# PCIe Gen5 GPU Solution

Empowering Generative AI with Flexibility and Efficiency


Revolutionizing Generative AI: PCIe Gen5 GPU Solution Meets the Growing Demands

The PCIe Gen5 GPU solution revolutionizes the field of generative AI, satisfying the escalating demands by enabling the independent deployment of GPUs at a larger scale. Achieved through a composable architecture and the innovative approach of composable disaggregation, this solution significantly impacts data-intensive tasks. It enhances transfer speeds and overall performance while providing flexibility and scalability by isolating GPUs from the server. The simplified hardware management in independent GPU expansion further contributes to cost savings. Overall, the PCIe Gen5 GPU solution meets the substantial GPU demands in generative AI, delivering enhanced performance, flexibility, and cost-effectiveness.
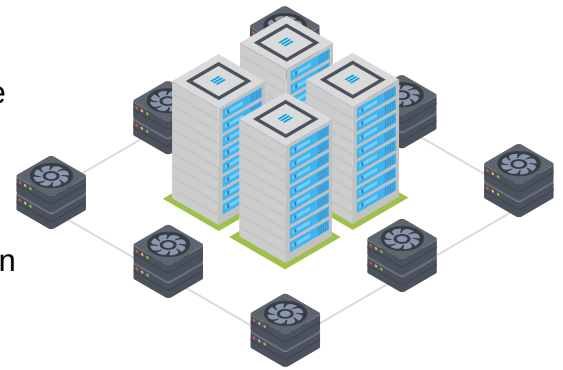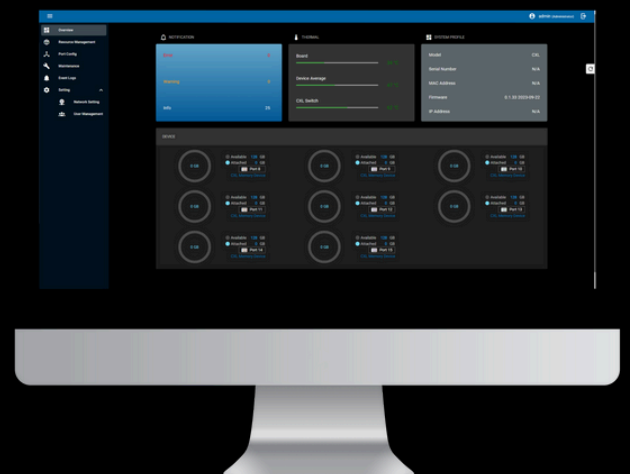

Next-Level System Dynamics: Breakthrough GPU Functionalities
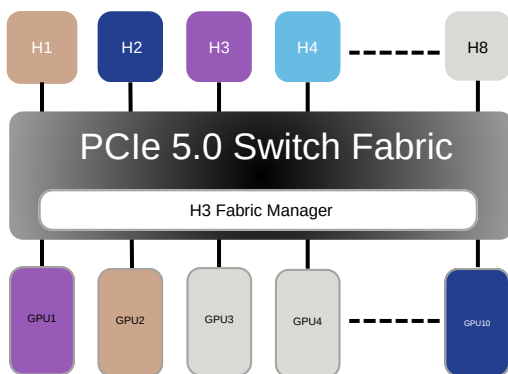
Revolutionizing system dynamics, Falcon 5012 introduces advanced GPU management, enabling up to 8 hosts to dynamically share 10 devices on-demand. This streamlines resource allocation, saving significant setup time. The solution allows seamless removal and addition of GPU resources between hosts, ensuring intelligent rearrangement. IT administrators benefit from the cohesive hardware and software solution, focusing more on gaining insights and less on component management. Falcon 5012 makes your workflow more efficient and effective.

Comprehansive Resource Management with RESTful API and GUI

Falcon 5012 ensures optimal system performance through strategic resource management. Actively monitoring GPU health, real-time performance, and errors in both host and device ports, it facilitates quick issue identification and resolution for stable operation. Introducing flexible switch cascade topologies and configurable host virtual trees enhances resource management flexibility. The solution also provides intuitive resource allocation insights through the Host and GPU views, enabling intelligent resource deployment.

H3 Platform

Software Features	**Features**	- GPU composability - GPU hotplug - GPU peer to peer (GPU P2P) - PCIe port configuration to host/device ports - Power control of a single PCIe slot - GPU health information - Performance and error monitoring of host and device ports - Configurable host virtual tree - Host view of assigned GPU - GPU view of assigned host - Temperature threshold setup - Device overheated protection - Firmware secure boot - Link capability and status of devices - User authentication management 256bit encryption of sensitive data - Sha128 encryption - ELK central log server integration
	Management Interface	Redfish®, RESTful API, GUI
	System Management	- H3 management center - Real-time GPU cluster topology - Dashboard for GPU utilization, performance and other information - Predictive health monitoring - Role-based authentication and access control
Hardware Features	**Model Name**	Falcon 5012
	BMC	- AST 2600, fully IPMI V2.0 compliant - AST 2500
	mCPU	Intel ATOM X86 CPU
	PCIe Switch	Broadcom PEX 89144 PCIe 5.0 switch
	Device	- GPU, FPGA, NIC, and other standard PCIe device (add-in card) - Dual-slot width and 10.5" length - Support up to 700W GPU
	Host Interface - Standard mode	- One (1) PCIe 5.0 x16 port for host connection - Need one HBA (host bus adapter) card on server host 1m CDFP cable
	Host Interface - Advanced mode	- Up to Four (4) PCIe 5.0 x16 ports for host connections - Need HBA (host bus adapter) cards on server hosts 1m CDFP cables
	Operating temperature	10°C ~ 35°C (50°F ~ 95°F)
	Power	Four (4) pieces 3200W (2+2 or 3+1 redundant), back removable, hot-swap
	Fan	Hot-swappable 8 80x80mm dual rotor fans
	Ethernet Ports	RJ45 of mCPU for GPU management/BMC for chassis management
	Dimension	4U; 175(H) * 448(W) * 650 (D)mm