

# Inspur HPC Update

Vangel Bojaxhi  
Global AI&HPC Director  
E-mail: [VBojaxhi@inspur.com](mailto:VBojaxhi@inspur.com)

1 Inspur Introduction

---

2 Inspur HPC Products and Solutions

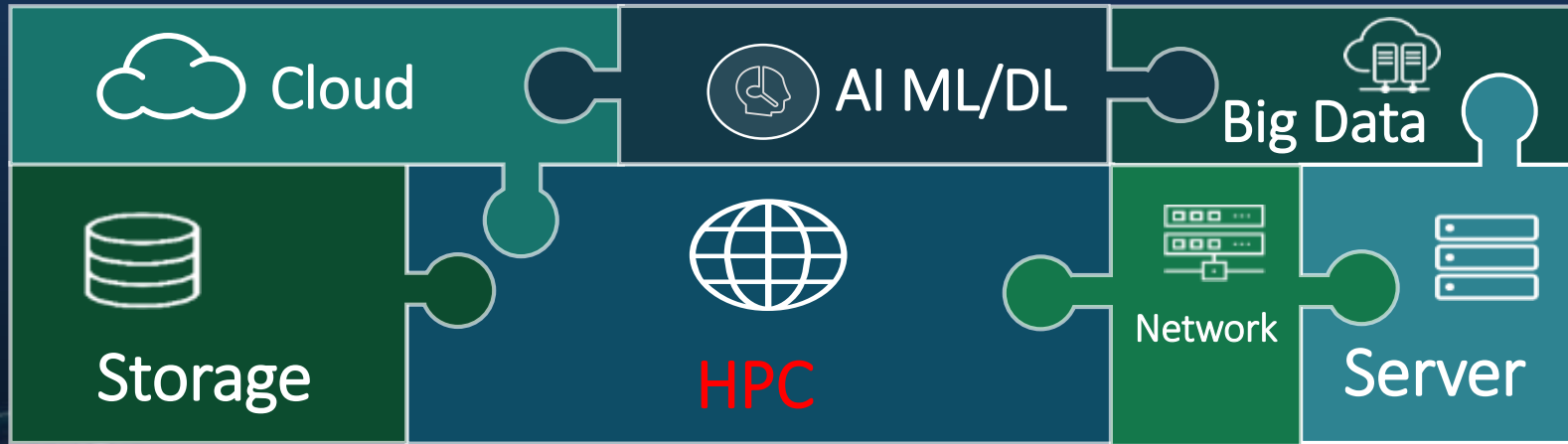
---

# Inspur Company Name



Leading HPC & AI Computing  
E2E Total Solution Provider

# Inspur Product Portfolio



# Inspur Position in the Global Market

Total 68 systems, 13.6%



2017

Top 3 Server Vendor



2017

Top 1 Server Vendor



2022

# Inspur Global Presence



113  
Countries and regions



2  
Global Call Centers



50+  
Service Partners



600+  
Service Engineers



**8**  
Worldwide R&D  
Center

**6**  
Worldwide  
Manufacture

**113**  
Worldwide  
Delivery

**113**  
Worldwide  
Service



## EU OFFICE LOCATIONS

-  **Frankfurt**  
(EU Headquarters)
-  **Stuttgart**  
(EU technical center)
-  **London sales office**
-  **Czech Inspur manufacturing**

# Inspur Europe



- Field Engineer Location
- Field Engineer & Logistics Location
- European parts and delivery center (under construction)

- **Germany, UK, France, Switzerland, Austria, Ireland, Poland, Spain, Romania, Bulgaria:**

- 7\*24 helpdesk
- 3 years standard warranty: NBD onsite service, SBD 4H is available
- Support: field engineers support

- **Other EU countries:**

- 7\*24 helpdesk
- 3 years standard warranty: NBD onsite service
- Support: nearest service bases

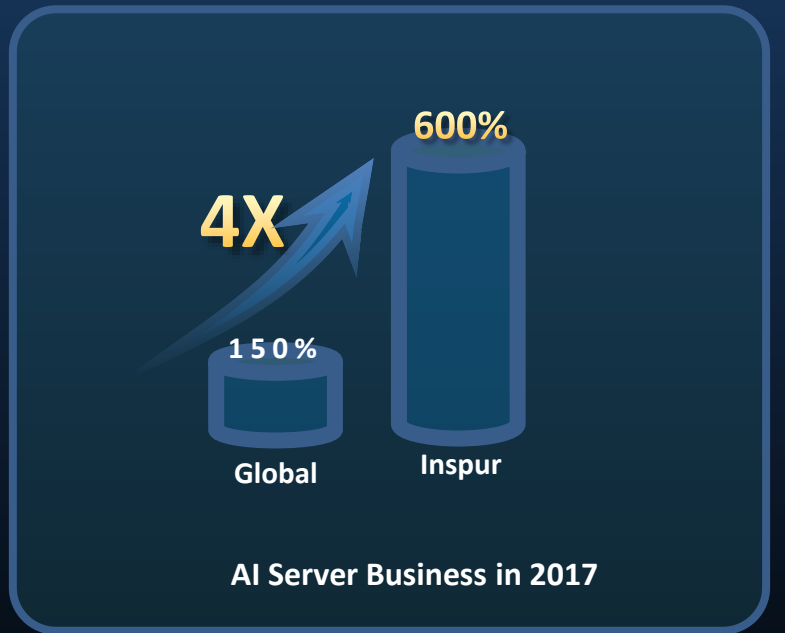
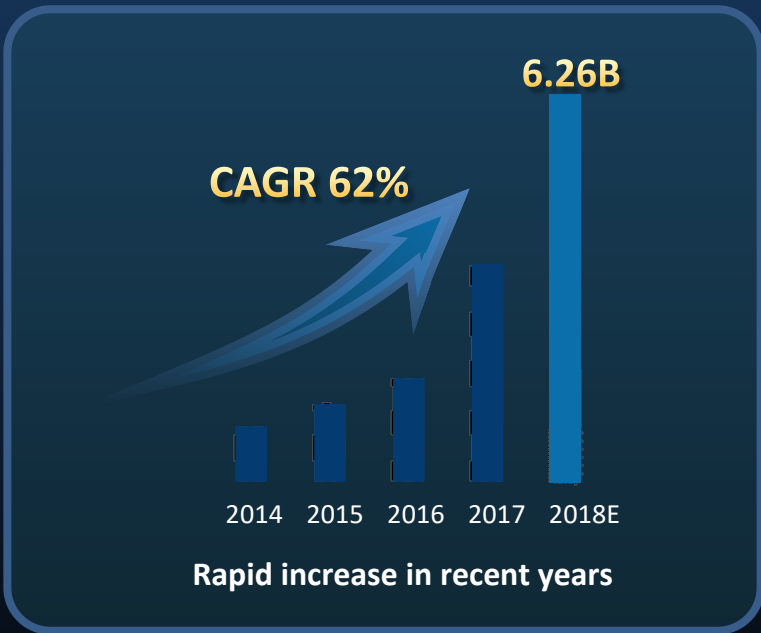
- Partnerships: NEC, Clustervision
- HPC systems deployed in two EU large automakers

