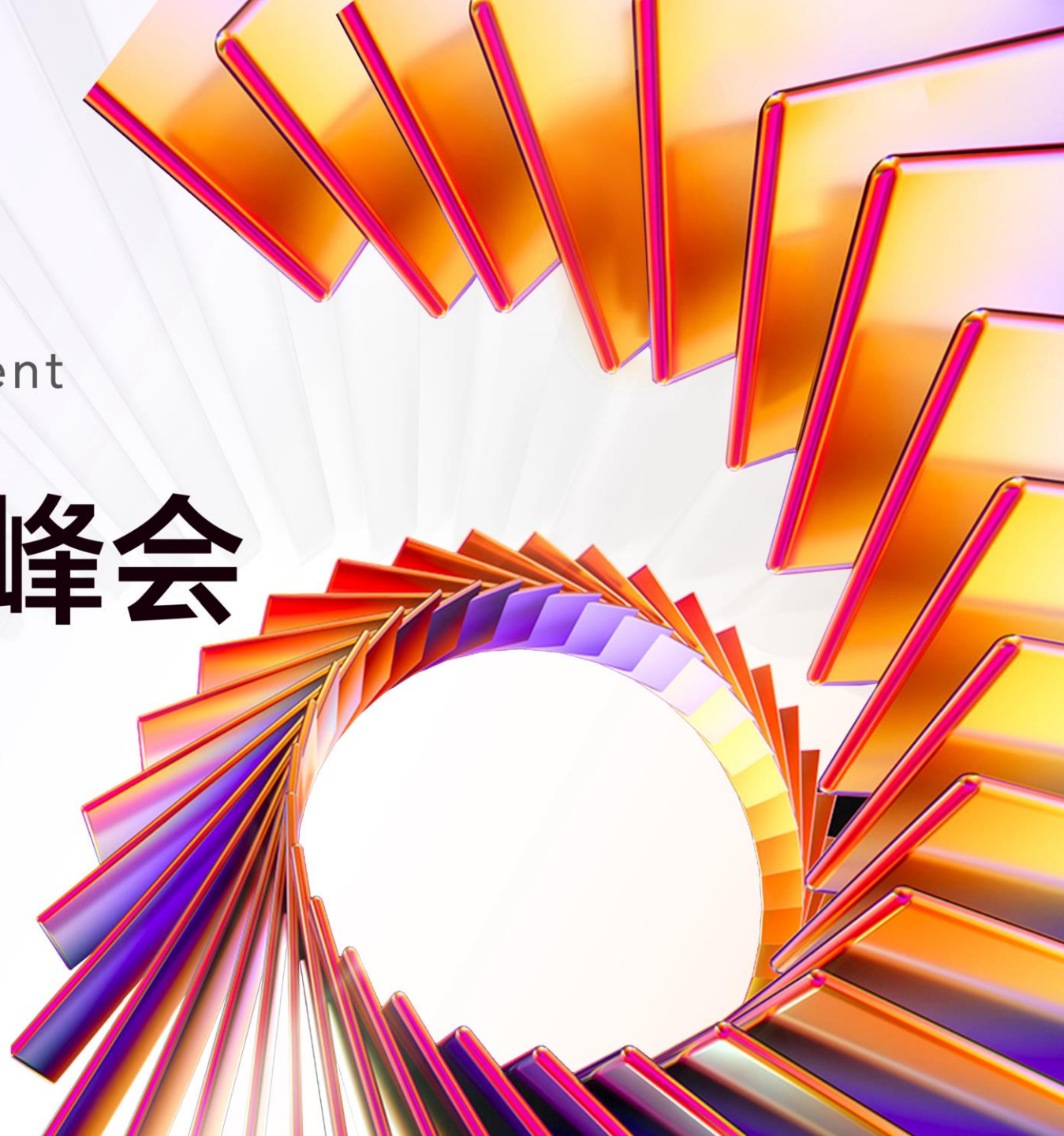
AiDD 7th 2025

第7届 AI+ Development Digital Summit AI+研发数字峰会

拥抱AI 重塑研发

8月8-9日 | 北京站



AiDD 8th 2025

第8届AI+研发数字峰会

拥抱AI 重塑研发 AI+ Development Digital Summit

查看会议详情

下一站预告 11/14-15 | 深圳站 12/19-20 | 上海站

深圳站论坛设置

智能装备与机器人

超越“编程 Copilot”

下一代知识工程

智能网联与汽车智能化

AI 测试工具开发与应用

AI 基础设施和运维

数据智能及其行业应用

可信 AI 安全工程

大模型和 AI 应用评测

多 Agent 协同框架

从智能测试到自主测试

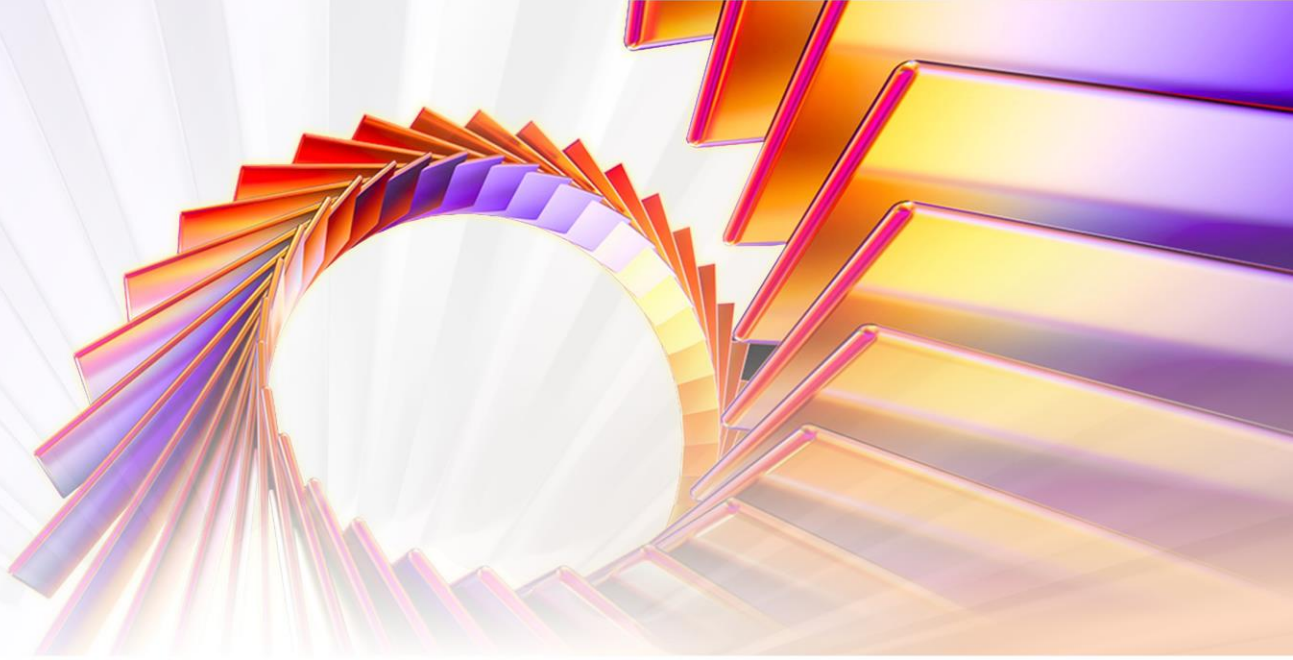
大模型推理优化

多模态 LLM 训练与应用

智能化 DevOps 流水线

上下文工程

AiDD 7th 2025 | 8月8-9日 | 北京站
第7届 AI+ Development Digital Summit
AI+研发数字峰会
拥抱AI 重塑研发



基于用户视角的算力服务及算网服务

陈健 博士
中国计算机学会、北京并行科技股份有限公司

陈 健

中国计算机学会（CCF）副理事长

北京并行科技股份有限公司（BJ839493）董事长

CCF高专委常委，CCF人工智能专委执委，CCF YOCSEF主席（2019-2020），TEEC清华企业家协会北京分会副主席，清华航院校友会常务副会长。

1993-2002年于清华大学获流体力学学士和博士学位，期间在荷兰TUDelft访问学者一年；

2016-2021年为清华五道口金融学院GFD全球金融博士生、日内瓦大学财富管理博士生；

2002-2005年，任联想集团高性能服务器事业部方案处经理、副主任工程师；

2005-2010年，任英特尔中国高性能计算架构师、资深性能优化工程师；

2010年起，历任并行科技CTO、CEO、董事长；

2011年作为创始合伙人，与中国科学院计算机网络信息中心、北京市怀柔区政府共同筹建北京超级云计算中心，2020年北京超级云计算中心A分区荣登中国超算Top100排行榜第三名，紧随先后登顶世界第一的天河二号和太湖之光之后，助力中国科研和科技发展，截止2023年连续四年蝉联中国超算通用CPU算力第一。2023年北京超级云计算中心成为北京市通用人工智能产业创新伙伴计划成员名单（第一批）唯二算力伙伴之一。

目录
CONTENTS

PART 01

算力行业发展现状与服务运营思路

算力服务四类市场

	尖端超算	通用超算	业务超算	智能超算
客户画像	**万核以上的应用**，追求极大规模、极致性能；高端超算的从业人员。攻坚型科研，国家级客户，各行业顶级研究机构；对超算硬件系统要求非常高	**万核以下的应用，绝大多数是千核以下的应用**，需要优质服务，关注性价比；海量无超算资源用户的日常需求；当前自主建设中小微超算系统	**单核到几千核应用**，行业实际业务，关注服务，关注性能和性价比；超算只是业务中一个环节，需要实现完整业务上云，需要保证业务运行稳定性、可靠性	**以GPU算力为主，单卡到万卡的应用**，计算量极度密集，需要优质服务，关注性价比；算力投资大，自建较少，主要使用租用智算算力资源
产品定位	**“塔尖上的明珠”，“国之重器”；**计算、访存、通信、I/O都非常出众，性能设计很平衡的高端超级计算机；需要国家集中力量办大事，需要国家投入，不能核算性价比；**国家超级计算中心**	海量用户需求聚类，应用运行特征分析，针对不同类型应用，动态随需增长式建设最高性价比超算服务计算资源；帮助用户从自建中解脱出来，租用超算服务；**超级云计算中心**	面向行业，按照行业业务需求设计完整的云上业务流程，保证用户业务各环节能够快速、高效、动态实现，弹性、高性能、高稳定性、高可靠性、高可维护性；**公有云/超算云+专业超算服务商**	解决大模型训练需求的超算中心模式，和解决推理等需求的云计算模式，应用运行特征分析，动态随需增长式建设最高性价比智算算力资源；帮助用户从自建中解脱出来，租用超算服务；**智算中心**
计算类型	**FP64/FP32精度计算** • 典型领域：工业仿真、气候模式、计算化学、分子动力学CFD流体力学等高精度科学计算领域 • 典型应用：WRF、Lammps、Gromacs、OpenFOAM、 Ansys、Comsol、Abaqus			**BF16/FP8/FP4精度计算** • 典型领域： NLP、文生图、文生视频 • 典型应用：DeepSeek、LLaMa、YOLO、SDXL **FP32/FP16半精度计算** • 典型领域：**具身智能** • 典型应用：**Isaac Sim**

2021年-2026年，复合增长率：智算52.3%，通用18.5%

图2 中国智能算力规模及预测，2019-2026

来源：IDC，2022

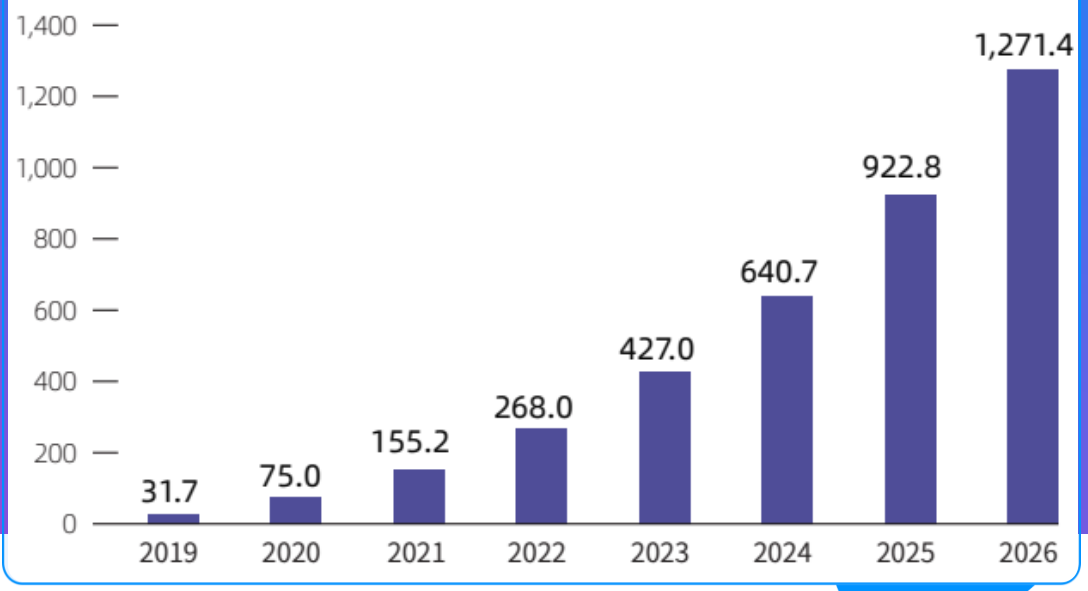
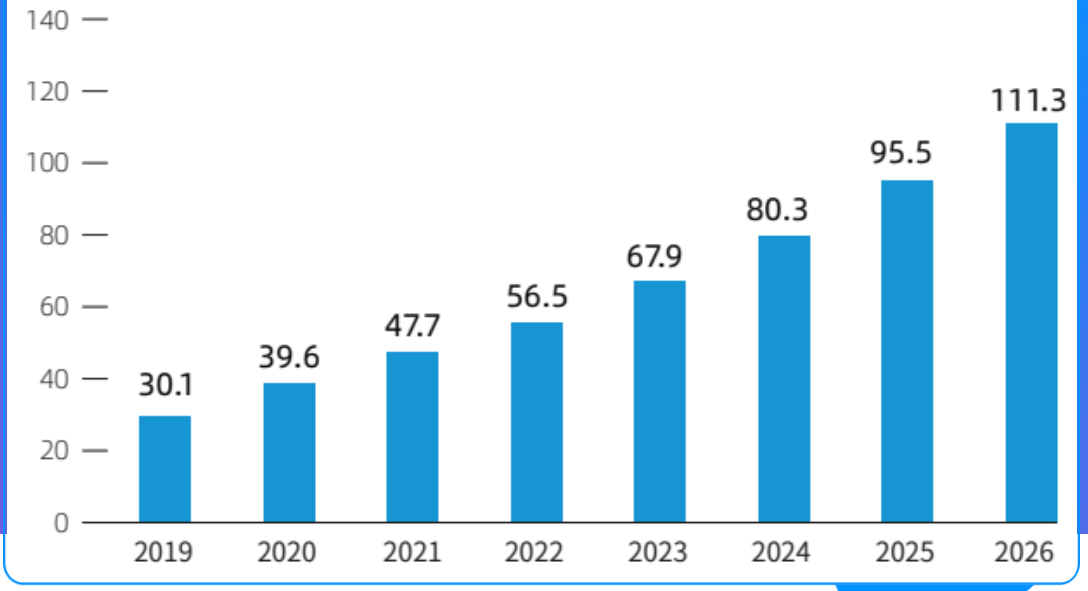


