inspur

CLOUD COMPUTING AND AI TOTAL SOLUTIONS PROVIDER

# Powering The Open Data Center Revolution

# About Inspur

- Top 3 Server Vender Worldwide
- Leading Full Stack AI Solution Provider
- Leading Open Data Center Products and Solutions Provider

## JDM Model

**J** = Joint digital collaboration with customers & partners

**D** = Design/Development/Delivery

**M** = Manufacture
Online orders,
Fast customization

## AI Full Stack Solution

| | | | |
|---|---|---|---|
| AI Applications Acceleration | GPU/MIC/FPGA | | |
| Optimized AI Framework | Caffe-MPI | | |
| AI Management Software | AIStation | T-Eye | |
| AI Hardware Products | AGX-5 | AGX-2 | FPGA |

## Platinum Member of OCP

ODCC OCP Open19

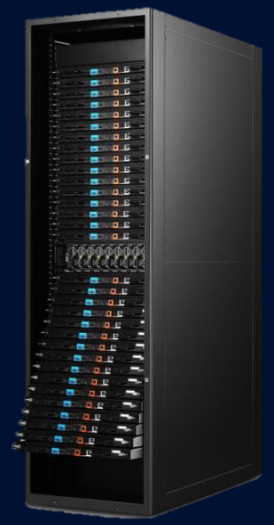
inspur

# Open Hardware Projects for the Hyperscale Datacenter

International Open

Datacenter Platforms

**Inspur is the only server vendor in the world to support all open source hardware projects.**



ODCC Rack

OCP Rack

Olympus Rack

Open19 Rack

InCloud Rack with Intel® RSD

| | | | |
|---|---|---|---|
| OPEN19 Foundation — Lead Vendor | OPEN Compute Project — OCP Platinum Member Solution Provider | ODCC Solution Provider | intel — Intel® Rack Scale Design |
| OpenPOWER™ | Microsoft Project Olympus | Inspur is a Key Member in Open Platform Communities. | openstack FOUNDATION GOLD MEMBER |

inspur

# Open Platform: Making Ecosystem More Open

## Embrace Open

### Open Hardware

**Member:** Facebook, Inspur, Google, Microsoft

Over billion deployment

**Members:** Baidu、Tencent、Alibaba、Intel、Inspur

Millions of nodes deployed

OpenPOWER

OpenPower

### Open Software

OpenBMC

DMTF Redfish

Rack Level Management

## Deliver Value

Open Design Platform

- Accelerate Innovation
- $$ Decrease TCO
- eco Manage Power Consumption
- X Lower Failure Rate
- Easy to Join

## Open to Customers

Baidu 百度 · 阿里巴巴 Alibaba.com · 360 www.360.cn · Tencent 腾讯 · LinkedIn · amazon · 網易 NetEase www.163.com · 搜狐 SOHU.com · LeTV 乐视网 · facebook · UBER · 华数 wasu · sina 新浪网 sina.com.cn · 中国铁路 · Microsoft · Google · IBM

inspur

# Goal for Open Platform

## Contribute IP

Contribute Designs and IP to the OCP Community

## JDM

Jointly Design with Customers

Alibaba Microsoft IBM LinkedIn Tencent UBER

## Design Base

- Architecture
- Compute + Storage + AI Schematic Design
- Interconnection & Pooling
- Rack Management

## Use Cases & Technology Experience

**No.1 Server supplier for Alibaba**
**Alibaba e-Commerce Platform is running on Inspur ODCC Rack**
Peak of system transaction reached 175,000 orders per second

