

FPGA Accelerator Card

▷ F10A: HHL Extreme Density FPGA Accelerator Card



- Powered by Arria10 Chip, deliver 1.366 TFlops of superb performance with low latency
- Support OpenCL framework, dramatically drive AI development efficiency with mature ecosystem
- Suited for compute-intensive applications like AI inferencing, data compression, image encoding, video transcoding, and more

Product Specification	
Model	F10A
Chip	Intel® Arria® 10 GX1150
Performance	1.366 TFlops (Peak)
Dimensions	Half-height, Half-length
High-Speed Interface	Dual port SFP+ 10GbE, PCIe 3.0 x8
Featured Flash	32bit data interface; 1G Flash;
On-board DIMM-DDR4 SDRAM	Support 2 DDR4 SODIMM with 4-16GB of storage, 2133Mbps; Configured with 16GB
Power Supply	Powered by 12v via PCIe3.0 interface, no external power supply is required
Power Consumption	45W (Peak), 35W (Average)
Cooling	Positive/Passive Cooling (optional)

▷ F37X: Industry 1st FPGA Accelerator Card with On-Chip HBM2



- Deliver 28.1 INT8 TOPS of superior performance with low latency in full-height half-length form factor
- Feature 8GB integrated on-chip HBM2, offering 460GB/s of ultra-bandwidth
- Support C/C++, OpenCL & RTL, enabling flexible development and migration of AI algorithms and applications
- Ideal for AI inferencing, video transcoding, image recognition, natural language processing, genome sequencing analysis and more

Product Specification	
Model	F37X
Chip	Xilinx VU37P
Performance	28.1 INT8 TOPS
Dimension	Full-Height Half-Length
HBM DRAM	HBM2 8GB
DDR	Supports 3 channels 72bit-DDR4 24GB memory in total
System Interface	PCIe 3.0 x16
High-speed Network Interface	Supports 100GE dual-port QSFP28+
Debugging Interface	USB debugging interface
Power Supply	12V @ 75W via PCIe, plus 12V @ 75W via external Aux
Power Consumption	150W (Peak), 75W (typical AI applications)
Cooling	Dual slot passive cooling

AI Development & Management Suite

▷ AIStation: Agile Deep Learning Development Platform



Inspur AIStation is an agile AI workflow management tool. It is designed to provide enterprises customers with unified management and scheduling of AI computing resources, a complete AI development software stack and development process, accelerating AI research and development innovation.

Key Features:

- Unified Management of AI Computing Resources
- Easy Setup of AI Development Environment
- System-level Performance Optimization

▷ AutoML Suite: Easy-to-Use AutoML Toolset

Inspur's AutoML suite is an easy-to-use AutoML toolset, offering fast parallel modeling, flexible on premise & cloud deployment, and graphical user interface (GUI). Create deep learning models in 4 simple steps: data management, model training, generation and deployment, from modeling to enterprise scenario.

AutoML Suite

Key Features:

- Offer Fast Parallel Modeling
- On-Premise & Cloud Deployment
- Support Graphical User Interface

▷ T-Eye: AI Application Profiling and Performance Tuning Tool

Inspur T-Eye is management tool used to analyze AI applications performance features of hardware and system resources running on GPU clusters, revealing the running features, hotspots and bottlenecks of these applications.



Key Features:

- Runtime Performance Monitoring
- Identify Critical Index with Radar Chart
- Comparison Analysis to Facilitate Optimization

