How QCT Edge Platforms Accelerate the AI Workloads

Michael Quan
Director Sales Engineering
Quanta Cloud Technology (*QCT*)


QCT
1010

QCT – Quanta Cloud Technology

- A global datacenter solution provider that combines the efficiency of hyper-scale hardware with infrastructure software from a diversity of industry.
- Product lines include hyper-converged and software-defined datacenter solutions as well as servers, storage, switches, integrated racks.

Quanta – Quanta Computer Inc.

- Is the parent of QCT
- Support QCT in designing, engineering, manufacturing, system and rack integration and supply chain support through the Quanta global network. All under one roof.

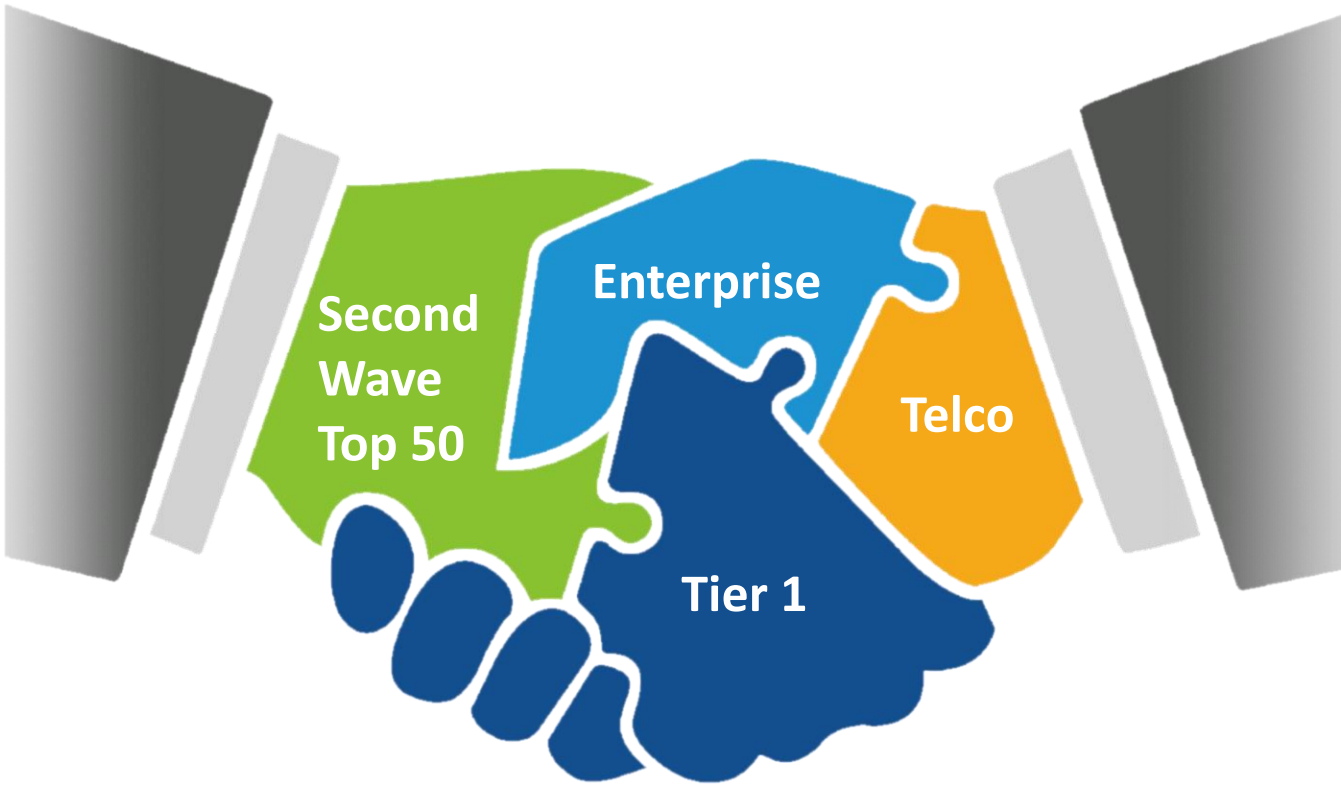
QCT Global Footprints

Established Offices:

- USA: Silicon Valley, Seattle
- Taiwan: Tao Yuan
- China: Beijing, Hangzhou, Chongqing
- Japan: Tokyo
- Korea: Seoul
- Germany: Düsseldorf

QCT NOW POWERS MOST OF THE GLOBAL TIER 1 HYPERSCALE DATACENTERS, TELCOS AND LARGE ENTERPRISES

QCT's Customer Extending from HyperScale to Enterprise and Telco



QCT, With 3S1R Complete Product Line Under One Roof to Fulfill All Your Needs

- OCP-Open Rack
- OCP-OCS
- Project Olympus

Racks

- 1G Switch
- 10G Switch
- 25G Switch
- 40G Switch
- 100G Switch
- 200G Switch
- 400G Switch

Network Switch

- Cloud Server
- High Density Server
- GPU Server

Server

- Storage Server
- JBOD/JBOF
- NVMeoF

Storage

End-to-End Workloads for 5G and AI

CPE: Customer Premise Equipment
SD-WAN: Software-defined Wide Area Networking
RAN: radio access networking ; MEC: mobile edge computing
EPC: Evolved packet core BNG: broadband network gateway;
CDN: content delivery networking; NFV: Network function virtualization

5G Everywhere	**IoT**	**Outdoor Edge**	**Indoor Edge**	**Regional Data Center**	**Data Center**
End to End Workloads	CPE SD-WAN Firewall	RAN MEC	RAN MEC	CPE/BNG/EPC NFV/CDN	Video Apps
AI	**Inferencing**				**Training**
QCT Accelerated Platform		 **S1K** 1S- Intel Xeon D 1x FPGA (Proprietary)	 **SD2H-1U** 1S- Intel Xeon D Front Access 2x GPU T4, FPGA CPU: GPU = 1:2  **D52Y-2U** 2S- Intel Xeon SP Front Access 2x GPU V100, 4x T4 or 2x FPGA support CPU: GPU = 1:2	 **D52BE-2U** 2S- Intel Xeon SP 2x GPU T4, FPGA CPU: GPU = 1:1  **X11C-8N** Coffee Lake-S, 8N Front Access 8x GPU T4 CPU: GPU = 1:1 or 1:2	 **D52BV-2U** 2S- Intel Xeon SP 4x GPGPU CPU: GPU = 1;2  **D52G-4U** 2S- Intel Xeon SP 8x V100 w/Nvlink CPU : GPU = 1:4

Mainstream GPU Servers

Mezzanine **V100 NVLink**	**D52G-4U**		**8x SXM2**
Double Width **V100 PCIe & RTX Quadro**	**D52G-4U**		**8x GPU**
	D52BV-2U		**4x GPU**
	D53XQ-2U **D43KQ-2U**		**2x GPU**

Training

HPC Rendering

HPC Training Inference

VDI Rendering

HPC Inference

VDI

AI Model Complexity

H

L

INFRASTRUCTURE OF THE FUTURE, NOW!!

Full Product Line to Support Versatile T4

	20x	(2) Xeon	D52G-4U
	6x	(2) Xeon	D53XQ-2U
	5x	(2) Rome	D43KQ-2U
	4x	(2) Xeon	D52BV-2U
	3x	(2) Xeon	D53X-1U
		(2) Rome	D43K-1U
	2x /node	(1) Xeon E3/node	X11C-8N
		(1) Rome /node	S43CA-4N
	2x	(1) Rome	S43KL-1U
		(2) Xeon	D52B-1U
		(2) Xeon	D52BQ-2U
		(4) Xeon	Q72D-2U
		(2) Xeon	D52Y-2U
		(1) Xeon-D	SD2H-1U

Edge

2x No. of GPU per node

- Both Scale up & Scale out design
- Different CPU Platform
- Various CPU to GPU ratio server model

INFRASTRUCTURE OF THE FUTURE, NOW!!