# Inspur EU strategy

World-Class Quality Products & Solutions Tailored for Diverse Industries/Application in EU



# Inspur Server Business Growth



# HPC Further Frontiers



## Scientific Calculations

Computational Physics & Chemistry  
/ Cosmology / Material Science  
Weather forecast / Climate research  
Security / Defense / Life Science /  
Molecular Dynamics / Medical Health  
/ Genetic Sequencing



## Engineering Simulation

Earth Science / Energy / Petroleum  
Exploration / Manufacturing / Design /  
Simulation / CAE / Computation Fluid  
CFD / Electronic Design Automation EDA  
/ IC Design / Aerospace / Aircraft Design  
/ Financial Transactions / Risk Analysis



## AI

Big Data analysis  
Deep Learning  
Image / speech recognition  
Self / assisted driving

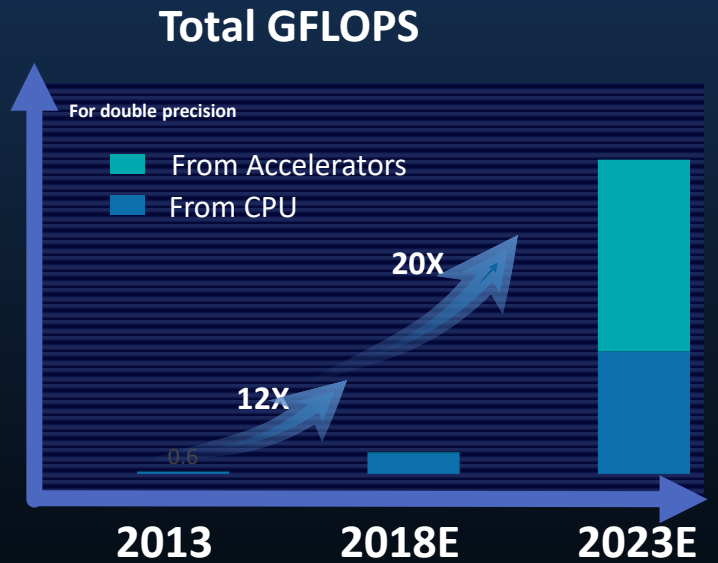
# HPC & AI Convergence

The revolutionary accelerators enable [multi-precision computing](#) that fuses the highly precise calculations to tackle the challenges of HPC with the efficient processing required for AI/DL.



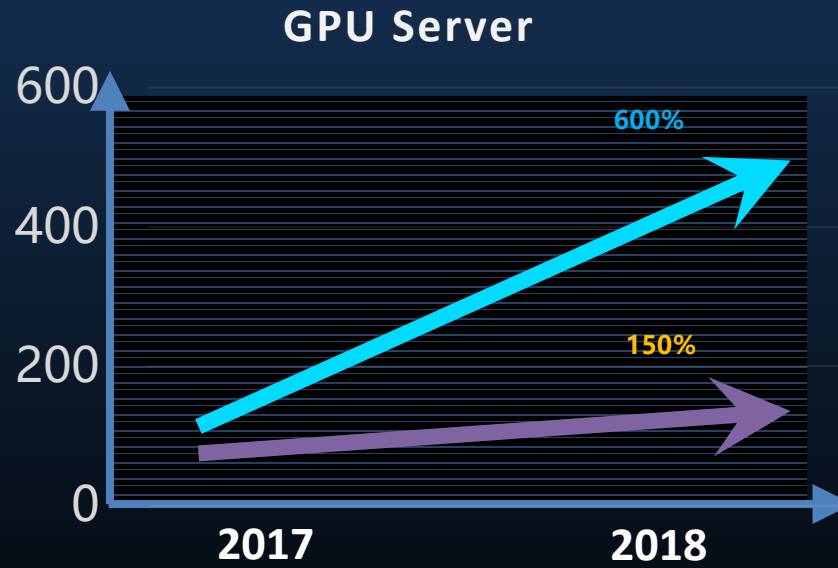
# Inspur GPU Server Market

## Computational Power Forecast



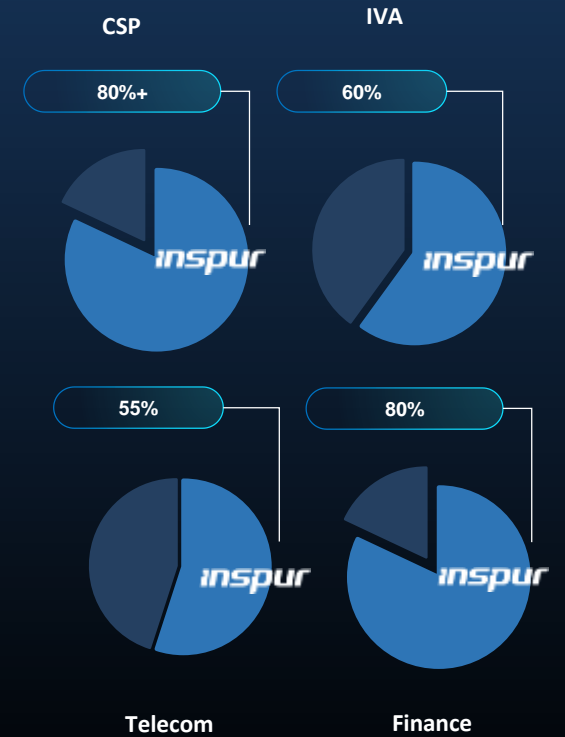
Data source: Inspur Research Institute

## Inspur GPU Server Growth



— Inspur GPU Growth  
— Global GPU Growth

## Inspur AI China Market Share



1 Inspur Introduction

---

2 Inspur HPC Products and Solutions

---

# Pervasive HPC



## Traditional HPC

- Modeling & Simulation
- More iterative methods (stochastic, parametric, ensemble)
- More SMEs