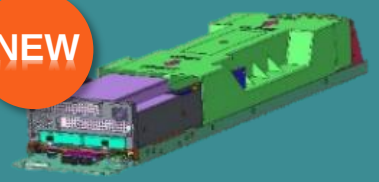
**Over 80% of Baidu's hyperscale rack-level servers Deploying 10,000 nodes per day**

inspur

# Inspur OCP Rack Scale Product Overview

## Compute

NEW



### I/O Balance

- 2x2.5”HDD or NVMe
- 2x M.2
- 1x FHFL(x16) : CPU0
- 1x FHHL(x16) : CPU1

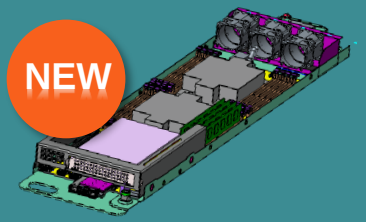
NEW

OCP ACCEPTED

### High Expansion

- 4x2.5”HDD or NVMe
- 1x M.2
- 3x HHHL(x8) : CPU0

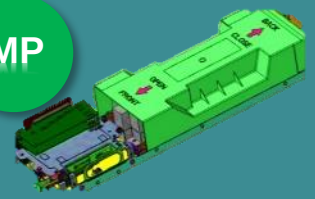
NEW



### Density Computing

- 2x M.2
- Expansion Slot:
- 1x FHHL(x16) : CPU0

MP



### Energy Efficiency

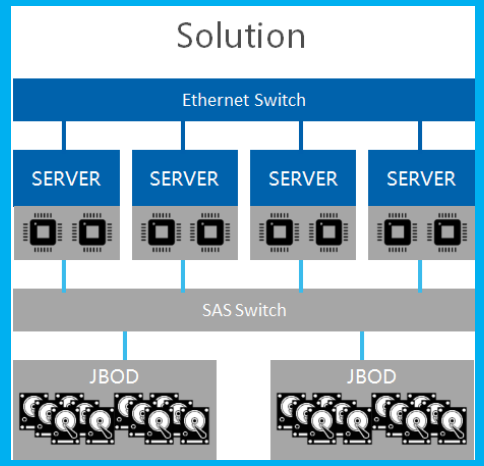
- 1x3.5”HDD
- 1x M.2
- 1x FHHL(x16) : CPU0
- 1x FHHL(x16) : CPU0

## Storage

NEW

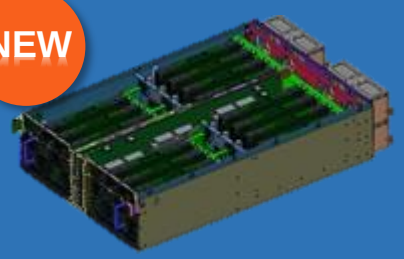
### JBOD

- 2OU form factor
- Two drive tray design
- 34*HDDs or SSD
- Each drive supports hot-plug



