Inspur SC19 OCP AI Solution Presentation

Gavin Wang Customer Product Manager
About Inspur

- **Inspur Server Retains Fastest Growth in 1HY19**
  - World No.3 in shipment and revenue for 5 consecutive quarters
  - Server Revenue 60%+ CAGR in past 5 years based on Inspur’ financial report
  - The only OEM that retains positive growth in 1H19

- **AI Market Leadership**
  - 2018 AI Server Market Share by Vendor
    - Inspur 50.1%
    - H3C 5.7%
    - Huawei 14.5%
    - Lenovo 10.7%
    - Others 4.0%
    - Dell 6.6%

  - #1 Inspur AI server shipment in China, with MS 50.1%, growth by 103.6% vs 2017

  - 2019/1- Inspur named as SPEC ML 1st Chairman

Worldwide Server Unit Shipments in Q2, 2019

<table>
<thead>
<tr>
<th>Ranking</th>
<th>Vendor</th>
<th>2Q19/2Q18 Unit Growth</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Dell</td>
<td>-16.8%</td>
</tr>
<tr>
<td>2</td>
<td>HPE</td>
<td>-5.8%</td>
</tr>
<tr>
<td>3</td>
<td>Inspur</td>
<td>14.6%</td>
</tr>
<tr>
<td>4</td>
<td>Lenovo</td>
<td>-19.2%</td>
</tr>
<tr>
<td>5</td>
<td>ODM direct</td>
<td>-7.3%</td>
</tr>
</tbody>
</table>

Worldwide Server Vendor Revenue in Q2, 2019

<table>
<thead>
<tr>
<th>Ranking</th>
<th>Vendor</th>
<th>Growth Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Dell</td>
<td>-13%</td>
</tr>
<tr>
<td>2</td>
<td>HPE</td>
<td>-3.6%</td>
</tr>
<tr>
<td>3</td>
<td>Inspur</td>
<td>32.3%</td>
</tr>
<tr>
<td>4</td>
<td>Lenovo</td>
<td>21.8%</td>
</tr>
<tr>
<td>5</td>
<td>ODM Direct</td>
<td>-22.9%</td>
</tr>
</tbody>
</table>
Powerful AI Servers

**Training**

- **AGX-2**
  - World’s highest density 2U8GPU server with NVLink-Empowered

- **FP5295G2**
  - OpenPower CPU & GPU NVLink Tight Coupling

- **NF5488M5**
  - World’s first NVSwitch-empowered 4U8GPU Server, Optimized for Parameter Transfer

- **AGX-5**
  - One of the Most Powerful AI Server 2 PetaFLOPS Tensor Computing Performance

**Inference**

- **NF5468M5**
  - Optimized for AI Inference
  - Support up to 20*T4/MLU100/ NNP-I
  - Extreme storage of 384TB

- **I48**
  - Optimized for High Density Inference Computing accelerated by latest VNNI;
  - Up to 448 CLX CPU cores in 4U chassis

**Edge**

- **NF5260M5**
  - Edge AI Computing Support 2*FHFL/ 4*FHHL PCIe GPU Wall-mounted
Inspur Meta Brain Eco-Plan: Join to from great power and make difference

Meta Brain

- "Algorithm Partner" – AI companies able to develop core AI capabilities
- "Solution Partner" – SIMISV able to deliver total solution for industries

Ecosystem

Efficient Innovation
AI Computing Platform
- Powerful AI servers
- FPGA card

Agile Collaboration
AI Resource Platform
- AI Station (One-stop AI development platform)
- T-Eye (AI performance profiling & tuning tool)

Instant Delivery
AI Algorithm Toolkits Platform
- AutoML Suite
- Open-source TF2 FPGA tool kit
- Inference cluster
- LMS (Large model support for GPU system)
Cluster Cloud Empowered by Inspur AI Station Solutions

AIStation Case Study: Video Processing

60+ Training Servers, 500+ GPUs; 6 Algorithm Teams, 70+ Members

AI Station Development Platform

Task-Balance Strategy: S-3 H-10; Resource-Quota Strategy: 24
Resource-Group Strategy: P100, V100; GPU Sharing Strategy: 6

Sharing Strategy: 32 GPUs support simultaneous development and testing for 70 people;

Task Queuing: nights and holidays are taken advantage of, obtaining 20% improvement in utilization;

Dynamic Allocation: Developers work on 4-5 tasks simultaneously, reducing the development period to 1/3;

Resource Utilization: improves from 70% to 90%.