QCT Product Lineup with AMD EPYC™ 7002 Series

Product Positioning and Transition Guide

1U General Purposed

QuantaGrid D43K-1U

Dual AMD EPYC™ 7002 Processors

- 32x DDR4 DIMM slots
- **Up to 12x NVMe SSD**
- OCP 3.0 PCIe x8 up to 100GbE
- Up to 5 PCIe 4.0 expansion slots
- Up to 3x NVIDIA T4 GPU

PCI EXPRESS 4.0

2U General Purposed

QuantaGrid D43KQ-2U

Dual AMD EPYC™ 7002 Processors

- 32x DDR4 RDIMMs
- **12x 3.5” HDDs or 24x 2.5” U.2**
- OCP 3.0 PCIe x8 up to 100GbE
- Up to 10 PCIe 4.0 expansion slots
- Up to 2x NVIDIA V100 GPU
- Up to 5x NVIDIA T4 GPU

PCI EXPRESS 4.0

1U Workload Optimized Compute Server

QuantaGrid S43KL-1U

Single AMD EPYC™ 7002 Processors

- 16x DDR4 DIMM slots
- **Up to 12x NVMe SSD**
- OCP 3.0 PCIe x8 up to 100GbE
- Up to 5 PCIe 4.0 expansion slots
- Up to 2x NVIDIA T4 GPU

PCI EXPRESS 4.0

QCT

INFRASTRUCTURE OF THE FUTURE, NOW!!

QCT Product Lineup with AMD EPYC™ 7002 Series

Product Positioning and Transition Guide

Density Optimized 2U4N server

QuantaPlex S43CA-2U

Single AMD EPYC™ 7002 Processors

- 16x DDR4 DIMM slots
- **2U4N Server**
- ***Aggregated networking***
- Single/Dual 25G port per node
- Up to 1x NVIDIA T4 GPU per node

PCI EXPRESS 4.0

All NVMe Storage Server

QuantaGrid D42A-2U

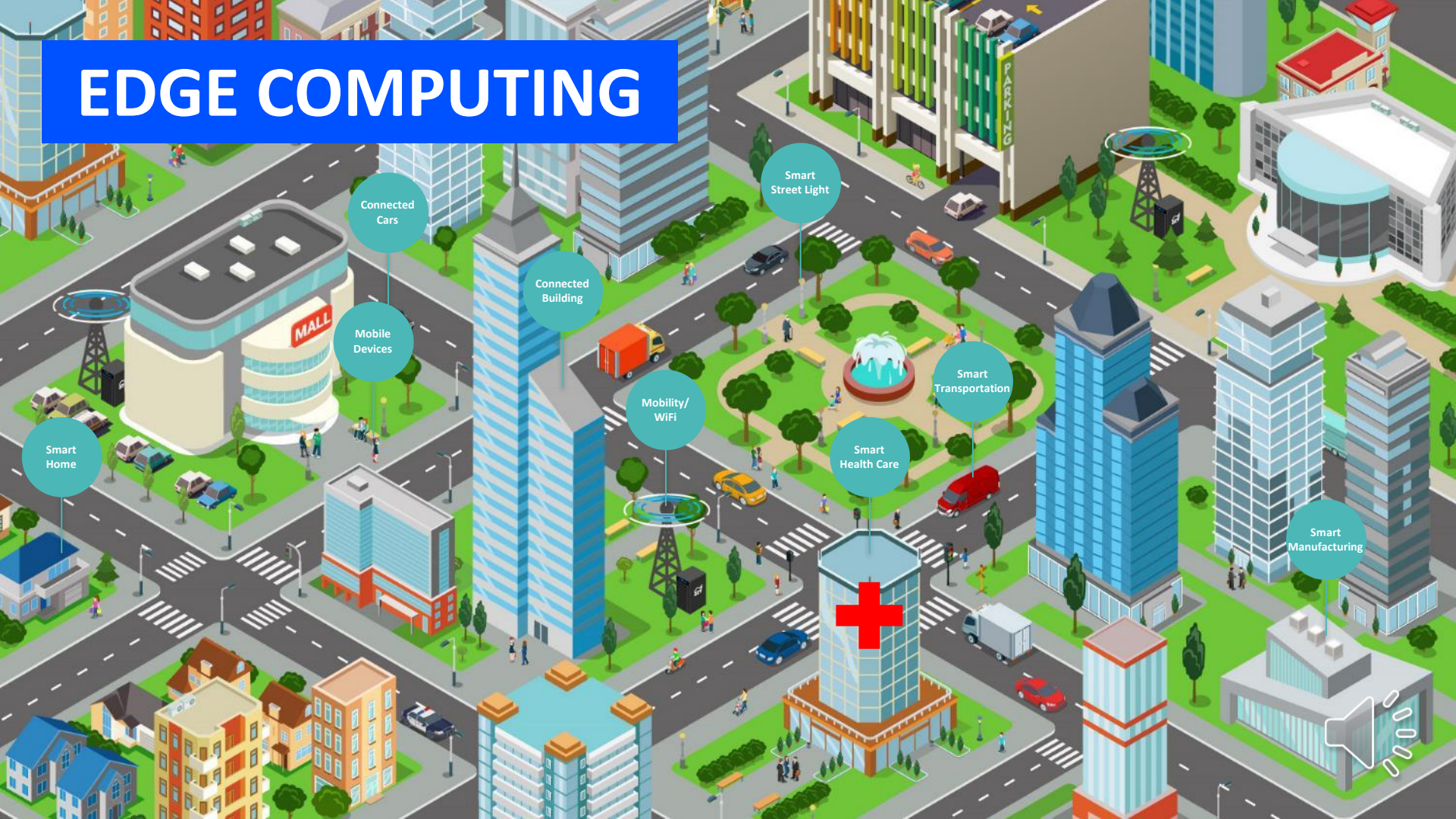
Single AMD EPYC™ 7001/7002 Processors

- 16x DDR4 DIMM slots
- **24x NVMe Drives**
- Up to 2x 100GbE networking

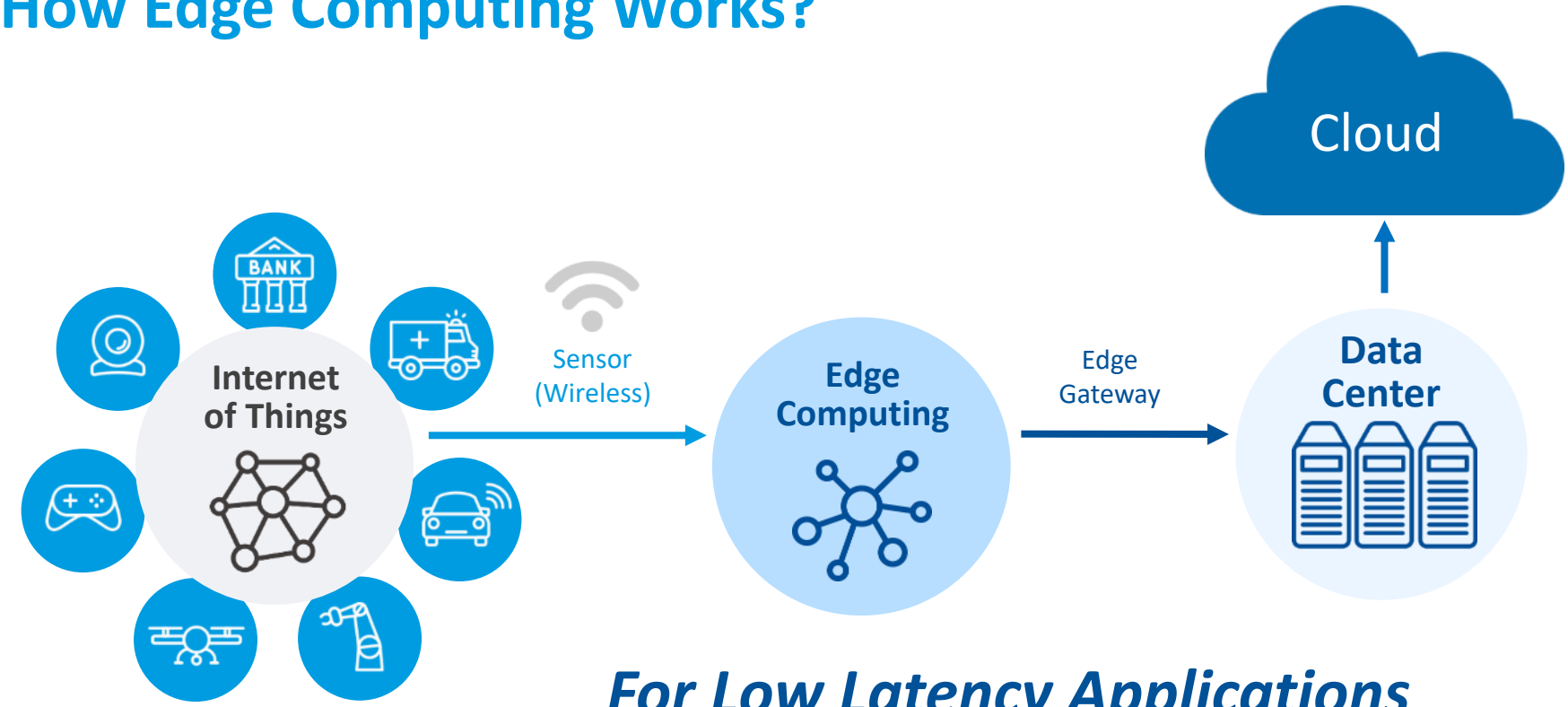
PCI EXPRESS 3.0

QCT

INFRASTRUCTURE OF THE FUTURE, NOW!!