## High Performance Data Analytics

- Today: Knowledge Discovery, BI/BA, Anomaly Detection, Marketing
- Emerging: Precision Medicine, Cognitive, AI, IoT



## HPC Anywhere

- On-Premise
- Cloud (Public, Private, Hybrid)
- Private Hosted

# Inspur HPC Innovation and Partnerships

## Innovation



Purpose Built Infrastructure



Open Source Contribution



Roadmap Alignment

## Strategic Partnerships



NVIDIA®



End-2-End Solutions



Center of Excellence

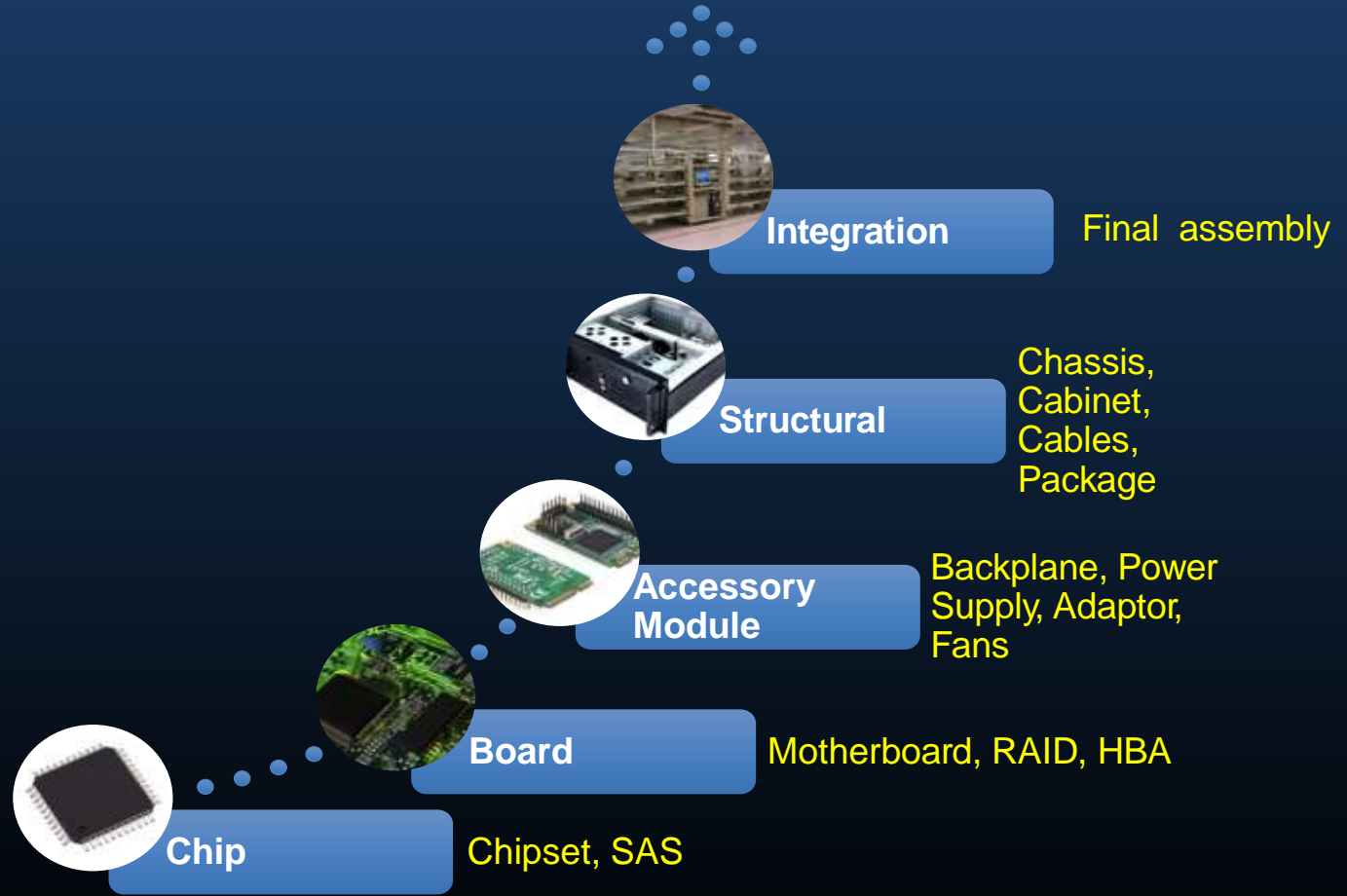


Workload Optimized

# Inspur HPC E2E IP Design & Manufacturing



Top quality ASIC



# Inspur Mainframe Unix Series

- Independent Unix OS: **Inspur K-UX 2.0**
- Successful 70,000 Unix certification tests / benchmarks.