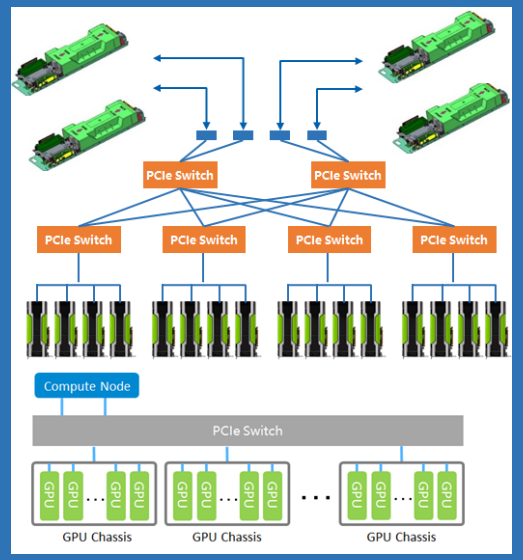

## AI

NEW



### GPU Box

- 4OU form factor
- 16 PCIE GPU cards




inspur

# The Industry and App Trend of AI

Image Recognition

Financial Enterprise

Medical Enterprise

Weather forecast

AI + CSP in 2011

AI+ Traditional Enterprise in 2015

AI + Scientific Computing in 2017

Speech Recognition

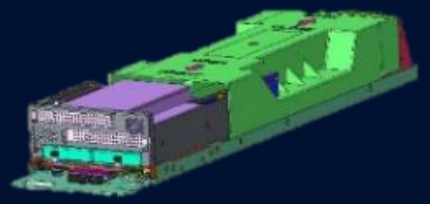
Car Companies

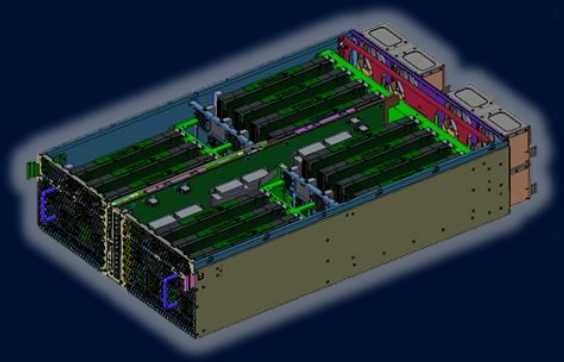
Security enterprise

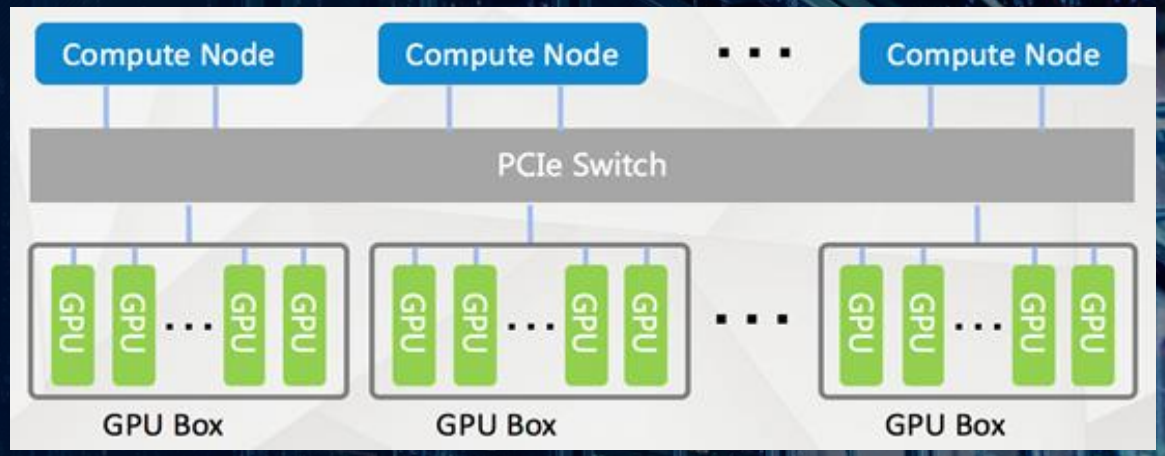
Oil & Gas

inspur

# Innovative Hardware Portfolio – OCP-AI Rack







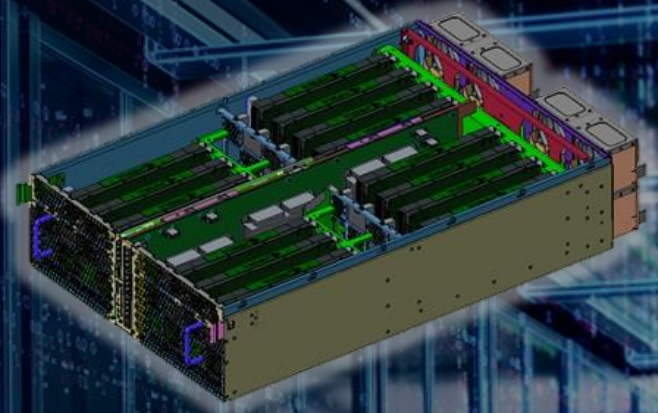

Advantages of Heterogeneous Pooling:

