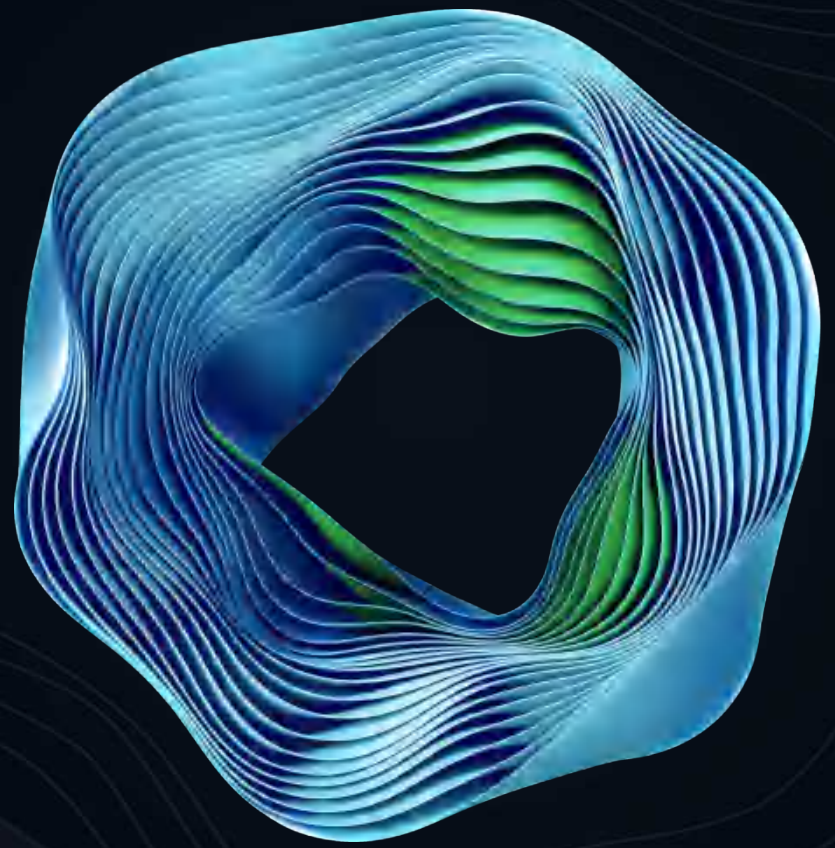
紫光集团 核心企业 | H3C 数字化解决方案领导者

新华三
智慧IT成就AIGC时代智造算力引擎

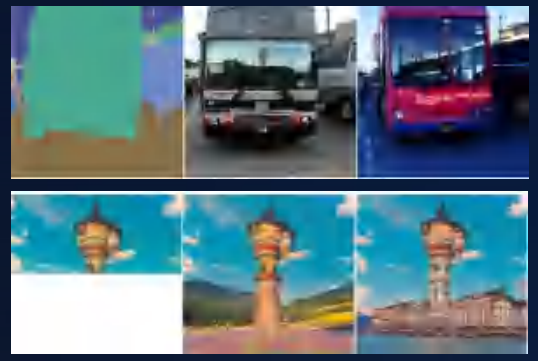


刘铮
新华三集团

ChatGPT与AIGC

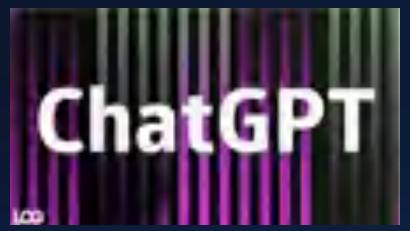
紫光集团 H3C
核心企业 数字化解决方案领导者

AIGC可以生成图像、文本、视频、音频等多种内容模态



图像

草图到图像
图像补全
图像编辑
3D图像生成
……



文本

文本创作
代码生成
对话问答
…

输入文本

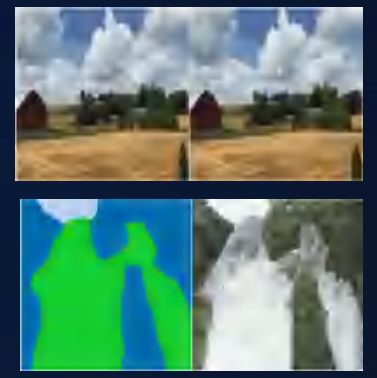
1. 一只逼真的泰迪熊正在旧金山的海里游泳
2. 泰迪熊潜入水中
3. 泰迪熊和五颜六色的鱼一起在水里游来游去
4. 一只熊猫正在水下游泳

多模态

文本到图像
文本到视频
文本到3D
……

视频

视频草图到视频
视频预测、视频编辑
视频画质增强、风格迁移
……

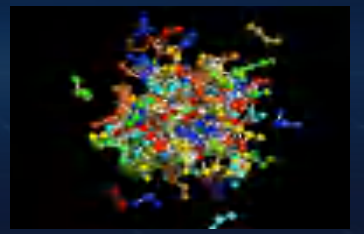


音频

文本合成语音
语音克隆
音乐生成
…

其他

科研如蛋白质生成
TRDiffusion模型



Gartner预测，2023年将有20%的内容被生成式AI所创建；
2025年，Gartner预计生成式AI产生的数据将占所有数据的10%，而今天这个比例不到1%

大模型应用覆盖企业全业务流程

H3C
数字化解决方案领导者

生产

- 产品/**工业设计**，产品工业设计、3D模型快速生成
- **软件开发**，代码补全与生成、BUG检查修复、代码注释生成
- **智能流水线**，智能化响应和调整生产计划