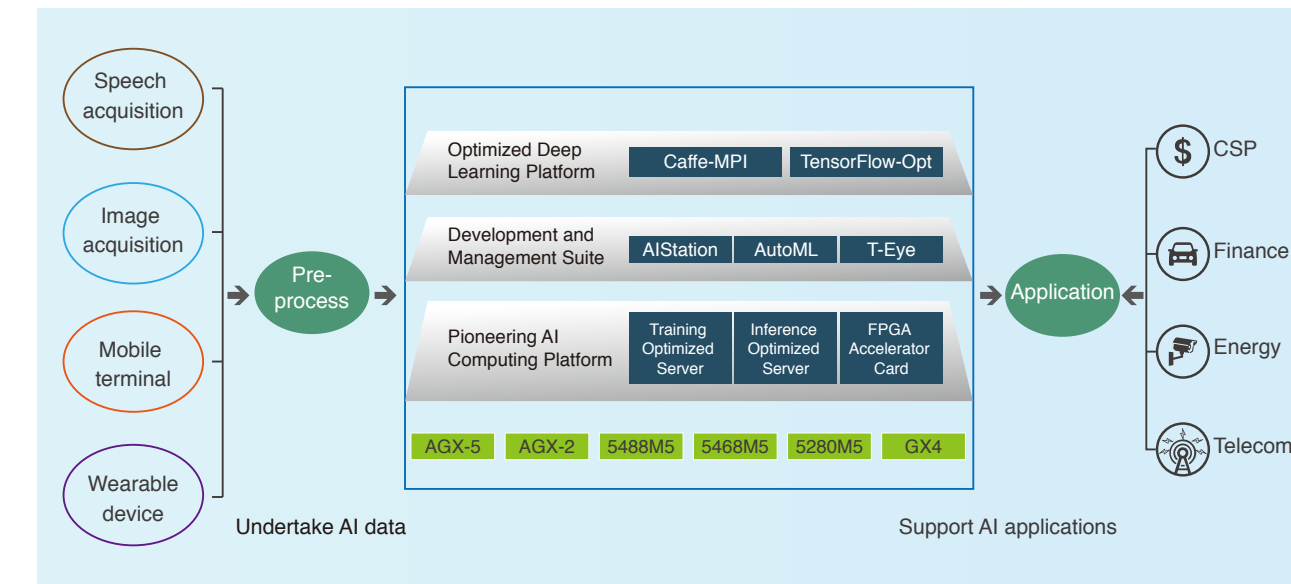
Optimized Deep Learning Framework

▷ Caffe-MPI: Cluster Parallel Deep Learning Framework

- The Inspur-developed Caffe-MPI is cluster parallel version of Berkley Caffe deep learning framework. It can run in high performance cluster systems with superb parallel scalability. Find open source Caffe-MPI 2.0 : <https://github.com/Caffe-MPI/Caffe-MPI.github.io>
- Inspur also provides the best encapsulation of the mainstream deep learning frameworks, (like Caffe, TensorFlow, CNTK, etc.), with its own code and third-party libraries, in deployable images that can be quickly and simply implemented in the clients' platforms.

End-to-End AI Solution

Inspur provides a wide range of end-to-end AI solutions from medical imaging, power device examination, financial service, voice recognition to AI cloud and more. A complete process of how Inspur helps to accelerate AI workloads from sample data pre-processing, model training and inference to model application is shown below.



This precipitates in omni-directional services in all aspects of AI industry, including AI system and application design, application code optimization and application platform evaluation.





Inspur AI

Products & Solutions

- **Leading AI Server Provider**
- **Full-Stack AI Capabilities**
- **Business Partner of Global AI Customers**

Inspur Full-Stack AI Capabilities

As world's leading AI computing provider, Inspur has the industry's most comprehensive AI portfolio. This four-layer AI portfolio includes pioneering AI computing platform, agile deep learning development suite, optimized deep learning frameworks, and the capability to deliver a variety of end-to-end AI solutions for its industry customers. Offering computing edge for global customers through innovative design, Inspur has become a business partner widely recognized and respected by many leading companies worldwide.