**K1 950**

32-Socket  
256 Cores



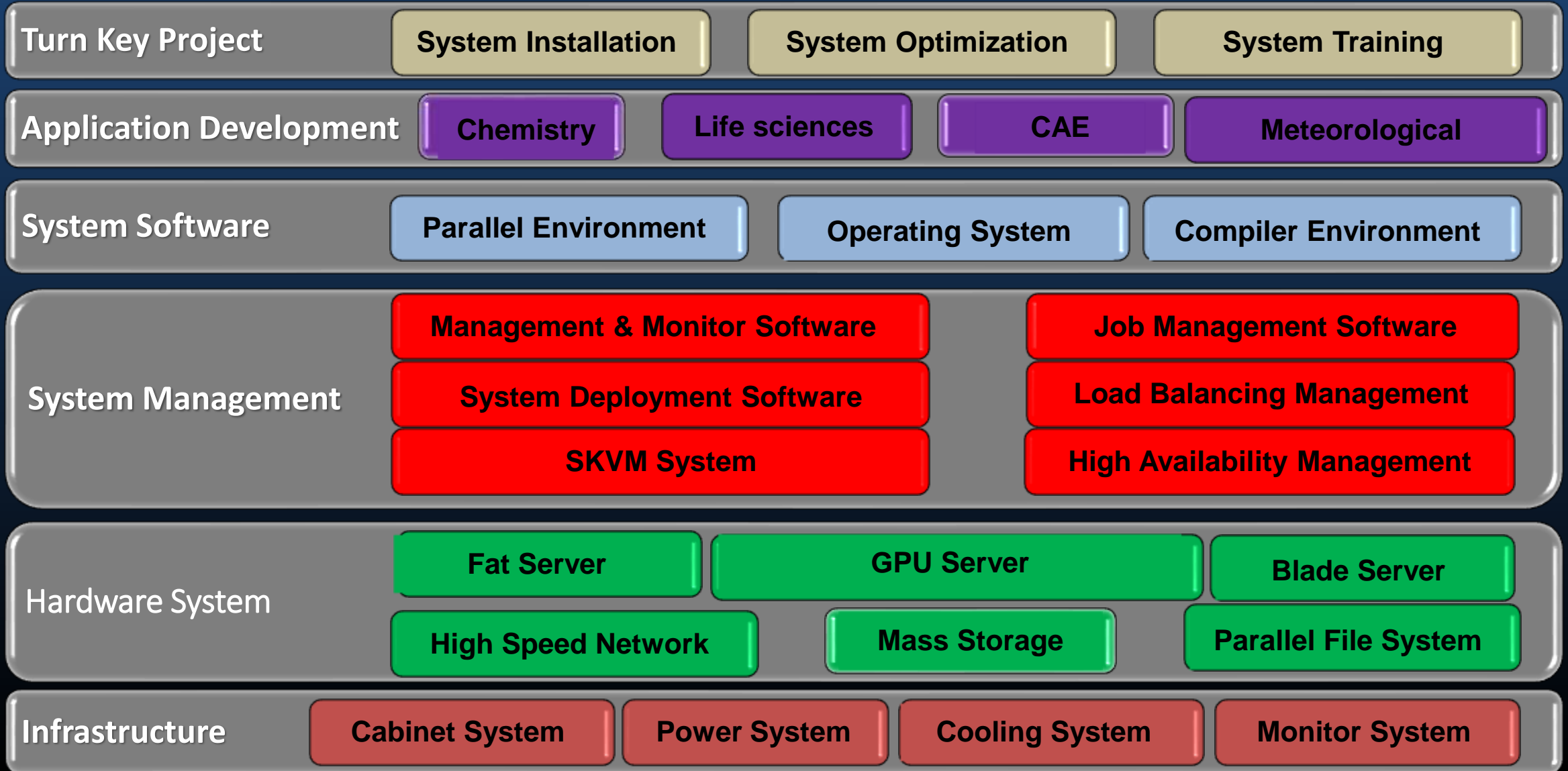
**M13**

64-Socket  
1152 Cores









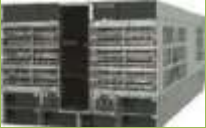





Operating Systems with UNIX Certification	
IBM	IBM AIX 5L/6
HP	HP-UX 11i
Oracle	Oracle Solaris 10/11 FCS
Apple	Mac OS X 10.8
<b>Inspur</b>	<b>Inspur K-UX 2.0</b>



# Inspur HPC Solution



# Inspur HPC Product Stak

<b>Computing</b>	 <b>InCloud Rack</b> High density converged 		 <b>Standard Rack</b>  	
<b>Storage</b>	 <b>IEEL solution</b> <ul style="list-style-type: none"> <li>• Intel HPC enterprise file system</li> <li>• High performance hundreds of GB/s</li> </ul>		 <b>BeeGFS</b> <ul style="list-style-type: none"> <li>• Server Architecture</li> <li>• HA Mirror &amp; BeeOND</li> </ul>	
<b>Network</b>	 <b>FDR/EDR InfiniBand</b>	 <b>Omni-path</b>	 <b>Management</b> <ul style="list-style-type: none"> <li>• 10Gb/1Gb</li> </ul>	
<b>Management</b>	 <b>Cluster Engine</b> Cluster management, configuration, monitoring, alarms, job scheduling & check stop; user account & billing		 <b>TEYE</b> Application run time monitoring & tuning	
<b>Infrastructure</b>	 <b>Water-cooling cabinet</b> <ul style="list-style-type: none"> <li>• Closed cooling cycle, variable speed fan</li> <li>• PUE &lt; 1.1 40KW per Rack</li> </ul>		 <b>Rack Cabinet</b> <ul style="list-style-type: none"> <li>• 42U standard</li> </ul>	

# Inspur Edgy HPC Products



ODCC OCP Open19

The **ONLY** vendor that meets **ALL** Open Standards



i48  
4U8Nodes

**HIGHEST** density for General Computing



NF5486M5  
4U106Disks

**HIGHEST** density for Storage



AGX-5  
8U16GPU

**HIGHEST** Performance / Density AI Computing

# Inspur AI Servers

GTC2017 · San Jose  
**AGX-2**



**AI Training**

2U 8 GPU, NVLink

World's highest density 2U server of 8 highest performance GPUs.

GTC2018 · San Jose  
**NF5280M5-V**



**AI Video**

2U8 P4

specialized optimized for intelligent video analysis.

IPF2018 · Beijing  
**NF5468M5**



**AI Cloud**

4U8 V100/4U16 P4

Elastic GPU server designed for AI cloud.