交易

- **助力营销沟通**，会议实时提词，沟通话术辅助改善
- **营销内容生成投放**，语料、海报等快速生成，差异化投放
- 采购**&营销策略优化**，智能响应调整优化
- 同报表生成，AI风险评估及合规审查

服务

- **智能售前咨询**，基于产品语料，专业话术，高效答疑
- **数字人服务**，数字人提供导览、播报、讲解等服务
- **智能售后客服**，问题响应范围提高，质量全天候保障

运营

- **办公模式革新**，文本写作及信息处理，工作指令自动执行
- **人力管理进化**，职位自动发布、筛选和面试，AI培训与问答
- 财务**&法务提速**，合同报表生成，AI风险评估及合规审查

全域AI应用，赋能企业全流程生产力进化

训练大模型需要大规模算力做支撑

H3C 数字化解决方案领导者

典型大模型的训练和部署，对AI算力消耗巨大

GPT-3模型
- **1750亿个参数，45TB**训练语料
- 训练175B的PPO-ptx模型需要**60pflops/s-days**
- 训练GPT-3算力消耗约**3,640pflops/s-days**
- **ChatGPT**按1300万/天访问量，估算需要**3万多张A100 GPU**

谷歌PaLM模型
- **5400亿个参数**
- 2.5亿个图像文本对数据集
- 2.56*E24FLOPs
- 消耗算力**29600pflops/s-days**

DALL.E模型
120亿个参数

2.5亿个图像文本对的数据集上训练

1*E22FLOPs

Stable Diffusion模型
- **8-9亿个参数**
- 25亿个图像文本对数据集（LAION-5B）
- **4000块** A100 GPU上训练

针对超大模型训练，Nvidia GPU是主力，国产卡在全力追赶

算力相对较强的国产卡如寒武纪和壁仞，均受到限制无法继续生产

*另外，大规模的AI计算，对**绿色能效、液冷方面**有很强需求*

训练大模型需要大规模算力做支撑

H3C 数字化解决方案领导者

GPT-3模型所需的GPU显存

- 参数量（FP16精度）：350GB（175B*2bytes）
- 梯度（FP16精度）：350GB（175B*2bytes）
- 优化器状态（FP32精度，包括权重、均值、平方）：2100GB（175B*12bytes）
- 总计：**2800GB**（350GB+350GB+2100GB）
- 满足上述显存的GPU数量：**35张A800 80G**

GPT-3模型所需的GPU算力

- 每次迭代所需的算力（参考 https://arxiv.org/abs/2104.04473）：

$$96BSlh^2[1+S/(6h)+V/(16lh)]=4.5\ \text{ExaFLOPs}$$

B: batch size, S: sequence length, l: transformer layer number, h: hidden size, V: vocabulary size

- 迭代~95000次所需的算力：**430 ZettaFLOPs**（1 ZettaFLOP=1024 ExaFLOPs）
- 在**128个HGX A800 8GPU模块**（1024张A800 80G），所需训练时间：

430*1024^3/(2496*50%)/3600/24/128=34天

-已知1个HGX A800 8GPU模块可提供2496 teraFLOPS算力，假设在集群训练中能发挥50%性能

- ChatGPT模型的训练时间可按如下公式估算:

$$\text{Training time(s)}\approx 8TP/nX$$

P: parameters，模型的参数量

T: Tokens，数据集的数据量大小

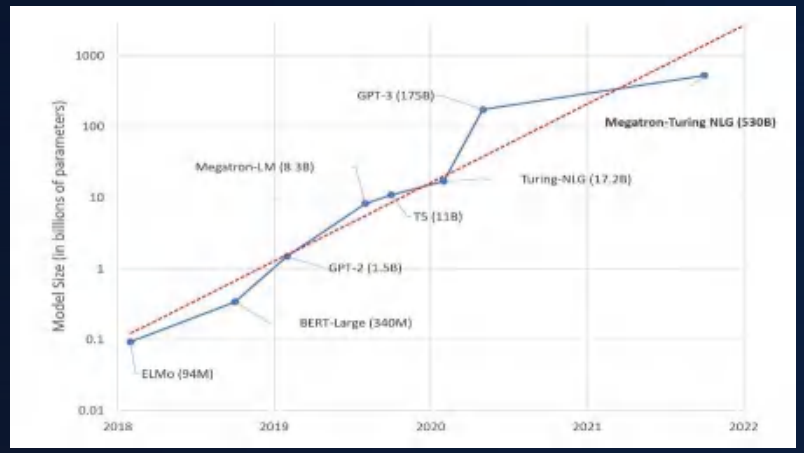
n: GPUs，GPU的数量

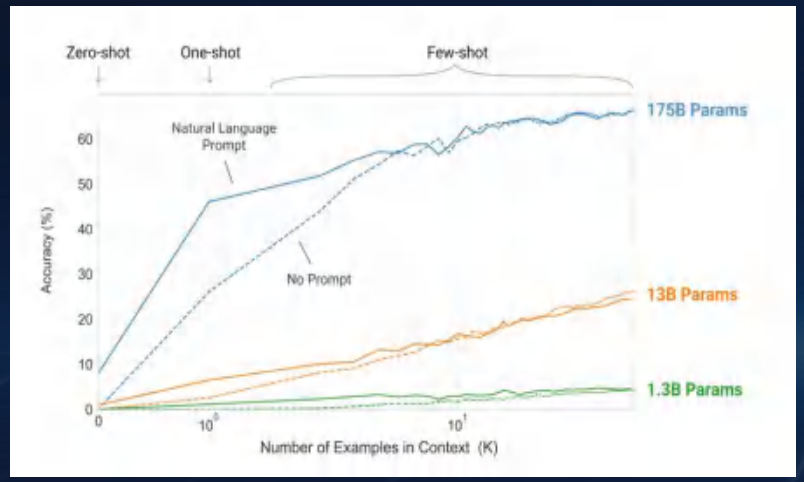
X: Tflops，GPU实际能达到的算力

- 此外，针对GPT模型参数量的大小，Nvidia也给出了一个大概的GPU集群规模

Model Size	Number of A800 GPUs	Number of DGX Servers
1.7B	32	4
3.6B	64	8
7.5B	128	16
18B	256	32
39B	512	64
76B	1024	128
145B	1536	192
310B	1920	115
530B	2520	315
1T	3072	384