图3 中国通用算力规模及预测，2019-2026

来源：IDC，2022

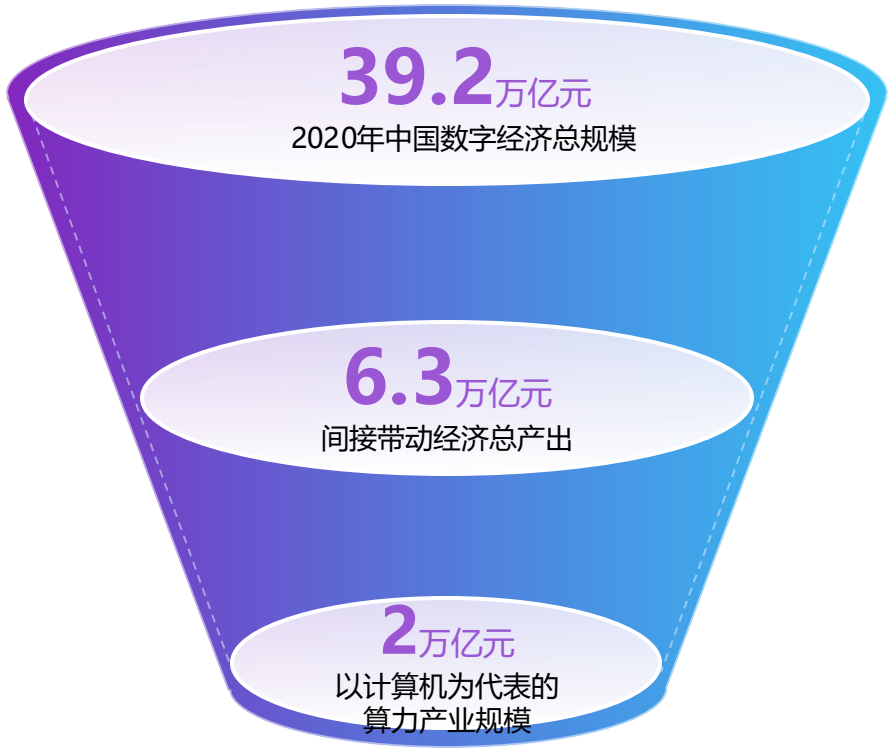


算力：数字经济时代的新生产力

数据、算力和算法是数字经济时代的关键资源
数据是新的生产资料，
算力是新生产力，算法是新的生产关系

生产要素	生产力	生产关系
土地	产业数字化	**科研范式** 算法创新
劳动力	算力生产 / 算力释放 / 算力聚合 / 算力调度	**生产方式** 算法创新
资本		**生活方式** 算法创新
技术		**治理方式** 算法创新
数据	数字产业化	

在算力中每投入1元，带动**3-4元**经济产出；
算力发展指数每提高1点，GDP增长约**1293亿元**。



来源：《中国算力发展指数白皮书》，信通院

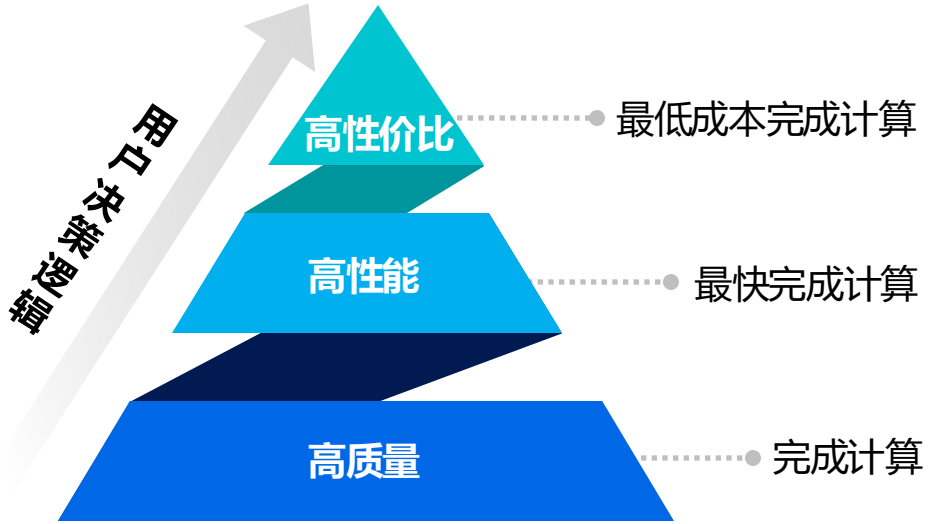
算力行业的本质

不是 资源业

是 服务业

算力行业	类比	核心
算力租赁	租车	整租、非按需，利用率风险由租赁方自己控制
算力服务	专车	专车、豪华车，有司机，优质服务，【按需使用】
算力运营	运营平台	复杂的调度平台（多类型算力资源/多类型网约车）
算力网络	全国调度平台	【算力网】调度全国所有运营平台的算力资源（服务车辆）

算力服务运营思路：从客户/用户需求视角，做业务/产品梳理



调度算力?
还是调度算力需求？

最终用户 不关心 功能
（管理员/软件厂商关心）

最终用户 关心 收益
（使用者/云服务厂商关心）

用户需求	产品形态	市场供求关系
超大规模大模型训练	裸金属	供不应求/长期供不应求
常规大模型训练/微调	裸金属/集群模式	供大于求/长期供大于求
推理	云主机	供大于求/需求稳定增长

智算芯片生态适配工作量

生态 适配工作量	跑起来/ 移植完毕	跑正确/ 运行正确	性能好/ 速度快/加速比高	性价比高/ 运行任务成本低
英伟达	工作量小	工作量小	工作量小	工作量中
AMD	工作量中	工作量中	工作量中	工作量中
Intel	工作量中	工作量中	工作量中	工作量中
华为	工作量中	工作量中	工作量中	工作量大
海光	工作量中	工作量中	工作量大	工作量大
其他国产芯片	工作量大	工作量大	工作量大	工作量大

智算用户应用需求和产品设计

用户应用类型	单用户需求量	总用户数量	产品设计	算力网络服务策略
基础大模型训练	极大	几个	超大规模的高端训练超级计算机	单点聚集海量算力资源（算海计划）
行业模型	中	几百	常规中高端训练集群/国产算力集群	算力网络服务模式 裸金属租用
场景模型	中小	几万	各种算力/国产算力均可用	优质算力服务 【按需使用】
学术/科研	中小	几万	通用算力为主，应对高频迭代开发	聚合各种算力资源的算力服务平台
主流大模型推理服务	极大	几十/几百	海量算力资源池 (通用+国产算力)	算力网络服务模式 整租+弹性按需
常规通用推理服务	中小	几十万	算力服务平台 (通用+国产算力)	聚合各种算力资源的算力服务平台
常规推理服务	中小	几万	通用算力为主，应对各种复杂场景	聚合各种算力资源的算力服务平台

国产GPU发展现状

国产GPU产品仍处在起步阶段，缺乏应用场景，产品性能与英伟达、AMD产品有一定差距，软件和生态较难竞争

企业名称	成立时间	最新融资/上市情况	核心技术	主要产品	应用领域
华为	1987年09月	-	全自主研发达芬奇芯片架构、完善的软硬件配套体系	910	信创、能源、交通、金融、人工智能等领域
芯动科技	2007年10月	-	LPDDR5X显存技术、低功耗	风华1号、风华2号	办公上网、娱乐游戏、工程制图、GIS等领域
上海兆芯	2013年04月	A轮	业内第一款集成CPU、GPU、芯片组的SoC单芯片国产通用处理器	KX-6000中的C-960 GPU	高性能桌面、便携终端、嵌入式等领域
天数智芯	2015年12月	10亿元 C轮&C+轮	自主架构的规模量产的7nm/2.5DGPGPU云端芯片	天垓100	自动驾驶、智慧医疗、智慧金融、智慧教育等领域
登临科技	2017年11月	A+轮	自主架构的规模量产的GPGPU高性能通用人工智能加速器	Goldwasser-UL(MXM)、Goldwasser-UL、Goldwasser-L	自动驾驶、智慧城市、生物医疗、AI计算等领域
瀚博半导体	2018年12月	16亿元 B轮	云端推理DSA架构 视频处理技术	SV100系列云端推理芯片、VA1通用AI推理加速卡	深度学习、AI推理、云计算等领域
智绘微电子	2018年12月	千万元 Pre-A轮	自主的指令集、自主的编译器和自主的架构	IDM919	图形图像显示、科学计算、人工智能等领域
壁仞科技	2019年09月	超47亿元 B轮	自主原创的GPU大芯片架构	BR100GPU	人工智能、云计算、图形渲染、大数据等领域
芯瞳半导体	2019年12月	Pre-A轮	统一渲染GPU架构，具有高度可扩展的互联结构和计算阵列	GenBu01GPU	信创、能源、交通、金融、人工智能等领域
摩尔线程	2020年06月	20亿元 A轮	3D图形计算和高性能并行计算技术	MTTS60、MTTS2000、MTTS10	物理仿真、AI计算、游戏娱乐、自动驾驶等领域

国产GPU发展现状

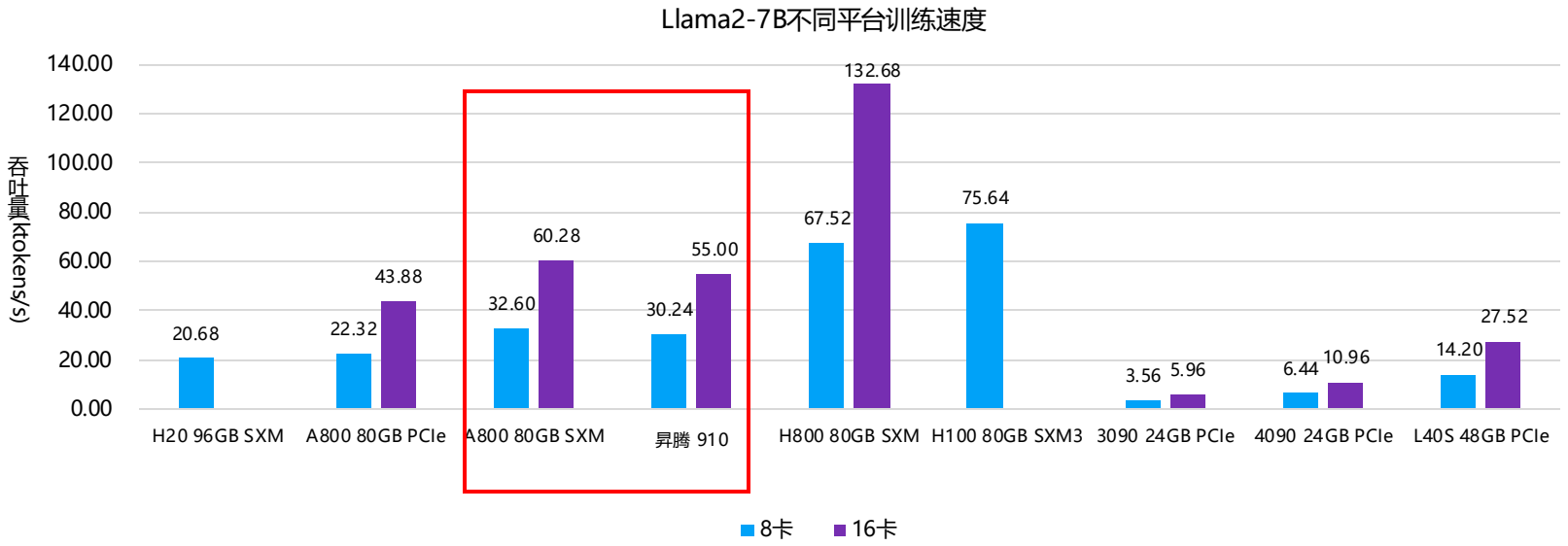
国产GPU产品仍处在起步阶段，缺乏应用场景，产品性能与英伟达、AMD产品有一定差距，软件和生态较难竞争

企业名称	成立时间	最新融资/上市情况	核心技术	主要产品	应用领域
燧原科技	2018年03月	18亿 C轮	全自主研发的芯片架构	S60	信创、人工智能、AI推理、视频解析等
沐曦集成电路	2020年09月	10亿 Pre-B轮	自主研发GPU芯片架构、兼容国际主流生态的完整软件栈	MXN、MXC、MXG	深度学习、数据分析、物理仿真、云游戏等领域
深流微智能	2021年05月	Pre-A2轮	拥有下一代超级流处理XST架构	XST-G01、XST-E01、XST-C01	视觉计算、人工智能、高性能计算等领域
励算科技	2021年08月	过亿元 Pre-A轮	自研盘古架构、AI Powered GPU	TrueGPU图形芯片	端、云、边、车等领域
芯原股份	2001年08月	688521.SH	GPU IP供应厂商，GPU(含ISP)市场占有率排名全球前三位，2020年全球市场占有率约10.2%	ArcturusGC8800、GC8400、GC8200、Gc8000、GPU Nano IP	小型物联网MCU、人工智能等领域
景嘉微	2006年04月	300474.SZ	支持国产CPU和国产操作系统的自主知识产权GPU	JM5400、JM7201、JM92系列	军用工业、人工智能、云计算金融、教育等领域
龙芯中科	2008年03月	688047.SH	自研的统一渲染架构	7A2000桥片的集成GPU	金融、政务办公、网安、教育、通信、医疗等领域
海光信息	2014年10月	688041.SH	海光DCU属于一种GPGPU,深算一号DCU产品已实现商业化应用	深算一号DCU产品	人工智能训练等领域
寒武纪	2016年03月	688256.SH	AI训练GPU新品，搭载双芯片四芯粒封装的思元370,集成寒武纪MLU-Link多芯互联技术	MLU370-X8	人工智能训练等领域

国内AI算力实测现状

	910	A800
峰值FP32算力	94TFLOPS	19.5TFLOPS
峰值FP16算力	376TFLOPS	312TFLOPS
HBM内存容量	64GB	80GB
芯片峰值功耗	400W	400W
CPU-NPU PCIE规格	PCIe4.0*16 64GB/s	PCIe4.0*16 64GB/s
NPU-NPU带宽	392GB/s(HCCL)	400GB/s(NVLink)
RDMA出口带宽	200Gb/s RoCE	100Gb-200Gb/s IB

