人工智能加速卡

F10A: 半高半长、极致密度FPGA加速卡

- 配备 Intel®Arria®10 芯片，计算性能高达 1.366TFlops，超低延迟
- 支持 OpenCL 编程框架，成熟的生态体系大幅提升 AI 开发效率
- 适用于 AI 推理、数据压缩、图像编码、视频转码等计算密集型应用场景

产品规格	
型号	F10A
芯片	Intel® Arria® 10 GX1150
计算性能	1.366 TFlops (Peak)
板卡规格	半高半长；主动散热单槽，被动散热双槽
高速接口	2个SFP+ GE/10GE接口，PCIe3.0 x8
配置Flash	32bit数据接口；1Gbit Flash；
板载SODIMM	支持2条DDR4 SODIMM，每条容量4~16GB，2133Mbps，板载标配16GB
板卡供电	PCIe接口供电
板卡功耗	45W(Peak)，35W(Average)
板卡散热	被动散热或主动散热（可选）

F37X：业界首款集成HBM2的FPGA 加速卡

- 全高半长设计，提供 28.1TOPS 的 INT8 卓越计算性能
- 集成 8GB HBM2 高速缓存，提供 460GB/s 极致带宽
- 支持 C/C++, RTL 和 OpenCL 开发环境、SDAceel 开发工具
- 可灵活开发和迁移不同的 AI 算法和应用，适用于 AI 推理、视频转码、图像识别、自然语言处理、基因组测序分析等应用场景

产品规格	
型号	F37X
芯片	Xilinx VU37P
计算性能	28.1TOPS（INT8）
板卡规格	全高半长；双槽位
HBM DRAM	HBM2 8GB
DDR	3 通道 72bits DDR4; 最大支持24GB板载内存
高速接口	PCIe3.0 x16
网络	2路100G QSFP28+
调试接口	USB调试接口
板卡供电	PCIe 插槽12V@75W供电+外部Aux供电12V@75W
板卡功耗	最大系统功耗150W，典型应用功耗75W
板卡散热	被动散热
BMC管理	智能BMC管理，板卡电源控制及板卡信息读取（温度，功耗，内存信息）；能够读取板卡SN
升级	支持PCIe在线升级固件

人工智能资源平台

AIStation: 敏捷的AI资源平台

浪潮 AIStation 面向深度学习开发场景，整合计算资源、数据资源以及 AI 开发环境，实现计算资源统一分配调度、训练数据集中管理并加速、模型流程化开发训练 ,为 AI 研发构建敏捷高效的一体化平台。

功能特性:
- 统一管理深度学习计算资源
- 秒级部署深度学习开发环境
- 系统级性能瓶颈优化

T-Eye: AI 应用特征分析和性能调优工具

T-Eye 是浪潮自主研发的 AI 应用特征分析和性能调优工具，用于分析 AI 应用程序在 GPU 集群上运行时对硬件及系统资源占用的情况，反映出应用程序的运行特征、热点及瓶颈，从而帮助用户最大限度的在现有平台挖掘应用的计算潜力。

功能特性:
- 实时监控系统运行性能，发现瓶颈
- 生成特征雷达图，读取关键指标
- 支持不同模型或算法特征的对比

算法工具平台

AutoML Suite: 高效自动机器学习平台

AutoML Suite 是浪潮自主研发的自动机器学习平台，支持图形化用户界面，操作简便。通过数据上传、模型搜索、模型训练、模型部署四个步骤的可视化操作，实现一站式自动生成模型。

功能特性:
- 快速高效开发 AI 模型
- 支持本地化和云端双模式部署
- 支持图形化用户界面

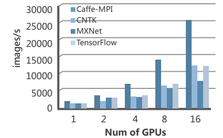
AutoLabel: 智能标注平台

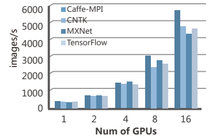
AutoLabel 是浪潮自主研发的智能标注平台，通过后台深度学习算法和高精度模型对样本进行自动标注。标注人员仅需对错标样本修正 ,后台利用纠正数据迭代更新模型，实现高效便捷的数据标注全流程。

功能特性:
- 支持批量审核纠正，自动迭代优化模型
- 支持定制化场景，可按需增加工作场景
- 支持图形化用户界面，易于使用

Caffe-MPI: 集群并行版深度学习框架

- Caffe-MPI 是浪潮主导开发的全球首个集群并行版 Caffe 框架，新版本 Caffe-MPI 2.0 在 4 节点 16 块 GPU 卡集群系统上训练性能较单卡提升 13 倍
- 开源地址：https: //github.com/Caffe-MPI/Caffe-MPI.github.io





端到端人工智能解决方案

结合实际应用业务场景 ,浪潮提供智能视频分析、医疗影像、电力设备巡检、金融汇率预测、语音识别、AI 云等端到端人工智能解决方案 ,包含应用场景咨询与系统方案设计、应用代码移植优化、计算加速部件性能横向评测等。