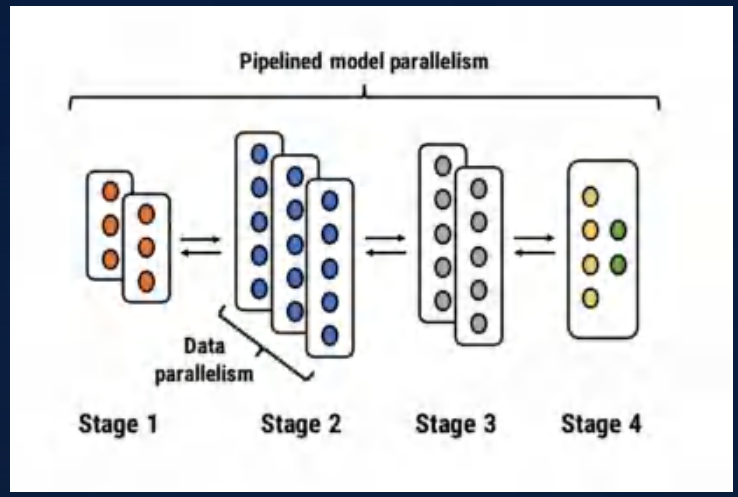
AI大模型对计算资源的要求

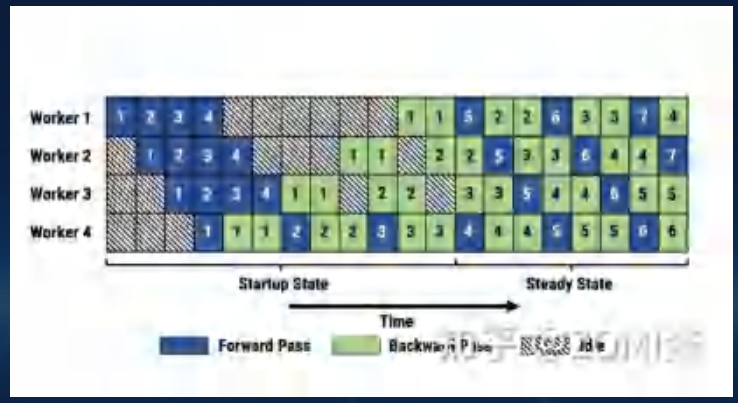




模型参数逐年增长，训练所需的算力水涨船高

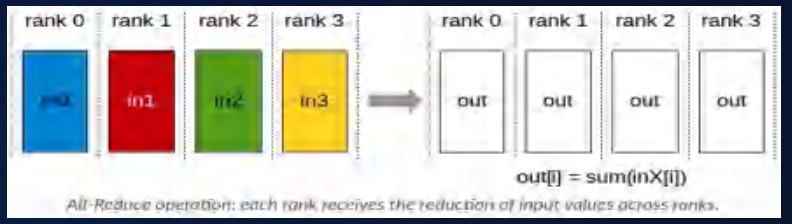
- NLP属于预训练模型，为保证算法准确，参数逐年指数增长
- Zero Shot Prompting需要通过大模型保证输出效果
- 模型训练时实际内存占用是参数的**5倍**以上



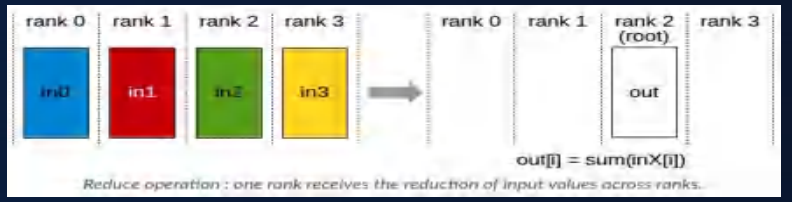


需要通过多机多卡GPU集群保证算力需求

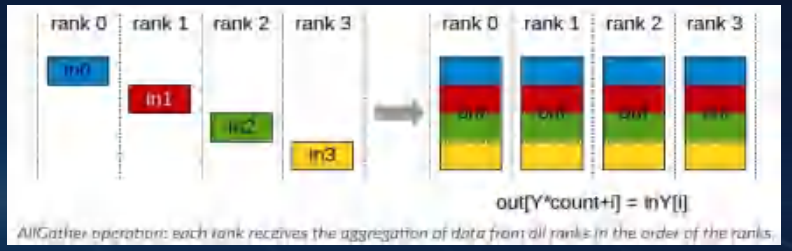
- 流水线并行需要对模型和数据进行分片，涉及GPU间fullmush和串行通信
- 提升了GPU利用率，但对数据集**加载速度**有较高要求



AllReduce，计算所有GPU数据后保存到所有GPU



Reduce，计算所有GPU数据后保存在单个GPU



AllGather，汇聚所有GPU中数据后保存到所有GPU

需要通过高速通信保证GPU集群训练效率

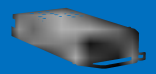
- 训练期间GPU通信存在P2P（1对1）和Collective通信（1对多或多对多）
- 其中Collective需要高速通信以匹配GPU的算力