ISC2017 · Frankfurt  
**GX4**



**PCI-E Pooling**

2U 4GPU BOX

Flexible Expansion, available for 2-16 GPU cards

SC2016 · Salt Lake City  
**F10A**



**AI Inference**

1.5 Tflops

higher density and better performance/watt FPGA

# Inspur Server Line Portfolio

NP5540M5  
NF5290M5  
NF5280M5  
NF5240M5  
NP3020M5  
NF5180M5

General Server

I9000  
NF8260M5  
NF8460M5  
TS860M5

Mission Critical Server

i48  
i24  
NF5486M5  
SN3410M5  
NF5466M5  
SA6224M5  
SA5212M5  
NF5288M5

Optimized Servers

OCS Rack  
OCP Rack  
InCloudRack  
Open19 Rack  
ODCC Rack  
I9000

Converged Rack Servers

# Inspur Server Functionality Classification



## Multi Function Node

- Management
- Login
- Bigdata
- I/O

## Computing Node

- Numerical weather forecast
- Pneumatic fluid calculation
- First principle calculation

## Computing Node

- Protein folding
- Oil seismic data processing
- Hadoop

## GPU Node

- Molecular dynamics
- Monte Carlo simulation
- Deep-learning

## Fat 8Socket Node

- Genomic splicing
- Oil exploration data interpretation
- Implicit Finite Element Analysis

# NF5280M5: Multi Function, Flexible, Open Standard

- High performance
- Large storage
- Large I/O expansion
- OCP network
- AC/DC efficient power supply



Module	NF5280M5
<b>CPU</b>	2 Intel® Xeon® Scalable processors, Maximum TDP 205W
<b>Memory</b>	24 memory slots , Max DDR4-2666 12 NVdimmm
<b>RAID</b>	Hardware-level protection for NVME SSD drives
<b>Storage</b>	Front HDD : max 24*2.5" , or 12*3.5" or 24*NVMe Built-in HDD : 4*3.5"HDD and 2*M.2 SSD BACK HDD : 4*3.5"HDD and 4*2.5"HDD
<b>GPU</b>	Support 4*GPU,P100/P40 etc
<b>I/O</b>	Max 10*PCIe slot
<b>NIC</b>	1*OCP
<b>Power</b>	High-efficiency power supplies with 80 PLUS Platinum and Titanium certifications.

# Inspur Density Optimized Server: i24



- Chassis
  - 2U 4Node
  - Support 24 SFF Or 12 LFF
  - 80 PLUS Platinum and Titanium certifications.
- Node
  - Two Intel® Xeon® SP Processors ,TDP 165W
  - 16 DDR4 2666MHz DIMMs, and NVDIMM support
  - Two M.2 SSD and 2x micro SD cards
  - Two PCIe, EDR and 1 OCP NIC
  - Support 1Gb/10Gb/25Gb/50Gb OCP NIC

# Modular & Scalable High Density Server: i48



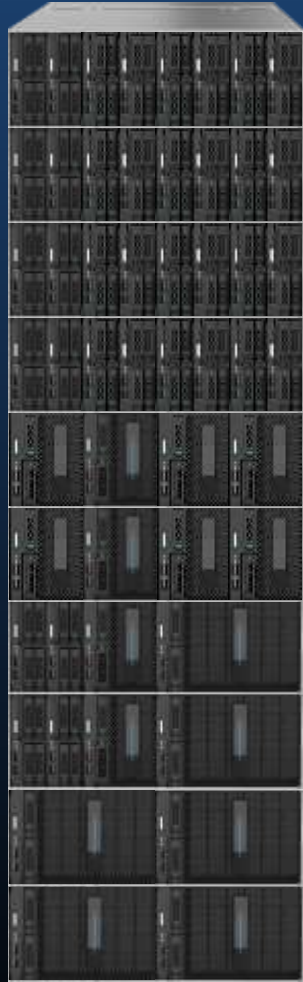
## Flexible node configurations optimized for different applications

- High Density Compute: 8 nodes in 4U, 2S, 16 Dimms, 2 SFF, 1 OCP, 2 PCIE
- Balanced Configuration: 4 nodes in 4U, 2S, 16 Dimms, 12 LFF, 1 OCP, 2 PCIE
- Storage Node : 2 nodes in 4U, 2S, 16 Dimms, 2 SFF & 72 LFF, 1 OCP, 2 PCIE

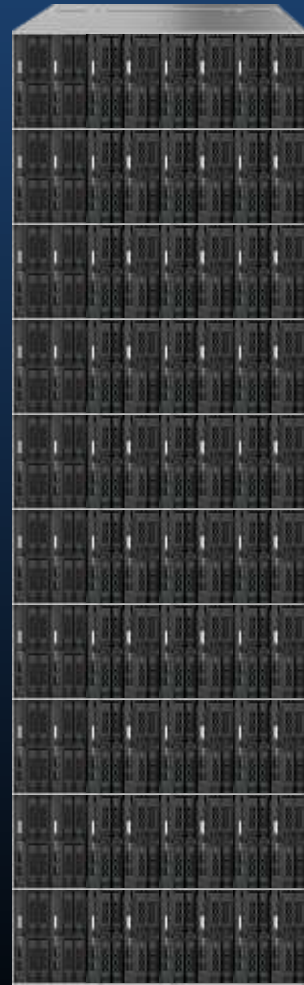
## Flexible expansion and hybrid deployment

- Multiple unit mixed deployment, integrated management, 5X increase of deployment efficiency
- Flexible expansion, significant reduction in data centre infrastructure initial investment.

