AI Solutions and Use Cases Up Close

Dolly Wu, Vice President/GM
Alfie Lew, Solution Architect
Inspur Quick Overview

4 Business Groups

- Inspur Business has expanded to more than 104 countries and regions, and growing rapidly.
- Inspur now owns 8 R&D Centers, 7 Manufacturing Centers, 26 Branch Offices, 12,000+ Partners worldwide.

$10.2 Billion
FY16 Revenue

27,000+
Employees

Inspur HQ @ Jinan, China

Server Manufacturer Worldwide
NO. 1
Server Brand in China

Inspur HQ @ Jinan, China
### Artificial Intelligence & Robotics – Global Market Size

Source: Capital Group

<table>
<thead>
<tr>
<th>Segment</th>
<th>2018</th>
<th>2020</th>
<th>2025</th>
<th>2030</th>
<th>2020-2030</th>
</tr>
</thead>
<tbody>
<tr>
<td>Military</td>
<td>$7.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Drones</td>
<td>$14</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agriculture</td>
<td>$16.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Home &amp; Entertainment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Health Care</td>
<td>$19.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Industrial</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$55</td>
</tr>
<tr>
<td>AI-Based Analytics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$70</td>
</tr>
<tr>
<td>Transportation</td>
<td></td>
<td></td>
<td></td>
<td>$87</td>
<td></td>
</tr>
<tr>
<td>Financial</td>
<td></td>
<td></td>
<td>$255</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Global AI and robotics market size estimates by segment (financial represents assets under management), based on forecasts between 2018 and 2030 (in billion U.S. dollars).
AI Cumulative Revenue, Top 15 Use Cases, World Market: 2016-25

(Source: Tractica)
AI Revenue by Technology: 2025

(Source: Tractus)
Inspur’s Leadership in AI Solutions: 3 Case Studies
Case Study 1: Alibaba

Inspur is the No.1 Server supplier for Alibaba

Inspur ODCC SR Rackscale infrastructure and AI Solutions supported the massive amount of data analytics, images search processing and floating point computation with single precision to handle on Single Day 2016:

- 175,000 orders per second
- 120,000 payment transactions per second
- Delivery to 235 countries and regions
- 680 million+ logistic orders

Single Day 11/11/2016 eCommerce
Revenue: $17.8B USD for one day.

- Precise Recommendation
- Anti-Fraud Analysis
- Intelligent Warehouse
- Intelligent Logistics
Case Study 2: Baidu

- Baidu invested a significant amount of effort in developing artificial intelligence and became the world leader in voice, image and autonomous driving technology. Baidu Brain TTS is currently being requested 250 million times per day and delivers a speech recognition rate of 97% and a face recognition rate up to 99.7%.

- The robot Xiaodu, based on Baidu AI technology, passed 3 rounds at a stretch in the reality TV show called "Super Brain" and entered the final of "King of the brain".

- Inspur is the #1 Server Supplier to Baidu and has cooperated with Baidu for years to establish data centers and jointly develop computing architecture for AI and develop various AI products.

- Jointly developed and released the world’s first 4U16 GPU BOX (SR-AI, for Baidu Driverless Car Project.)
Inspur AI Solutions enable BMW to bring new technologies for vehicle digitalization (IOV – Internet of Vehicle) for the car to become a new living space for customers. BMW Smart Cars taking IoT on the road to make life easier and more efficient. Standard production BMW cars will be able to partially or fully drive itself by 2021.
Practical use of AI in China is adopted much faster than USA

- 560 million users
- Millions of merchants
- 8.08 billion red packets / day

- Resident Situation
- Identity Characteristics
- Consumption Power
- Contractual Capacity
- Travel

Credit Assessment

- BAD CREDIT
- GOOD CREDIT

Inspur is the leading provider of AI solutions
Market share more than 60% in China.
Inspur: China AI Computing Market Performance

Inspur as Preferred AI Server Partner for Large CSPs in China

Inspru as Priority Supplier

BAT GPU Server Vendor Share

Inspur 90%

Others 10%
Practical Use of AI in China: Example 1

- More than 150 cities
- More than 10,000,000 bikes
- 10TB data per day
- More than 120 million users
- 30,000,000 rides per day

Way of travel: time
Traffic: location
Weather: ......

