
Open. Together.
OCP
SUMMIT

[HPC/GPU/FPGA]

Inspur Optimized Architecture for AI in OCP

[Gavin Wang & Han Wang, System Engineer, Inspur]

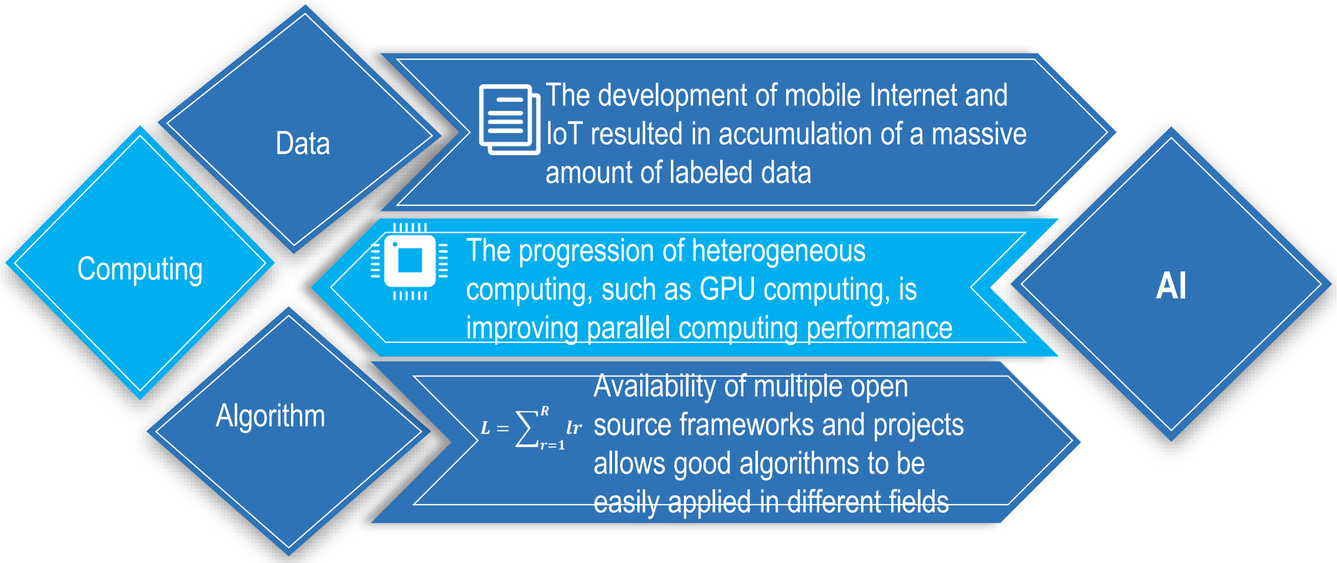
OPEN Compute Project

SOLUTION PROVIDER®

OPEN PLATINUM™

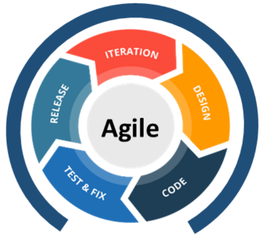
What Led to the Rapid Growth of AI Industry