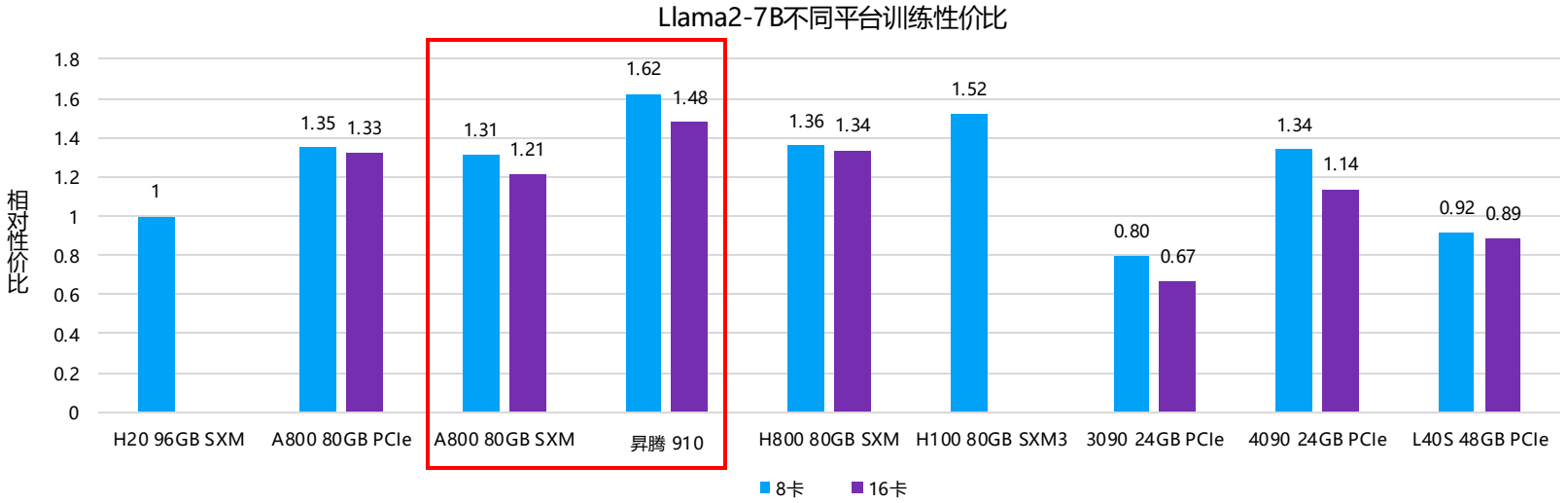
Llama2-7B训练不同平台性能与性价比



92.8%

910 vs A800 性能

昇腾910大模型训练性能，经过优化，可达A800 80GB SXM的92.8%。



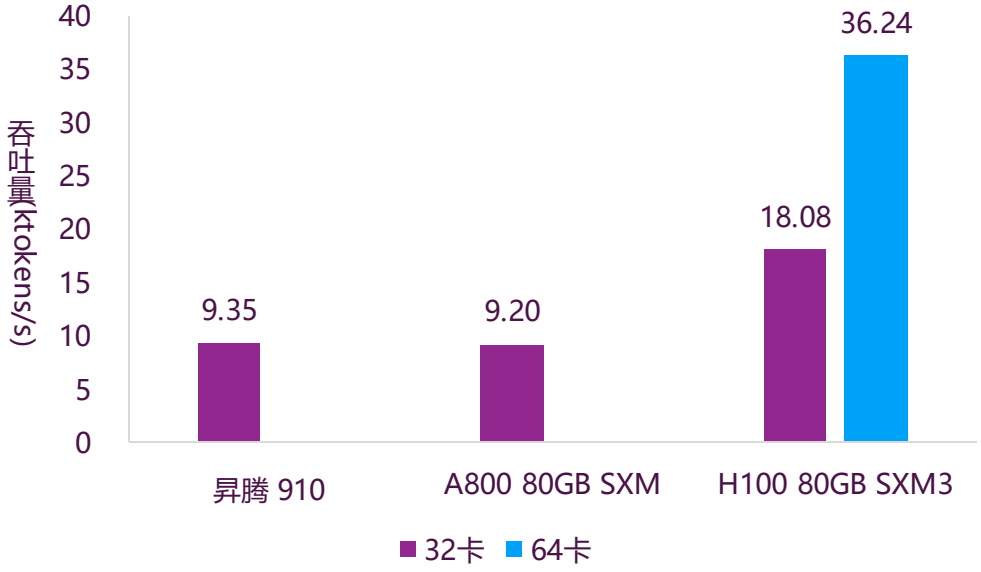
1.23

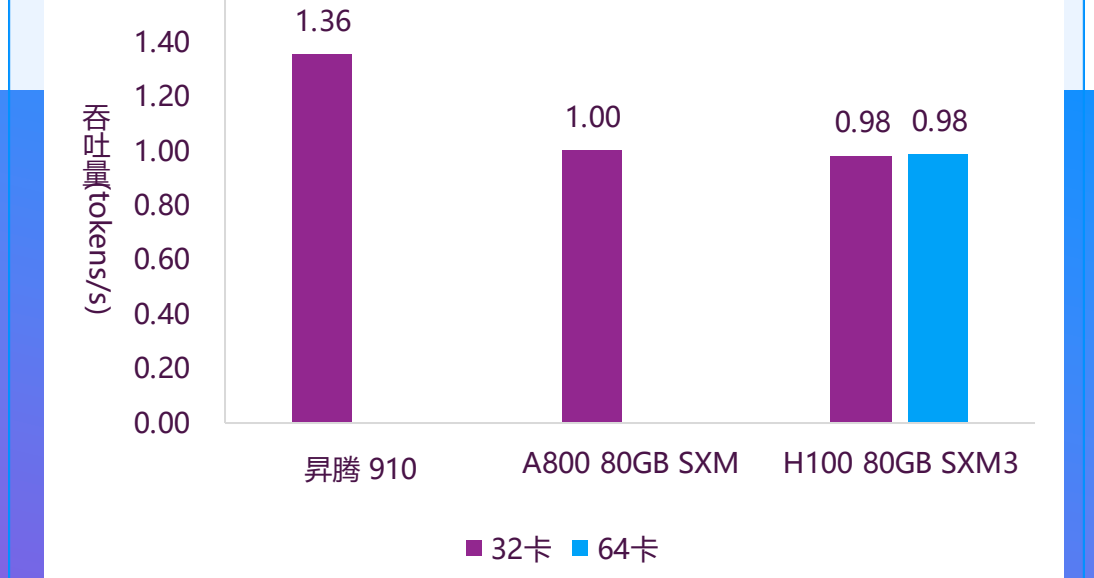
910 vs A800 性价比

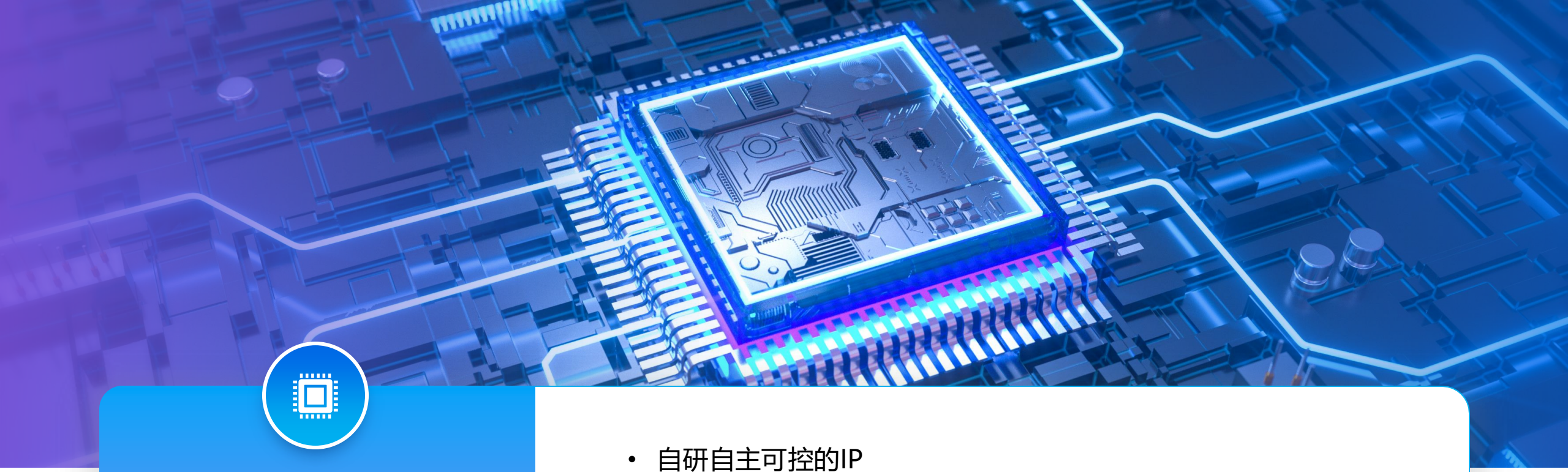
昇腾910性价比可达A800 80GB SXM的1.23倍，领先当前各个NVIDIA平台。

Llama3-70B 不同平台训练速度及性价比

Llama3-70B训练，昇腾910性价比可达A800的1.36倍







国产GPU面临的挑战

- 自研自主可控的IP
- 与国产CPU同频共振，实现控制调用
- 解决产能问题
- 应用生态仍待进一步丰富，需要海量真实业务的计算任务验证
- 获得行业用户的认可和信任

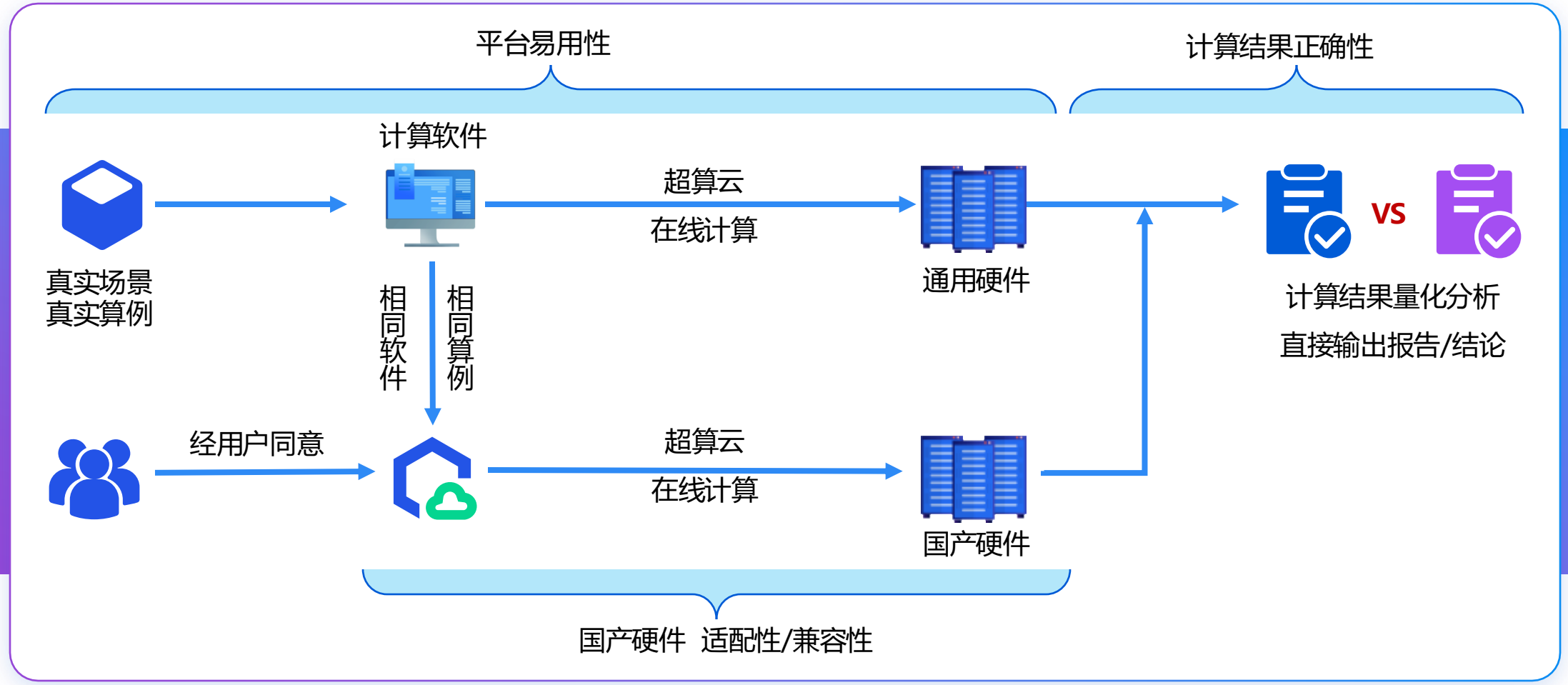
杀虫剂悖论

杀虫剂悖论

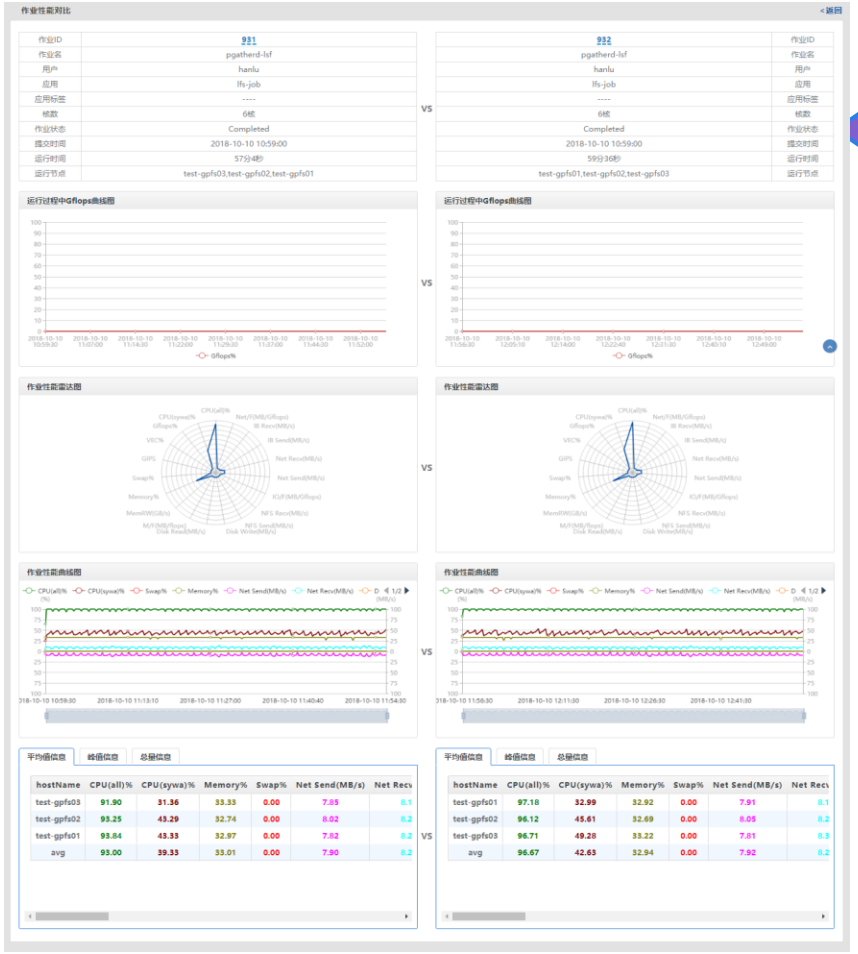
用同样的测试用例，**多次重复进行测试**，最后将不再能够发现新的问题

- 测试用例需要定期评审和修改
- 不断增加新的不同测试用例，来测试硬件和软件的不同表现

国产算力硬件测试验证


平台易用性
计算结果正确性
计算软件
超算云
在线计算
通用硬件
真实场景
真实算例
相同软件
相同算例
VS
计算结果量化分析
直接输出报告/结论
经用户同意
超算云
在线计算
国产硬件
国产硬件 适配性/兼容性

测试结果实时展示



编号A0001	xx大学	CFDXXX	1000核	结果：正确
编号A0002	xx大学	CFDXXX	1000核	结果：正确
编号A0003	xx大学	CFDXXX	1000核	结果：正确
编号A0004	xx大学	CFDXXX	1000核	结果：正确
编号A0005	xx大学	CFDXXX	1000核	结果：待确认
编号A0006	xx大学	CFDXXX	1000核	结果：正确
编号A0007	xx企业	CFDXXX	1000核	结果：正确
编号A0008	xx企业	CFDXXX	1000核	结果：正确
编号A0009	xx企业	CFDXXX	1000核	结果：正确
编号A0010	xx企业	CFDXXX	1000核	结果：正确

国产算力适配评估框架

硬件兼容性评估

- **显存需求分析**：初步判断国产硬件是否满足新模型运行的显存要求，考虑是否需要对模型进行压缩或者对硬件进行扩展。
- **指令集兼容分析**：关注模型运行所需的精度以及算子国产GPU是否支持。