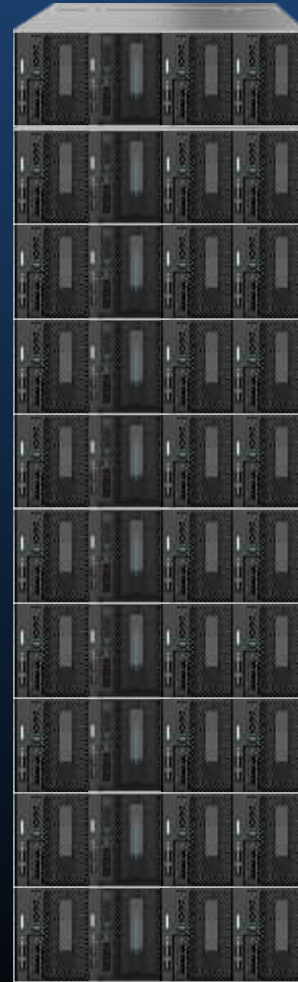
# HPC Data Center on i48



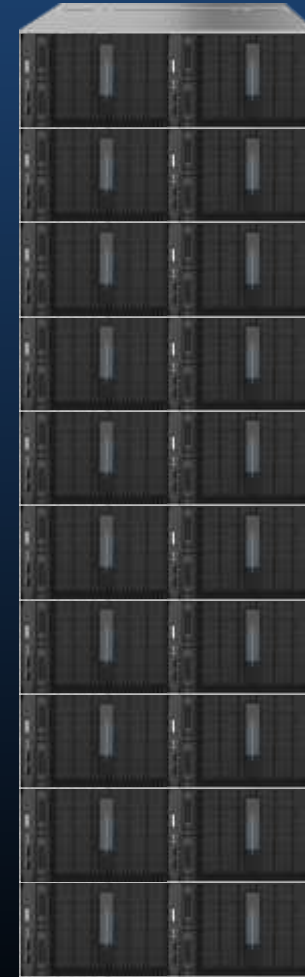
**Mixed**



**Compute**



**General**



**Storage**

# Inspur 8 Socket Sever TS860M5



- 8 CPUs in 4U
- 96 Memory slots, UP to 12 TB
- Up to 1.5X Memory Bandwidth

Module	TS860M5
CPU	8 Intel® Purley Skylake 61xx&81xx , Maximum TDP 205W
Memory	96 memory slots , Max DDR4-2666
RAID	providing hardware-level protection solution
Storage	Up to 24 SFF , SATA/SAS/U.2
I/O	Max 12*PCIe slot
NIC	1*OCP
Power	High-efficiency power supplies with 80 PLUS Platinum and Titanium certifications.

# Inspur Server AGX2



- 8 GPUs in 2U
- Purley system and support NVlink
- Adjustable PCIE interconnect topology, Adapt to different application
- Up to 4 100G Remote Direct Memory Access (RDMA) NIC

## Specifications

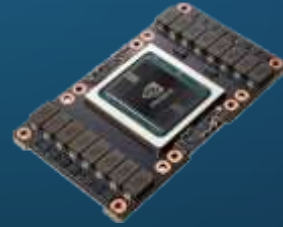
<b>Model Number</b>	AGX-2
<b>GPU</b>	8*NVIDIA® Tesla® NVLink™ V100/P100 or 8*PCle P100/P40/P4
<b>Tensor TFLOPS / TOPS on INT8</b>	960 / 376
<b>CPU</b>	2*Intel® Xeon® Scalable Processors
<b>Memory</b>	16*DDR4-2666
<b>Storage</b>	8*2.5" U.2/SAS/SATA 2*M.2 PCIe & SATA on Board
<b>Network</b>	4*10G Ethernet on board Up to 4*100G RDMA NIC for NVIDIA® NVLink™ GPU Up to 2*100G RDMA NIC for PCIe GPU
<b>Cooling</b>	Redundant Hot Swap System Fans Air cooling /Air-Liquid Hybrid cooling
<b>Power Supply Unit</b>	2*3000w PSU 80plus Titanium

# Inspur Server AGX5



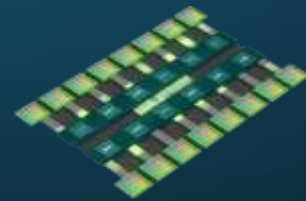
**AGX5**

**The Most Powerful / Dense AI 8U Server**



## Ultimate Performance

- 2 PetaFlops AI computing performance
- 16× Tesla V100 GPU
- Near linear speedup



## Fast GPU Interconnection

- Most advanced NVSwitch™
- 2.4TB/s full-chip cluster high-speed interconnects
- 48-channel NVlink non-blocking communication

# Inspur OpenPOWER9 Server Key Features



## Modular design optimized for AI acceleration and scale-out deployments

- 2 OpenPOWER9 CPUs
- Up to 4 NVIDIA P100 GPUs
- Optimized for big data workloads,
- Gen4 PCIe is 2x faster than Gen3

**Inspur PowerAI native platform, FP5280, FP5290, FP5295 both hardware platform and software stack are ready.**

# Inspur OpenPOWER9 Server Tech Specs & Customization

<b>Model</b>	<b>OpenPOWER9 Server</b>
<b>Form Factor</b>	2U, 2-socket rack server
<b>Processor</b>	OpenPOWER9®Sforza CPU , 2 Sockets
<b>Memory</b>	16 DDR4 DIMM slots, up to 2666MT/s
<b>Drive Bays</b>	<ul style="list-style-type: none"> <li>● Front drive bays: 12 x 3.5 ' or 24 x 2.5 ' SAS/SATA HDD drives</li> <li>● Near drive bays: 4 x 3.5 ' SAS/SATA HDD drives</li> </ul>
<b>Network</b>	1x OCP NIC1 ports 25Gb
<b>PCIe Expansion</b>	Up to 8 PCIe-Gen4 slots <ul style="list-style-type: none"> <li>●4 slots HHHL PCIe4 x8 and 2 slots FHHL PCIe4 x16</li> <li>●8 slots PCIe 4 x8 for FHHL standard card</li> <li>●4 PCIe x16 for GPU</li> </ul>
<b>RAID</b>	RAID 0/1/5/6/10/50/60
<b>Ports</b>	1 x VGA, 2 x USB 3.0, 1 x UID、 1xVGA, 1x serial, 2 x USB 3.0, IPMI
<b>Power</b>	Platinum 550W, 800W, 1200W, 1600W, 2 Hot plug
<b>OS</b>	Red Hat Enterprise Linux、 Ubuntu、 NeoKylin
<b>Operating Temperature</b>	5°C to 45°C (41°F to 95°F)
<b>Chassis</b>	435mm×87mm×780mm



