1. Inspur Introduction

2. Inspur HPC Products and Solutions
Inspur Company Name

**SPUR**

verb. give an incentive or encouragement to (someone).

**INSPIRE**

verb. fill (someone) with the urge or ability to do or feel something, especially to do something creative.

Leading HPC & AI Computing
E2E Total Solution Provider
Inspur Product Portfolio

Cloud

AI ML/DL

Big Data

Storage

HPC

Network

Server
Inspur Position in the Global Market

Total 68 systems, 13.6%

Top 3 Server Vendor: 2017

Top 1 Server Vendor: 2022
Inspur Europe

EU Office Locations

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

- **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

- **Industry Vertical / Application**
  - Automobile
  - Telco & CSP
  - Enterprise
  - HPC
  - AI

- **Products**
  - i24
  - NF5280M5
  - NF5288M5
  - I9000
  - GX4(4X4GPU Box)
  - SN 3410M5(4*16GPUBox)
  - FPGA Card

- **Value Proposition**
  - Cost efficiency, High Demands
  - Optimal performance
  - Energy-saving
  - Co-design
  - End to End fast delivery
Inspur Server Business Growth

- **Inspur Server Business in 2017**
  - CAGR 62%
  - 6.26B
  - Rapid increase in recent years

- **Inspur Oversea Business in 2017**
  - 3.5X

- **AI Server Business in 2017**
  - 4X
  - 600%
  - 150%

Global

Inspur
HPC Further Frontiers

Fundamental Research

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

Business Innovation

Engineering Simulation
- IC Design / Aerospace / Aircraft Design / Financial Transactions / Risk Analysis

Machine Learning

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

Total GFLOPS

- For double precision
- From Accelerators
- From CPU

2013 2018E 2023E

Inspur GPU Server Growth

GPU Server

- 2017
- 2018

600% 150%

Data source: Inspur Research Institute

Inspur AI China Market Share

- CSP
  - 80%+
- IVA
  - 60%
- Telecom
  - 55%
- Finance
  - 80%

Computational Power Forecast

- Computational Power Forecast
  - 0.6
  - 12X
  - 20X

Inspur GPU Server Growth

- Inspur GPU Growth
- Global GPU Growth

2013 2018E 2023E

Data source: Inspur Research Institute
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
- Intel
- NVIDIA
- DDN
- Mellanox

End-2-End Solutions
Center of Excellence
Workload Optimized
Inspur HPC E2E IP Design & Manufacturing

Top quality AISC

Chip
Chipset, SAS
Board
Motherboard, RAID, HBA
Accessory Module
Backplane, Power Supply, Adapter, Fans
Structural
Chassis, Cabinet, Cables, Package
Integration
Final assembly

Insipur Mainframe Unix Series

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

<table>
<thead>
<tr>
<th>Operating Systems with UNIX Certification</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM</td>
</tr>
<tr>
<td>HP</td>
</tr>
<tr>
<td>Oracle</td>
</tr>
<tr>
<td>Apple</td>
</tr>
<tr>
<td>Inspur</td>
</tr>
</tbody>
</table>

**K1 950**
- 32-Socket 256 Cores

**M13**
- 64-Socket 1152 Cores
Inspur HPC Solution

**Turn Key Project**
- System Installation
- System Optimization
- System Training

**Application Development**
- Chemistry
- Life sciences
- CAE
- Meteorological

**System Software**
- Parallel Environment
- Operating System
- Compiler Environment

**System Management**
- Management & Monitor Software
- System Deployment Software
- SKVM System
- Job Management Software
- Load Balancing Management
- High Availability Management

**Hardware System**
- Fat Server
- GPU Server
- Blade Server
- High Speed Network
- Mass Storage
- Parallel File System

**Infrastructure**
- Cabinet System
- Power System
- Cooling System
- Monitor System
InCloud Rack
- High density converged

IEEL solution
- Intel HPC enterprise file system
- High performance hundreds of GB/s

FDR/EDR InfiniBand

Omni-path

Cluster Engine
- Cluster management, configuration, monitoring, alarms, job scheduling & check stop; user account & billing

BeeGFS
- Server Architecture
- HA Mirror & BeeOND

Management
- 10Gb/1Gb

Application run time monitoring & tuning

Water-cooling cabinet
- Closed cooling cycle, variable speed fan
- PUE < 1.1 40KW per Rack

Rack Cabinet
- 42U standard

Computing

Storage

Network

Management

Infrastructure

InCloud Rack

Standard Rack

InCloud Rack

Standard Rack

Inspur HPC Product Stak

Inspur HPC Product Stak
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
specially 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 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