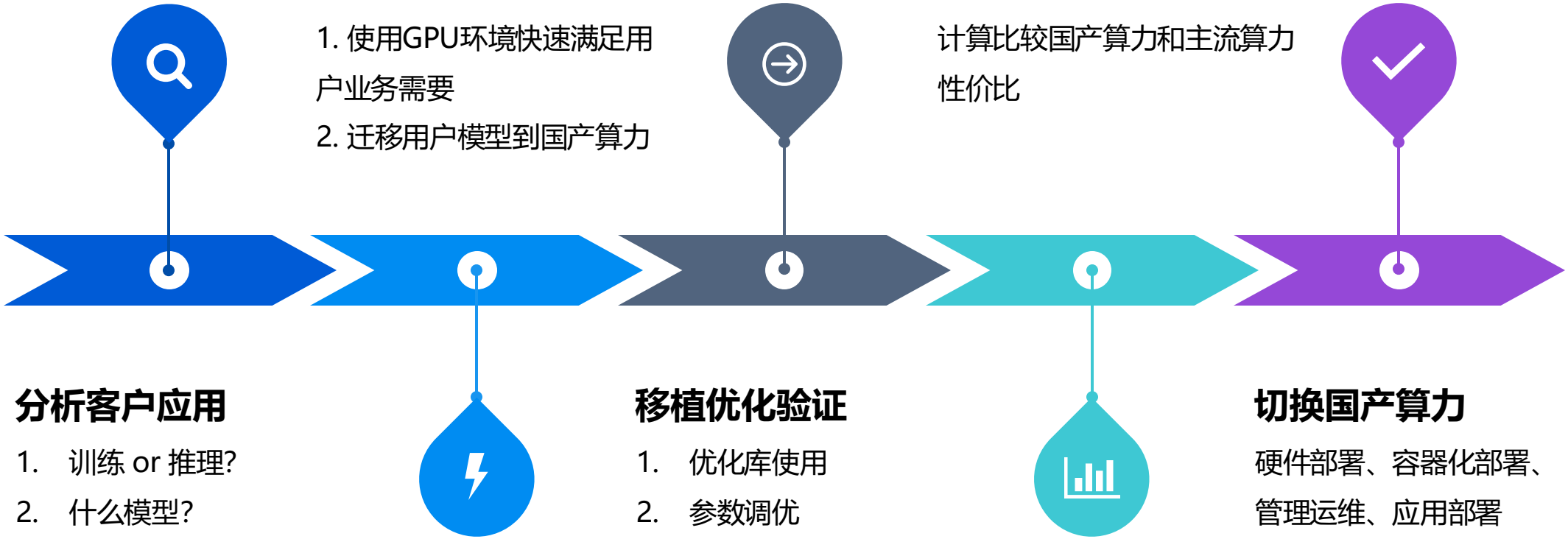
软件适配性评估

- **操作系统及软件工具链适配评估**：检查国产硬件所搭载的操作系统是否与新模型的开发环境和运行环境兼容。
- **深度学习框架适配评估**：确定新模型所依赖的深度学习框架是否支持国产硬件。此外，还需要关注深度学习框架的版本兼容性问题。

性能评估

- **ParaSelect性能预测**：基于ParaSelect对模型在新平台的性能进行预测。在具备条件时，进行实测。
- **性能优化成本估算**：评估为了达到新模型在国产硬件上预期的性能表现所需的技术，包括算子开发、推理引擎优化等，从而评估需要进行的优化工作量和相关成本。

国产算力支持案例



分析客户应用

1. 训练 or 推理?
2. 什么模型?
3. 什么框架?

快速实现可用

1. 使用GPU环境快速满足用户业务需要
2. 迁移用户模型到国产算力

移植优化验证

1. 优化库使用
2. 参数调优

性价比计算

计算比较国产算力和主流算力性价比

切换国产算力

硬件部署、容器化部署、管理运维、应用部署

国产算力支持案例

某客户从N卡切换到昇腾的过程：

01

平台精度对齐

用户对昇腾平台训练精度有疑问。通过比较训练loss曲线，证明昇腾精度可与N卡对齐。

02

框架精度对齐

用户对ModelLink训练精度有疑问。通过比较训练loss曲线，证明ModelLink精度可与torch原生对齐。

03

部署推理服务

对于华为已经官方支持的Llama模型，使用MindIE部署；对于华未支持的Gemma2模型，使用trans-formers框架部署。

04

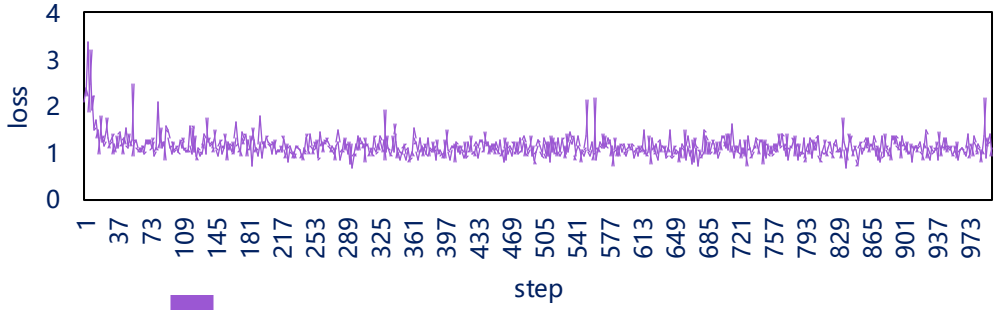
切换国产算力

用户正式切换到国产算力，帮助用户进行存储、网络搭建，基础软件环境搭建，数据迁移。

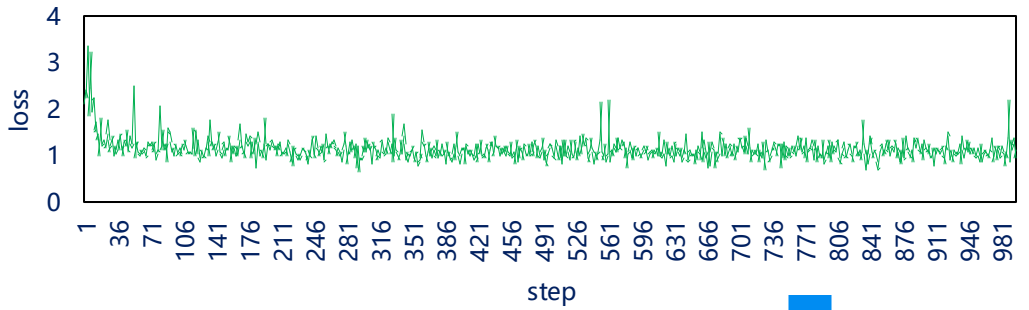
国产算力支持案例

比较训练loss曲线，昇腾精度可与N卡对齐

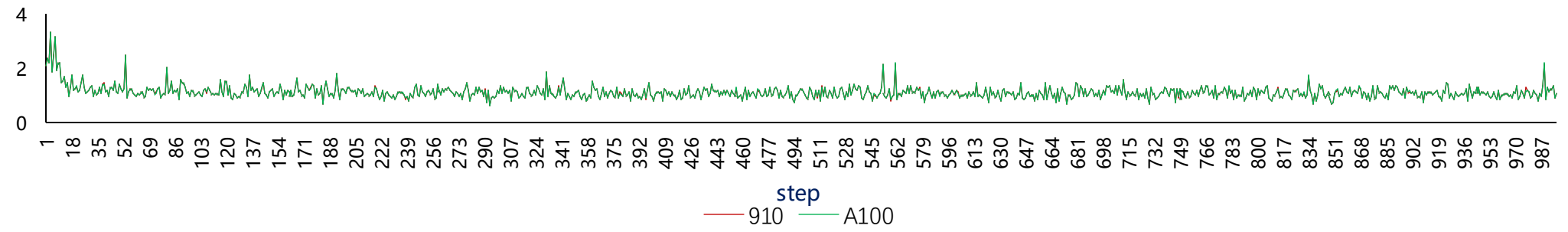
Llama3-8B在Stanford_Alpaca数据集上使用LlamaFactory框架
基于Zero3分布在910上训练的loss曲线



Llama3-8B在Stanford_Alpaca数据集上使用LlamaFactory框架基
于Zero3分布在A100上训练的loss曲线



Llama3-8B在Stanford_Alpaca数据集上使用LlamaFactory框架基于Zero3分布在910和A100上训练的loss曲线



以真实场景与结果，增强用户信心

国产硬件测试验证推广服务一体化平台

研发团队

- 真实算例
- 高可复用性
- 覆盖率容易度量
- 激励团队士气
- 积累推广数据

最终用户

- 无需参与测试过程
- 国外超算硬件可替代
- 增强国产超算硬件信心
- 测试低成本或无成本
- 计算结果获得“双”验证

硬件推广

- 精准覆盖行业用户
- 省去采购、建设等环节
- 国产硬件云化，加速赶超
- 快速构建应用软件生态
- 突破单一应用领域和范围

PART 02

大模型应用特征分析/选型/优化与案例

打造“算力买手”模式

凭借专业的选型能力，从众多的供应商/算力资源池中，挑选出最适合的算力产品或服务，并将其整合后提供给最终用户。

业务需求确认

1. 了解业务运行的需求边界条件 （如：用量规模、环境依赖、时长等）
2. 了解获取业务运行效率【可用/好用】的评判标准

了解业务运行的基础条件

1. 规模指标：算力总量（卡）、存储总量（TB）……
2. 环境指标：OS、驱动……
3. 时效指标：供给时间、供给时长……

……

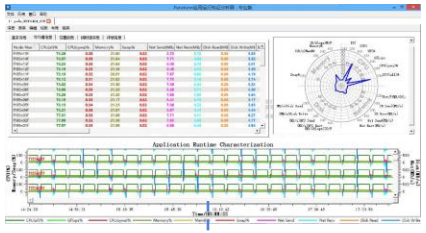
获取特定业务运行效率的评判标准

1. 性能指标：输出速度（Tokens……）、延迟（ms）……
2. 效率指标：利用率、吞吐量、并发处理能力……
3. 扩展指标：线性扩展能力……
4. 稳定指标：容错能力、恢复能力、

……

应用运行特征分析

1. 在典型资源上，仅需运行一次业务，即可完成应用运行特征分析
2. 呈现业务特点【计算密集型、访存密集型、通信密集型、IO密集型等】以及资源依赖程度



应用运行特征分析报告

基于性能数据梳理影响计算效果的核心要素

1. 精度分析：FP64、FP32、FP16、FP8等
2. 性能分析：GPU利用率、CPU利用率等
3. 吞吐分析：内存带宽、显存带宽、IO吞吐等
4. 通信分析：NVLink、计算通信、PCIe通信等
5. 瓶颈分析：内存瓶颈、显存瓶颈等
6. 规模分析：核数、卡数、并发量、加速比等

ParaSelect资源选型

1. 利用ParaSelect方法，评估业务在不同类型算力型号上的性能表现
2. 输出资源选型方案（涵盖：最高性能方案、最高性价比方案）

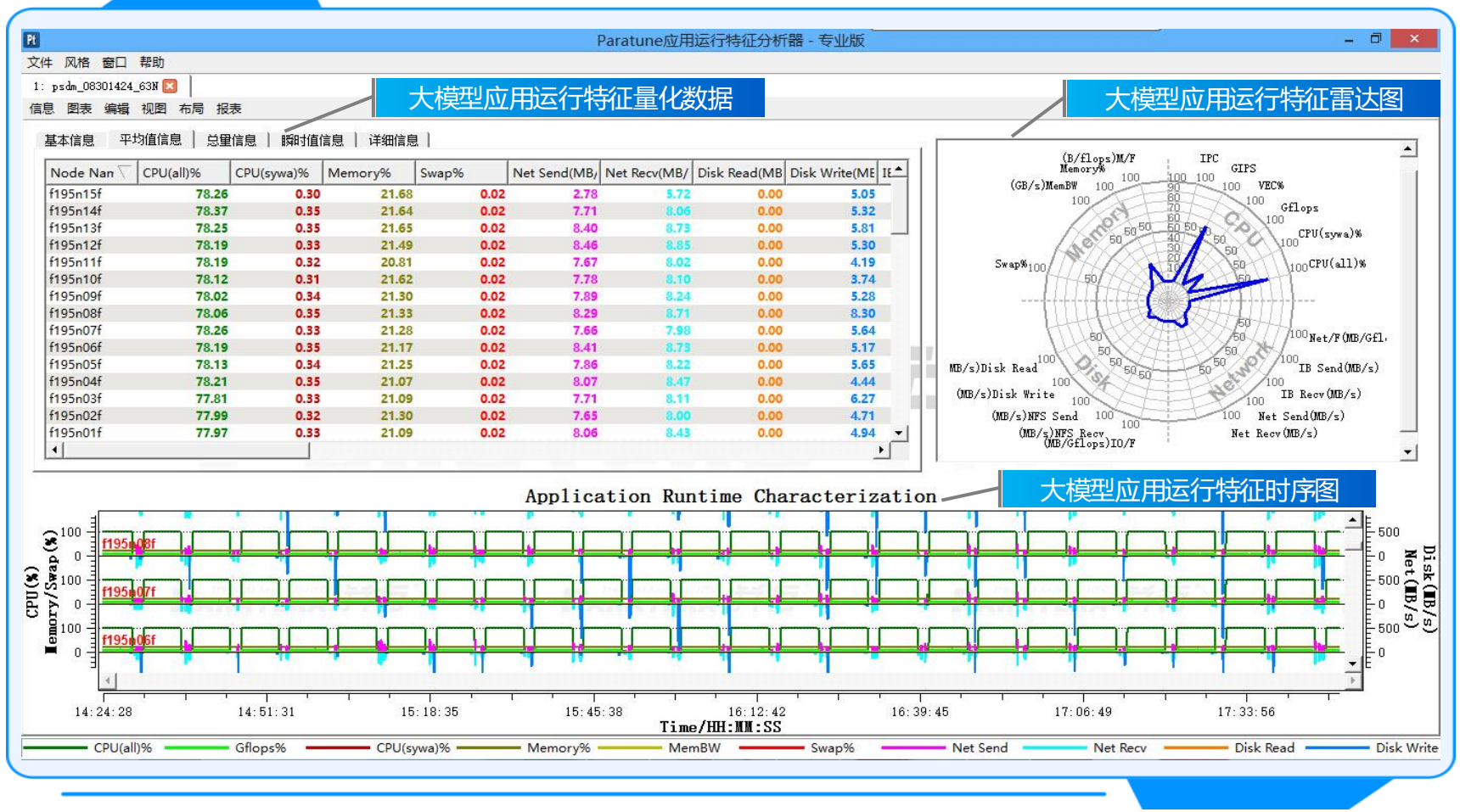
$$speed_{target} = speed_{h800} / \left(k_1 * \frac{TFLOPS'_{h800}}{TFLOPS'_{target}} + k_2 * \frac{gmem\ bw_{h800}}{gmem\ bw_{target}} + k_3 * \frac{PCIe\ bw_{h800}}{PCIe\ bw_{target}} + k_4 * \frac{nvlink\ bw_{h800}}{nvlink\ bw_{target}或PCIe\ bw_{target}}\right) * speedup(batchsize)$$

基于ParaSelect的算力选型方案

	成本与性能信息					交付信息		
方案1	配置	总价	项目时长	性价比	性能	可用x卡	供给时间	有效期
方案2	配置	总价	项目时长	性价比	性能	可用x卡	供给时间	有效期
方案3	配置	总价	项目时长	性价比	性能	可用x卡	供给时间	有效期