# Inspur Storage System

Inspur  
Validated

IEEL 3.1 Stable version



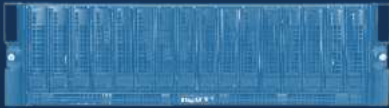
Deployed on EDR & OPA



Enterprise  
implementations

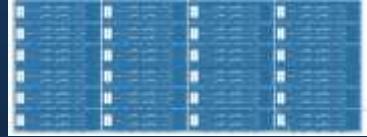


### Tstor2000



- IO+FC Architecture
- Lustre
- High Bandwidth

### Tstor3000



- Storage Server
- BeeGFS
- High IOPS
- BeeOND



Inspur  
Validated



Rigorous stress test



Code-level tuning



Redundancy protection



Multi-network adapter

Global Partnership with DDN for SFA series

Inspur NF5485M5 high-density storage server



Biology



Education



Ocean



CAE



Oil&Gas



IVA



Material

# Inspur High Performance Network



## Mellanox EDR

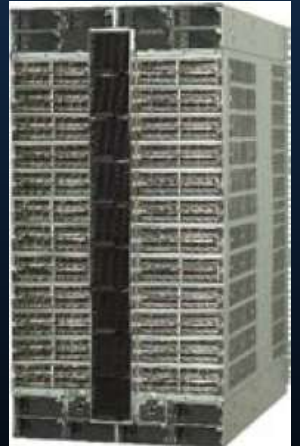
- 36 Ports EDR/FDR
- 108 Ports EDR/FDR
- 216 Ports EDR/FDR
- 324 Ports EDR/FDR
- 648 Ports EDR/FDR

Inspur Validated



## Intel Omni-path

- 24 Ports OPA
- 48 Ports OPA
- 192 Ports OPA
- 768 Ports OPA



Inspur Validated

Low Latency 700ns

High Bandwidth 100Gb/s

# Inspur HPC Management & Service - Cluster Engine

Product Performance Dashboard

TEYE

- Comprehensive management
- Job scheduling - management
- User friendly - remote access
- User account management
- Resource utilization trend analysis
- Application runtime analysis
- Hot spot - bottleneck alleviation
- Result checkpoint & restart

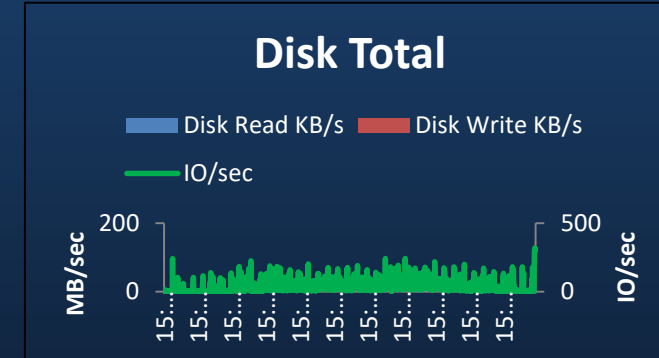
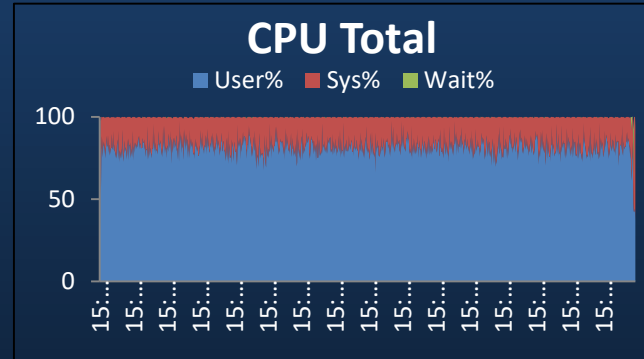


# T-Eye: Performance Profiling and Tuning Tool

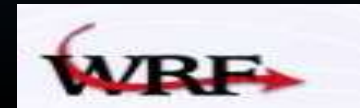
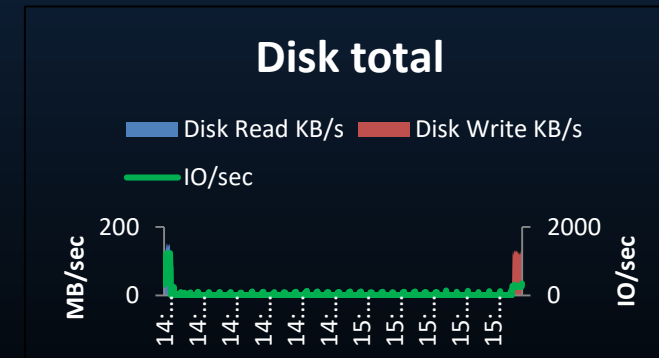
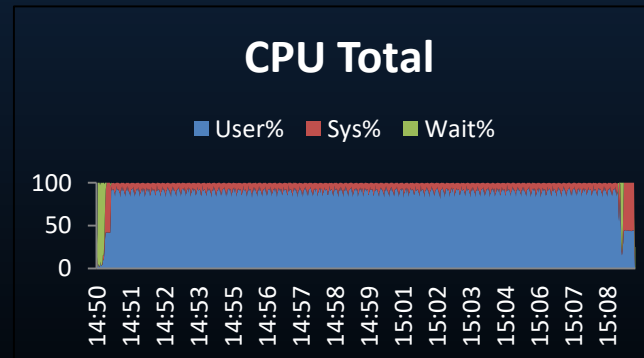
- ❖ Computing (CPU) intensive
- ❖ Memory-constrained
- ❖ Storage with high IOPS
- ❖ Network-intensive
- ❖ Scalability

- Balance computing resources
- High-performance scalable system
- Maximize application performance