H3C AIGC应用场景产品一览

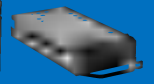
H3C
数字化解决方案领导者

数据中心交换机

H3C S9820-8C
8 Slot



*16*100G*



*4*400G*

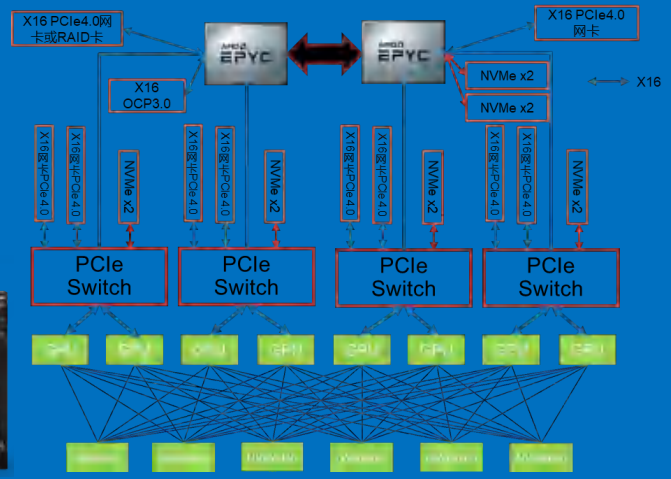
H3C S9825-64D
*64*400G*

S9855-24B8D
*24*200G + 8*400G*

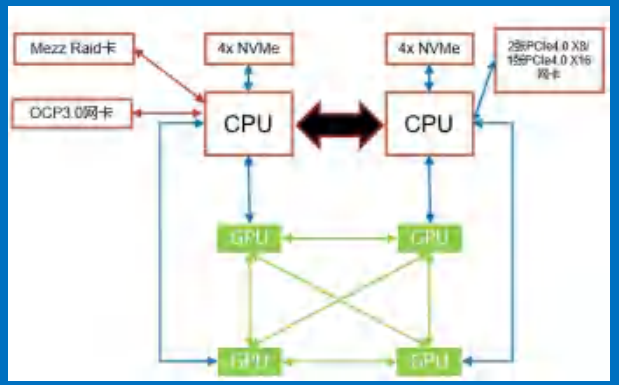
H3C S12516CR/R
*48*200G*16 Slot*
*36*400G*16 Slot*

服务器-训练场景

H3C R5500 G5



H3C R5300 G5



GPU服务器基础设施 — 从推理到训练全场景覆盖 持续创新的G6新品

紫光集团 核心企业 | H3C 数字化解决方案领导者

多样算力先锋
H3C UniServer R4950 G6

192核 处理器
10个 PCIe 5.0槽位
512TB 存储空间

专业算力旗舰
H3C UniServer R4900 G6 Ultra

120核 处理器
4张 PCIe5.0 双宽GPU
400Gb/s IB网络

混合算力引擎
H3C UniServer R5350 G6

10块 双宽GPU
400Gb/s 极速网络
4种 智能拓扑

智能算力旗舰
H3C UniServer R5500 G6

7倍 AI算力提升
400Gb/s 极速网络
12块 NVMe SSD

推理/小规模训练
1~4 A800-AI训练/HPC
1~4 A30-AI训练/推理
1~4 A10-AI推理
1~8 T4-AI推理/智能监控

大规模推理/训练
HGX A800 4-GPU-AI训练/HPC
4~8 A800-AI训练
4~8 A30-AI训练/推理
4~8 A10-AI推理
8~16 T4-智能监控