EDGE COMPUTING
Connected Cars
Mobile Devices
Smart Home
Connected Building
Smart Street Light
Mobility/ WiFi
Smart Health Care
Smart Transportation
Smart Manufacturing
MALL
PARKING

How Edge Computing Works?

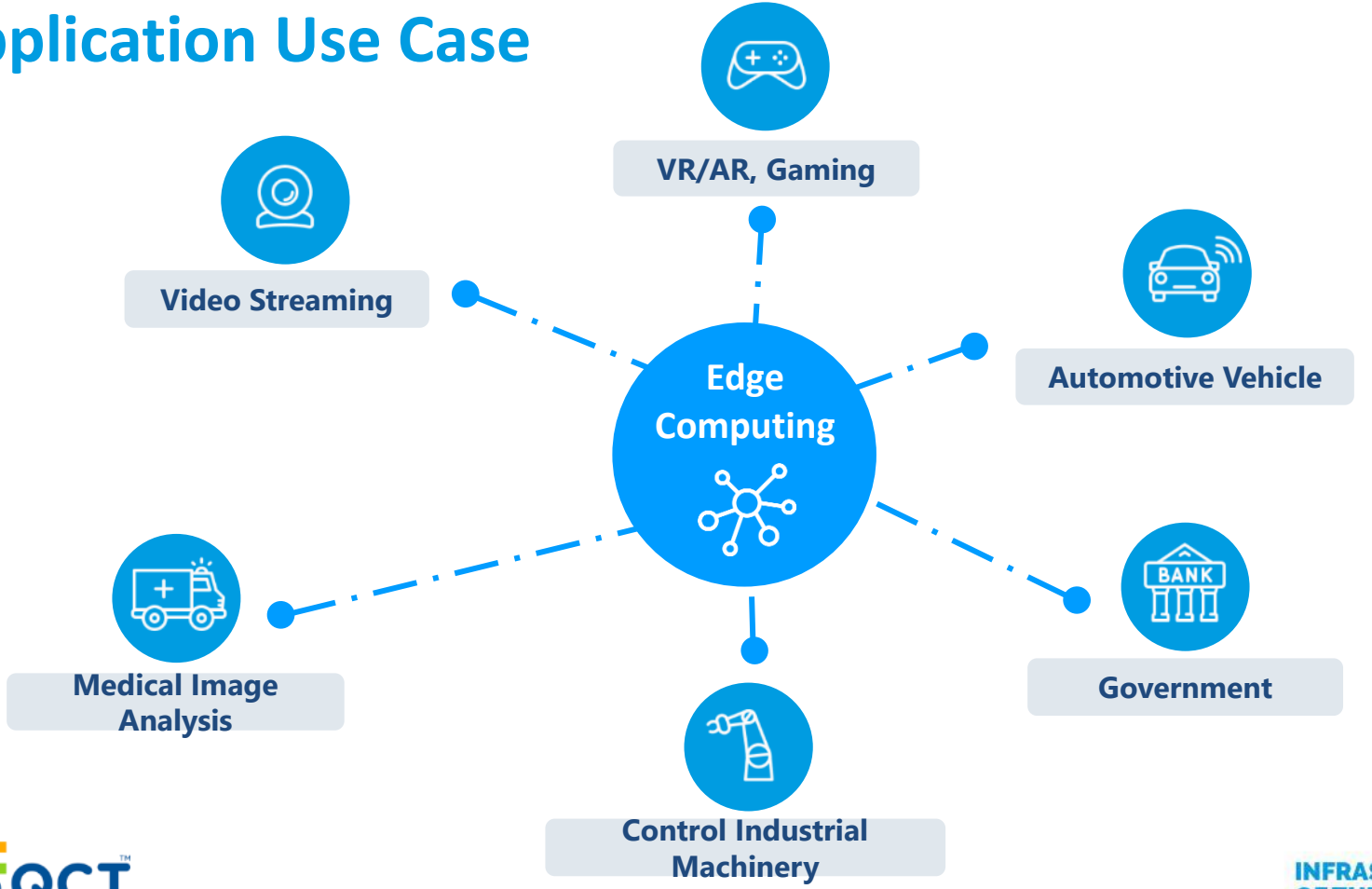


For Low Latency Applications

QCT

INFRASTRUCTURE OF THE FUTURE, NOW!!

Application Use Case



QCT

INFRASTRUCTURE OF THE FUTURE, NOW!!

QCT Telco Product Portfolio

Customer	Outdoor Edge	Indoor Edge	Central Office	Data Center

uCPE

Performance
Intel Skylake/CascadeLake SP
1U Dual sockets

Mid-Range
Intel Skylake-D
4/8/16 core

Entry
Intel Denverton/SnowRidge
4-8 core

Sub-Entry
Intel Denverton/Snow Ridge
2 core

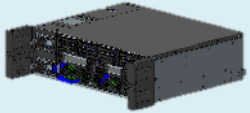
Outdoor MEC / vRAN_Pole Mountable

S1K-1U
Outdoor Broadwell-DE
NEBS Level 3
IP65 water and dust proof
Operation temperature -40 to 85°C

Indoor MEC / vRAN_400mm Depth

S3H-1U
Skylake- D
Front Access
Flexible I/O design
-48DC PSU support

S5Y-2U
Skylake- SP
Front Access
Flexibile I/O design
8x SATA SFF
-48VDC PSU support



S6L-3U
Dual Intel Cascade lake SP
2x Dual width GPU or 8x NVMe SSD

Central Office/ NFVI Optimized_780mm Depth

D52BQ-2U 3UPI
(NUMA Balance)
Skylake- SP
12xLFF/24xSFF

T42S-2U (4-node)
Skylake-SP
16x DIMM per node
12xLFF/24xSFF

D52B-1U
Skylake- SP
12xSFF(NVMe opt.)
4xLFF + 4xNVMe

Central Office/ NFVI Optimized_600mm Depth

D52BE-2U
2U dual socket
2xSFF

SD2H-1U

Ultimate I/O Flexibility for vRAN/MEC Edge Server

400mm chassis depth

Flexible Front I/O module design

Up to **2x** PCIe expansion slots for FPGA / GPU / LAN

Up to **2x** U.2 for cache acceleration

Support **-48V DC** PSU

Support both **Front & Rear** power inlets

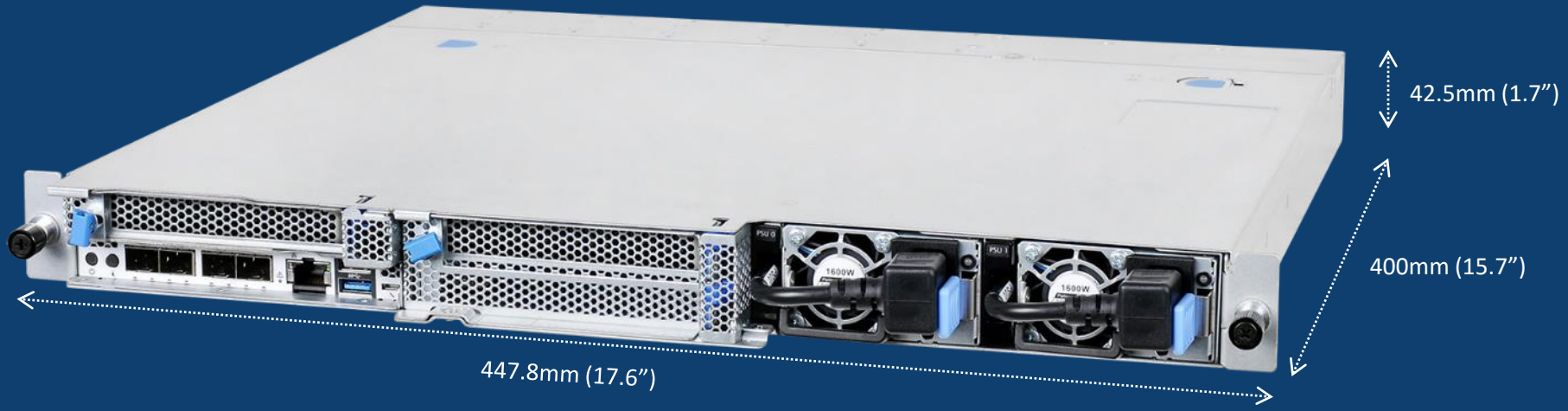
Innovative Performance-Optimized Infrastructures for Telecom Ecosystem

QCT™

INFRASTRUCTURE OF THE FUTURE, NOW!!