……

大模型应用运行特征方法论（选型 优化 降本 增效）



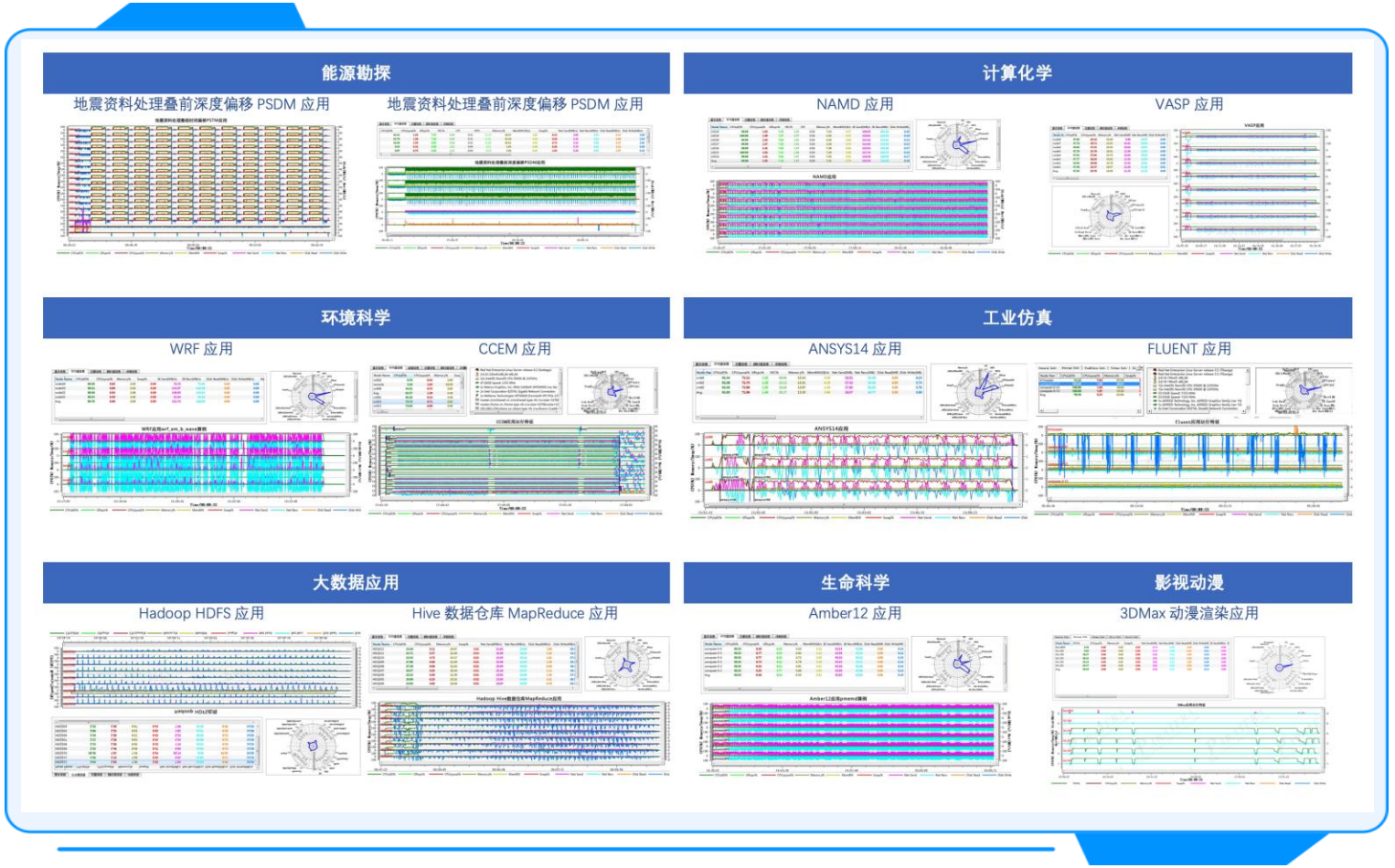
采集应用运行特征

- Tensor Core利用率
- FP32利用率
- FP16利用率
- 显存利用率
- 显存带宽利用率
- PCIe利用率
- NVLink利用率
- 硬盘读写速率
- GPU利用率
- IB读写速率

分析应用瓶颈

- 计算密集型
- 访存密集型
- 通信密集型

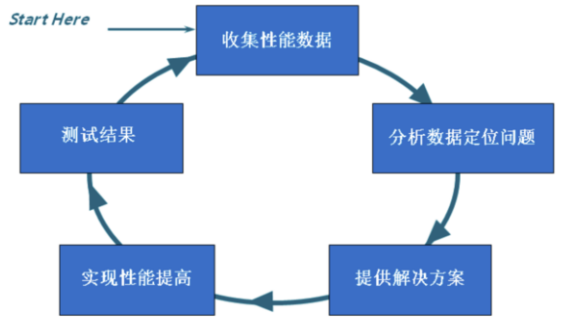
多行业应用运行特征图谱



针对运行在特定硬件上的大规模应用软件进行运行特征分析

通过数据收集和可视化定量统计分析方法

全面了解大规模应用软件在不同硬件平台上运行的全过程和局部细节



ParaSelect – 性能预测与智能选型

用应用运行特征分析方法，快速确定选型的最高性能和最高性价比

采集应用特征

计算，访存，卡间通信

节点间通信，显存容量

性能预测

性能最佳平台

性价比最佳平台

01 → 02 → 03 → 04 → 05

明确应用场景

训练 or 微调 or 推理，大模型

or 小模型，优化前 or 优化后

分析最适平台

大显存 or 小显存，GDDR or HBM

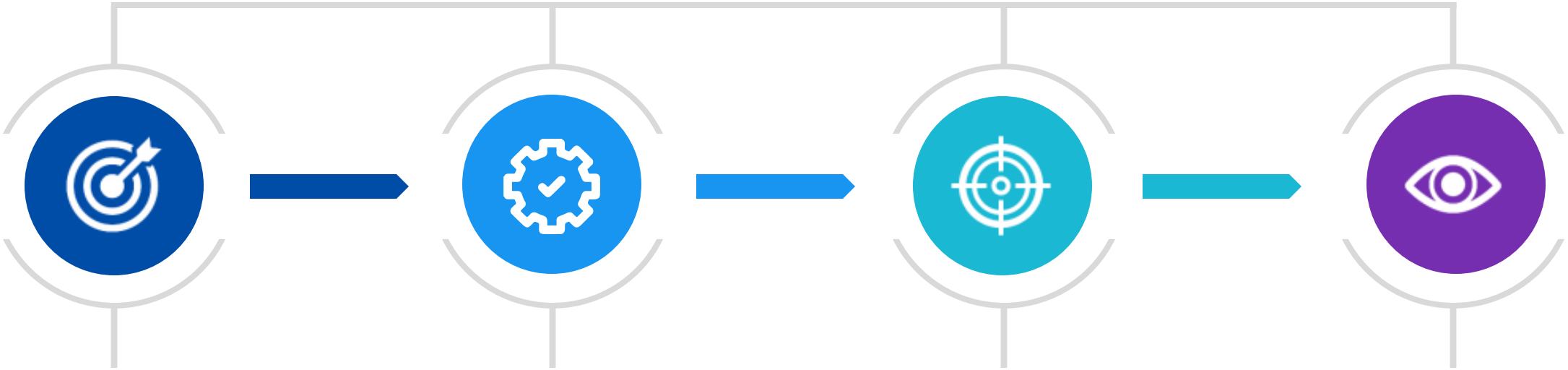
PCIe or NVLINK，RoCE or IB

性能验证

性能实测

性价比计算

ParaSelect – 性能预测流程



特征设计

以GPU平台应用最重要的4个参数（Tensor Core、显存带宽、PCIe带宽、NVLink带宽）作为输入性能预测公式的特征

数据采集

收集已有平台的应用运行性能数据及应用运行特征

回归训练

以均方根误差作为损失函数，最小化损失函数，得出权重系数

性能预测

预测应用运行性能（好用性）

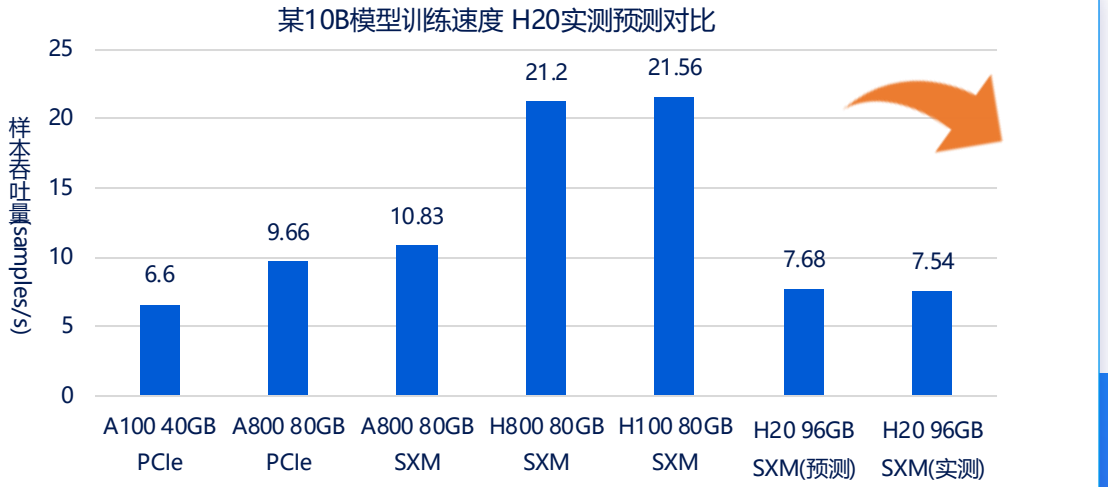
预测应用显存占用（可用性）

ParaSelect – 新平台性能预测公式

$$speed_{target} = speed_{h800} / \left(k_1 * \frac{TFLOPS'_{h800}}{TFLOPS'_{target}} + k_2 * \frac{gmem\ bw_{h800}}{gmem\ bw_{target}} + k_3 * \frac{PCIe\ bw_{h800}}{PCIe\ bw_{target}} + k_4 * \frac{nvlink\ bw_{h800}}{nvlink\ bw_{target}或PCIe\ bw_{target}} \right) * speedup(batchsize)$$

$speed$	训练速度
$TFLOPS'$	经过修正的理论TFLOPS
$gmem\ bw$	理论显存带宽速率
$PCIe\ bw$	理论PCIe带宽
$nvlink\ bw$	理论nvlink带宽
$speedup(batchsize)$	batch size 对应的加速比
k_1, k_2, k_3, k_4	常数参数，由回归获得