Traffic Planning
Intelligent Operation
Product Optimization
Prediction of Rides
Practical Use of AI in China: Example 2

- >22K KMs
- >23000 trains
- >1000 sensors/train
- >1.4 billion passengers/year

Temperature, Humidity, Traffic, Electricity, Pressure, Speed, ……

Intelligent Control
Intelligent Maintenance
Travel Prediction
Practical Use of AI in China: Inspur Intelligent Factory

Delivery cycle shortened from 15 days to 5-7 days
Production efficiency increased by 30%
Reduced human workforce by 60%
✓ 6 integrated core information systems: ERP enterprise resource management system, MES precise manufacturing system, WMS warehouse management system, WCS warehouse control system, AGV and RGV system.
✓ 600+ RFID, 4000 sensors+, 50 equipment controllers, 330 sets of smart devices
✓ 30 working procedures, 120 information collection points
✓ 25 strict product tests, product yield rate reaches 99.6%

Inspur Intelligent Factory

https://youtu.be/Ojo-j9uzKRI
End-to-end Optimized AI Platform

- **E2E Vertical Solution**
- **Optimized Framework**
- **Comprehensive Management Suite**
- **Cutting-edge Hardware Product**

- AI Solution
  - GPU/Phi/FPGA
- Caffe-MPI
- AIStation
- T-Eye
- GPU Server
- Xeon Phi Server
- FPGA

**Inspur AI Platform**
# Inspur AI GPU Server Family

The most comprehensive AI GPU server portfolio with cutting-edge design

<table>
<thead>
<tr>
<th>Density</th>
<th>Rack-mount / Blade Server</th>
<th>AI Box</th>
<th>Hyper-Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>4GPU/U</td>
<td>AGX-2(NF5288M5) 8x PCIe GPU Oct. 2017</td>
<td>AGX-2(NF5288M5) 8x V100/P100 NVLink GPU Oct. 2017</td>
<td>16 GPU BOX Planning SR-Al Rack 16x PCIe GPU Apr. 2017</td>
</tr>
<tr>
<td>2GPU/U</td>
<td>NF5280M5 4x PCIe GPU Sep. 2017</td>
<td>NF5288M4 4x PCIe GPU</td>
<td>GPU BOX or 1S GPU SV 4x PCIe GPU Oct. 2017</td>
</tr>
<tr>
<td>1-1.5 GPU/U</td>
<td>NF5280M4 3x PCIe GPU</td>
<td>NF5568M4 4x PCIe GPU</td>
<td>NX5460M4 &amp; 2xGPU for InCloudRack</td>
</tr>
</tbody>
</table>
AIStation: Deep Learning Cluster Management Platform

• Full workflow management, including sample data pre-processing, DL Training, Result Analysis and Inference
• Supporting multiple frameworks: Caffe, TensorFlow, CNTK, Mxnet, etc.

**Key Value**

- Deployment of DL environment
- Management of Training tasks
- GPU Resources Management
- Monitoring of CPU/GPU running state
- Cluster resources reports

**Image management**  **Task scheduling**  **Statistics & report**  **Monitor**  **Application analysis**
## T-Eye: AI Application Performance Tuning Tools

- Identify/analyze runtime features, hotspots and bottlenecks of AI application
- Support CPU/GPU Cluster analysis with visual graph
- Use less than 1% of system resource.

<table>
<thead>
<tr>
<th>Tool</th>
<th>CPU Utilization</th>
<th>GPU Memory Utilization</th>
</tr>
</thead>
<tbody>
<tr>
<td>TensorFlow</td>
<td><img src="image1" alt="CPU Graph" /></td>
<td><img src="image2" alt="GPU Mem" /></td>
</tr>
<tr>
<td>Caffe</td>
<td><img src="image3" alt="CPU Graph" /></td>
<td><img src="image4" alt="GPU Mem" /></td>
</tr>
<tr>
<td>MXNet</td>
<td><img src="image5" alt="CPU Graph" /></td>
<td><img src="image6" alt="GPU Mem" /></td>
</tr>
</tbody>
</table>
AI Application Optimization

Kmeans Algorithm
Search by Image

DNN
Speech Recognition

Neural Network
Network Traffic Identification

- Speed up CPU vs. GPU
- Speed up 1GPU vs. 16GPU
- Speed up CPU (Python) vs. CPU (C+MKL)
Inspur AI End-to-End Solution

Smart Voice  Smart Image  Smart Traffic  Smart Medical
Smart Finance  Smart Aerospace  Smart Education  Smart House

AI Computing Hardware  System Design  Solution Validation  Performance Optimization  Application Delivery
Thank You

Inspur Systems, Inc.:  
www.inspursystems.com  
Dolly Wu, VP/GM, dollywu@inspur.com  
Alfie Lew, Solution Architect, alfie@inspur.com