NovaChip AI
鸿芯智算

GPU WorkStation for Ai Agent

Ai 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

鸿芯智算 · 桌面型 GPU工作站

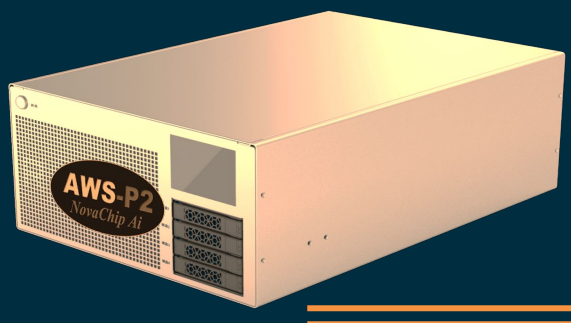
国产智算新势力，释放AI算力无限可能

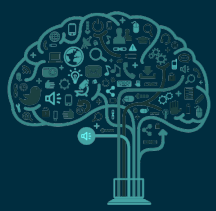
在AI时代，算力就是生产力。鸿芯智算推出全新桌面型GPU工作站，以紧凑设计承载强大性能，重新定义国产智算设备的边界。

- **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、代码编程助手等多场景的本地化部署智能体工具之需求。
- **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：** 兼容几乎全部的国产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。

鸿芯智算坚持自主可控路线，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。

技术规格	
产品信息	桌面型GPU Ai工作站、AWS P2、AWS P4两个型号
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、标准2.5寸数据磁盘分别支持4颗或8颗的P2与P4；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颗或4颗标准PCIe GPU智算卡（单/双宽），已适配广泛的厂商如：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。 配合国产GPU可覆盖广泛≤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）
调度平台	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！

NovaChip AI
鸿芯智算

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！

- 鸿芯智算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！

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；

NovaChip AI
鸿芯智算

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。诺亚鸿云团队在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！

NovaChip AI

- 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。

鸿芯智算专为大规模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、国产天数智芯-训/推GPU
- 同等性能，落地成本仅为竟商 60%

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

1个节点 vs. 2个节点
DeepSeek-671B-W8A8

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。

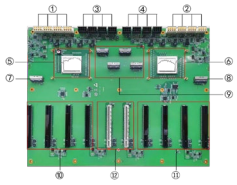
HCP-48 智算私有云节点

致力于交付一套整体的人工智能基础建设方案
通算＋智算一体机、构建安全隔离的智算私有云节点

HCP-48
4U 机架式
8 颗PCIe GPU

S160
通算加速卡

A100
GPU 加速单元



EXT-1
GPU – 基板

HCP-48 智算私有云平台，最初是为了满足某运营商的人工智能算力需求同时也要兼顾敏感业务而设计，其核心技术理念为：

- 确保存量的英伟达GPU智算卡，不会受到驱动与固件植入后门影响，同时规避被远程操控停机的风险，提供一套物理隔离的技术措施
- 通过硬件中间件方式，加速国产处理器性能，使其能像x86体系那样可承载私有云的能力，让配套人工智能的应用程序和系统以批量虚拟化或容器的方式部署后，提供“一机一站”式交付

业务与智算之间安全隔离	高密度与兼兼容性	智算环境采用加固隔离策略	全国产化体系自主可控的智算	具备信创认证的GPU加速卡	专为国产CPU设计的通算加速卡
处理器、内存/通算与GPU智算之间两段化设计通过独立的基板物理隔离隐患	最高可扩展8颗通用型PCIe GPU、没有厂商限制：天数智芯、沐曦、燧原、登临、壁仞或英伟达与AMD	GPU智算卡的驱动、算子框架与整套智算环境通过硬件构建的沙箱环境运行，规避隐私泄露、驱动程序后门、造成的隐患	采用标准化的E-ATX主板，可选择飞腾、海光、龙芯自主可控体系隔离智算，而无需担心是否兼容英伟达与	可配置2颗自主研发的A100用于4颗一组GPU智算卡的加速能力，A100之间同过自研的交换体系可突破500GB吞吐	S160 PCIe通算加速卡可有效的卸载CPU在处理云计算的性能开销，例如：存储IO、存储网络、分布式磁盘之间的纠删与复制等，让处理器更专注业务