大规模训练
HGX A800 8-GPU-AI训练
*液冷解决方案已在研，请各位关注GPU液冷需求，现有可支持方案为R5500 G5 INTEL平台

H3C UniServer R5500 G6 — 智能算力旗舰

双平台CPU
2颗第四代英特尔®至强®可扩展处理器/2颗AMD EYPC™处理器

强劲算力
32 PFLOPS

灵活拓扑
2种GPU拓扑

模块化设计
灵活选择CPU/AI平台

强劲算力，助力AI业务高效运转

适配最新一代NVIDIA企业级GPU模组，性能对比上一代产品提升3.4倍，为AI业务提供强劲算力

灵活拓扑，适配不同AI场景需求

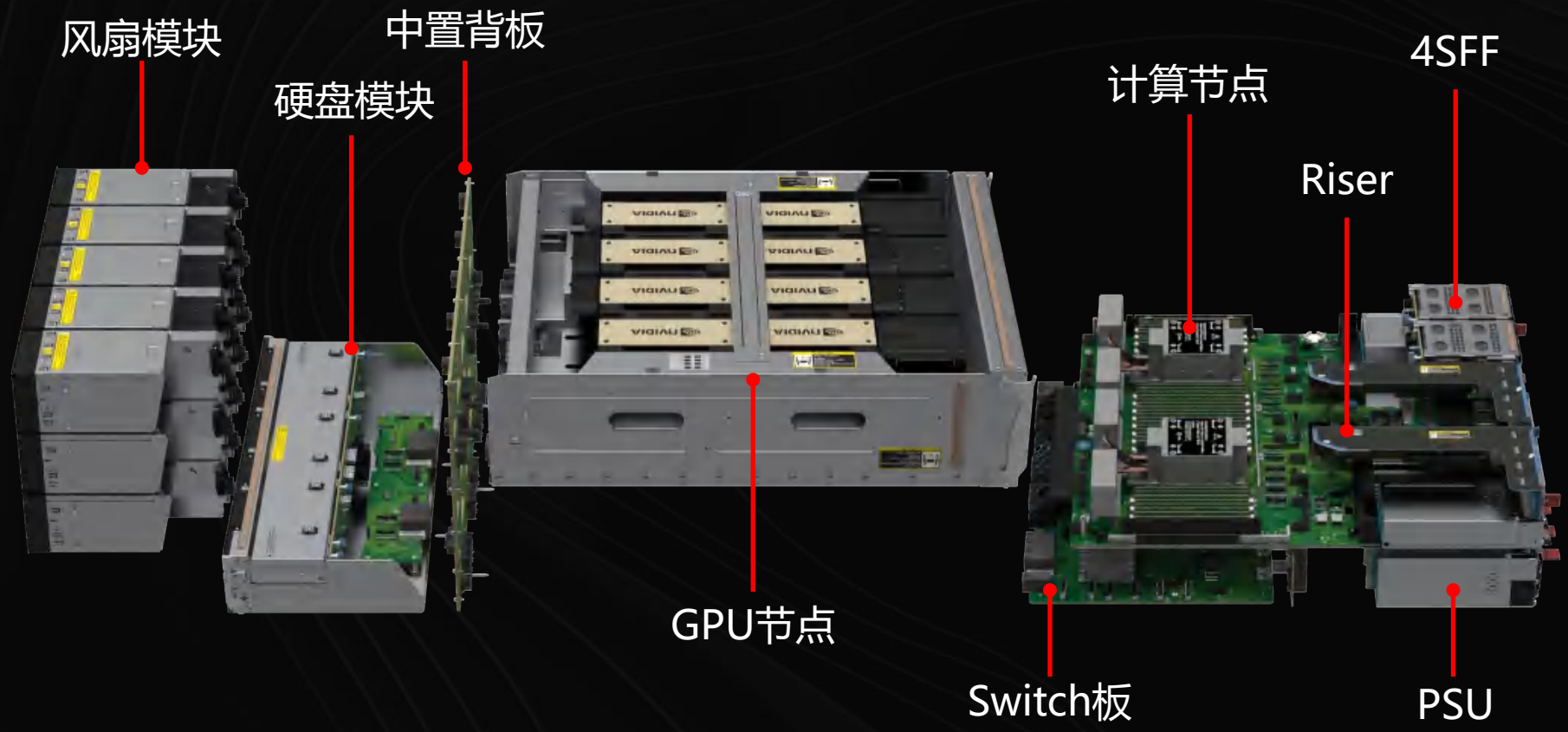
2种GPU拓扑设计，双平台CPU设计，灵活适配客户不同AI场景需求

模块化设计，轻松简便运维

系统解耦，模块化设计，无需下架可进行运维；
GPU与计算节点分开独立供电，保障业务稳定运转

[注] 1. 32 PFLOPS=3958 TFLOPS*8(H800 FP8) 2. 3.44=67/19.5 (FP32)