<table>
<thead>
<tr>
<th>Module</th>
<th>NF5280M5</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU</td>
<td>2 Intel® Xeon® Scalable processors, Maximum TDP 205W</td>
</tr>
<tr>
<td>Memory</td>
<td>24 memory slots, Max DDR4-2666</td>
</tr>
<tr>
<td></td>
<td>12 NVdimm</td>
</tr>
<tr>
<td>RAID</td>
<td>Hardware-level protection for NVME SSD drives</td>
</tr>
<tr>
<td>Storage</td>
<td>Front HDD: max 24<em>2.5”, or 12</em>3.5” or 24*NVMe</td>
</tr>
<tr>
<td></td>
<td>Built-in HDD: 4<em>3.5”HDD and 2</em>M.2 SSD</td>
</tr>
<tr>
<td></td>
<td>BACK HDD: 4<em>3.5”HDD and 4</em>2.5”HDD</td>
</tr>
<tr>
<td>GPU</td>
<td>Support 4*GPU, P100/P40 etc</td>
</tr>
<tr>
<td>I/O</td>
<td>Max 10*PCIe slot</td>
</tr>
<tr>
<td>NIC</td>
<td>1*OCP</td>
</tr>
<tr>
<td>Power</td>
<td>High-efficiency power supplies with 80 PLUS Platinum and Titanium certifications.</td>
</tr>
</tbody>
</table>
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.
Inspur 8 Socket Sever TS860M5

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

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

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

Ins pur Server AGX5
AGX5
The Most Powerful / Dense AI 8U Server
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

<table>
<thead>
<tr>
<th>Model</th>
<th>OpenPOWER9 Server</th>
</tr>
</thead>
<tbody>
<tr>
<td>Form Factor</td>
<td>2U, 2-socket rack server</td>
</tr>
<tr>
<td>Processor</td>
<td>OpenPOWER9® Sforza CPU, 2 Sockets</td>
</tr>
<tr>
<td>Memory</td>
<td>16 DDR4 DIMM slots, up to 2666MT/s</td>
</tr>
</tbody>
</table>
| Drive Bays | - Front drive bays: 12 x 3.5” or 24 x 2.5” SAS/SATA HDD drives  
- Near drive bays: 4 x 3.5” SAS/SATA HDD drives |
| Network | 1x OCP NIC1 ports 25Gb |
| PCIe Expansion | Up to 8 PCIe-Gen4 slots  
- 4 slots HHHL PCIe4 x8 and 2 slots FHHL PCIe4 x16  
- 8 slots PCIe 4 x8 for FHHL standard card  
- 4 PCIe x16 for GPU |
| RAID | RAID 0/1/5/6/10/50/60 |
| Ports | 1 x VGA, 2 x USB 3.0, 1 x UID, 1xVGA, 1x serial, 2 x USB 3.0, IPMI |
| Power | Platinum 550W, 800W, 1200W, 1600W, 2 Hot plug |
| OS | Red Hat Enterprise Linux, Ubuntu, NeoKylin |
| Operating Temperature | 5°C to 45°C (41°F to 95°F) |
| Chassis | 435mm×87mm×780mm |
Inspur Storage System

IEEL 3.1 Stable version
Deployed on EDR & OPA
Enterprise implementations

Global Partnership with DDN for SFA series

Inspur NF5485M5 high-density storage server

Tstor2000
- IO+FC Architecture
- Lustre
- High Bandwidth

Tstor3000
- Storage Server
- BeeGFS
- High IOPS
- BeeOND

Rigorous stress test
Code-level tuning
Redundancy protection
Multi-network adapter

Inspur Validated

Biology
Education
Ocean
CAE
Oil&Gas
IVA
Material
Inspur High Performance Network

MSX6025F/OPA 100SWE48 Fat tree

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

Intel Omni-path
- 24 Ports OPA
- 48 Ports OPA
- 192 Ports OPA
- 768 Ports OPA

Low Latency 700ns
High Bandwidth 100Gb/s

Inspur Validated

2018/10/2
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

Inspur HPC Management & Service - Cluster Engine
T-Eye: Performance Profiling and Tuning Tool

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

Before optimization

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

CPU Total

After optimization
Inspur HPC Application Development

Inspur Domain Expert Organization

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

CPU Total

- User%
- Sys%
- Wait%

System Platform Optimization
- CPU/Mem/IO/Net/OS/Mgt/Schd

Technical Architecture Innovation
- MIC/GPU/FPGA

Theory / Algorithm Innovation

Innovation
Inspur Significant HPC Contributions

Qingdao National Laboratory - Supercomputer Center

- China's fastest supercomputer in marine research

SKA Project, the world's largest radio telescope

- Inspur main provider for supercomputer solution

Inspur-Intel China Parallel Computing Joint Lab

- The world's first MIC book

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

Exascale Application Development: International cooperation

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.
Thank You

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