浪潮（北京）电子信息产业有限公司
网址：www.inspur.com
技术支持与服务热线：400-860-0011或0531-88546554
购买咨询热线：400-860-6708或800-860-6708或0531-88933376

让计算
创造无限可能

inspur

浪潮人工智能

产品&解决方案

- 人工智能计算领导品牌
- 全栈式人工智能解决方案
- 中国人工智能服务器市场份额超50%

浪潮人工智能

浪潮是人工智能计算的领导品牌，从计算平台、开发套件、框架优化、应用加速四个层次致力于打造敏捷、高效、优化的人工智能基础设施。中国市场份额始终保持 50% 以上，并与人工智能领先科技公司保持在系统与应用方面的深入紧密合作，帮助 AI 客户在语音、图像、视频、搜索、网络等方面取得数量级的应用性能提升。

	端到端人工智能解决方案	垂直行业解决方案		
	先进的算法工具平台	AutoML Suite	AutoLabel	Caffe-MPI
	敏捷的人工智能资源平台	AIStation	T-Eye	
	领先的人工智能计算平台	面向训练的人工智能服务器	面向推理的人工智能服务器	AI加速卡

人工智能计算平台

AGX-5: 极致计算性能 AI 服务器

- 配置 16 颗支持NVSwitch高速互联的NVIDIA® Tesla® V100 GPU，提供每秒 2 千万亿次的 AI 计算性能
- 内置 2 颗 28 核心的强大CPU，提供顶级通用计算性能
- 6TB 持久内存，可提供超大数据高速访问
- 面向最具挑战性的AI和HPC应用

产品规格	
型号	AGX-5 (NF5888M5)
GPU	支持16颗通过高速NVSwitch无阻塞全互联的最新一代NVIDIA® Tesla® SXM3 V100 Tensor Core 32GB GPU
处理器	2颗Intel® Xeon® Scalable处理器, 3*UPI
性能	2 petaFLOPS
内存	12通道，支持24条2666 MT/s DDR4 ECC内存，12个内存通道
存储	8块2.5英寸SATA硬盘 (可兼容最大支持4个2.5英寸NVMe)，内置2*M.2
PCIe	4*PCIe3.0 x16 for 100G NIC, 1*PCIe3.0 x8 for 50G NIC/NVMe
散热	冗余热插拔系统风扇 风冷散热
电源	(2+2)*2 冗余模式，12KW供电设计
机箱	8U（宽*高*深：448mm*351.6mm*850mm）

AGX-2: 极致计算密度 AI 服务器

- 2U内配置8颗支持NVLink的NVIDIA® Tesla® V100 GPU, 提供极致计算密度和强劲性能
- 支持风冷散热，实现数据中心简易部署，显著降低PUE
- 专为AI和HPC应用场景优化，可实现灵活配置

产品规格	
型号	AGX-2（NF5288M5）
GPU	支持8颗 NVIDIA® Tesla® NVLink™ V100 GPU或 支持8颗 NVIDIA® Tesla® V100/P100/P40 GPU
处理器	2颗 2nd Generation Intel® Xeon® Scalable处理器
内存	支持16条2933MT/s DDR4 ECC内存，12个内存通道
存储	8块2.5英寸U.2/SAS/SATA硬盘 内置2块M.2 PCIe & SATA硬盘
PCIe	可选：4个PCIe x16 扩展槽
散热	冗余热插拔系统风扇 风冷散热
电源	2个3000w 80plus铂金级电源
机箱	2U（宽*高*深：448mm*87.5mm*899.5mm）

NF5488M5: 业界首款支持NVSwitch高速互联的4U8GPU AI 服务器

- 4U 空间内集成 8 颗 NVSwitch高速互联的NVIDIA® Tesla® V100 GPU，AI 计算性能可达每秒一千万亿次
- 与内置较少GPU的单节点构成的GPU集群相比，能够以更低成本实现更高效的AI训练
- 4U 尺寸、6KW 供电设计，特别适合于功耗受限的机柜上架场景
- 广泛适用于深度学习和HPC应用场景

产品规格	
型号	NF5488M5
GPU	支持8颗通过高速NVSwitch无阻塞全互联的最新一代NVIDIA® Tesla® SXM3 V100 Tensor Core 32GB GPU
处理器	2颗Intel® Xeon® Scalable处理器, 3*UPI
性能	1 petaFLOPS
内存	支持24条2666 MT/s DDR4 ECC内存，12个内存通道
存储	8块2.5英寸SATA硬盘 (可兼容最大支持4个2.5英寸NVMe), 内置2*M.2
PCIe	4*PCIe3.0 x16 for 100G NIC, 1*PCIe3.0 x8 for 50G NIC/NVMe
散热	冗余热插拔系统风扇 风冷散热
电源	2+2冗余模式, 6KW供电设计
机箱	4U（宽*高*深： 448mm*175.5mm*850mm）