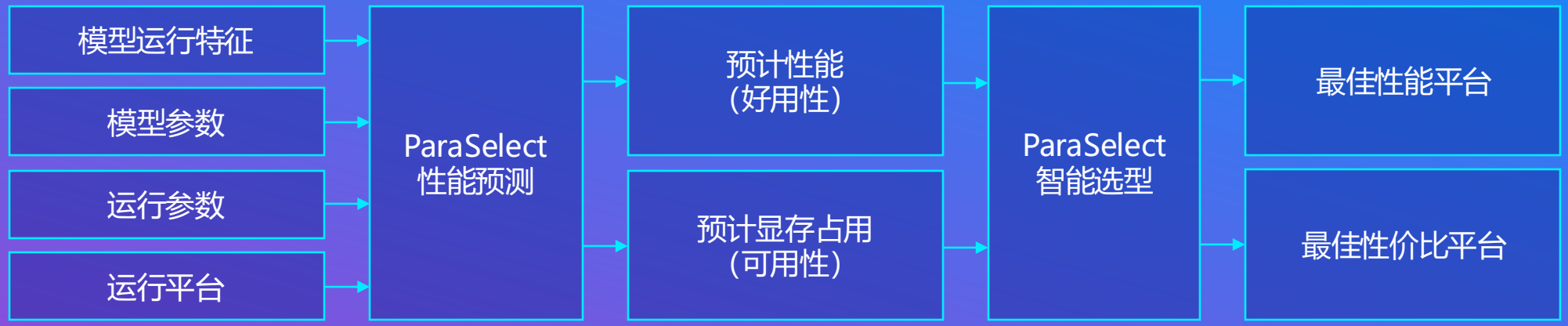
ParaSelect – 性能预测与智能选型



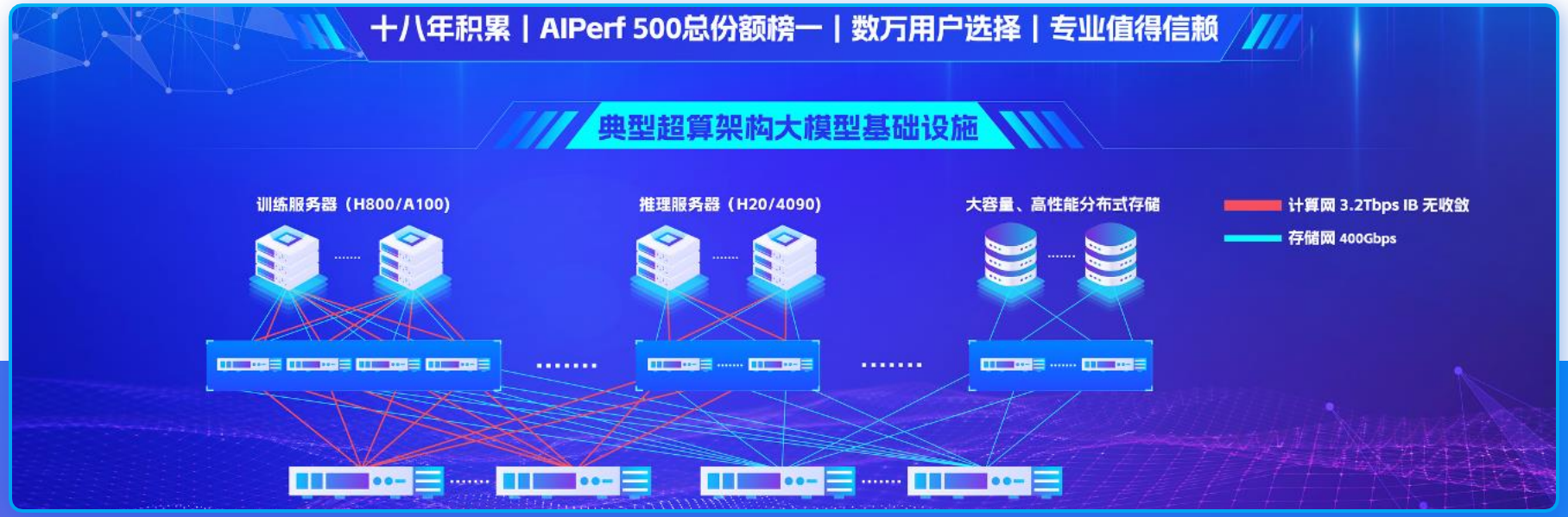
1.9%

新平台性能预测误差

根据某10B模型运行特征、模型参数、运行参数，使用ParaSelect对该大模型在H20平台训练速度进行预测，误差仅为1.9%。



大模型训练是超算应用，性能、加速比是核心关键要素

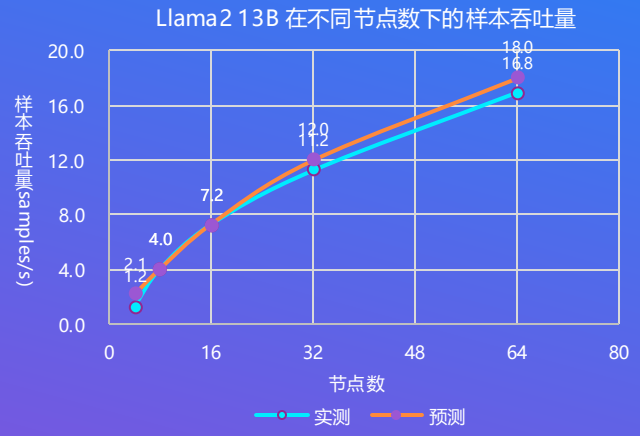


高性能

高端配置，最新架构GPU卡，高带宽内存，高速互联，采用最佳GPU算子，软硬结合，充分释放性能。

高加速比

基于超算架构，节点内、节点间采用高速互联，支持动态资源调度，采用高效并行算法，让模型训练能够无限扩展。

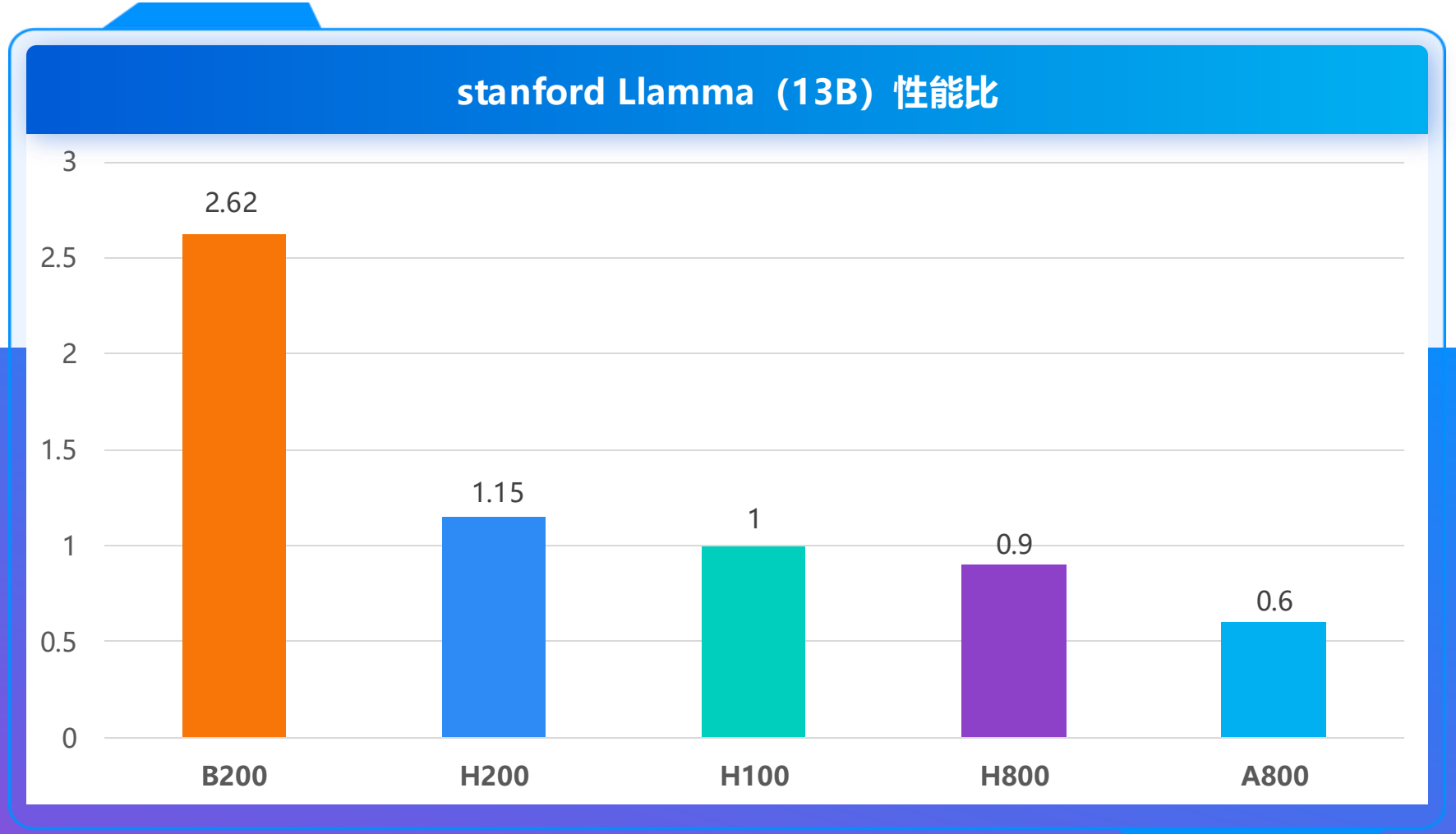


7.1%

多节点性能预测误差

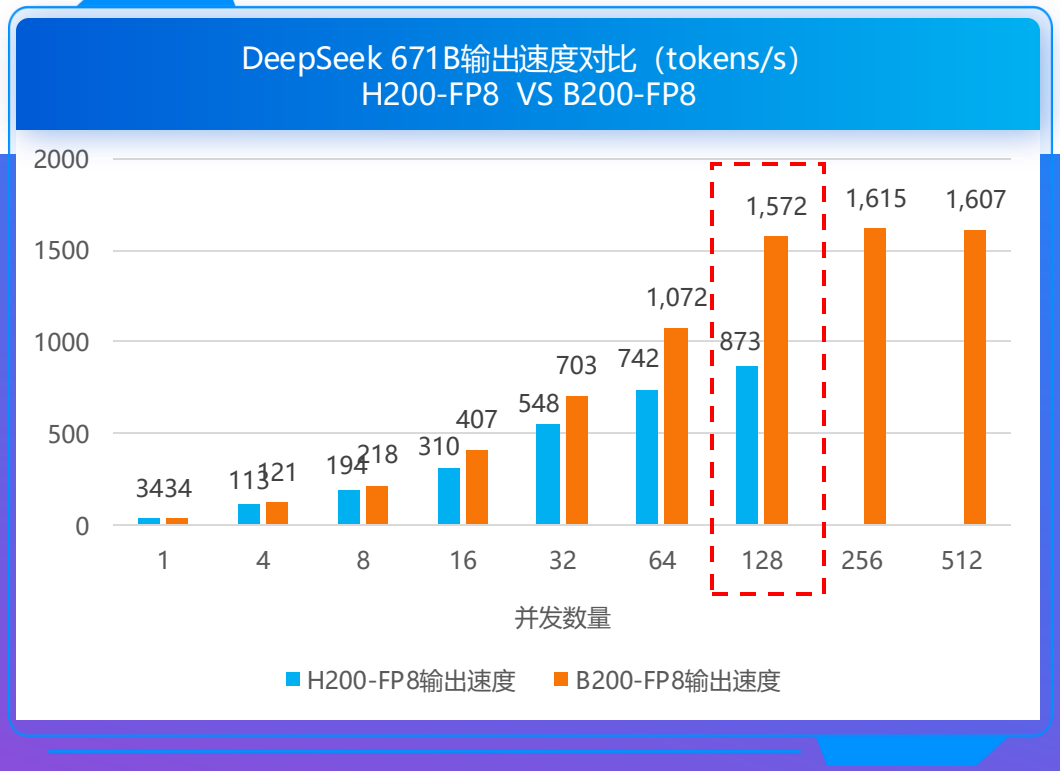
根据Llama2-13B运行特征、模型参数、运行参数，使用ParaSelect对该大模型在A100-40GB-PCIe、400Gbps RoCE平台的多节点训练速度进行预测，误差仅为7.1%。

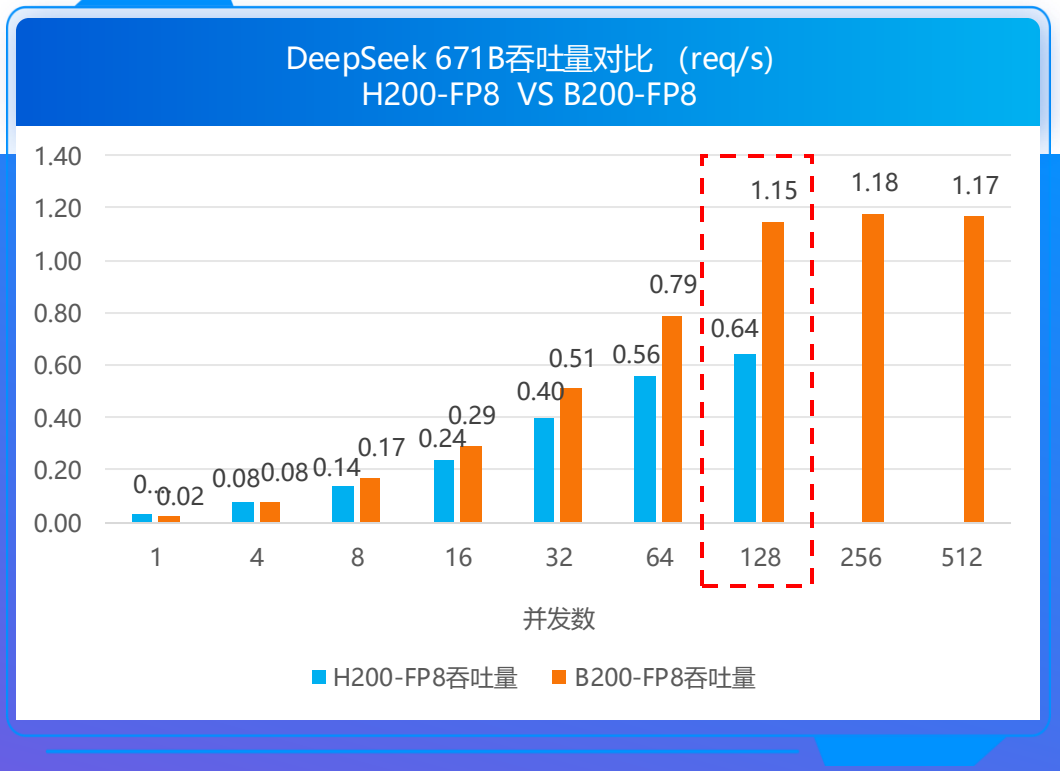
B200在典型的训练场景下性能较H100提升162%



DeepSeek 671B模型推理(H200-FP8 vs B200-FP8)

- 在128并发时，B200-FP8的输出速度是H200-FP8的约**1.8倍**
- 在128并发时，B200-FP8的吞吐量是H200-FP8的约**1.8倍**



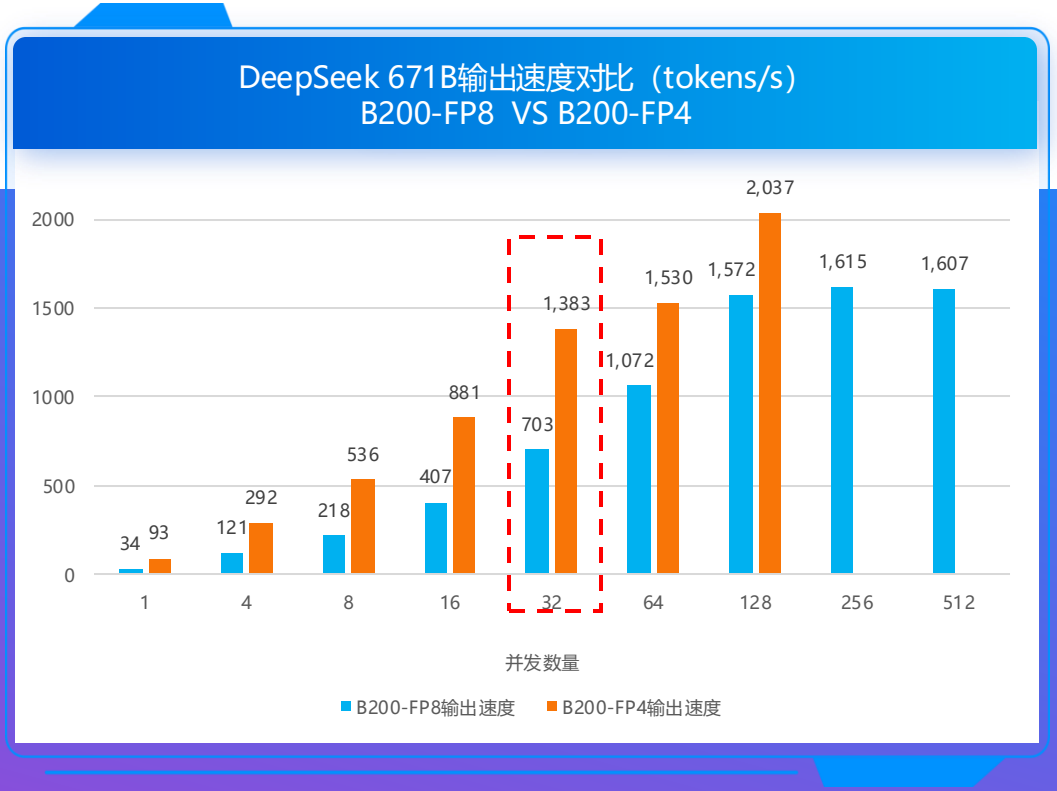


注：单节点测试，H200在256/512并发时，暂时无法运行。

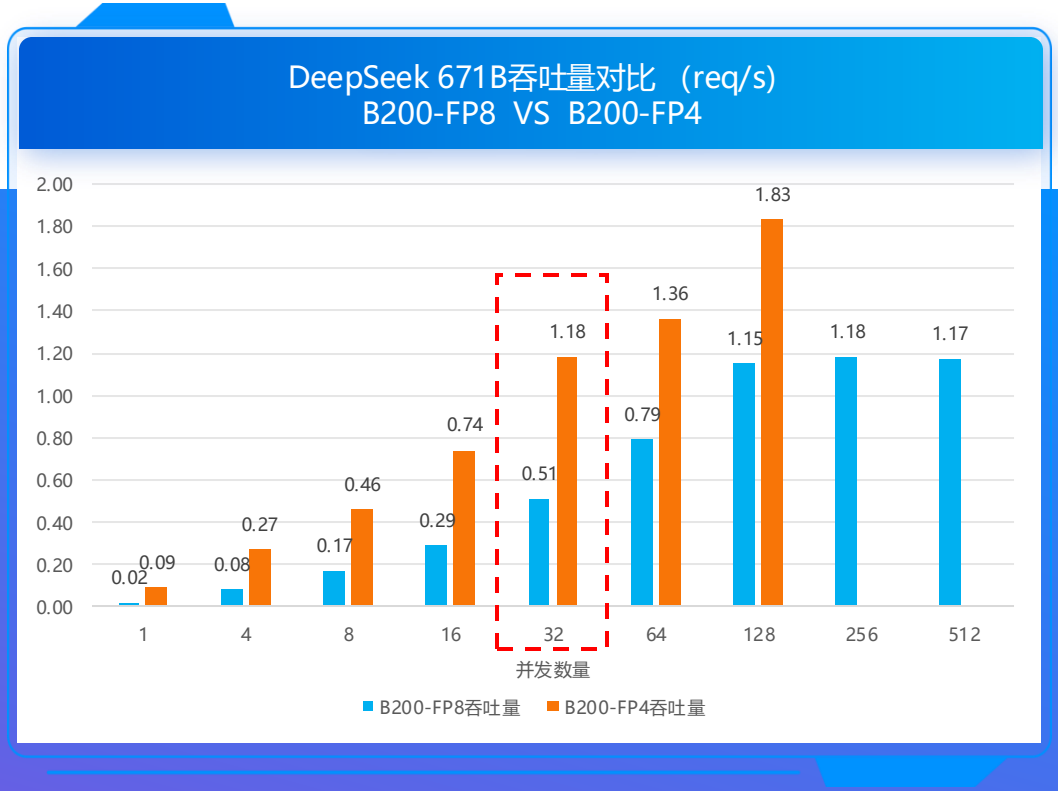
DeepSeek 671B模型推理(B200-FP8 vs B200-FP4)