- Communication between GPUs only depends on PCI-e, the latency can be reduced 50%+
- GPU expansion does not need to synchronize high-value IT resources (e.g. IB Switch), the cost can be optimized 5% +
- More GPUs can be supported, deployment density can be increased by 4 times

inspur

# Inspur OCP AI Product

## Scalable and Leading Compute Architecture

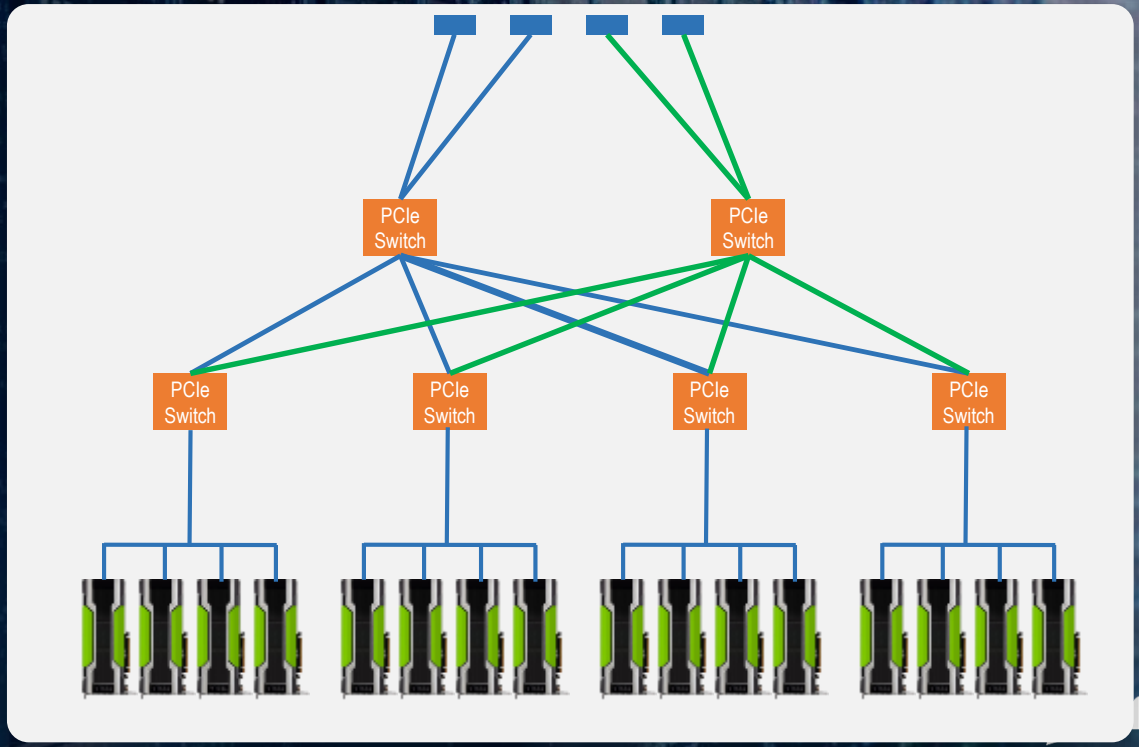


### Ultral High Density GPU

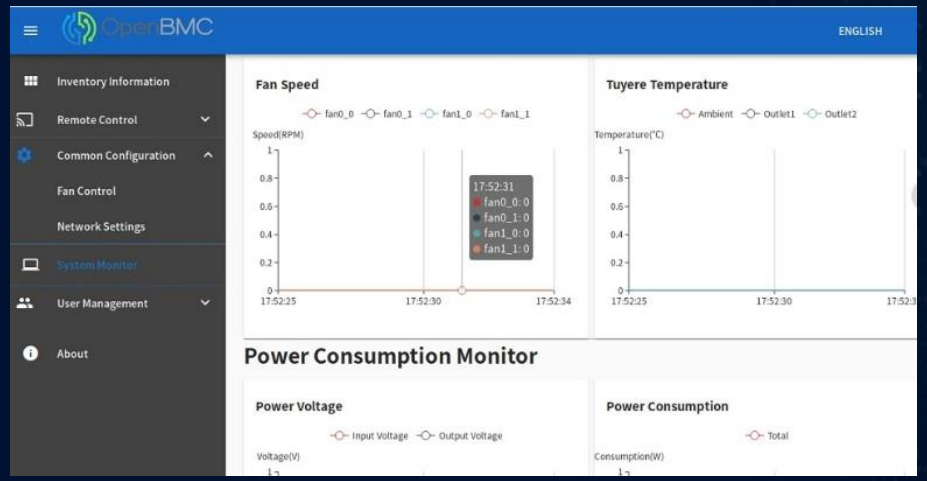
- 4OU 16*GPU for training or 16*FPGA for inference
- Improve efficiency of system's computation expansion