SD2H-1U

Chassis Overview



*Ultra Short Chassis **400mm** depth*

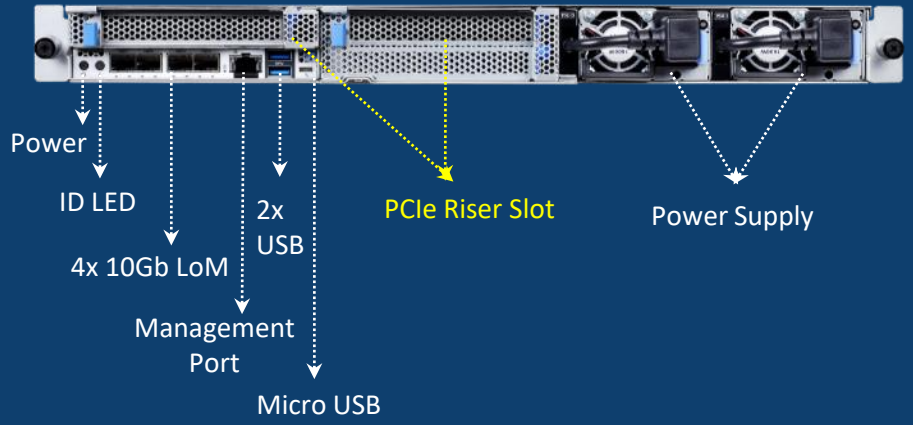
QCT

INFRASTRUCTURE OF THE FUTURE, NOW!!

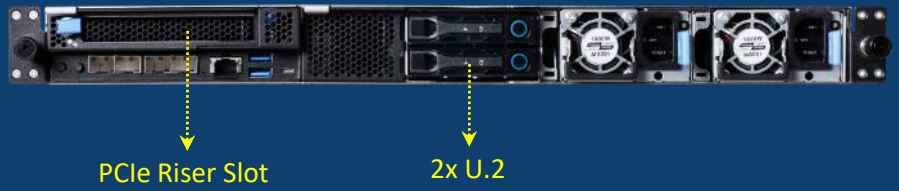
SD2H-1U

Front View

Networking/Inference SKU



Cache Accelerator SKU



SD2H-1U
Rear View

- Support ***rear AC socket*** for flexible PSU cable routing
- Support ***hot swappable easy service fans***
- Support **WiFi / LTE modules** for uCPE application

2x rear AC socket

Antenna:
2x Wifi
2x LTE

6x 4028 single rotor fan

SD2H-1U

Indoor Edge Server for MEC / vRAN

Flexible Module Assembly

Module Design
2x 2.5” 15mm U.2

PCIe Module slot Design Options

GPU (Dual width) / FPGA (FHHL)

Item	Module
GPU	Intel VCA2
	Nvidia Tesla T4
FPGA	Intel Vista Creek

NIC Card Module*

Infiniband Card Module*

SAS Mezz Module**

* Details please refer to CCL
** Need to have extra adapter card for SAS Mezz

SD2H-1U

(Networking/Inference SKU)

Dimensions (WxHxD): 447.8mm x 42.5mm x 400mm

CPU: 1x Intel SKL-D processors (up to 110W TDP)

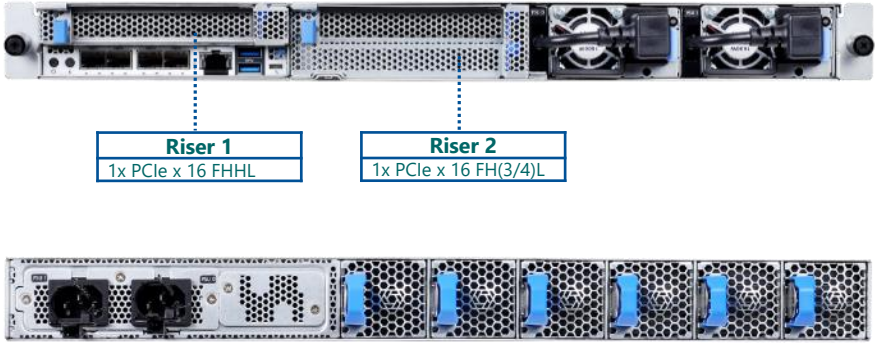
DIMM slot: 8x DDR4 RDIMM

Storage: 2x PCIE x4 G3 2280M.2

Expansion:

1x PCIe x16 G3 riser slot 1 , FHHL

1x PCIe x16 G3 riser slot 2 , FH(3/4)L



QCT

INFRASTRUCTURE OF THE FUTURE, NOW!!

SD2H-1U

(Cache Accelerator SKU)

Dimensions (WxHxD): 447.8mm x 42.5mm x 400mm

CPU: 1x Intel SKL-D processors (up to 110W TDP)

DIMM slot: 8x DDR4 RDIMM

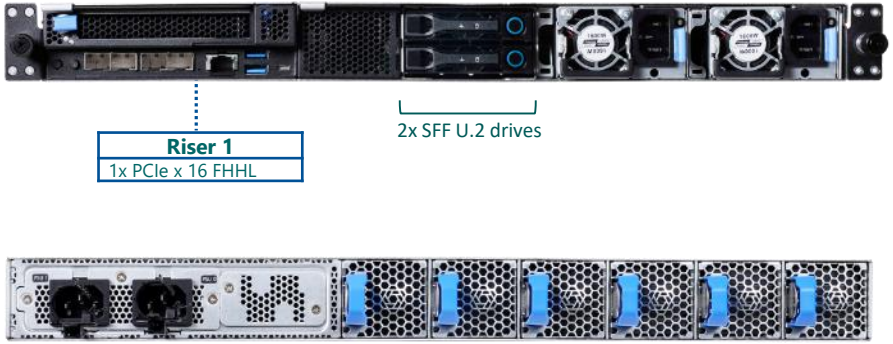
Storage:

2x 15mm U.2

2x PCIE x4 G3 2280M.2

Expansion:

1x PCIe x16 G3 riser slot 1 , FHHL



QCT

INFRASTRUCTURE OF THE FUTURE, NOW!!

D52Y-2U

NUMA-Balanced Infrastructures for Telecom Ecosystem

400mm chassis depth

Up to **10x** NVMe SFF for local storage

Flexible Front I/O module design

Support **-48V DC** PSU

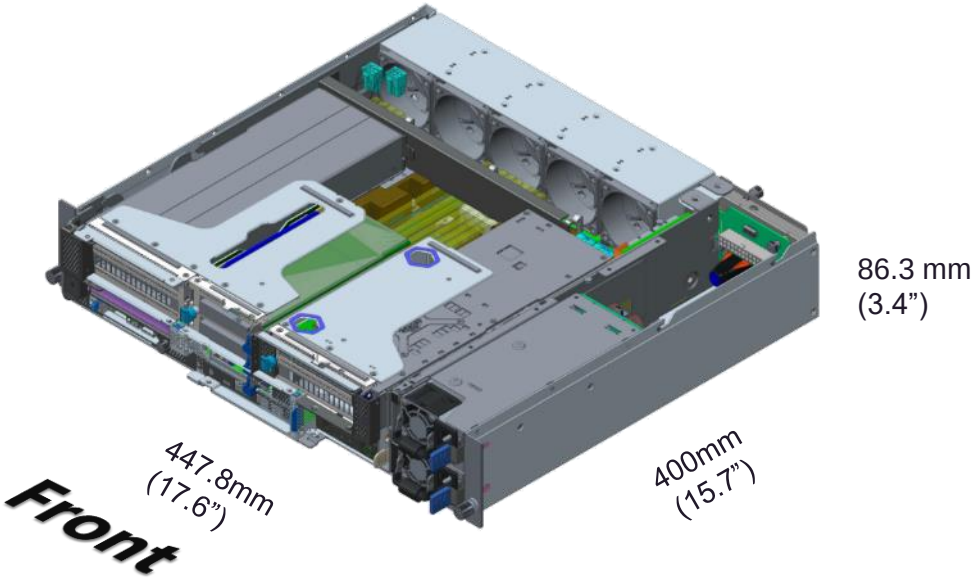
Support both **Front & Rear** power inlets

Performance Upgrade for vRAN/MEC Edge Server

QCT

INFRASTRUCTURE OF THE FUTURE, NOW!!