B200 GPU的FP4特性是其显著的性能优势之一，在FP4精度下的峰值算力可达18P（PetaFLOPS，稀疏）

当并发数量增加到32，FP4的输出速度明显高于FP8（**性能提升约95%**）。



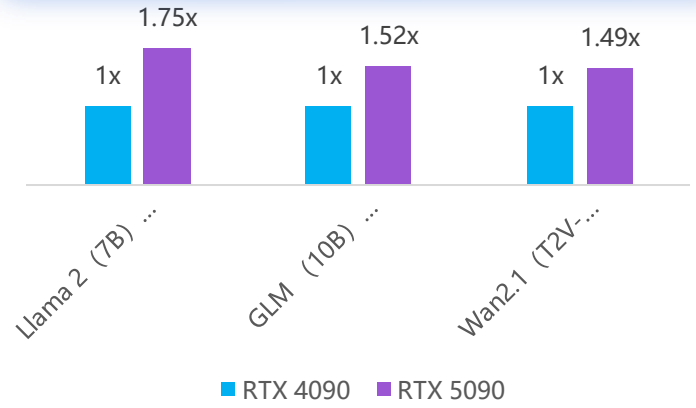
当并发数量增加到32，FP4的吞吐量明显高于FP8（**性能提升约131%**）。



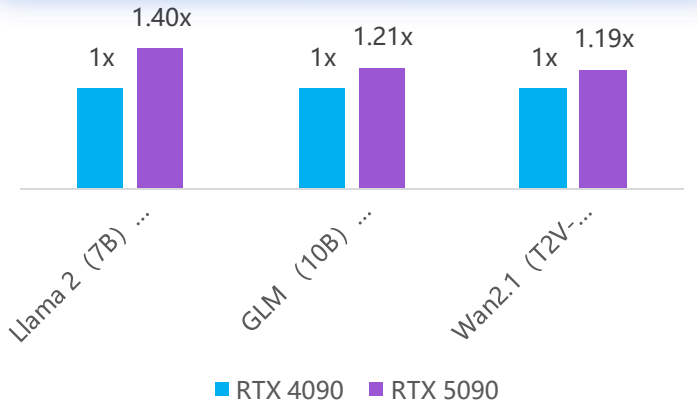
注：单节点测试，B200-FP4在256/512并发时，暂时无法运行。

5090：性价比之王

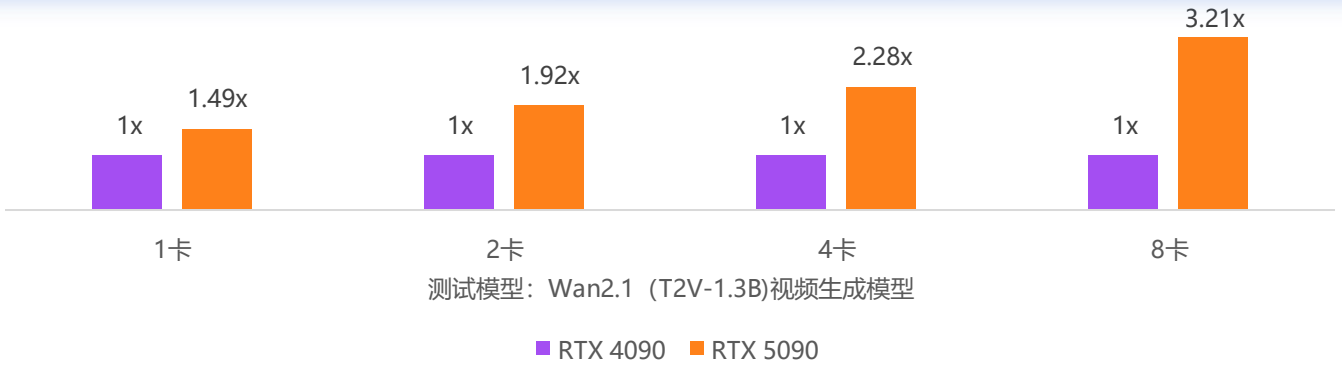
5090性能是4090的1.49-1.75倍



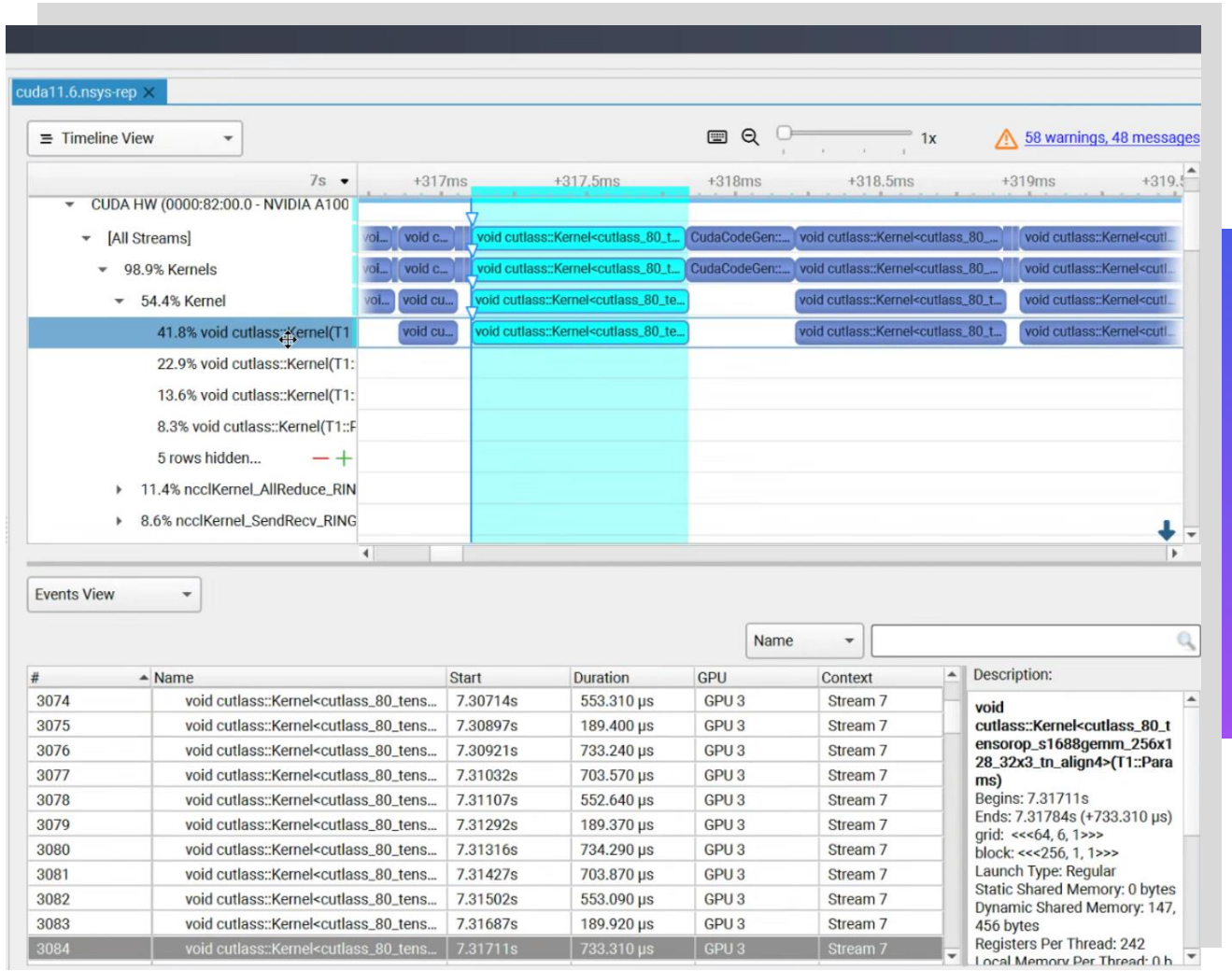
5090性价比是4090的1.19-1.4倍



5090多卡扩展性良好，相同卡数的性能是4090的1.49-3.21倍



GPT2模型性能优化



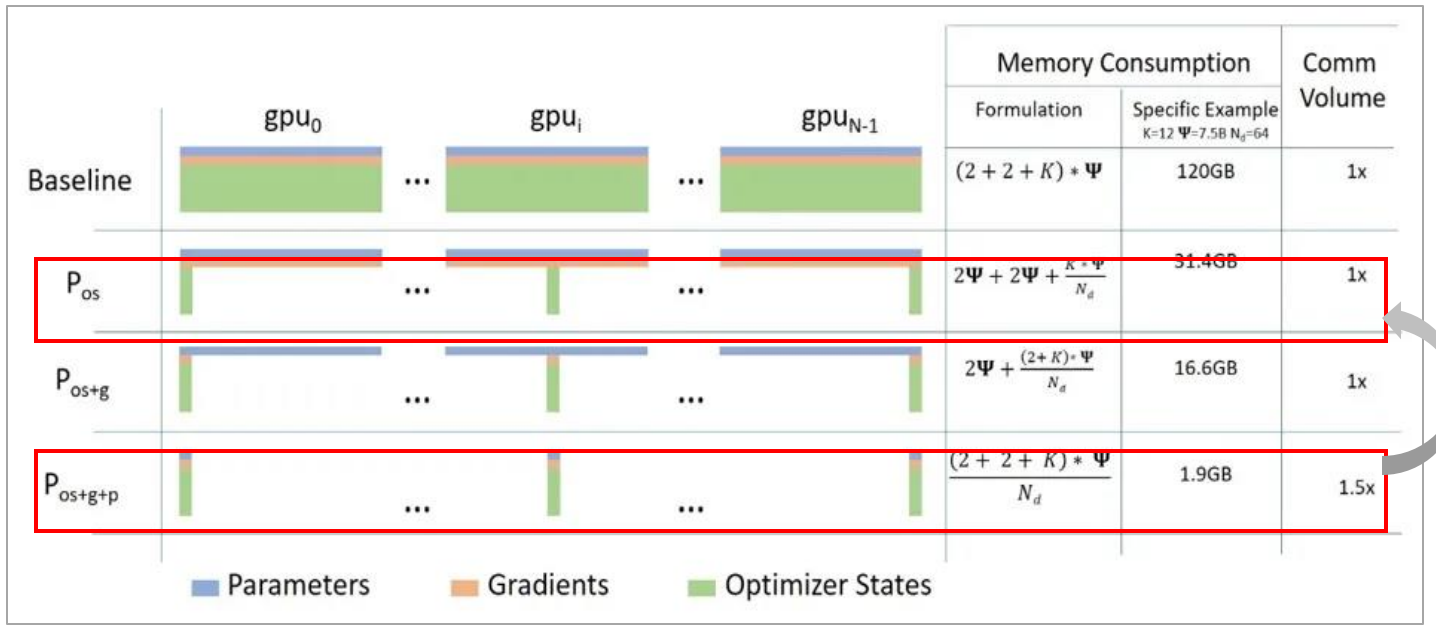
通过定位程序热点函数，发现热点函数是cublas，调整cuda版本后调用性能更优的cutlass函数

性能大幅提升

某用户模型的数据存储模式优化

应用性能中，时间与空间平衡是个永恒的话题。在显存空间允许情况下，合理选择数据存储模式，提高显存利用率、降低应用通信开销，提高应用计算效率。

DeepSpeed提供多种数据存储方式以应对不同硬件情况下的数据存储模式，不同模式也带来差异巨大的卡间、节点间通信开销。



	Memory Consumption		Comm Volume
	Formulation	Specific Example K=12 Ψ=7.5B N_d=64	
Baseline	$(2+2+K)*\Psi$	120GB	1x
P_{os}	$2\Psi+2\Psi+\frac{K*\Psi}{N_d}$	31.4GB	1x
P_{os+g}	$2\Psi+\frac{(2+K)*\Psi}{N_d}$	16.6GB	1x
P_{os+g+p}	$\frac{(2+2+K)*\Psi}{N_d}$	1.9GB	1.5x

数据存储模式优化后：充分利用显存空间，约**1000小时完成**

性能大幅提升

优化前：大约1800小时完成

某用户1300亿参数模型应用运行特征分析（优化前）

75%
GPU利用率

- 右图为上页某1300模型应用运行特征中所截取一段时间内的运行特征数据
- 分析：**在该应用运行过程中GPU利用率75%左右，NVLink通信平均速度在2500MB/s，IB通信发送接收速度为1419MB/s，可见该应用最大瓶颈为GPU利用率有待提高。**
- 针对性**优化措施：分析应用计算资源负载设计，降低不必要的CPU系统态负载，解决CPU的资源争抢，提高GPU计算资源利用率。**

大模型应用运行特征数据

Card Name	GPU%	GMem	Tensor%	NVLink(MB/s)	IB Send(MB/s)	IB Recv(MB/s)	CPU(sywa)%	Swap%	Net Send(MB/s)	Net Recv(MB/s)	Disk Read(MB/s)	Disk Write(MB/s)	NFS Send(MB/s)	Lustre Write(MB/s)	F/M(flops/B)	MIC%	F/Net(Gflop/MB)	Power(W)	Gflops	VEC%	AVX
g0004_gpu7	76.53	97.9	5.03	2543.04	1419.33	1419.36	0.00	0.00	0.00	0.00	0.00	0.00	N/A	0.00	0.00	N/A	0.00	N/A	0.00	N/A	N
g0004_gpu6	72.47	97.8	5.07	2497.18	1419.33	1419.36	0.00	0.00	0.00	0.00	0.00	0.00	N/A	0.00	0.00	N/A	0.00	N/A	0.00	N/A	N
g0004_gpu5	73.94	97.9	5.08	2543.04	1419.33	1419.36	0.00	0.00	0.00	0.00	0.00	0.00	N/A	0.00	0.00	N/A	0.00	N/A	0.00	N/A	N
g0004_gpu4	74.78	97.8	5.07	2497.18	1419.33	1419.36	0.00	0.00	0.00	0.00	0.00	0.00	N/A	0.00	0.00	N/A	0.00	N/A	0.00	N/A	N
g0004_gpu3	76.30	97.9	5.08	2543.04	1419.33	1419.36	0.00	0.00	0.00	0.00	0.00	0.00	N/A	0.00	0.00	N/A	0.00	N/A	0.00	N/A	N
g0004_gpu2	74.50	97.8	5.07	2497.18	1419.33	1419.36	0.00	0.00	0.00	0.00	0.00	0.00	N/A	0.00	0.00	N/A	0.00	N/A	0.00	N/A	N
g0004_gpu1	75.74	97.9	5.03	2543.04	1419.33	1419.36	0.00	0.00	0.00	0.00	0.00	0.00	N/A	0.00	0.00	N/A	0.00	N/A	0.00	N/A	N
g0004_gpu0	76.91	97.8	5.07	2497.18	1419.33	1419.36	0.00	0.00	0.00	0.00	0.00	0.00	N/A	0.00	0.00	N/A	0.00	N/A	0.00	N/A	N
g0003_gpu7	74.83	97.9	5.08	2543.04	1419.33	1419.36	0.00	0.00	0.00	0.00	0.00	0.00	N/A	0.00	0.00	N/A	0.00	N/A	0.00	N/A	N
g0003_gpu6	74.46	97.8	5.03	2497.18	1419.33	1419.36	0.00	0.00	0.00	0.00	0.00	0.00	N/A	0.00	0.00	N/A	0.00	N/A	0.00	N/A	N