HPC



Open

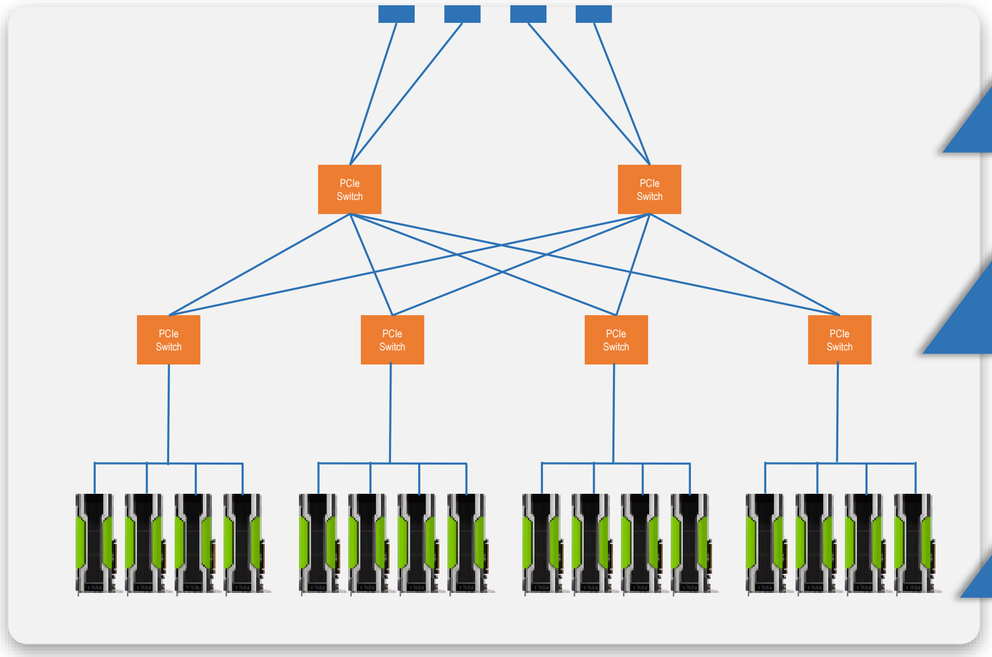
Converge



Agile

Ecosystem

16 GPU Ecosystem



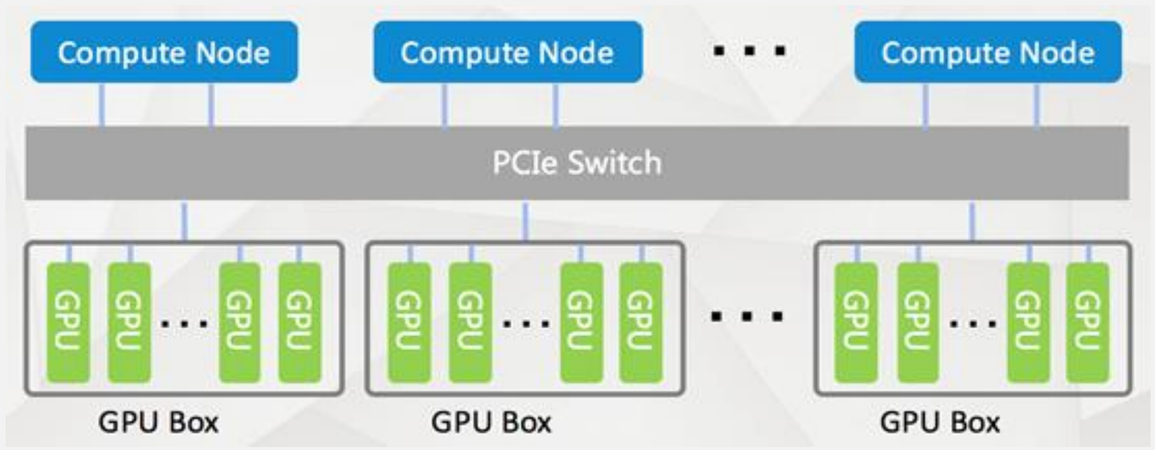
16×GPU in same domain

Centralized thermal, PSU, Management

Flexible GPU Pooling for 1-4 Host Server

HPC

Heterogeneous Pooling



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

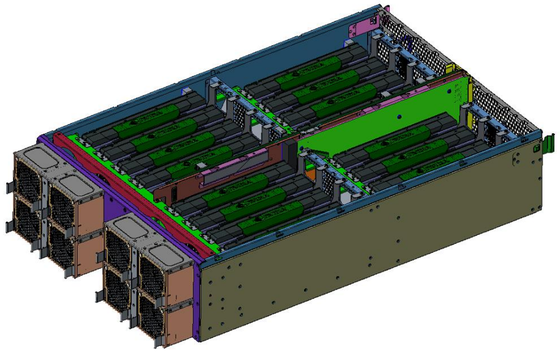
HPC

Inspur 16 GPU Product Info