D52Y-2U System Overview (GPU SKU)

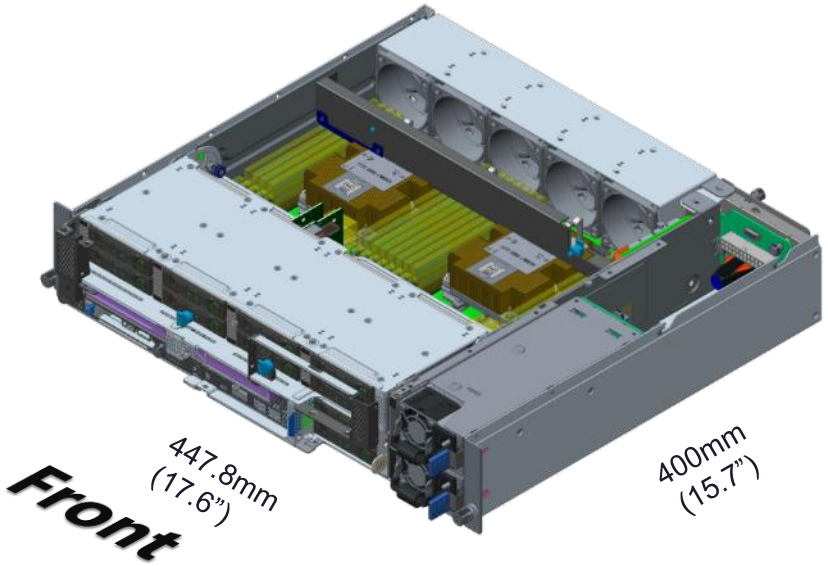


Ultra Short Chassis **400mm** depth

QCT

INFRASTRUCTURE OF THE FUTURE, NOW!!

D52Y-2U System Overview (NVMe SKU)



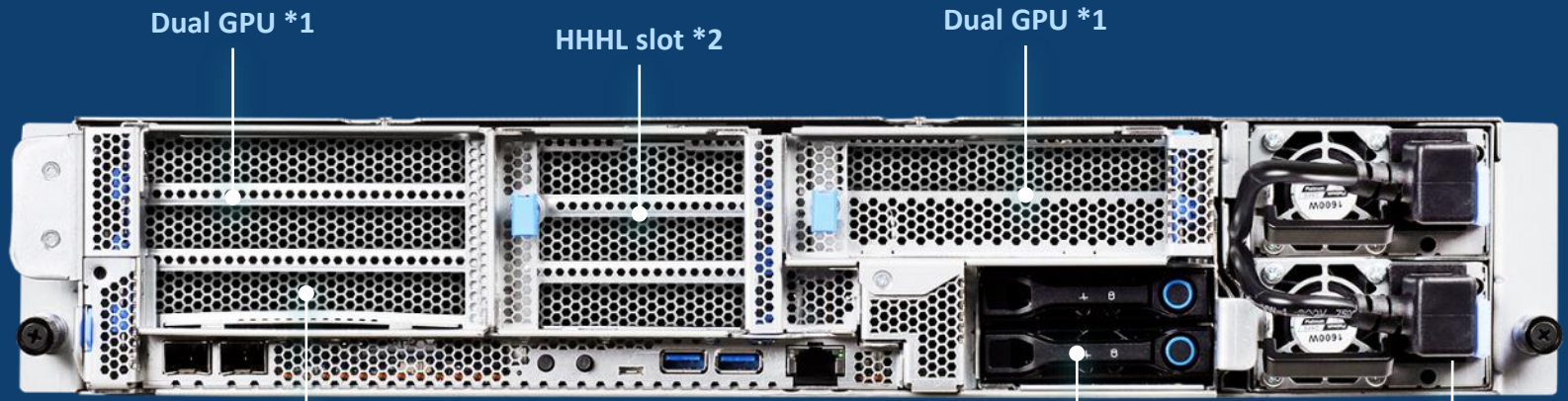
Ultra Short Chassis **400mm** depth

QCT

INFRASTRUCTURE OF THE FUTURE, NOW!!

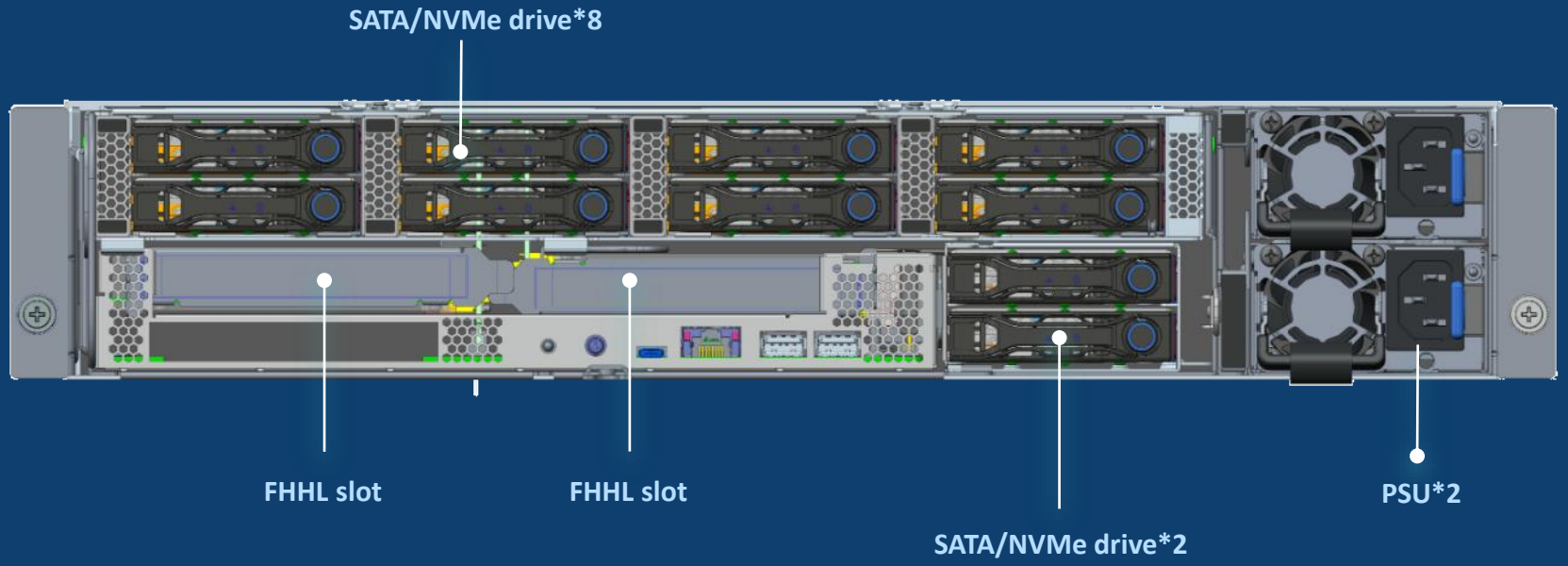
D52Y-2U

Front View



D52Y-2U - NVMe SKU

Front View



D52Y-2U

Rear View

- Support ***rear AC socket*** for flexible PSU cable routing
- Support ***hot swappable easy service fans***

D52BE-2U (S5BE)

Computing in the Edge

Top shelf Xeon® P processor[1]

579mm chassis depth

Up to 7x PCIe expansion slots

Up to 3TB memory capacity[2]

1. With limited conditions
2. With specific CPU

QCT

INFRASTRUCTURE OF THE FUTURE, NOW!!

