Inspur OAI Product Introduction
Jan 2021
# Inspur Full-Stack AI Portfolio

## E2E AI Solution

- Internet
- Telecom
- Finance
- Education
- Smart City
- Media
- Manufacture
- Medicine

## AI Algorithm Toolkit Platform

- **AutoML Suite**
  - Cloud& On-premise Deployment
  - Automatic modeling
  - Automatic tuning
  - Automatic cropping

- **TF2**
  - Lossless model in accuracy
  - Speed up FPGA development

- **LMS**
  - Self-developed AI model computing framework, supporting GPU large-scale training

- **TensorFlow-opt**
  - Optimized TensorFlow framework with the fastest AI training speed on the public cloud, 512 GPU expansion efficiency 90%

- **Caffe-MPI**
  - One of the first parallel versions of Caffe framework

## AI Resource Platform

- **AIStation (Training Platform)**
  - Model Development
  - Model Deployment
  - Application Development

- **AIStation (Inference Platform)**
  - Compatible with multiple DL frameworks
  - Support AI model online testing and evaluation
  - Multi-model deployment and weighted calculation

## AI Computing Platform

- **Accelerator Card**
  - F10A
  - F37X
  - N10X
  - N20X
  - F10S
  - F07V

- **Server**
  - Training
  - Inference
  - Edge

- **General Server and Open Compute**

## AI Application and Framework Feature Analyzer

- **T-Eye**
  - AI application and framework feature analyzer
Driving AI Ecosystem---Metabrain

Efficient Innovation
AI Computing Platform
• Industry’s Most Comprehensive AI Server Portfolio
• General Server 2U/4U/6U
• Open Hardware Compute and OAI
• M5 AI Servers, FPGAs, ASIC Cards....

Agile Collaboration
AI Resource Platform
• AIStation: One-stop AI development platform, efficient and flexible computing resource scheduling; easy to deploy AI dev environment
• T-Eye: AI performance profiling and tuning tool, empower AI application optimization

Time to Delivery
Algorithm Toolkit
• AutoML Suite: On-Premise & Cloud deployment; Parallel Acceleration; Effortless Model Generation
• Caffe-MPI: 1st Parallel Version of Caffe
• TensorFlow-Opt: Scale-out TensorFlow on public cloud, optimization on cloud RoCE

Solution Partner
-SI&ISV able to deliver total solution for industries

Algorithm Partner
-Al Companies able to develop core AI capabilities

Metabrain
Inspur is a Key Member in Open Platform Communities.
One-Stop AI Development and Deployment Platform

Data

Computing Resources Utilization

Training Time

40% → 80%

2 days → 4 Hours

Low Model Development Efficiency

Low Utilization of Computing Resource

Easily deploy AI development environment and development process, significantly improving development efficiency

Efficient and flexible platform, obtain AI computing resources on demand to speed up model training efficiency

Model Development and Training

Model Deployment and Inference

The deployment complication to deploy the trained model into production

The deployment complication to get the trained model into production

Seamless connection between model development and deployment, shorten the time of scaling to production

Unified management of multiple models, Centralized scheduling of computing resources, Dynamic allocation, Elastic expansion

Multi-application load balancing and resource elastic scaling

AI Service

PC

Mobile

Manufacture

Robot

IOT
• SAS Switch for pooling HDDs, improving storage flexibility
• NVMe pooling, more nimble resource
• NVMe over Fabric, high Perf storage pooling
• PCIe Switch for pooling GPU, GPU acceleration ratio increases linearly
• GPU/FPGA over Fabric, heterogeneous acceleration remote expansion
**MX1 System Features**

**MX1: World’s First OAI Reference System**

<table>
<thead>
<tr>
<th>Product Model: MX1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Chassis</strong></td>
</tr>
<tr>
<td><strong>Dimensions</strong></td>
</tr>
<tr>
<td><strong>Connection with Compute node</strong></td>
</tr>
<tr>
<td><strong>OAM</strong></td>
</tr>
<tr>
<td><strong>Power without OAM</strong></td>
</tr>
<tr>
<td><strong>PCIe Switch</strong></td>
</tr>
<tr>
<td><strong>PCIe re-timer</strong></td>
</tr>
<tr>
<td><strong>Phy re-timer</strong></td>
</tr>
<tr>
<td><strong>Expansion slots</strong></td>
</tr>
<tr>
<td><strong>BMC</strong></td>
</tr>
<tr>
<td><strong>I/O</strong></td>
</tr>
<tr>
<td><strong>Ambient Working Temperature</strong></td>
</tr>
</tbody>
</table>

---

**OAI Reference System MX1**

- **OCP Certificated 2S Compute Node ON5263M5 (San Jose)**
- **High Density Whisper Cable**
- **4 x HHHL PCIe Expansion**
- **Front I/O connector**
- **1570W without OAMs**
• Solutions for the implementation of the rack Mgmt based on node level
• Southbound manages system resources; northbound presents Info
• Meet the needs of Mgmt encryption and resource pooling

• Relying on vendors maintenance for traditional BMC code base
• Complex to modify the traditional BMC code for new HW
• Poor readability of IPMI tool binary code
Open OAI Rack Level Reference System

48VDC Open Rack
- 1 pairs 48V Bus Bar
- 1 shelf per Rack

Power Shelf
- 33KW(12xPSU)
- 40V-58V
- 93mm (H, 2OU) x 537mm (W, 21”) x 586 (D) mm

System Devices
- Inspur 3OU OAI systems x4
- Inspur 2OU compute node x4

Open RMC + Open Hardware
- RMC WEBUI
- USER TOOLS
- BMC
- Redfish

• RMC Web Server：基于OpenBMC的Rack Manager控制器服务
• RMC Web UI：资源收集及服务配置文件
• 南向接口：支持Redfish RESTful API
• 北向接口：支持Redfish RESTful API并丰富了服务配置文件