模块化设计—轻松简便运维



高效散热系统

15颗超高性能风扇配合智能温控
保障系统性能持续稳定输出

模块化设计

系统解耦，用户可以灵活选择
CPU/AI平台

多重冗余电源

多重冗余电源设计，GPU与计算节点
分开独立供电，为系统提供多重保障

混合算力引擎UniServer R5350 G6

紫光集团 核心企业 | H3C 数字化解决方案领导者

192核
2颗第四代AMD EPYC 9004系列处理器

混合算力
支持GPU、NPU、XPU等多种AI加速卡

400Gb/s
极速网络

400TB
存储空间

卓越性能

适配最新一代NVIDIA企业级GPU，性能对比上一代产品最高提升5倍，高效算力，助力智能时代

智能拓扑

4种GPU拓扑配置，灵活匹配客户不同应用场景，智能调配资源，驱动算力的高效运转

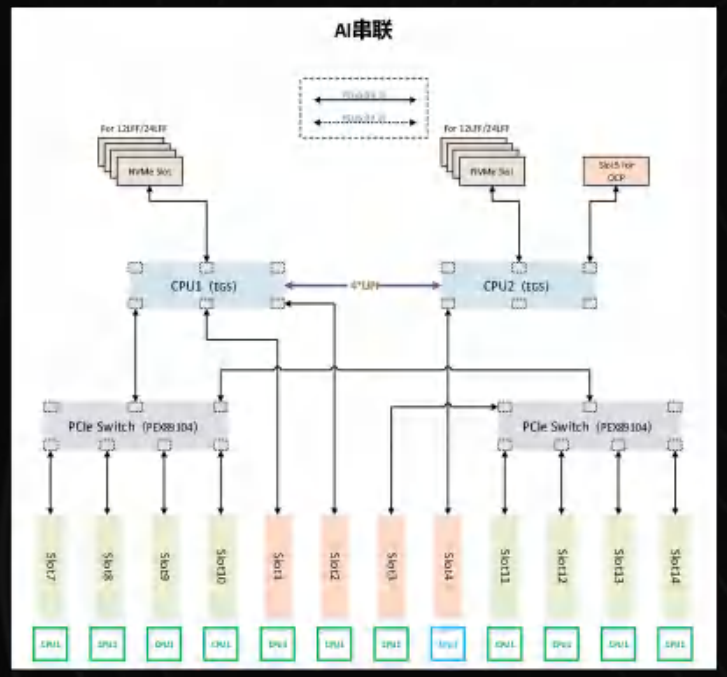
算存一体

灵活适配AI加速卡和智能网卡，集训练，推理能力于一身。海量存储，满足人工智能数据对于存储空间的需求

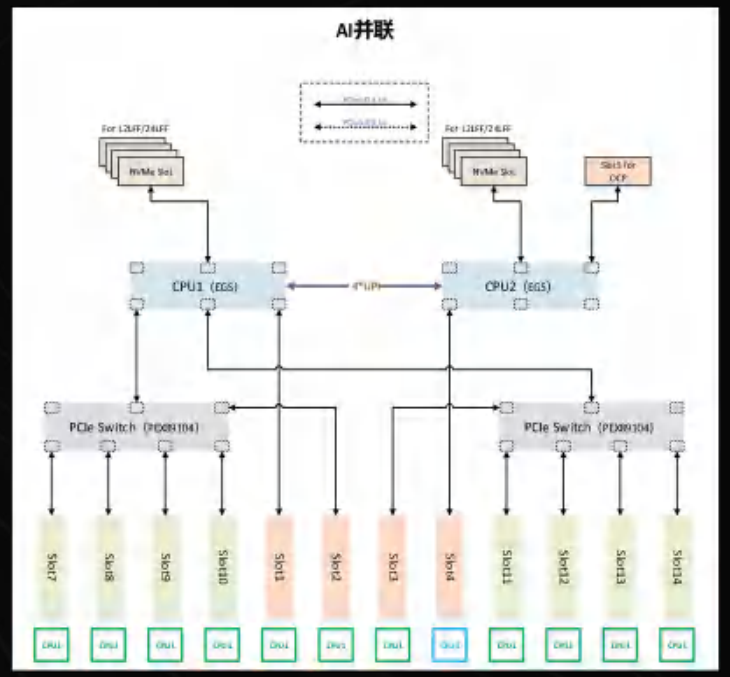
已通过信通院基础性能测试、算力算效测试以及节能减碳测试，并荣获“卓越产品”称号。

灵活拓扑—适配不同应用场景需求

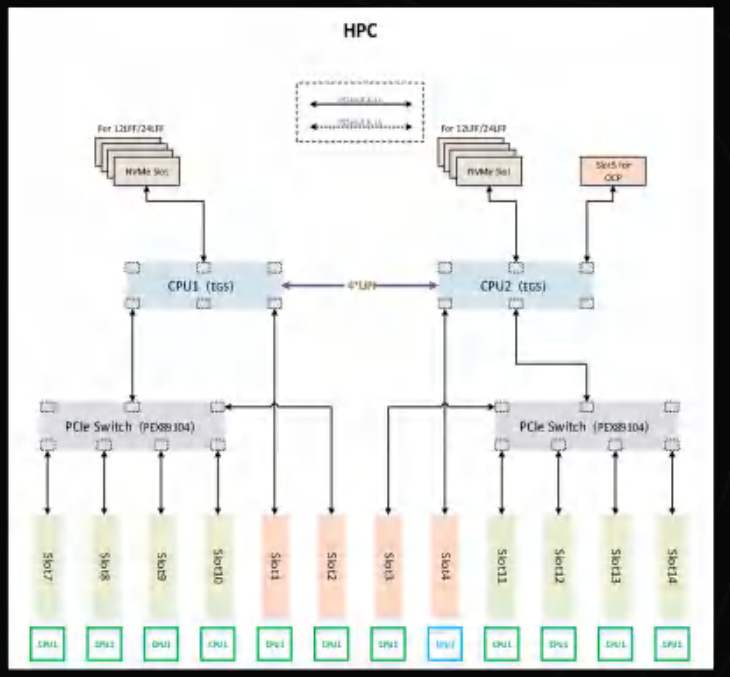
8卡Switch型



- 所有GPU均挂载在同一CPU下，CPU和GPU间共用一条x16 Lane
- 任一GPU都可通过PCIe Switch通信，GPU间通信效率最优



- 所有GPU均挂载在同一CPU下，CPU和GPU间占用两条x16 Lane
- 每组4个GPU都通过PCIe Switch通信，组内GPU通信效率高
- GPU与CPU间的并发带宽较高



- 两组GPU分别挂载在两个CPU下，负载均衡，CPU算力较高
- 每组4个GPU都通过PCIe Switch通信，组内GPU通信效率高
- GPU与CPU间的并发带宽较高

MLPerf性能测试领先

国际权威AI基准评测组织MLPerf™公布最新AI训练（Training）榜单，在本次基准测试中，新华三集团全新一代AI服务器H3C UniServer R5300/5350 G6，凭借卓越性能、智能拓扑、算存一体等优势，一举夺得25项训练任务第一。