D52BE-2U

Chassis Overview

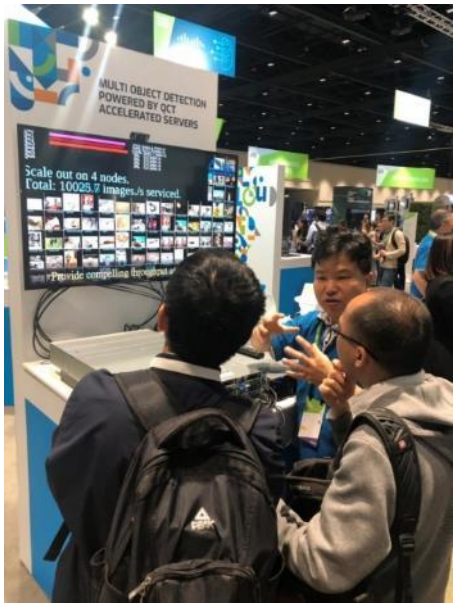
QCT

INFRASTRUCTURE OF THE FUTURE, NOW!!

SC18, OCP Summit 19, GTC 19 US

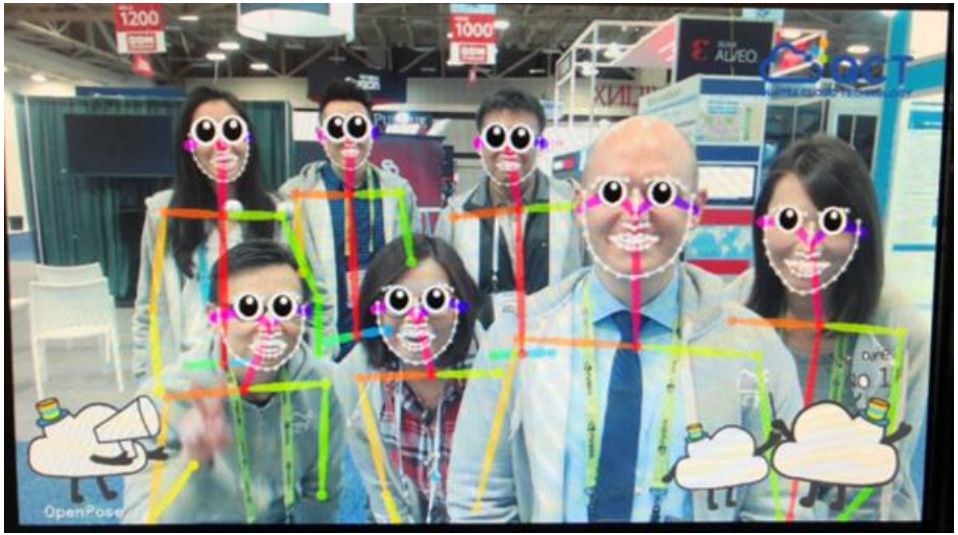
X11C-8N

Multi-Object Detection



D52BV-2U

Multi-Object Detection

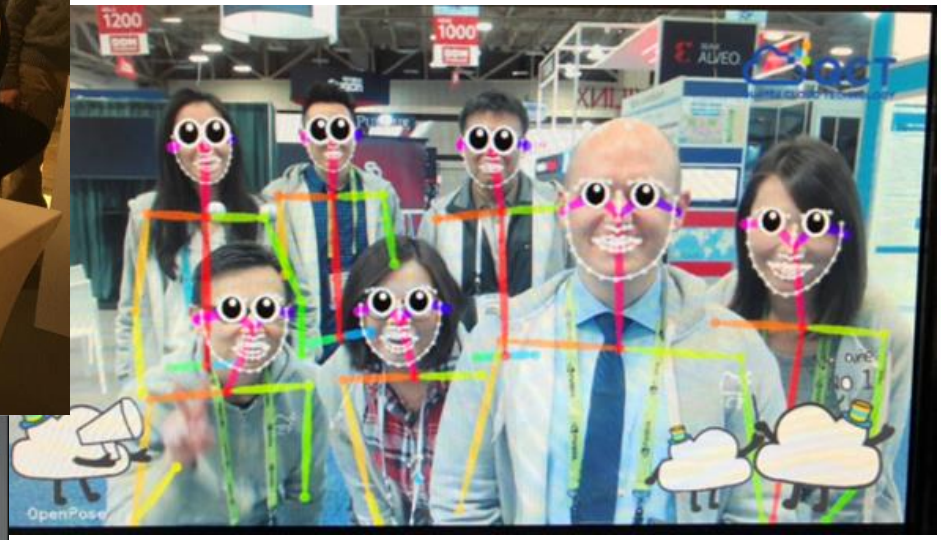


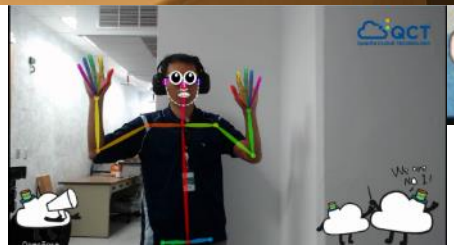
QCT

INFRASTRUCTURE OF THE FUTURE, NOW!!

D52BV + P100 - Open Pose

Real-time multi-person keypoint detection





QCT

INFRASTRUCTURE OF THE FUTURE, NOW!!

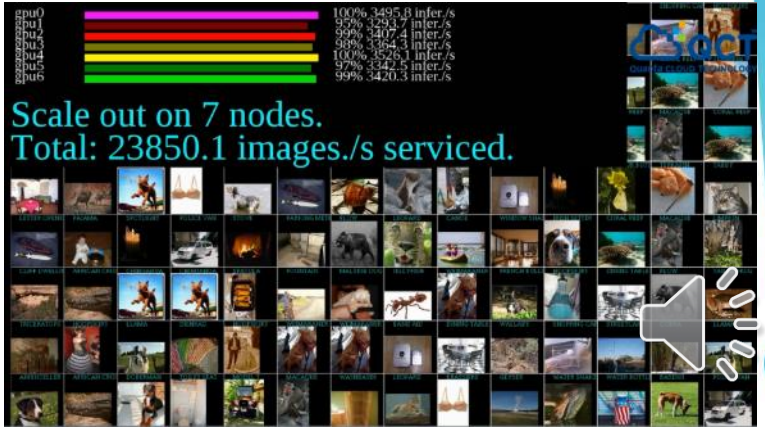
X11C-8N + T4 – Multi Object Detection

Linear Performance on ResNet-50

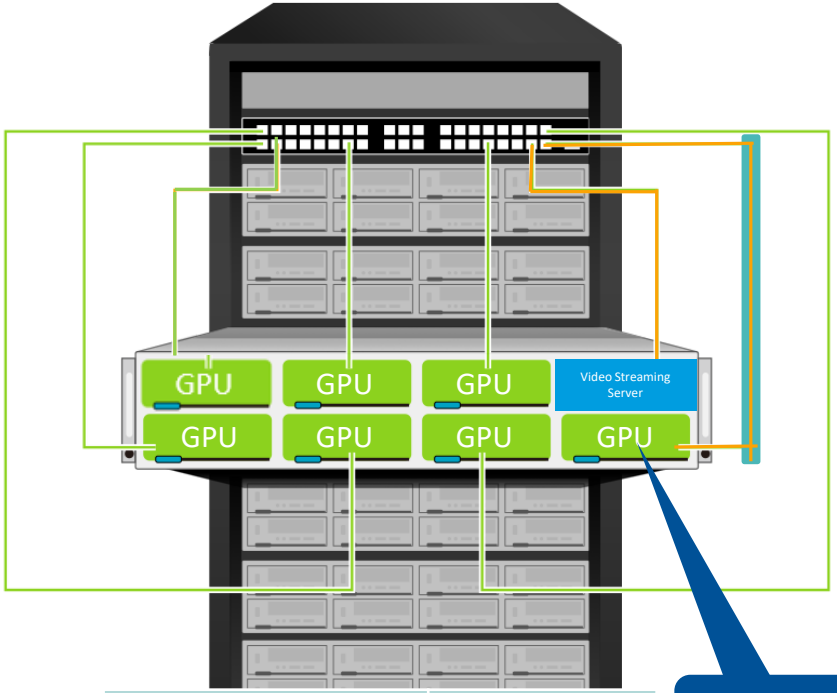
.....

.....

.....



