Multi-GPU Platform for Large-Scale

智算大模型一体机 MGP-820ls

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！

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！

H-Cloud THE AUTHORITY 技术革新

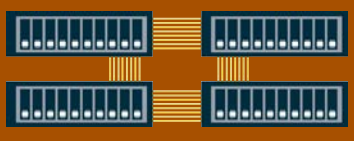
- 革命性的GPU-热拔插特性，使通用GPU可像硬盘一样在线热拔插，提供极致的维护便利性
- 每个GPU提供独立的嵌入式电源管理模块，能够在高吞吐Tokens与低频访问之间平衡电力消耗
- 支持GPU异构，即：支持多个品牌国产GPU在同一个节点，服务同一个大模型，并提供丰富的任务调度策略“GPU池化”特性，多个GPU可聚合成算力池专为：气象分析，数学运算，高精度计算，生物分析，基因工程与算法等，迫切需求大规模算力的单一任务需求
- 创新型GPU热备特性，类似硬盘的RAID技术，促使GPU在本地具备N+1冗余特性。一旦GPU故障“Hot-Spare”GPU会立刻接管，避免关键推理中断，训练过程不可避免的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。

我们深知：
让科技企业保持竞争力的条件之一、便是具备“持续创新”的能力。
这导致我们不断探索与量化、每次“创新”对于客户产生的实际效益!

诺亚鸿云：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、已适配国产CPU体系和广泛的国产GPU厂商、让企业在GPU选型谈判方面掌握更多的主动权。

10颗GPU扩展的MGP-410、20颗GPU扩展的MGP-820ls
可支持以上的关键特性、采用相同的开放式体系结构设计！

BF16/671B 仅需1台820ls*1024GB RAM
16颗48GB国产GPU，768GB 显存
同等性能，成本仅为竞商 40%



FP16/BF16最小配置4台（2048GB显存）
INT8最小配置2台（1024GB显存）*700GB RAM。
并发192，1911Token/s;

671B、BF16精度
4个节点 vs. 1个节点