H3C UniServer R5300 G6，智能时代的多元算力旗舰

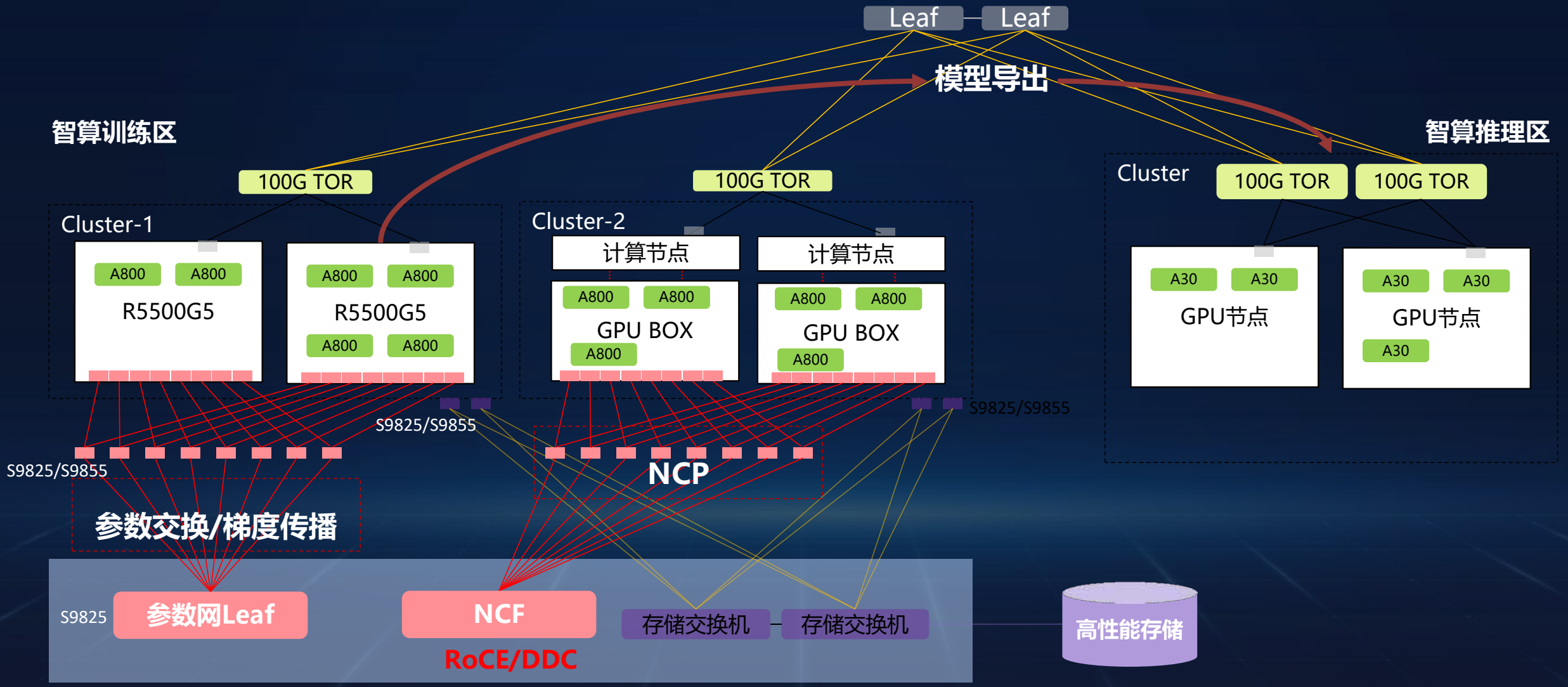
- 在本次AI推理基准测试中，R5300 G6服务器性能表现优异，一举夺得数据中心场景、边缘场景**23项同配置第一**，以及**1项绝对配置第一**，实力诠释了其对于大规模、多元化、高复杂度算力场景的强大支撑能力。
- 在ResNet50模型赛道，R5300 G6每秒可对282029张图片进行实时分类，提供高效准确的图像处理和识别能力。
- 在RetinaNet模型赛道，R5300 G6每秒可完成对5268.21张图片中的目标进行识别，为自动驾驶、智能零售、智能制造等场景提供算力底座支撑。
- 在3D-UNet模型赛道，R5300 G6在99.9%精度要求下，每秒可处理26.91张3D医学影像的分割，辅助医生快速诊断，提升诊疗效率和质量。

H3C UniServer R5350 G6，智能时代的混合算力引擎

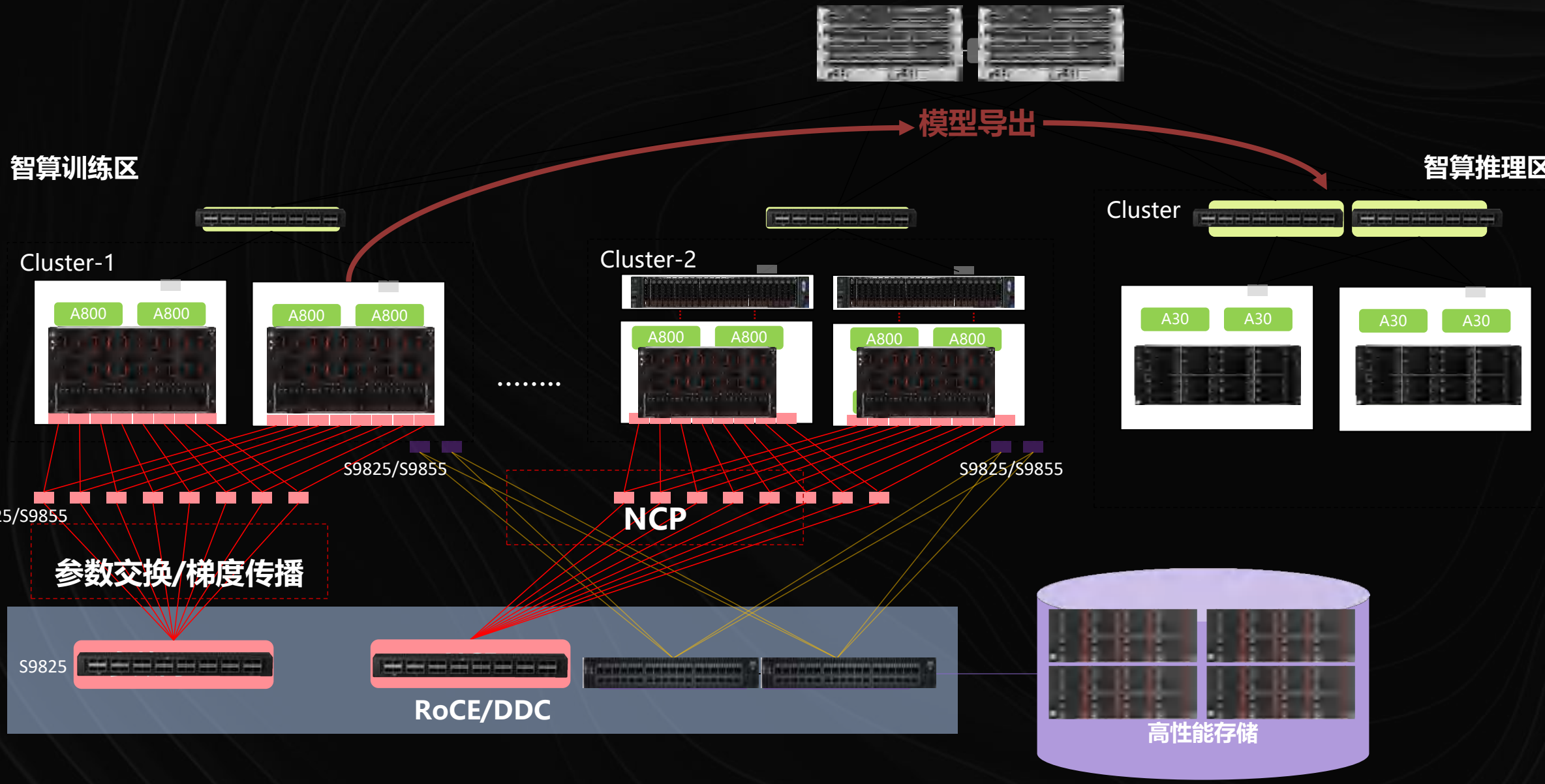
- 与此同时，凭借领先的AI系统设计和全栈优化能力，R5350 G6服务器在本次基准测试中，获得了ResNet50（图像分类）评测任务同配置第一。