技术规格

产品信息	4U、最多支持8颗PCIe GPU智算卡、可提供智算隔离的算力平台
通用算力	E-ATX标准主板，支持：飞腾、海光、龙芯等自主可控的指令集
智算支持	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。
智算加速	A100 GPU加速卡可对4颗PCIe GPU进行加速，A100 加速卡之间通过自主研发的交换体系，为模型大量参数交互提供500GB带宽的吞吐
通算加速	S100/S160 可有效的卸载处理器用于基础设施的算力消耗，具备自组存储网络的能力而需消耗CPU对于存储协议的性能开销，嵌入式存储系统可构建高性能分布式存储，S160更是通过内置RAID 芯片，强化了对于数据的自主可控能力。
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；

NovaChip AI
鸿芯智算

鸿芯智算

C2C Powers the Future of AI Computing

C2C 系列提供丰富的智算负载场景

基于颠覆性互连体系结构，重塑智算产业的“芯秩序”、联合-东芯半导体打造下一代创新型人工智能基础设施！

Tiny 2 – 是我们打造的 AIPC全场景算力负载产品，灵动小巧，内置2颗GPU算力芯片：

- **24 Core、32线程、64G RAM**
- **WiFi、蓝牙、2 TypeC、1 OcuLink、2 CDFP**
- **智能算力：**
  - 2 GPU Chips 、＞500TFLOPS (FP8)
  - 32GB 显存
- OLED 实时显示负载信息
- 推理、渲染、图形、仿真，或通过旋钮切换为显卡扩展坞
- 最多3台通过 CDFP端口组成超节点，提供300%算力与显存
- 180mm*260mm*90mm

配置高密度芯片的企业级智算系统

16C – 可移动智算一体机

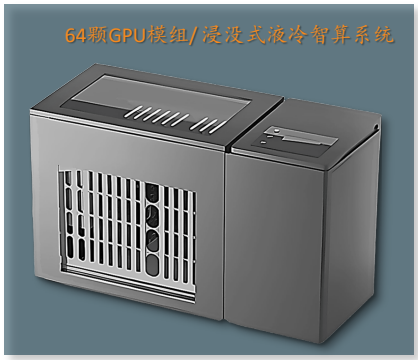
- 算力超过 ＞5PFLOPS(FP16)、
- 16颗GPU模组、768GB 显存可选
- 类CUDA、DirectX、OpenGL
- 经过特殊设计的48V直流供电
- 符合信创：CPU 16*Core、128线程、最高2TB 内存

冗余电源设计

- 经过减震优化的万向轮可移动算力单元
- 适用苛刻的运行环境而无需机柜即可获得算力
- 适用于特殊行业的野外算力调度

32C/64C –机架智算一体机

- 配置（一）东芯 8120*64颗：1024 GB 显存 ＞12 P(FP8)
- 配置（二）东芯 8140*64颗：1024 GB 显存 ＞17 P(FP8)
- 基于创新“C2C架构”构建的，具备颠覆性创新的新型高密度智算系统
- 高性能，东芯8000系列全Mesh 芯片互联，超低的Tokens转换成本
- 多场景负载能力：人工智能，图形图像渲染，仿真与图计算



64L –机柜式智算一体机

- FP 16 ＞10 PFLOPs
- FP 8＞ 20 PFLOPs
- 64 Chips of 8180、Up 3TB RAM
- 8 个200Gbps、智算组网、RoCE 协议
- 4 个100Gbps 存储网络及业务网络
- 1810mm*800mm*1300mm（长宽高）
- 电源 50KW
- 可满足广泛的算力负载场景：类CUDA、支持DirectX、OpenGL

高性能
首屈一指的
高密度算力整机

维护性
行业仅有的
可在线维护智算

高可用性
内部组件冗余
提供板卡级的热备

适用性
推理/训练/多模态
应对复杂的智算服务

高级特性

- **专为智能算力平台研发的智能BMC监控系统**
- **板卡级别的“热拔插”特性，提供极具便利的维护性**
- **高度冗余：具备“GPU全局热备”，最大限度保障业务连续性**