	E2E Vertical Solution	End-to-End AI Solution		
	Optimized Deep Learning Framework	Caffe-MPI	TensorFlow-Opt	
	Agile AI Development and Management Suite	AIStation	AutoML Suite	T-Eye
	Leading AI Computing Platform	Training Optimized Server	Inference Optimized Server	FPGA Accelerator Card

AI Computing Platform

AGX-5: Powerful Scale-up AI Super Server

- Unprecedented 2 petaFLOPS of computing power in one server, accelerated by 16 NVSwitch-empowered V100 GPUs
- Feature two 28-core processors to boost general-purpose computing
- Up to 6TB of persistent memory for high-speed data access
- Designed to tackle some of the toughest AI&HPC workloads



Product Specification	
Model	AGX-5 (NF588M5)
GPU	NVIDIA® Tesla® Volta, Volta Next (SXM3)
Performance	2 petaFLOPS
Processor	Server: 2*Intel® Xeon® Scalable Processors, 3*UPI JBOG: 2x8*GPU or 16*GPU
Memory	24x DDR4 DIMM, 6 channel
Storage	8 x 2.5" SATA storage (Includes 4*2.5" NVMe), 2*M.2
Fan & Cooling	Redundant Hot swap System Fan, Air cooling Air Cooling or Air / Liquid Hybrid Cooling
PSU	(2+2)*2 Redundant, up to 12 KW total
Chassis	8U W*H*D 448mm*351.6mm*850mm

AGX-2: Extreme High Density AI Server

- 8 NVLink-empowered V100 GPUs in a compact 2U design, deliver uncompromising performance in maximum density
- Options of air cooling or air-liquid hybrid cooling, make it easy to deploy in Green Datacenters with lower PUE
- Power a variety of AI&HPC applications with flexible configurations



Product Specification	
Model	NF528M5
Storage	8*NVIDIA® Tesla® NVLink™ V100/P100 or 8*PCIe P100/P40
Processor	2*2nd Generation Intel® Xeon® Scalable Processors
Memory	Up to 2TB 2667MHz DDR4
Storage	8*2.5" U.2/SAS/SATA 2*M.2 PCIe & SATA on Board
PCIe	Optional 4* PCIe x16 slots
Cooling	Redundant Hot Swap System Fans Air cooling /Air-Liquid Hybrid cooling
PSU	2*3000w PSU 80plus Platinum
Chassis	2U W*H*D 448mm*87.5mm*899.5mm

NF5488M5: Industry 1st 4U 8GPU NVSwitch-Empowered AI Server

- 8 NVSwitch-empowered V100 GPUs in 4U, deliver 1 petaFLOPS of superb performance
- Enable faster AI training with lower cost compared to more servers with fewer GPUs each
- 6KW in 4U design, make it easy to deploy in power-constrained racks
- Designed for a wide range of deep learning and HPC applications



Product Specification	
Model	NF5488M5
HGX-2 Base Board	8*SXM3 GPU NVLink, up to 350-400W TDP, NVSwitch Full Connection NVIDIA® Tesla® Volta, Volta Next (SXM3)
Processor	2*Intel® Xeon® Scalable Processors, 3*UPI
Performance	1 petaFLOPS
Memory	24x DDR4 DIMM, 6 Channel
Storage	8 x 2.5" SATA storage (Includes 4*2.5" NVMe), 2*M.2
PCIe	4 x PCIe x16 for 100G NIC, 1x PCIe8 for 50G NIC / NVMe
Fan & Cooling	Redundant Hot swap System Fan, Air cooling
PSU	2+2 Redundant, up to 6 KW total
Chassis	4U W*H*D 448mm*175.5mm*850mm

NF5468M5: Popular Inference Server with 16*T4 GPU in 4U

- Up to 16 T4 GPUs in one server, turbocharge AI Inference to gain real-time insight
- Up to 8 NVLink-empowered V100 GPUs, supercharge large-scale training of AI models
- Provide perfect combination of performance and internal storage with 384 TB
- One click to change GPU topology with BMC, drive a broad range of applications from AI cloud, telecom to healthcare.