MLPerf™基准检验由图灵奖得主David Patterson联合顶尖学术机构发起，在全球AI领域极具影响力，其评测指标和AI领域的前沿应用紧密结合，测试成绩具有极大的应用参考价值，能够为用户衡量设备性能提供权威有效的数据指导。

ChatGPT基础设施总体解决方案



ChatGPT基础架构解决方案



AIGC引擎大规模部署

液冷解决方案助力构建AIGC数据中心

清洁能源，绿色发电

采用光伏发电，减少碳排放**2500吨/年**

整机柜方案 多维提升

部署密度提升**100%**，交付效率提升**10倍**

安全
绿色

智能
极简

极致液冷，智能温控

Pue降至1.1以下，减少碳排放1750吨/年

重构营维，智能安全

运维效率提升**35%**资源利用率提升**20%**

全面支撑“数字中国”+“低碳中国”战略

东数西算算力节点

智算中心

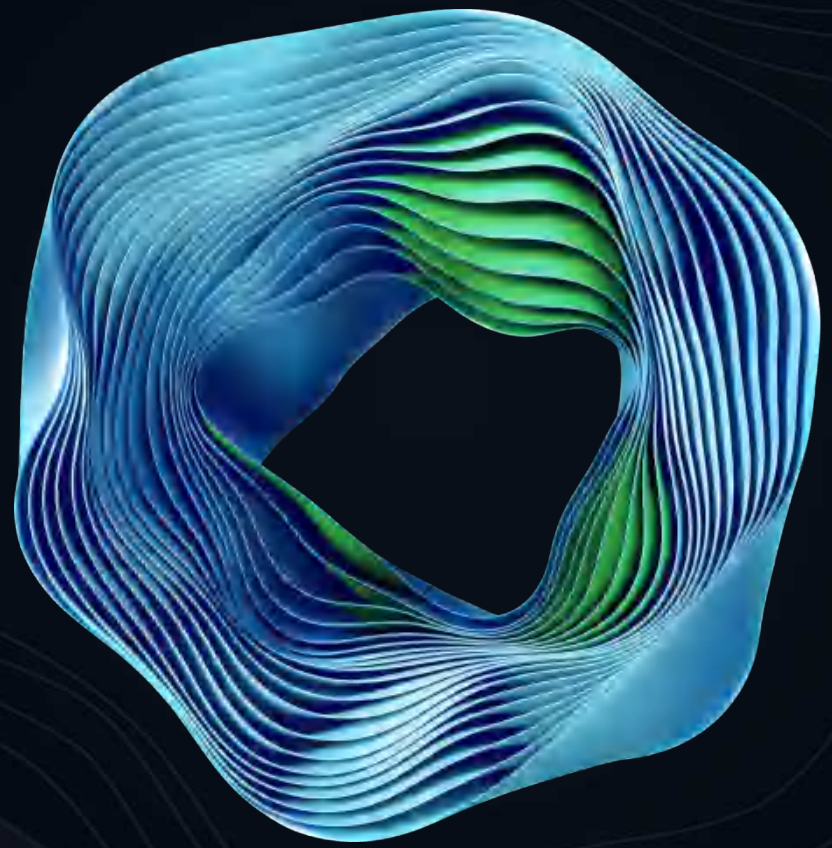
HPC算力中心

传统数据中心能效提升

数据中心模型：1500柜，8kW/柜，2N，负载率50%@北京，全球平均碳排放因子(kgCO2/kWh)：0.475

紫光集团 核心企业 | H3C 数字化解决方案领导者

AIGC成为人工智能实用化的拐点，
可能带来生产力的革命性发展，
新华三智慧IT成就AIGC智造算力引擎。



紫光集团 H3C
核心企业 数字化解决方案领导者

谢谢观看！