## Before optimization



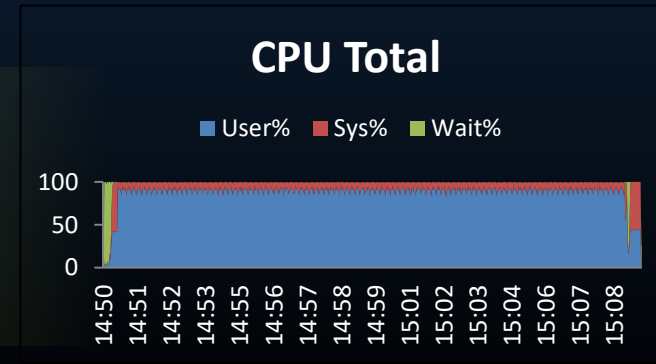
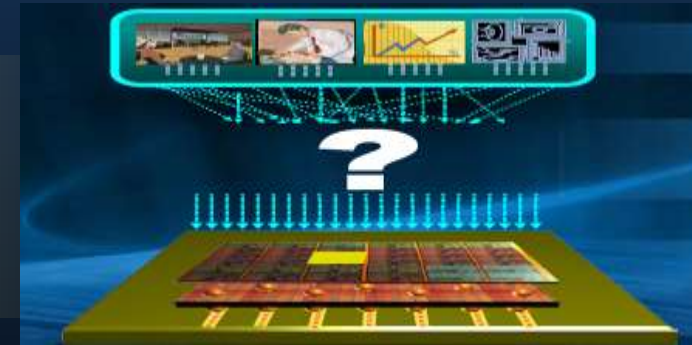
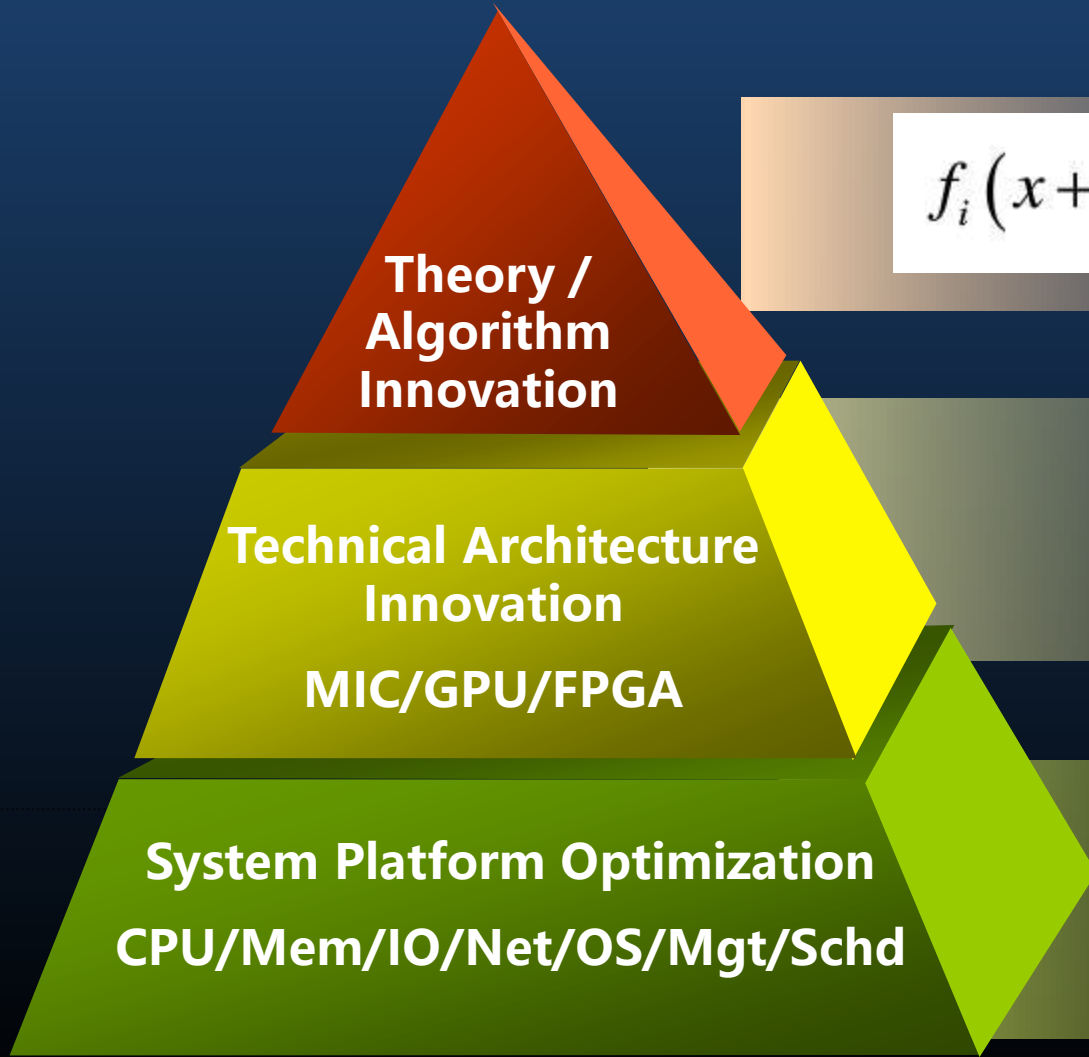
## After optimization



# Inspur HPC Application Development

## Inspur Domain Expert Organization

$$f_i(x + e_i \Delta t, t + \Delta t) - f_i(x, t) = -\frac{1}{\tau} [f_i(x, t) - f_i^{eq}(x, t)]$$

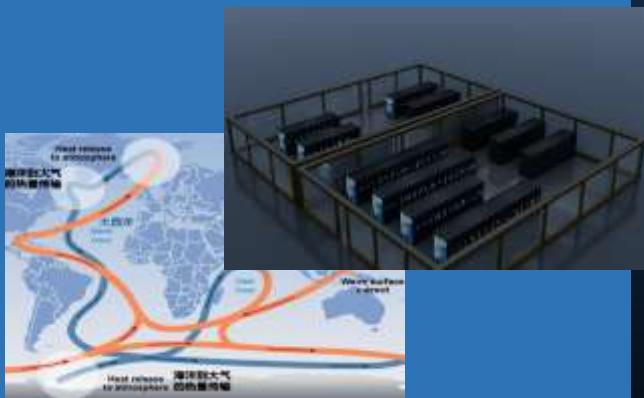


# Inspur Significant HPC Contributions

**inspur**

Qingdao National Laboratory -  
Supercomputer Center

China's fastest supercomputer in  
marine research



Exascale Application Optimization:  
Application development performance  
tuning, tools and analysis

SKA Project, the world's largest  
radio telescope

Inspur main provider for  
supercomputer solution



Exascale Application Development:  
International cooperation

Inspur-Intel China Parallel  
Computing Joint Lab

The world's first MIC book



Supercomputing architecture  
research collaboration

# Inspur HPC Services



## Deployment Services

- Rapid integration using best practices
- HPC Factory Integration
- Flexible services to meet the customer needs



## Remote Cluster Management Services

- Turnkey system management
- Remote system monitoring, administration & support
- Increased system utilization and uptime



## Support Services

- 24x7 hardware and software support.
- Onsite services
- Dedicated Engineer
- Partner support



## Optimization Services

HPC application code optimization to improve run-time performance.

# Asia Supercomputer Community



2011 – 2018

<https://www.asc-events.org/>



# Thank You

For more information, visit <http://en.inspur.com/>