Deep Learning Architectures with Deep Thinking

2019.03
Agenda

- Introduction to QCT
- AI Accelerated Server Product Lines
- From Scale-up Training to Scale-out Inference
- S5G Introduction
- S3H/S3HQ/S9C Introduction
- Demo
**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.
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
From Cloud, 5G to AI
QCT, With 3S1R Complete Product Line Under One Roof to Fulfill All Your Needs
19” General Purpose Server Lineup

- 1U Multi-node
- 2U Micro-Server
- GPU 4 CPU
19’’ Storage Product Lineup

Storage Server (1)
1U12

JBOD (1)
4U60

Storage Server (2)
4U78

JBOF (SSD)
2U72

JBOD (2)
4U24
Rackgo X – 21” Servers

Multi-Node

GPU Server
JBOG

Micro-Server
NVMeoF

JBOD
DGX-1 - World’s First AI Appliance

“250 Servers in-a-box @$129,000”
NVIDIA's reach across the AI ecosystem grew even broader this week with the addition of new servers for inferencing from Quanta. It's yet another example of how industry leaders are supporting our end-to-end artificial intelligence infrastructure.

Quanta – one of the world's largest enterprise server vendors – has expanded its server lineup to feature GPUs based on the NVIDIA Pascal architecture, which are purpose-built to be the engine of computers that learn, see and simulate our world.
Scale-Out Inference Server vs. NVIDIA T4

https://www.youtube.com/watch?v=Tom07EPbzrE&t=4134

@GTC China, 2018
QCT Big Sur & Big Basin

Engineering at Facebook

Facebook to open-source AI hardware design
The Computing Conference 2017

APSARA INTELLIGENCE
Accelerated Server Roadmap
End-to-End Workloads for 5G and AI

- **5G Everywhere**
  - IoT
    - CPE
    - SD-WAN
    - Firewall
  - Edge
    - Radio/Optical
    - RAN/Access
    - MEC
  - Central Office
    - CPE/DPI
    - BNG/EPC
    - CDN / IMS
  - Data Center
    - IMS
    - Video
    - Apps

- **End to End Workloads**

- **AI**
  - **Inferencing**
    - QCT GPU Accerlated Platform
    - S9C-2U
    - S3HQ-2U
    - S3H-1U
    - Big Sur
  - **Training**
    - S5G-4U
    - S5BV-2U
    - JG3 (JBOG)
    - S7D-2U (4- Socket)
    - Big Basin
    - JG3 (JBOG)
    - S5G-4U

- **Quanta CLOUD TECHNOLOGY**
  - In Development
  - Production
## Training at Datacenter

<table>
<thead>
<tr>
<th>GPU</th>
<th>GPU Server</th>
<th>JBOG</th>
</tr>
</thead>
<tbody>
<tr>
<td>V100 NVLink</td>
<td>S5G-4U</td>
<td>Big Basin</td>
</tr>
<tr>
<td></td>
<td>Big Sur</td>
<td>JG3 (4U)</td>
</tr>
<tr>
<td>V100 PCIe</td>
<td>S5G-4U</td>
<td>8x SXM2</td>
</tr>
<tr>
<td>V100 PCIe</td>
<td>S5BV-2U</td>
<td>8x GPU</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4x GPU</td>
</tr>
</tbody>
</table>

AI model complexity:
- H: High
- L: Low
### Inference Everywhere

#### GPU

<table>
<thead>
<tr>
<th>No. of GPU per node</th>
<th>Datacenter</th>
<th>Edge</th>
</tr>
</thead>
<tbody>
<tr>
<td>2x</td>
<td>GPU</td>
<td></td>
</tr>
</tbody>
</table>

#### GPU Options

- **P4, T4 (75W)**
  - **GPU Models**
    - S5G-4U: 16x
    - JG3 (480G): 20x
    - S5BV-2U: 4x
    - S7D-2U (4-Socket): 2x
    - S9C-8N: 4x
- **S5H-1U**
  - S3HQ-2U: 2x

#### No. of GPU per Node

- 2x (4 nodes)
- 1x (8 nodes)
S5G-4U

3-in-1 Accelerator Server

Training on baseboard with NVLink

Inferencing with 20x accelerators with power efficiency

HPC/DL on optimized topology
S5G-4U
3-in-1 Accelerator Server

➢ Extreme DL performance to shorten customer R&D development

➢ Run massive inference workloads with power efficiency concurrently.