大模型应用运行时序图

某用户1300亿参数模型应用运行特征分析（优化后）

95%
GPU利用率

- 右图为上页某1300亿模型应用运行特征中所截取一段时间内的运行特征数据
- 分析：**在该应用运行过程中GPU利用率95%左右，NVLink通信平均速度在7000MB/s，IB通信发送接收速度为3845MB/s**
- 结论：**随着GPU利用率的提升，NVLink和IB通信都得到了充分利用，整体效率提升了40%+**

PART 03

并行科技介绍

关于并行

BJ839493
股票代码

成立于**2007**年
高新技术企业，专精特新“小巨人”企业（国家级）
注册资本**5823**万元

近**500**
员工

7个
子公司

北京、天津、广州
长沙、宁夏、海南、上海

20+个
办事处与服务站

上海、武汉、西安、成都、南京、
无锡、青岛、深圳、绵阳、重庆、哈尔滨、长春、
郑州、杭州、沈阳、合肥、大连、苏州、厦门、
贵阳、昆明、太原、兰州……

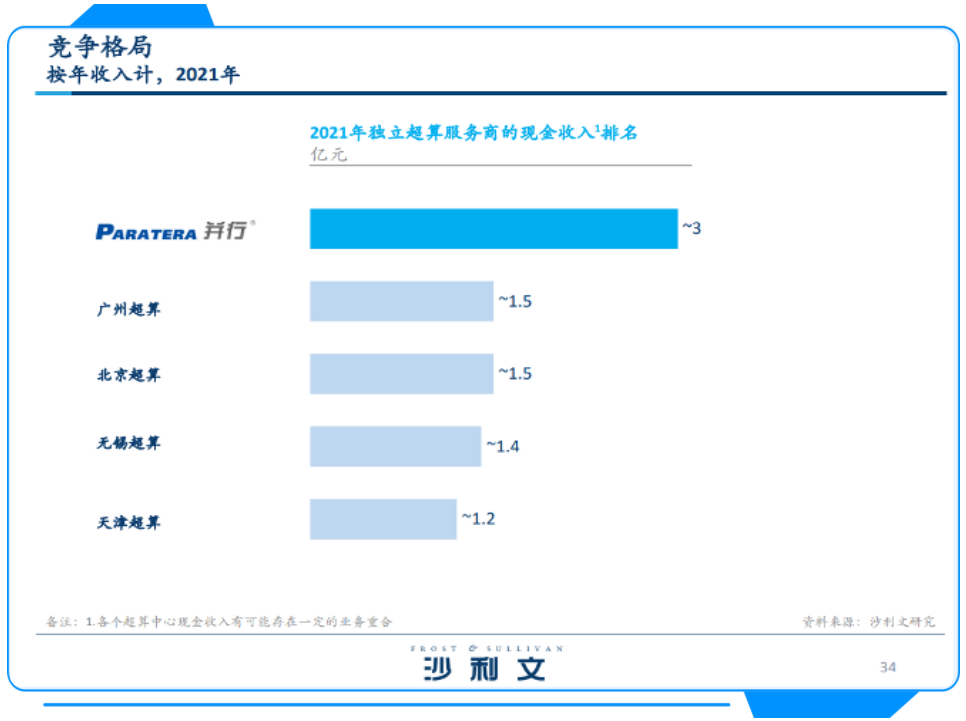
4个
技术研发中心

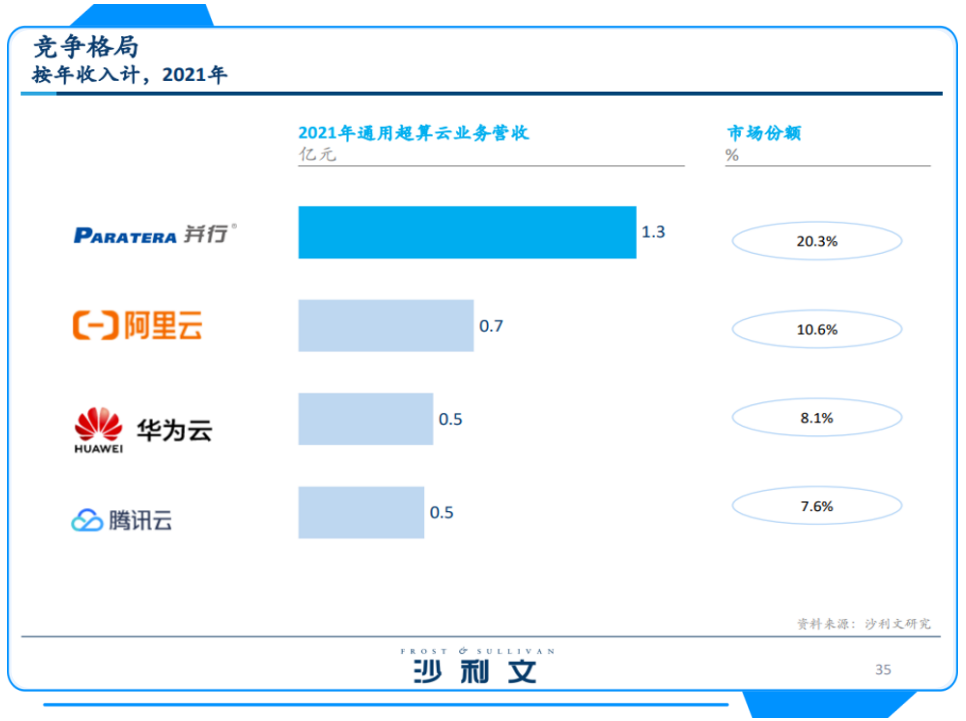
服务规模

18年	16,000+	5万+卡	200万+核
超算服务经验	用户	可调度GPU资源	可调度CPU资源

20+	200+款	300+	400+	600+
行业	应用软件SaaS化	科研机构	高校	企业

中国高性能计算（超算）市场算力服务格局



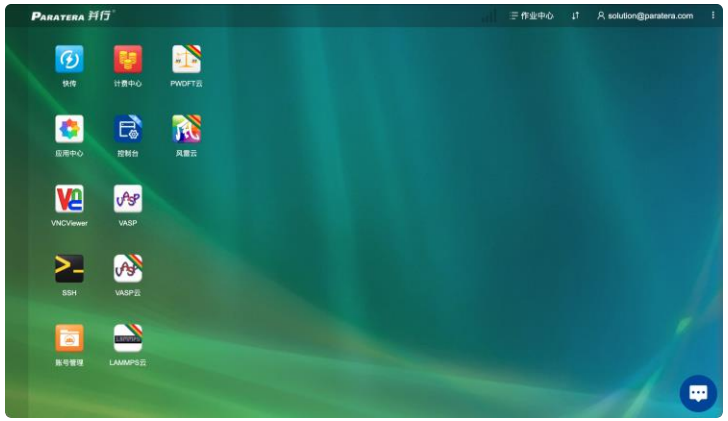


沙利文报告分析指出：并行科技是我国第一大独立的超算云服务商。

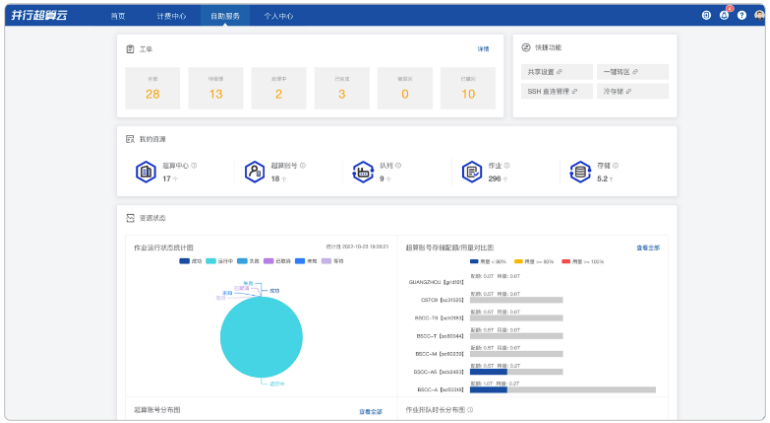
沙利文《中国超算云服务市场》独立研究报告下载链接：
http://www.frostchina.com/?p=17305

算力服务及运营能力

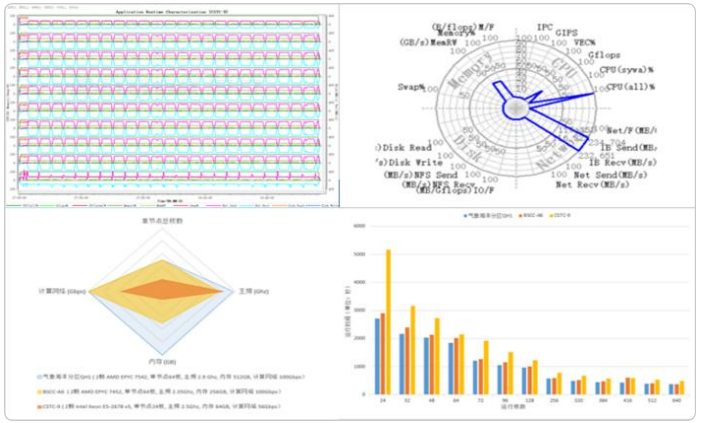
超算/智算云服务平台



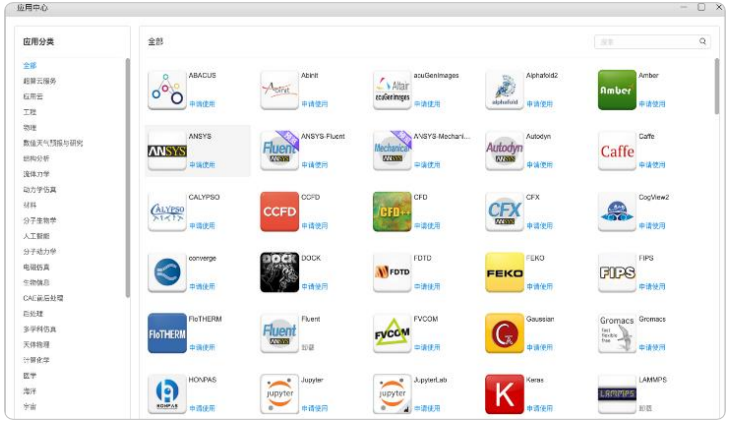
用户控制台



应用运行特征分析与算力评测

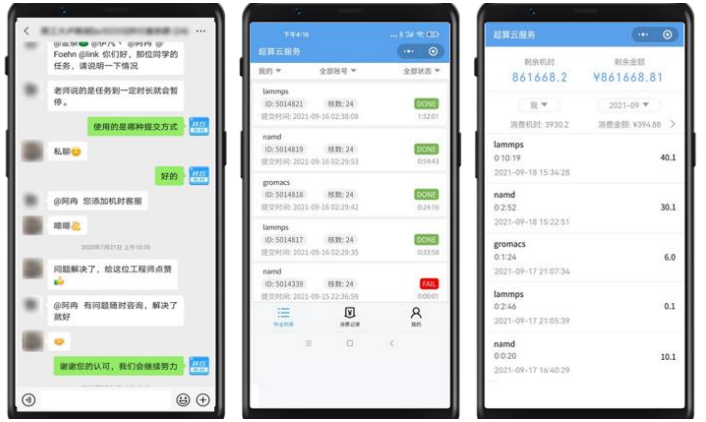


应用集成与多集群算力智能调度



运维、运营可视化

在线技术支持、移动端作业查询



并行科技算力网络

CPU核数:
2,000,000+

GPU卡数:
50,000+

存储容量:
1,000PB+

用户总数:
16,000+

用户行业分布

算网总调度量已接入**45**个智算中心，
15个超算中心，
可调度GPU资源**5万卡**以上，
CPU资源**200万核**+，
高性能存储容量达**1,000PB**。

基于超算架构为底座，以HPC和AI技术优势及海量算力资源池为依托，打造了
并行智算云、并行智造云、并行超算云、算力调度软件、算力运营服务等产品。

多样化选择，满足不同客户场景需求

50,000+卡

可调度GPU资源

2,000,000+核

可调度CPU资源

内蒙古枢纽 和林格尔集群	宁夏枢纽 中卫集群	长三角枢纽 浙江云谷磐石云数据中心	京津冀枢纽 北京超级云计算中心	海南海底算力枢纽 海南海底智算集群
内蒙古和林格尔新型算力基地	**宁夏中卫美利云数据中心**	**浙江云谷磐石云数据中心**	**北京超级云计算中心**	**新增海南陵水海底智算中心**
能源丰富：内蒙古是中国的主要能源基地之一，拥有丰富的煤炭和风能资源，可以提供稳定且成本较低的电力供应。 气候条件：内蒙古的气候对数据中心温湿度很有利，可以减少运维成本。 地理位置：位于中国北部，可以覆盖北方地区的算力需求。	政策支持：中卫作为宁夏的地级市，享有地方政府提供的优惠政策，如税收减免、土地使用优惠等。 气候适宜：中卫的气候条件适宜，有利于数据中心的稳定运行和节能。 网络基础设施：中卫拥有良好的网络基础设施，确保数据传输的高速和稳定。	经济发达：乐清位于浙江省温州市，靠近长三角经济圈，经济活跃，市场需求大。 人才集聚：乐清及附近地区教育资源丰富，可以吸引和培养所需的技术人才。 交通便利：乐清地理位置优越，交通便利，便于与周边城市进行数据和服务的快速交换。	首都优势：怀柔作为北京科技创新支撑的新区域，靠近国家政治、文化中心，具有政策和信息优势。 科研环境：怀柔聚集了众多科研机构和高校，有利于技术交流和创新合作。 高标准基础设施：作为首都的一部分，怀柔拥有高标准的基础设施，包括网络、交通等。	绿色低碳：借助海洋冷源和海上风电等清洁能源，实现低能耗、低排放的绿色算力。 独特区位：临近国际通信海缆，便于开展国际数据业务。 空间利用：利用海底空间，不占用陆地资源，减少土地成本。

并行科技合作内蒙古算力基地【算海计划】

总占地面积60亩；

共4000个20KW机柜，自建110KAV变电站；

2024年5月上线对外运营，联合共建6万高端卡单一超大集群

并行科技合作内蒙古算力基地【算海计划二期】

总占地面积100亩；

共6000个20KW机柜，自建2个110KAV变电站；

2025年10月上线对外运营，联合共建【10万】高端卡单一超大集群

AiDD 8th 2025

第8届AI+研发数字峰会

拥抱AI 重塑研发 AI+ Development Digital Summit

查看会议详情

下一站预告 11/14-15 | 深圳站 12/19-20 | 上海站

深圳站论坛设置

智能装备与机器人

超越“编程 Copilot”

下一代知识工程

智能网联与汽车智能化

AI 测试工具开发与应用

AI 基础设施和运维

数据智能及其行业应用

可信 AI 安全工程

大模型和 AI 应用评测

多 Agent 协同框架

从智能测试到自主测试

大模型推理优化

多模态 LLM 训练与应用

智能化 DevOps 流水线

上下文工程

AiDD

「深行·浅智」

Walk Deep, Think Light.

2025.11.16

AiDD首届麦理浩径徒步

KEYLINK ing

科技生态圈峰会 + 深度研习

——1000 + 技术团队的共同选择

AiDD峰会详情

上海站

AI+研发数字峰会

AiDD 峰会

时间：

2025.05.23-24

北京站

AI+研发数字峰会

AiDD 峰会

时间：

2025.08.08-09

深圳站

AI+研发数字峰会

AiDD 峰会

时间：

2025.11.14-15

香港站

「深行·浅智」

AiDD 首届麦理浩径徒步

时间：

2025.11.16

上海站

AI+产品创新峰会

AiDD × 联合峰会

时间：

2025.12.19-20

第7届AI+研发数字峰会
AI+ Development Digital Summit

感谢聆听！

扫码领取会议PPT资料