

联想万全异构智算平台

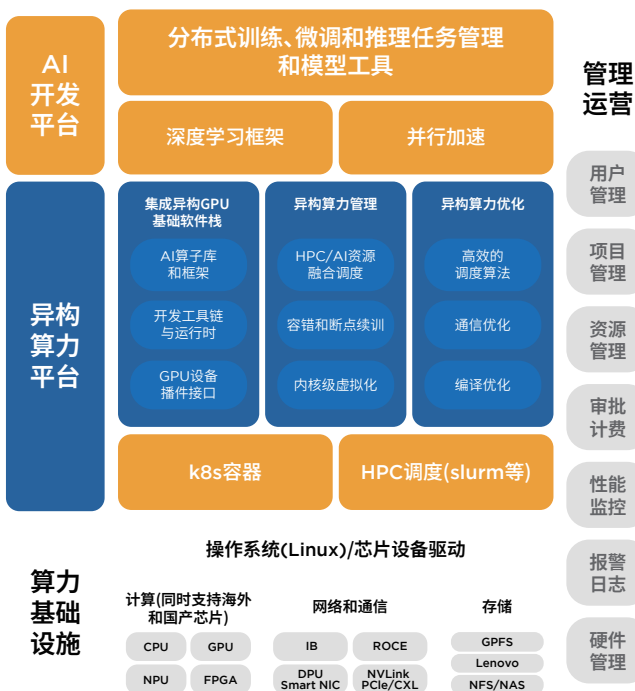
提供多元、稳定、高效的AI算力

产品概述

联想万全异构智算平台是一个能高度自动化完成AI全流程开发的平台，可输出高可用算力并不断突破计算效率瓶颈的利器。

用户可自动完成AI计算并发布模型或发布推理服务。此外，平台也为专业的AI开发用户留有手动深入调整计算过程的空间，包括工具和模型的选择，算力配置的调整，对任务的定制化监控等。

逻辑架构图



产品价值



用户智能匹配算力的算力匹配魔方。



逼近GPU算力极限的 GPU内核态虚拟化。



提升网络通信效率的 联想集合通信算法库。



减少AI训练中断时间的 AI高效断点续训技术。



AI与HPC集群 超级调度器。

如需更多信息

了解有关联想万全异构智算平台的更多信息，请联系 Lenovo 代表或业务合作伙伴，或访问biz.lenovo.com.cn



400-819-6776

以大模型为基础特征的AI 2.0时代已经到来，“大模型+大算力+大数据”成为新一代人工智能发展的基本范式。

智能计算较通用计算的投入成本、技术门槛更高，广大用户需要一个新的计算开发和管理平台。

挑战一

场景日益丰富，算力复杂多元，如何选择和匹配合适的算力？

目前已经有五种以上主流算法框架和十种以上算子库来支持不同的应用场景。同时，不同类型的GPU、服务器、存储和网络组成了上百种不同的集群配置，特别是场景、算法和集群之间的选择和验证，不但技术繁杂而且周期长，这都给用户带来了极大的压力。

挑战二

从千卡到万卡，如何减少故障恢复时间，提升集群的可用性？

以千卡训练集群为例，目前千卡集群每月至少有15次的故障断点，在常规的断点续训手段下，每次恢复训练需要几个小时，产生的额外费用超过百万元，而随着AI集群规模从千卡到万卡，故障中断次数及恢复所需时间呈指数级增长。

挑战三

智能算力基础设施的建设成本极高，如何改善和最大化AI算力利用率？

AI对算力的需求是无穷尽的，然而随着集群集成的芯片越来越多，我们不得不应对算法效率不高、计算资源不足、互联带宽受限等众多技术挑战。数据显示，当前业内MFU（模型算力利用率）普遍在30%左右，顶尖的集群利用率也只能做到50%，因此这就需要行业突破和解决GPU卡利用率、集群通讯效率、AI故障恢复、算法匹配度等系统性问题。

如需更多信息

了解有关联想万全异构智算平台的更多信息，请联系 Lenovo 代表或业务合作伙伴，或访问**biz.lenovo.com.cn**



400-819-6776