



Inspur OAI Product Introduction

Jan 2021

Compute Project®	Compute Project® Inspur Full-Stack Al Portfolio INSP					
E2E AI Solution	Internet 홆 Telecom 🔇 Finance	Education	Smart City	🖁 Media 🥳 Manufac	ture 🗭 Medicine	
Al Algorithm Toolkit Platform	AutoML Suite	TF2	LMS	TensorFlow-opt	Caffe-MPI	
	Cloud& On- premise Automatic Automati Automatic Deployment modeling c tuning cropping	Lossless Speed up model in FPGA accuracy development	Self-developed AI model computing framework, supporting GPU large-scaletraining	Optimized TensorFlow framework with the fastest Al training speed on the public cloud, 512GPU expansion efficiency 90%	One of the first parallel versions of Caffe framework	
Al Resource Platform	AlStation (Training Platform)	AlStation	T-Eye			
	Model Model Application Development Deployment Development	Compatible with multiple DL frameworks	Support Al model online testing and evaluation	Multi-model deployment and weighted calculation	Al application and framework feature analyzer	
Al Computing Platform	Accelerator Card	Server				
	F10A F37X N10X N20X F10S F07V	Training Ge	neral Server	Inference and Open Compu	Edge	



INSPU 浪潮



AI Computing Platform

- Industry' s **Most Comprehensive** Al Server Portfolio
- General Server 2U/4U/6U
- **Open Hardware Compute and OAI**
- M5 AI Servers, FPGAs, ASIC Cards....

AI Resource Platform

- AlStation: One-stop Al development platform, efficient and flexible computing resource scheduling; easy to deploy AI dev environment
- T-Eye: AI performance profiling and tuning tool, empower AI application optimization

Algorithm Toolkit

- AutoML Suite: On-Premise & Cloud deployment; Parallel Acceleration: Effortless Model Generation
- Caffe-MPI: 1st Parallel Version of Caffe
- TensorFlow-Opt: Scale-out TensorFlow on public cloud, optimization on cloud RoCE



Inspur Rack Scale Servers for Open Datacenter Projects

INSPUF 浪潮





OPER Compute Project* One-Stop Al Development and Deployment Platform



AiStation —

Data	Model Dev and Tr	elopment aining		Model D and I	eployment nference		Al App
	Low Model D Efficie	evelopment ency		The deployment complication to	The deployment complication to get		
Internet	Low Utiliz Computing	zation of Resource		deploy the trained model into production	the trained model into production		PC
Telecom				Coomless connection	Unified management		
3	Easily deploy Al development	Efficient and flexible platform, obtain Al		between model	of multiple models, Centralized scheduling		Mobile
Finance	environment and development process,	computing resources on demand to speed up		development and deployment, shorten the	of computing resources		
	significantly improving development efficiency	model training efficiency		time of scaling to production	Dynamic allocation, Elastic expansion		
Government	Data		Model			Al Service	Manufacture
transport					न्ने न्यू		(P
đ				On-premise Priva	ate Public		Robot
Manufacture				Deployment			
A	$2 \text{ days} \rightarrow 4 \text{Hours}$	40% → 8 <u>0%</u>		2 days → <u>5 mir</u>	Multi-applica	tion load	
MedicalScience	Training Time	Computing Resolutilization	urces	One-stop Model Deployment	balancing and resource elas	d tic	ΙΟΤ



Breaking physical boundaries Flexibly pooling the resource



• SAS Switch for pooling HDDs, improving

storage flexibility

- NVMe pooling, more nimble resource
- NVMe over Fabric, high Perf storage pooling
- PCIe Switch for pooling GPU, GPU

acceleration ratio increases linearly

• GPU/FPGA over Fabric, heterogeneous

acceleration remote expansion









INSPU 泉潮







OPER
Compute
Project*MX1 System Features
MX1: World' s First OAI Reference System

INSPUC 浪潮

Product Model: MX1				
Chassis	21" 30U Rack mount			
Dimensions	537W*141H*803D (mm)			
Connection with Compute node	Up to PCIe Gen4 x32			
ΟΑΜ	Support Max 8pcs 48~54V OAM(up to 450W each); Support Max 8pcs 12V OAM (up to 350W each)			
Power without OAM	1570W			
PCIe Switch	Support PCIe Gen4 (100lanes/chip)			
PCIe re-timer	Support PCIe Gen4 x16			
Phy re-timer	56Gbps PAM-4 or 10/28Gbps NRZ x16			
Expansion slots	Up to 4 x PCle Gen4 x16 low profile standard card			
ВМС	AST2520			
Ι/Ο	Dongle connector for dedicate NIC and UBS, UID/PWR Button with LED , QSFDDx8 for OAM scale out, micro USBx2 for OAM debug			
Ambient Working Temperature	5-35 ℃			









INSPUC 浪潮



INSPUR CONFIDENTIAL



Inspur Smart Datacenter Management Open Structure

INSPUC 浪潮

Solutions for the implementation of the

rack Mgmt based on node level

• Southbound manages system resources;

northbound presents Info

• Meet the needs of Mgmt encryption and

resource pooling

- Relying on vendors maintenance for traditional BMC code base
- Complex to modify the traditional BMC code for new HW
- Poor readability of IPMI tool binary code





Open RMC + Open Hardware

48VDC Open Rack

- ✓ 1 pairs 48V Bus Bar✓ 1 shelf per Rack

Power Shelf

- ✓ 33KW(12xPSU)
- ✓ 40V-58V
- ✓ 93mm (H, 2OU) x 537mm
 (W, 21") x 586 (D) mm

System Devices

- ✓ Inspur 3OU OAI systems x4
- ✓ Inspur 20U compute node x4





- RMC Web Server:基于OpenBMC的 Rack Manager控制器服务
- RMC Web UI:资源收集及服务配置文件
- 南向接口: 支持Redfish RESTful API
- 北向接口:支持Redfish RESTful API并丰 富了服务配置文件

INSPUR CONFIDENTIAL



Thank You

Jan 2021