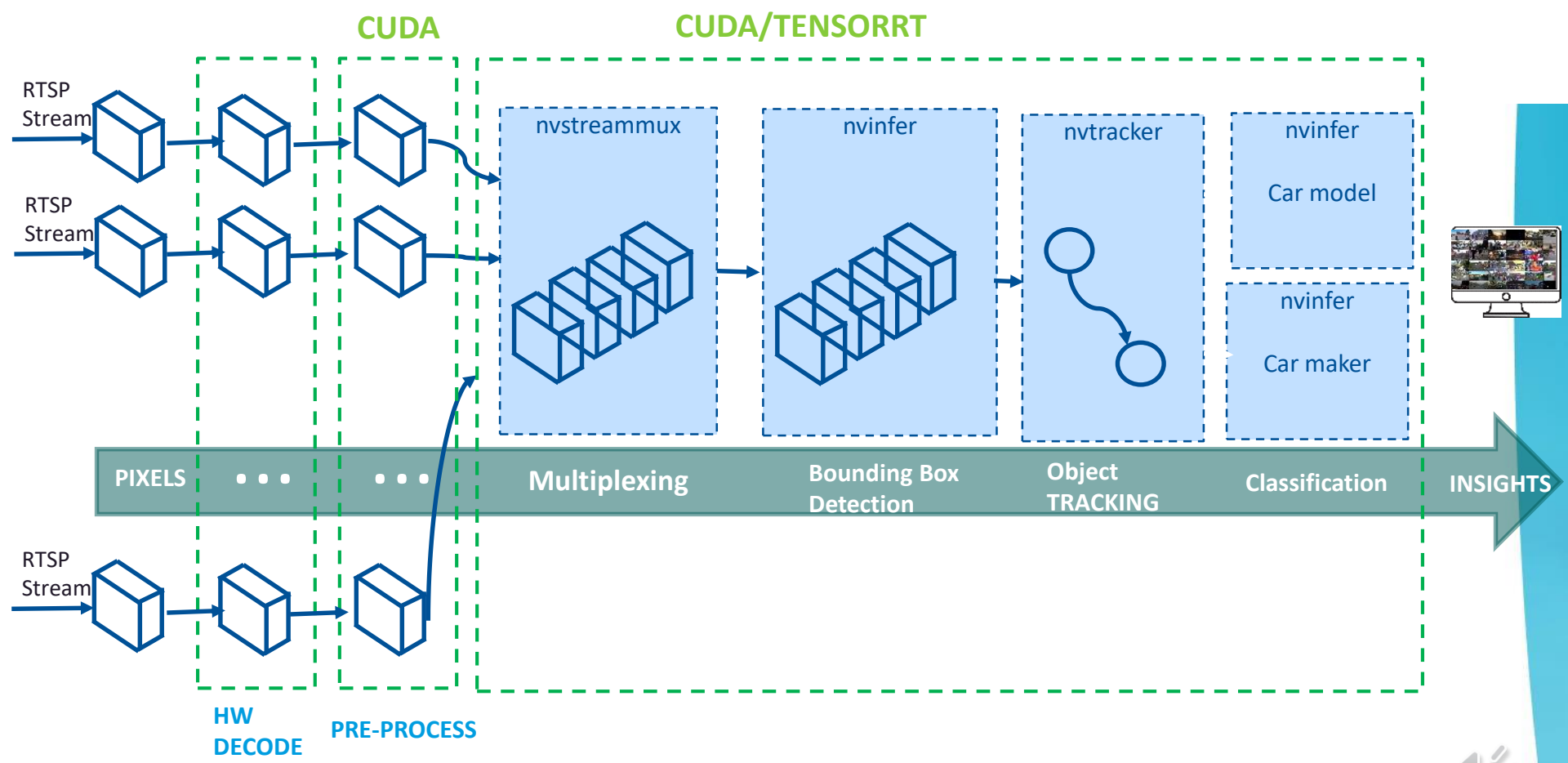
Multiple Video Streams AI Analytics Reference Design



SW	Version
Operating System	Ubuntu 18.04
NVIDIA Device Driver	418.87
CUDA Toolkit	10.1
TensorRT	5.1.5
Gstreamer	1.14.1
DeepStream SDK	4.0

X11C-8N(E3 CPU + Tesla T4 GPU)

1. Intelligent video analytics reference design
2. Decode 38 streams concurrently on single node
3. Accelerated DL inferences on NVIDIA DEEPSTREAM SDK 4.0



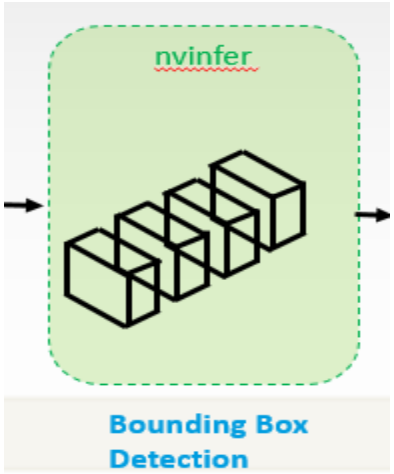
RTSP: Real Time Streaming Protocol

Model for Bounding Box Detector

- Resnet-10 pre-trained model to generate detected bounding box of 4 classes: Vehicle, RoadSign, TwoWheeler, Person

Primary Object Detector (P-gie)

Stream Multiplexer



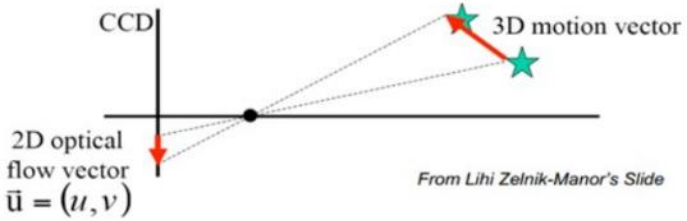
- Vehicle
- RoadSign
- TwoWheeler
- Person

Tracker

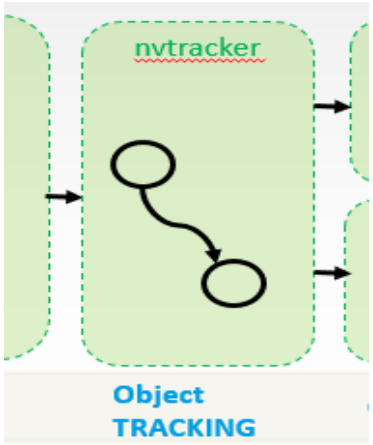
KLT Tracker

- Nvtracker: KLT (Kanade-Lucas-Tomasi) algorithm implementation.

  Making use of spatial intensity information to direct the search for the position that yields the best match. It is faster than traditional techniques for examining far fewer potential matches between the images



Primary Object Detctor



Secondary inference (S-gie1) Car model

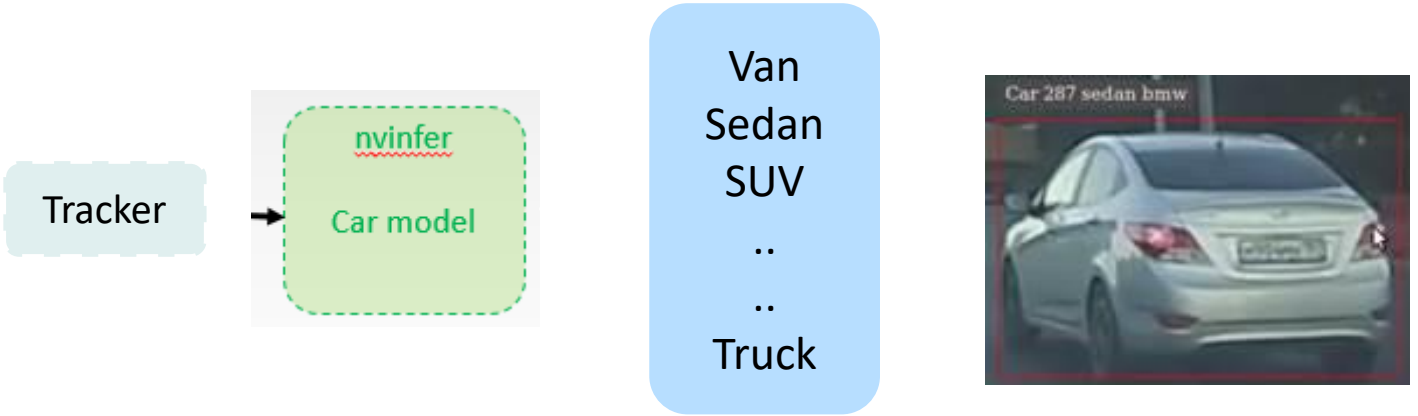
Secondary inference (S-gie2) Car make

QCT

INFRASTRUCTURE OF THE FUTURE, NOW!!