NF5468M5: 高密度推理服务器

- 4U空间内最多支持20颗NVIDIA® Tesla® T4 GPU，实现强劲高效的实时AI推理
- 最多支持8颗NVLink V100 GPU，实现AI模型大规模训练
- 支持384TB超大容量存储，实现内存与计算性能的绝佳组合
- 可支持多种拓扑切换，面向AI云、安防、金融、医疗等典型应用场景

产品规格	
型号	NF5468M5
GPU	支持8颗NVIDIA® Tesla® NVLink™ V100 或 8*NVIDIA® Tesla® V100/P100/P40 20颗NVIDIA® Tesla® T4 GPU
处理器	2颗2nd Generation Intel® Xeon® Scalable处理器
内存	支持24条2666 MT/s DDR4 ECC内存，12个内存通道
存储	24块2.5/3.5英寸 HDD硬盘（8块NVMe SSD硬盘）+ 2块M.2 SSD HDD硬盘 支持RAID 0/1/10/5/50/6/60
散热	冗余热插拔系统风扇
PCIe	最大支持20条 PCIe 3.0x16插槽
电源	2+2冗余模式，4个1600W/2000W/2200W 80Plus 铂金级电源
机箱	4U（宽*高*深： 435mm*175.5mm*830mm）

NF5280M5: 通用AI服务器

- 2U 空间搭载 4 颗 NVIDIA® Tesla® V100 GPU, 兼顾扩展与计算性能
- 专为 AI 应用优化的两路机架式人工智能服务器
- 极致的拓展性能，优化的散热设计，以及模块化的系统架构
- 面向一系列高要求的 AI 应用场景

产品规格	
模型	NF5280M5
GPU	4颗NVIDIA® Tesla® V100, P100, P40 GPU
处理器	2颗2nd Generation Intel® Xeon® Scalable处理器
内存	支持24条2933 MT/s DDR4 ECC内存，12个内存通道
存储	前置：25块 2.5英寸HDD硬盘 或 12块3.5英寸HDD硬盘 24块NVMe SSD硬盘 内置：4块3.5英寸HDD硬盘 和 2块M.2 SSD硬盘 后置：可支持4块3.5英寸 和 4块2.5英寸HDD硬盘
I/O拓展	8个 PCIe标准插槽，1个OCP 2.0
电源	220VAC/240VDC, 1+1 冗余钛金级电源
机箱	2U（宽*高*深：435mm×87mm×779.5mm）

GX4: GPU资源池化服务器

- 2U 空间搭载 4 颗 NVIDIA® Tesla® V100 GPU，4 台 GX4 可搭配 1 台双路服务器组成 16GPU 卡系统，实现高效的并行运算处理能力
- 实现 CPU 与 GPU 计算资源解耦合，实现 GPU 资源池化
- 更好的匹配计算平台和应用程序，为 AI 应用场景提供绝佳的计算性能支持

产品规格	
型号	GX4
GPU	4颗NVIDIA® Tesla® V100/P100/P40/GPU
管理	后置1个IPMI网口
散热	N+1冗余系统风扇
PCIe	支持一个PCIe x16插槽
电源	1600W 1+1冗余电源
机箱	2U（宽*高*深： 435mm*87.5mm*740mm）

NE5250M5: 边缘计算AI服务器

- 2U 空间内支持 2 颗 NVIDIA® Tesla® V100 GPU 或 6 颗 T4 GPU
- 加速 5G 边缘应用场景，如物联网、移动边缘计算和网络虚拟化
- 支持壁挂式和机架式安装，对于部署环境可以即插即用
- 适用于众多计算密集型 AI 应用场景，如自动驾驶、智慧城市等

产品规格	
型号	NE5250M5
GPU	支持2颗NVIDIA® Tesla® V100 或 6颗NVIDIA® Tesla® T4 GPU
处理器	2颗2nd Generation Intel® Xeon® Scalable处理器，TDP 205W
内存	最多支持16x DIMM w/ 2个 AEP
存储	2x M2 2280/22110 SSD (SATA/PCIe) 2x 2.5" HDD/SSD (SATA/NVMe)
网络	双万兆网口支持 NCSI 功能 (by PCH) 最高可支持 100G 网络连接
PCIe	最大支持 6 个 PCIe3.0 插槽 2 个双宽 PCIe x16 GPU (TDP 300W) + 2x FHHL PCIe x16 插槽 4 个 PCIex16 FH 3/4 L card+ 2x PCIex8 FHHL card
电源	2 个Slim PSU, 支持1+1 冗余
机箱	2U（19 英寸 /430mm）
	支持壁挂式安装和机架式安装
工作温度	长期运行 5℃-40℃； 短期运行-5℃-45℃