➢ Maximize utilization of IT investment by providing flexibility to optimize performance
S5G-4U
3-in-1 Accelerator Server

➢ Featuring with 300GB/s NVLink interconnect
➢ 24 SFF drive bays including up to 8x U.2
➢ Up to 16x single-width power efficient GPUs
➢ Single baseboard to multiple PCIe routing topology
➢ Up to 4x100Gb/s RDMA over GPU
S9C-8N: Scale Out AI Inference Server

- Flexible chassis design to accommodate different sleds:
  1. Compute Sled: E3/Xeon D/Xeon SP
  2. Compute Node & 1x Accelerator Sled
  3. 2x Accelerator Sled (Nvidia T4 Cloud GPU)
S9C-8N: Different SKUs

8x Nodes
CPU: Accelerator = 1: 0

8x Nodes
CPU: Accelerator = 1: 1

4x Nodes
CPU: Accelerator = 1: 2
### S9C-8N Product Specification

<table>
<thead>
<tr>
<th>Specification</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Form factor</strong></td>
<td>2U8Node</td>
</tr>
<tr>
<td><strong>Processor</strong></td>
<td>Intel® Xeon E Processor (Coffee Lake-S)</td>
</tr>
<tr>
<td><strong>Chipset</strong></td>
<td>Intel® C240 Series Chipset (Cannon Lake, C246)</td>
</tr>
<tr>
<td><strong>Memory</strong></td>
<td>(4) 2666MHz DDR4 UDIMM ECC</td>
</tr>
<tr>
<td><strong>Network</strong></td>
<td>Option 1: (2) 10Gbe BCM57412 LoM</td>
</tr>
<tr>
<td></td>
<td>Option 2: (1) 25GbE BCM57412 LoM*</td>
</tr>
<tr>
<td><strong>Storage</strong></td>
<td>(1) Fixed SATA SSD</td>
</tr>
<tr>
<td><strong>Onboard storage</strong></td>
<td>Option 1: (2) PCIe/SATA M.2</td>
</tr>
<tr>
<td></td>
<td>Option 2: (2) NF1</td>
</tr>
<tr>
<td><strong>Expansion slot</strong></td>
<td>(1) PCIe Gen3 x8 Expansion Slot (Optional)</td>
</tr>
<tr>
<td><strong>PSU</strong></td>
<td>(2) 1600Watt 73.5mm Platinum</td>
</tr>
</tbody>
</table>

*Note: The dimensions of the server are: 86.3mm (3.4’’), 447mm (17.6’’), and 761mm (29.9’’).*
**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 Nvidia T4 cards
- 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*
SD2H-1U System Front View
Flexible Front I/O Module Design

Inference SKU
Support up to 2x T4 GPU or 1x dual width GPU (V100)

Networking SKU
Support up to 2x NIC Card

Cache Accelerator SKU
Support up to 2x U.2

- Power Button*1
- ID Button*1
- 10GbE*4
- RJ45*1
- USB 3.0*2
- Serial*1
- PSU*2
Flexible Module Assembly

Module Design
2.5” 15mm U.2*2

FHHL Module slot Design Options

GPU (Dual width) / FPGA (FHHL)

Front Haul NIC with POE/POE+

NIC Card/Smart NIC
**SD2H-1U System Rear View**

- Support rear AC socket for flexible PSU cable routing
- Support hot swappable easy service fans
- Support wifi / LTE modules for uCPE application
QCT solution of SD2H-1U