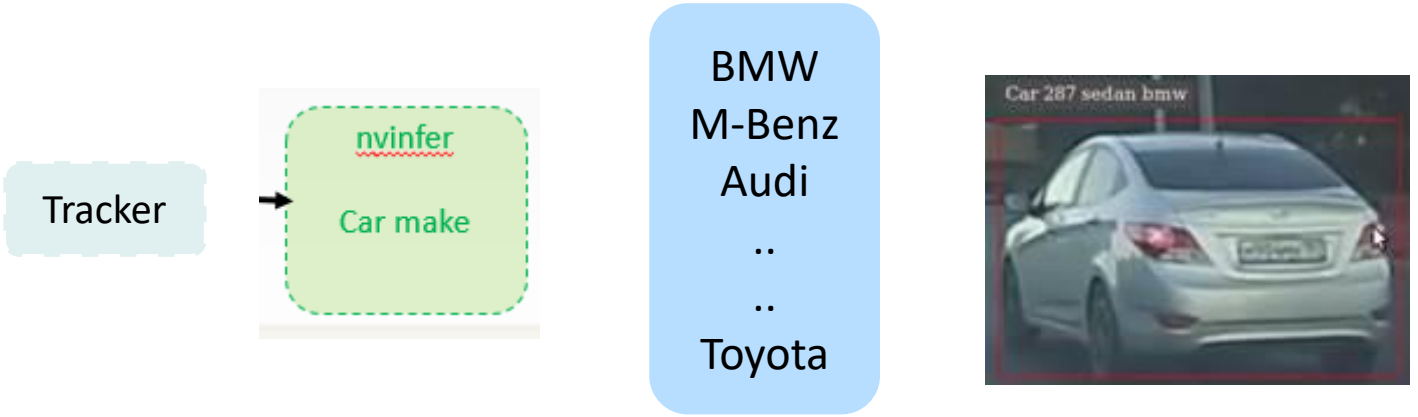
Classifier for Car Model

- Classifier of Resnet-18 network that trained on ImageNet. Works as secondary-gie to identify car model.

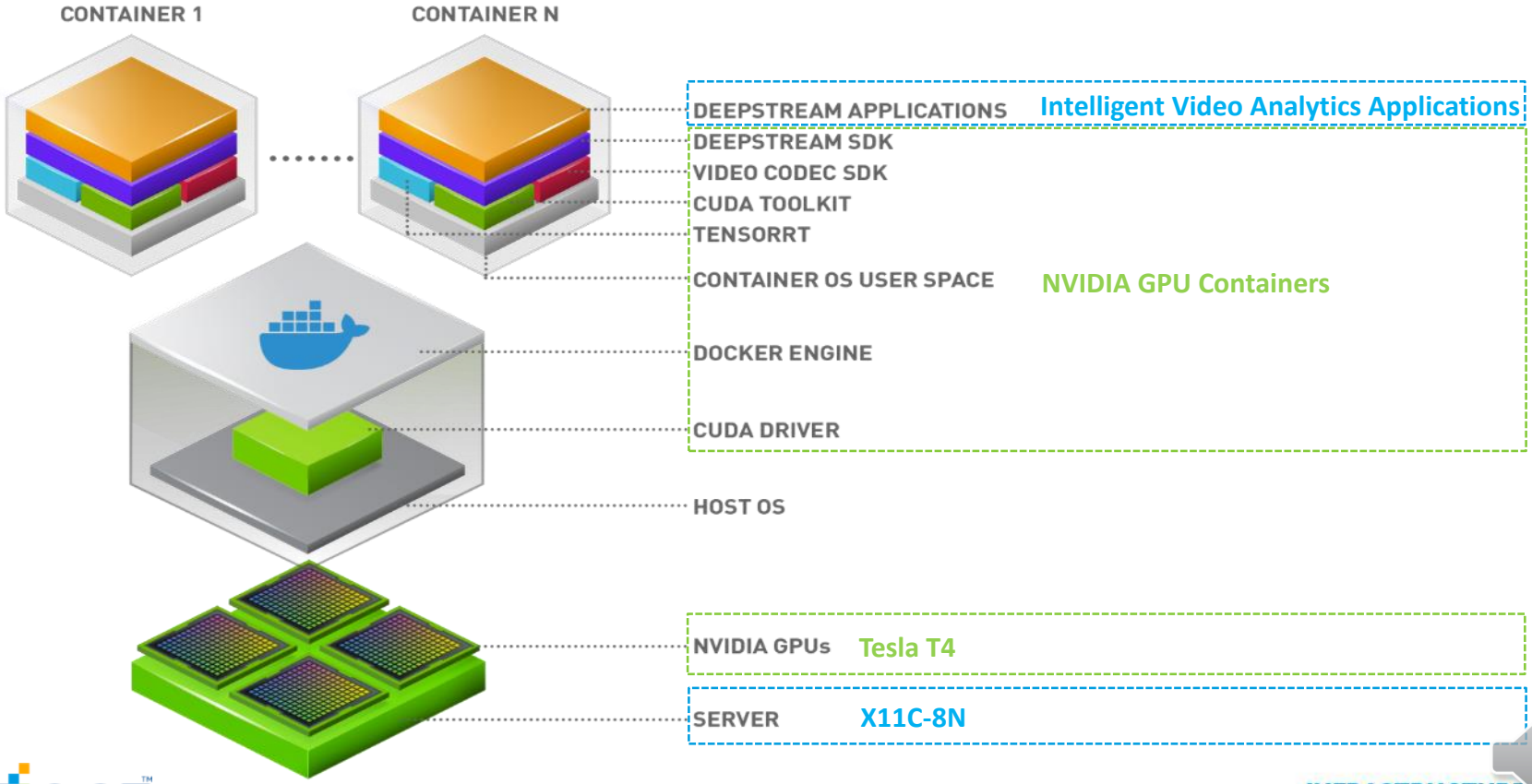


Classifier for Car Make

- Classifier of Resnet-18 network that trained on ImageNet. Works as secondary-gie to identify car maker.



Software Stack



https://developer.nvidia.com/deepstream-sdk


DeepStream
FPS: 22.90
Person 182
FPS: 29.87
Car 64 largevehicle jeep
Car 66
Car 98
Car 81 sedan hyundai
Car 99 sedan
Roadsign 179
NB in Atlanta
from Exit 231 to Exit 265
FPS: 29.87
Person 79
Person 115
FPS: 29.87
Person 649
Car 702
Bicycle 699
FPS: 23.10
Car 35
FPS: 29.87
anytime | anywhere | any light source
FPS: 25.03
Car 2
FPS: 22.74
Car 79
Car 147
FPS: 23.01
Car 116 largevehicle chevrolet
Person 114
FPS: 23.21
Car 19 ford
Person 72
FPS: 23.10
FPS: 22.80
Person 180
Person 169
Person 181
FPS: 23.46
Car 6 truck
Person 8
FPS: 23.33
FPS: 23.13
drivers when crossing
intersections
FPS: 22.80
Car 295
Car 297
Car 276
Car 315
FPS: 22.98
Person 55
Person 53
Person 56
FPS: 23.30
Car 40
Car 27
Car 36
Car 42
FPS: 23.10
Car 109
FPS: 23.30
Person 69
FPS: 23.46
FPS: 22.39
Car 99
Car 5 seda
Car 6 lexus
FPS: 23.10
Car 204
FPS: 23.10
Car 12
Car 25
Car 26
FPS: 29.87
FPS: 22.98
Person 497
Person 528
Person 523
Person 473
FPS: 23.18
FPS: 23.30
Car 93 van
FPS: 23.10
Car 14
Car 21 van
Car 19
FPS: 23.10
Person 139
Person 144
Person 138
FPS: 23.10
FPS: 23.50
Car 35
FPS: 22.90
Car 29 truck
Car 53 suv
Person 78
FPS: 22.70
Person 79
FPS: 29.47
Car 134 van mercedes
Person 137
FPS: 24.89
FPS: 23.10


QCT
CLOUD
Looking for
innovative cloud solutions?
Come to QCT, who else?
INFRASTRUCTURE
OF THE FUTURE, NOW!!
www.QCT.io


QCT
CLOUD
Thank You!
INFRASTRUCTURE
OF THE FUTURE, NOW!!
www.QCT.io