4@lenovo.com
  - Twitter: @Miss_Coco_Gao
  - OpenStack Cyborg Contributor
  - Lenovo research, staff researcher

- Zhi Chang
  - Email:changzhi@lenovo.com
  - Twitter: @changzhi1990
  - OpenStack Neutron Contributor
  - Lenovo research, Cloud computing researcher

- Shaohe Feng
  - Email:shaohe.feng@intel.com
  - IRC: Shaohe_feng
  - OpenStack Cyborg Contributor
  - Intel Cloud Team

Different is better

Lenovo™

intel®