<table>
<thead>
<tr>
<th>P100_share</th>
<th>P100</th>
<th>V100</th>
<th>2080Ti</th>
<th>1080Ti_dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quantity: 96GPU</td>
<td>Quantity: 120GPU</td>
<td>Quantity: 64GPU</td>
<td>Quantity: 120GPU</td>
<td>Quantity: 32GPU</td>
</tr>
<tr>
<td>Sharing: 3</td>
<td>Sharing: 0</td>
<td>Sharing: None</td>
<td>Sharing: None</td>
<td>Sharing: 6</td>
</tr>
<tr>
<td>Purpose: Training</td>
<td>Purpose: Training</td>
<td>Purpose: Training</td>
<td>Purpose: Robots, Face Recognition</td>
<td>Purpose: Development &amp; Testing</td>
</tr>
<tr>
<td>Quota: None</td>
<td>Quota: 8</td>
<td>Quota: None</td>
<td>Quota: 8</td>
<td>Quota: None</td>
</tr>
<tr>
<td>SSD Caching</td>
<td>SSD Caching</td>
<td>SSD Caching</td>
<td>SSD Caching</td>
<td>SSD Caching</td>
</tr>
</tbody>
</table>

© 2017 INSPIR. ALL RIGHTS RESERVED.
Leading OAI (Open Accelerator Infrastructure) Design

- Open Accelerator Module (OAM) spec released
- Universal Baseboard (UBB) spec in development
- Inspur showing first UBB today
- Prototype samples today at the Summit
- Inspur announces developer systems support

OCP CTO Bill Carter show Inspur UBB on OCP Global Summit

Baidu & Inspur
X-Man 4.0
Announced in OCP summit, Amsterdam

Inspur
21” OAM reference Platform

NF5488M6
6U 8x AI Acceleration Module
Inspur AI Server for Open Compute Object (OCP)

- 21-inch 3OU System
- Flexible JBOG Solution
- 8*OAM Modules with Universal Backplane
- Fully Connect Sale out
- 8* PCIE x16 High IO Expansion
- Deep Learning Training & Training Cluster
Inspur AI Server for Open Compute Project (OCP)

Mission Bay Project

- 21-inch 3OU JBOG System
- Flexible topology for different application
- 8* NVidia Tesla V100
- 6* PCIE3.0 x16 FHHL for IO Expansion
- Open Rack V2 Compatible
Inspur AI Server for Open Compute Project (OCP)

Whistler Project

- 19-inch 3U 4-Socket Head node
- Olympus Architecture Compatible
- 4* Xeon Processors with 24* DIMMs
- 10* PCIe3.0 x16 slots for IO Expansion
- 1600W PSU 2+2 Redundancy

Olympus 4S system can serve as head node to connect with 1, 2 or 4 GPU expansion boxes (such as HGX-1).

Bring up to 8* PCIe x16 links for CPU-GPU communications. Great scale-up capacity for large neural network models.

Support more CPU cores and memory capacity. Improved 42% training performance on certain deep learning framework comparing to 2S.
Inspur AI Server for Open Compute Project (OCP)

4S-8GPU Base Mode (4 links to 4 GPUs)

4S-8GPU VS Mode (8 links)

4S-16GPU Base Mode (8 links)

Flexible CPU-GPU interconnection topologies based on different workloads.
Thank you