Training / Inference SKU: Support up to 2x single width GPU or 1x dual width GPU
- Modular design for both single/dual width GPU

Networking SKU: 4x 10G SFP LoM + 2x extra NIC Card
- Flexible I/O module for networking scalability

Accelerator SKU: Support up to 2x FPGA
- Accelerate network traffic for high bandwidth connection up to 100G
- Improve capabilities and performance of the CDN
- Smart NIC offload CPU loading, reduce the latency

CDN: content delivery network
SD2HQ-2U
Storage Upgrade for vRAN/MEC Edge Server

- **400mm** chassis depth
- Up to **8x** SATA SFF for local storage
- **Flexible** Front I/O module design
- Support **-48V DC** PSU
- Support both **Front & Rear** power inlets

Innovative Storage-Optimized Infrastructures for Telecom Ecosystem
SD2HQ-2U System Front View

Support up to 8x SATA drives for storage extension
Support wifi / LTE module for uCPE application
QCT solution of SD2HQ-2U

Storage SKU:
Support 8x SATA drive, 2x NVMe drives
- Up to 64TB to support local storage

Back Haul Switch SKU: (in planning)
Support 1GbE/25GbE Switch
- Support Data switch
- Support Management switch

Front Haul Switch SKU: (in planning)
Support 1GbE/25GbE Switch with POE/POE+
- Support switch module
- Support 15.4W/ port
NCHC AI Cluster
Liquid Cooling on AI Servers for NCHC Taiwan

252x GPU servers
2016x NVidia Tesla V100

*NCHC: National Center for High Performance Computing
**TOP500 List - November 2018**

*R_{\text{max}}* and *R_{\text{peak}}* values are in TFlops. For more details about other fields, check the TOP500 description.

*R_{\text{peak}}* values are calculated using the advertised clock rate of the CPU. For the efficiency of the systems you should take into account the Turbo CPU clock rate where it applies.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Site</th>
<th>System</th>
<th>Cores</th>
<th>R_{\text{max}} (TFlop/s)</th>
<th>R_{\text{peak}} (TFlop/s)</th>
<th>Power (kW)</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>National Center for High Performance Computing, Taiwan</td>
<td><strong>Taiwania 2 - DCT QuantaGrid</strong> D52G-4U/LC, Xeon Gold 6154 18C 3GHz, Mellanox InfiniBand EDR, NVIDIA Tesla V100 SXM2 Quanta Computer / Taiwan Fixed Network / ASUS Cloud</td>
<td><strong>170,352</strong></td>
<td><strong>9,000.0</strong></td>
<td><strong>15,208.2</strong></td>
<td><strong>798</strong></td>
</tr>
</tbody>
</table>
Ranked #10 in Green500 List

Green500 List for November 2018

Listed below are the November 2018 Green500’s energy-efficient supercomputers ranked from 1 to 10.

Note: Shaded entries in the table below mean the power data is derived and not measured.

<table>
<thead>
<tr>
<th>Rank</th>
<th>System</th>
<th>Cores</th>
<th>Rmax (TFlop/s)</th>
<th>Power (kW)</th>
<th>Power Efficiency (GFlops/watts)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>Taiwania 2- QCT QuantaGrid D52G-4U/LC, Xeon Gold 6154 18C 3GHz, Mellanox InfiniBand EDR, NVIDIA Tesla V100 SXM2, Quanta Computer / Taiwan Fixed Network / ASUS Cloud / National Center for High Performance Computing / Taiwan</td>
<td>170,352</td>
<td>9,000.0</td>
<td>798</td>
<td>11.285</td>
</tr>
</tbody>
</table>

Datacenter PUE = 1.09 ~ 1.17

Energy Saved in 5 years (100% loading):

USD $8,464 per system
USD $78,993 per rack