### Optimized Architecture for AI

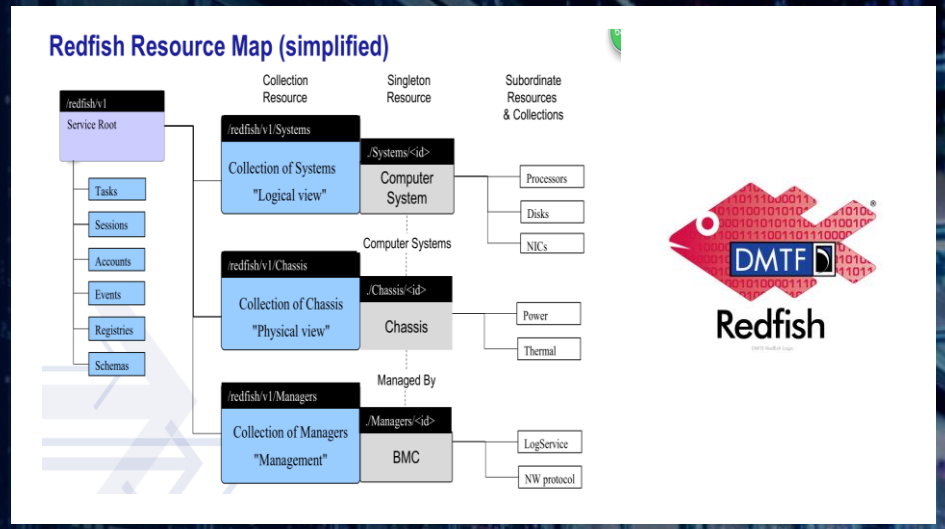
- Speed up the training process
- Flexible architecture to support multi-node based on application



