GPU WorkStation

支持标准PCIe GPU智算卡的桌面型智算工作站

全国产化体系打造、让高性能智算服务不再局限数据中心

瀚诚科技 · 桌面型 GPU工作站

国产智算新势力，释放AI算力无限可能

在AI时代，算力就是生产力。**瀚诚科技**推出全新桌面型GPU工作站，以紧凑设计承载强大性能，重新定义国产智算设备的边界。

- **高密度算力配置：** GPU支持最多2颗与4颗扩展的P2 /P4，满足企业推理/微调、知识库、政务公文、ChatBi、代码编程助手、科学计算等多场景的内网算力需求。
- **标准PCIe架构：** 兼容主流 PCIe GPU，灵活扩展。CPU与内存采用嵌入式设计，把办公室220V电耗极力留给GPU。
- **自主设计Switch芯片：** 核心互联不再依赖进口，PCIe交换芯片完全国产化，实现高速低延迟的数据通路。
- **国产+国际双支持：** 兼容 国产GPU 与 NVIDIA GPU，为用户提供更广泛的选择空间。
- **桌面级体积，数据中心级性能：** 小巧机身，轻松部署于办公桌面，释放本地算力潜能。

瀚诚科技坚持自主可控路线，从主板到互联芯片，全面国产化设计，保障数据安全与供应链稳定。为科研机构、政府单位、工业企业提供可信赖的算力基础。

项目	技术规格
产品信息	桌面型GPU智算工作站、Ai WorkStation：AWS P2、P4
通用算力	国产处理器算力芯片FPGA+ASIC嵌入式（PCIe卡式形态）；系统磁盘 480GB SSD、数据磁盘分别支持4颗和8颗的P2与P4，标准2.5寸磁盘，P4额外配置960GB SSD专为大模型载入的高速缓存。 配置1Gb以太网提供算力与模型调用，可选2个10Gbps。
智算支持	AWS P2 /P4分别支持2颗与4颗标准PCIe GPU智算卡（单/双宽），已经适配广泛的国产厂商如：天数智芯、摩尔线程、登临、寒武纪、沐曦、燧原、壁仞等等，以及AMD与NVIDIA。可覆盖≤FP16 /BF16 -70B尺寸模型。
电源	P2：1*1600W、P4：2*1600W
环境	5° C~35° C（41° F~95° F），符合ASHRAE Class A1/A2
尺寸	P2：L（423mm）*W（290mm）*H（150mm） P4：L（434mm）*W（444mm）*H（160mm）
增值软件	基于WebUI的GPU调度AIOS管理系统：GPU性能监控、本地模型上传、公网开源模型载入、模型参数设置与自助式创建API Key、主机管理与维护等统一管理操作；可配置本地磁盘、创建本地存储系统、设置共享存储、磁盘安全策略。可外接非结构化数据的第三方存储系统，支持广泛的NAS协议，支持Object协议；

提供完善的交付能力，从国产GPU驱动程序、算子适配、框架优化与调优、一站式交付、解决国产GPU市场化推广难题的最后一公里！

瀚诚科技 VASTOON

Multi-GPU Platform for Large-Scale

智算大模型一体机 MGP-410

我们不追求降本增效带来的锦上添花，而是富有激情的去探索科学边界，
通过一系列革新性的科技成果，提升IT组织与企业的生产力水平，
让自主可控更高效！

- **瀚诚**科技专为大规模算力构建的 MGP-410可支持高达10颗标准PCIe GPU的扩展能力。
- 通过国产体系CPU与GPU构建完全自主可控的智算堆栈，让算力高效运行的同时加安全可控。
- 已经深度适配主流的GPU厂商，让企业在GPU选型谈判方面掌握更多的主动权。
- 基于AGC体系构建，强调以GPU为核心的技术理念，迫使智算不在受限于CPU总线带宽瓶颈，即：通过1颗处理器可全速运行10颗标准尺寸的PCIe GPU。