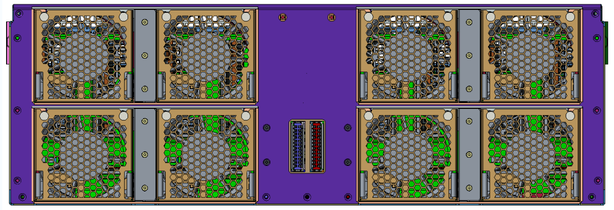
HPC

Ultral High Density GPU

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

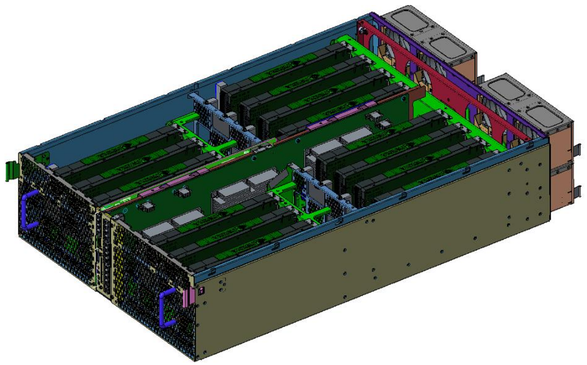


System rear view

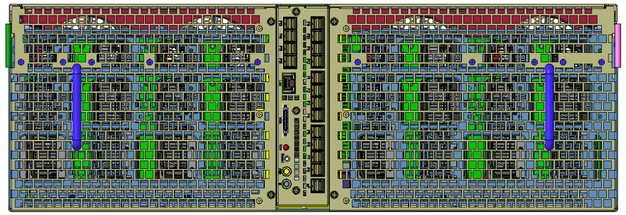


Optimized Architecture for AI

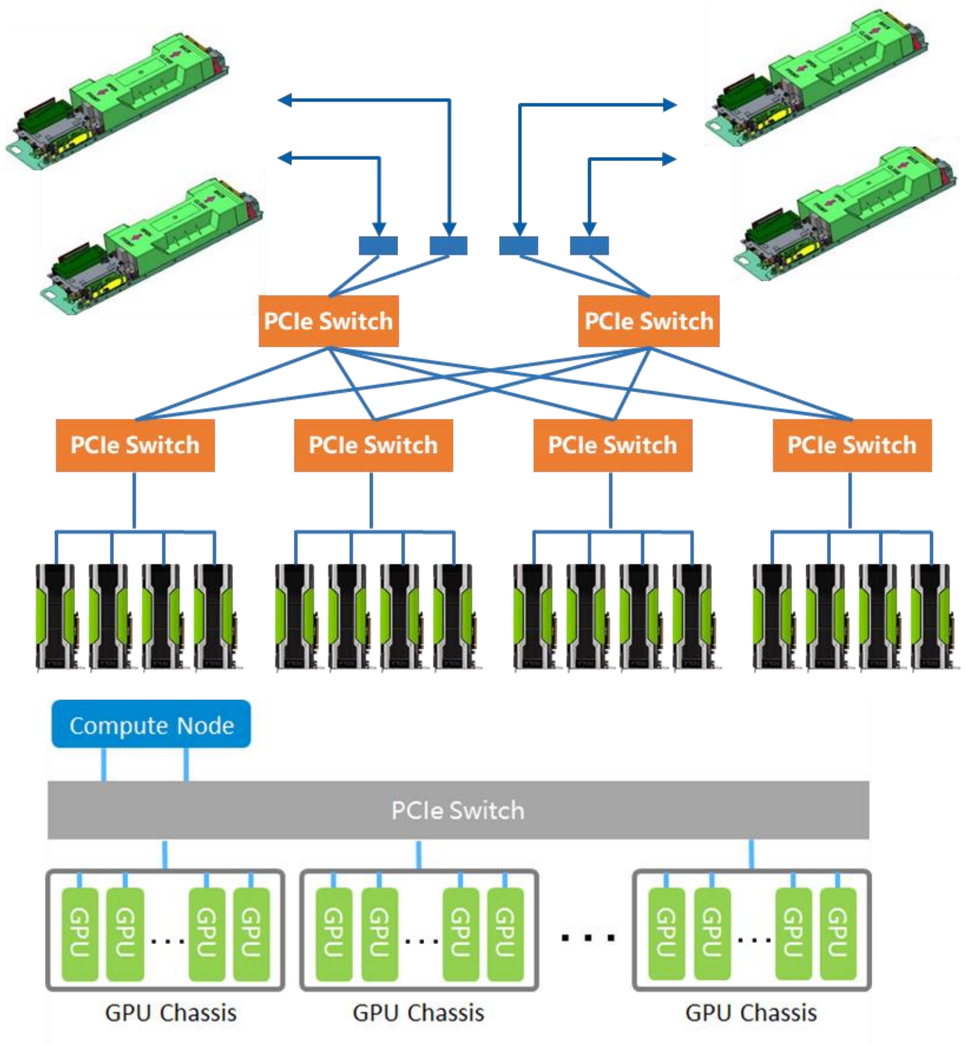
- **Speed up the training process**
- **Flexible architecture to support multi-node based on application**