Product Specification	
Model	NF5468M5
GPU	8*NVIDIA® Tesla® NVLink™ V100/P100/P40 or 16*NVIDIA® Tesla® T4 GPUs
Processor	2*Intel® Xeon® Scalable Processors
Memory	24 2666MHz memory slots, support DDR4 ECC
Storage	24* 2.5/3.5 HDD (8* NVMe SSD) + 2* M.2 SSD HDD Support RAID 0/1/10/5/50/6/60
Cooling	Redundant Hot swap System Fan
PCIe	Support 20* PCIe 3.0x16
PSU	2+2 Redundant, 4*1600W/2000W/2200W 80PLUS Certified Platinum Power Supply Modules
Chassis	4U W*H*D 435mm*175.5mm*830mm

NF5280M5: Optimized For AI Applications

- 2U 4*NVIDIA® Tesla® GPUs for high quality and performance
- Dual-socket rackmount server optimized for AI applications
- Superior scalability, with optimized cooling design and modular system architecture
- Suitable for a wide spectrum of demanding AI applications



Product Specification	
Model	NF5280M5
GPU	4*NVIDIA® Tesla® V100, P100, P40
Processor	2*2nd Generation Intel® Xeon® Scalable Processors
Memory	24 Memory Slots, DDR4 ECC,
Storage	Front: 24*front 2.5" HDD or 12*3.5" HDD 24*NVMe SSD Built-in: 4*3.5" HDD and 2 M.2 SSDs Rear: Up to 4*3.5" and 4*2.5" HDD
I/O Expansion	10*Standard PCIe Slots
PSU	220VAC/240VDC, 1+1 Redundant Titanium Power Module
Chassis	2U W*H*D 435mm*87mm*779.5mm

GX4: GPU Resource Pooling AI Server

- 2U 4*NVIDIA® Tesla® GPUs, 4*GX4 can compose an AI system support up to 16*NVIDIA® Tesla® GPUs
- Decouples CPU and GPU, enabling GPU resource pooling
- Equipped a dual-socket server as head node
- Deliver unparalleled compute power for AI applications



Product Specification	
Model	GX4
GPU	4*NVIDIA® Tesla® V100/P100/P40
Hard Disk Controller	NVMe
Storage	16*2.5" U.2
I/O Expansion	1*PCIe3.0 Slot, 4 mini PCIe 4-bit Lines
PSU	Support platinum/titanium power supply Single or Dual Power Supply options Support PMBUS
Chassis	2U W*H*D 435mm*87.5mm*740mm

NE5250M5: Edge Computing AI Server

- 2U 2*NVIDIA® Tesla® V100 or 6*NVIDIA® Tesla® T4 GPUs
- Accelerate 5G edge applications include IoT, MEC and NFV
- Support optimization design for the edge's harsh deployment environment
- Ideal for the most compute-intensive AI applications, including autonomous vehicles, smart cities and smart homes



Product Specification	
Model	NE5250M5
GPU	2U 2*NVIDIA® Tesla® V100 or 6*NVIDIA® Tesla® T4 GPUs
Processor	2*2nd Generation Intel® Xeon® Scalable Processors, TDP 205W
DIMM	16x DIMM w/ 2 AEP supported
Storage	2x M2 2280/22110 SSD (SATA/PCIe) 2x 2.5" HDD/SSD (SATA/NVMe) SATA Slimline x8 supported (from x8 QAT)
Data LAN	2x Integrated 10G SFP+ w/ NCSI (by PCH) Up to 100 Gb Ethernet connectivity for boosting performance
PCIe	Support up to 6x PCIeGen3 slots 2x Dual-width PCIe x16 GPU TDP 300W + 2x FHHL PCIe x16 4x PCIe16 FH 3/4 L card+ 2x PCIe8 FHHL card
PSU	1+1 1100W Slim PSU
Chassis	2U, 19", 430mm Support wall-mount and rack-mount
Operating Temperature	Long-term 5°C – 40°C and short-term -5° – 45°C