MGP-410
独领全球的创新型
智算平台已然就绪！

项目	技术规格
产品信息	中型与大型算力的GPU整机单元，具备10颗GPU扩展能力
通用算力	可支持标准的EATX通算主板，已适配：C86海光、飞腾ARM等自主可控体系； 同时可选Intel与AMD X86体系。 内存与磁盘依赖于CPU支持能力，可支持外接高性能的智算存储系统；
智算支持	支持10颗标准PCIe GPU智算卡（单/双宽），已经适配广泛的国产厂商如：天数智芯、摩尔线程、登临、寒武纪、沐曦、燧原、壁仞等等，以及AMD与NVIDIA。
基于AGC构建的高级特性	1）支持GPU异构，即：支持多个品牌国产GPU在同一个节点，并提供丰富的任务调度策略。 2）革命性的GPU-热拔插特性，使通用GPU可像硬盘一样在线热拔插，提供极致的维护便利性。 4）每个GPU提供独立的嵌入式电源管理模块，能够在高吞吐Tokens与低频访问之间平衡电力消耗。 5）创新型GPU热备特性，促使GPU在本地具备N+1冗余特性。一旦GPU故障“全局热备”GPU会立刻接管，避免关键推理中断，以及为了业务连续性而购买备用的GPU整机用于Standby。
电源	3x：3200W热插拔电源，2+1模式。 2x：1300W热插拔电源，1+1模式。
环境	5° C~35° C（41° F~95° F），符合ASHRAE Class A1/A2
尺寸	4U、机架式：L(1000mm)*W（444mm）*H（177.8mm）
增值软件	基于WebUI的GPU调度AIOS管理系统：GPU性能监控、本地模型上传、公网开源模型载入、模型参数设置与自助式创建API Key、主机管理与维护等统一管理操作；

瀚诚科技 VASTOON

Multi-GPU Platform for Large-Scale

智算大模型一体机 MGP-820ls

基于AGC体系结构（AI computer system with the GPU at its Core），
颠覆以往通算底座构建的智算体系、通过构建可扩展20颗PCIe GPU密集
智算单元、打造下一代创新型智算平台！

独领全球的创新型智算体系已然就绪！
专为大规模算力集群互联打造的整机智算单元

我们的人工智能科学家深知在各厂GPU与框架的适配工作耗时、耗力，同时需要用户具备丰富的AI知识储备。现在，可以省去这些复杂，繁琐的工作。瀚诚科技团队在Ai-模型一体机预置了广泛的模型，并且已经完成与各家GPU厂商的适配!

- 革命性的GPU-热拔插特性，使通用GPU可像硬盘一样在线热拔插，提供极致的维护便利性
- 每个GPU提供独立的嵌入式电源管理模块，能够在高吞吐Tokens与低频访问之间平衡电力消耗
- 多个GPU可聚合成算力池专为：气象分析，数学运算，高精度计算，生物分析，基因工程与算法等，迫切需求大规模算力的单一任务需求创新型GPU热备特性，类似硬盘的RAID技术，促使GPU在本地具备N+1冗余特性。一旦GPU故障“Hot-Spare”GPU会立刻接管，避免关键推理中断，训练过程不可避免的Check Point回退，甚至为了业务连续性而购买备用的GPU整机用于Standby。

瀚诚科技专为大规模GPU算力扩展的智算单元，让GPU算力底座高效运行的同时更加安全可控，可选配国产CPU体系和广泛的GPU体系，让企业在GPU选型谈判方面掌握更多的主动权。

仅需1台820ls*1024GB RAM
16颗64GB国产训/推GPU，
同等性能，落地成本仅为竞商 60%

DeepSeek-671B-W8A8
最小配置2台（1024GB显存）
4颗ARM CPU、2048GB RAM。
并发192，1911Token/s；

DeepSeek-671B-W8A8
2个节点 vs. 1个节点

产品信息： 8U、20卡、机架式。单机16颗GPU即可运行671B全尺寸。专为超大参数模型同时需求高精度设计，同时规避跨节点组网而带来的延迟和性能损耗。

通用算力：

- 可支持标准的EATX通算主板，已适配：C86海光、飞腾ARM等自主可控体系。同时可选Intel与AMD X86体系。
- 内存与磁盘依赖于CPU支持能力，可支持外接高性能的智算存储系统；

智算支持： 支持20颗标准PCIe GPU智算卡（单/双宽），已经适配广泛的国产厂商如：天数智芯、摩尔线程、登临、寒武纪、沐曦、燧原、壁仞等等，以及AMD与NVIDIA。