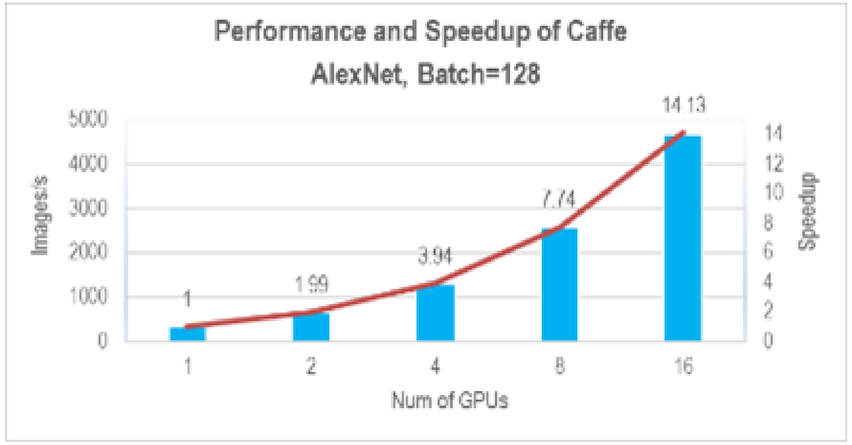


System front view

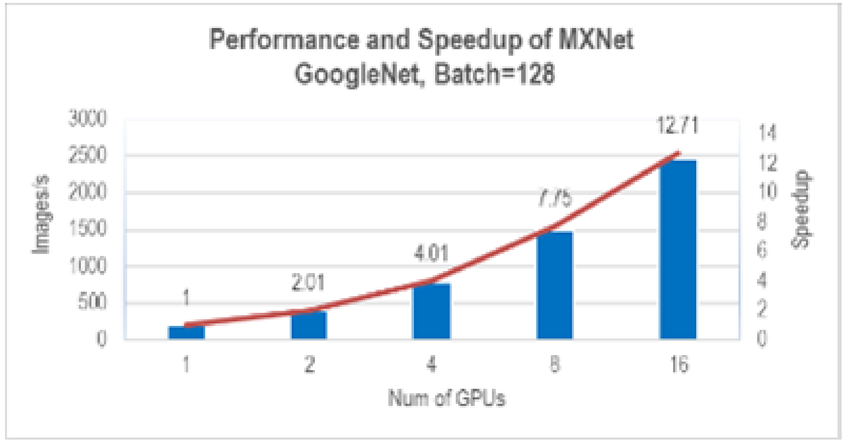


Inspur 16 GPU Product Info





high-speed/low-latency P2P GPU access needed



transfers gradients using GPU P2P

HPC


Open. Together.
OCP Global Summit | March 14–15, 2019
OCP
SUMMIT