# Inspur OpenBMC Key Features



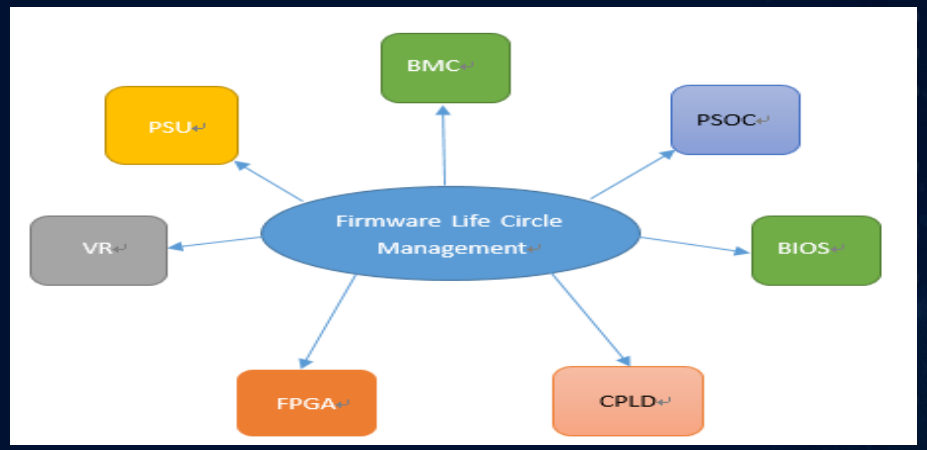

**WebUI** to monitor system status




**Redfish/Restful Function**, support Redfish OCP Baseline Profile

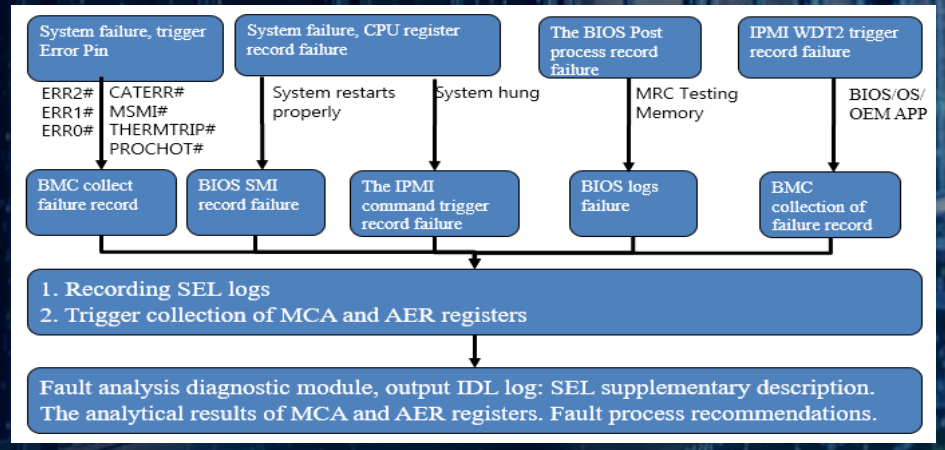
Component **Firmware Life cycle Management**

1. Version auto discovery
2. Intelligent update for BMC,BIOS,CPLD,FPGA etc.
3. Firmware rollback when error occurs.




**Fault Diagnosis**

1. Diagnosis system fault directly
2. Output the detailed fault records and recommendations
3. BMC subsystem fault diagnosis.

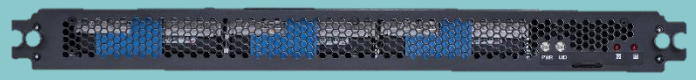



inspur

# Inspur Open19 Product Overview

## Current Product

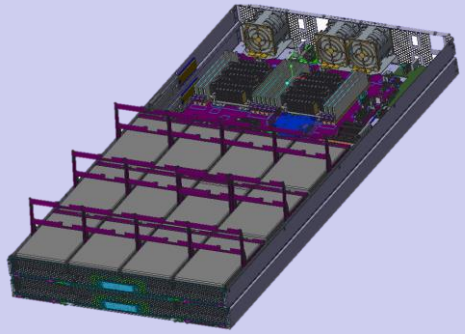
1U Half Brick Compute



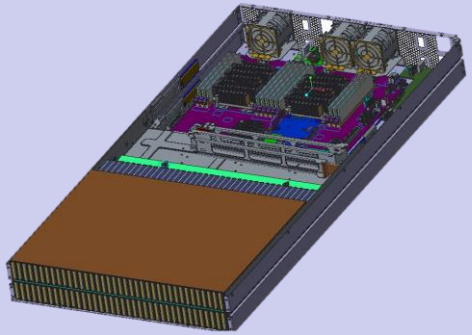
1U Full Brick Storage

Open19 12U Cage

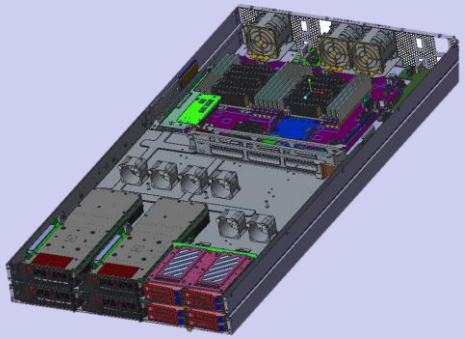
## Next Generation



2U Full Brick Storage



2U Full Brick Flash



2U Full Brick AI

# THANK YOU

Learn more about Inspur at www.inspursystems.com

inspur