Tencent 腾讯 | CSIG 云与智慧产业事业群

新一代大模型与国产化云平台的融合落地

2025.3

腾讯专有云 方天戟

个人简介

方天戟

- 腾讯专有云首席架构师，《大模型时代基础架构》作者，《云鉴》编写组成员，参与信通院与中国软件测评中心编制的多个行标与国标编制；
- 18年行业经验，曾服务于华为、新华三、Juniper等业界著名企业；
- 为宝马、航天科技、中国建筑、新加坡HBO Asia等业界头部客户设计过企业上云整体方案并落地；

微信号：cybernetics

个人微信公众号

DeepSeek的开放性与低成本，实现了大模型的普惠性

FlashMOA
（减少矩阵计算工作量）

MOE
（多专家库）

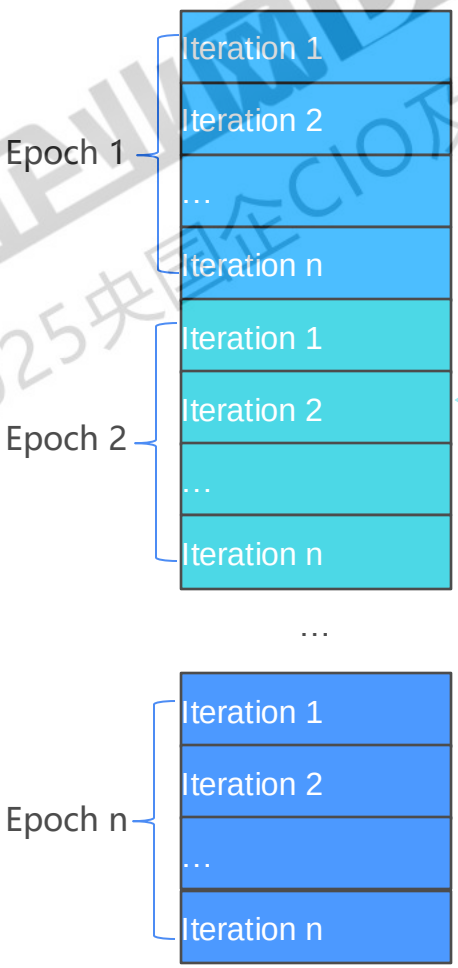
DeepGEMM
（矩阵计算优化）

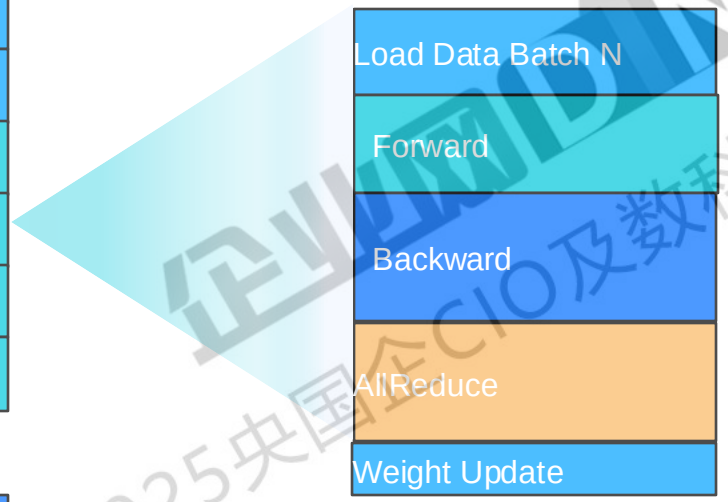
……

效率提升200倍+，百万Token成本从45 USD降低到0.2 USD

大模型训练关键要素：算力+存储+网络

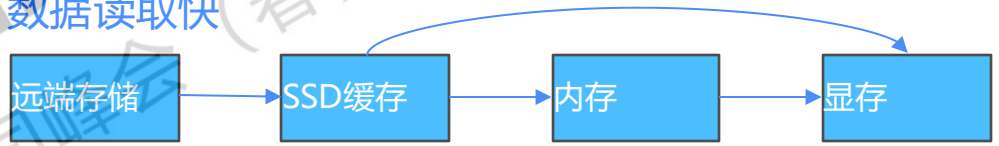
AI训练过程由多个迭代循环组成





AI训练需求

数据读取快



训练计算快

模型状态存储等　　训练计算主力

网络交换快

Nvlink(节点内)

GpuDirect RDMA(节点间)

高性能训练需要：高性能算力+高性能存储+高性能网络

大模型推理关键要素：RDMA网络与资源调度

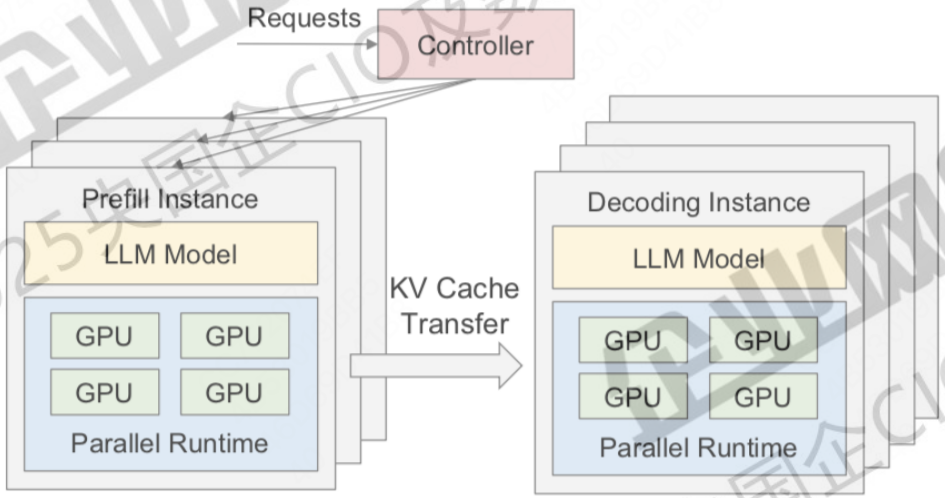


Figure 6: DistServe Runtime System Architecture

- P-D分离架构使得单个推理任务的不同阶段分布在多卡上进行，因此多机多卡间的高吞吐低延迟通信必不可少，RDMA网络成为分布式推理集群的刚需。
- P-D分离架构需要compute bound GPU（计算密集型）和memory bound GPU（访存密集型）等不同类型的GPU构建成一个分布式推理集群。因此广泛的GPU兼容与精细的调度策略能极大优化效率和成本。

大模型推理不再是单卡推理，P-D（Prefill、Decode)分离架构和分布式推理集群是必然趋势

常见的DeepSeek模型训练推理的硬件需求与方案

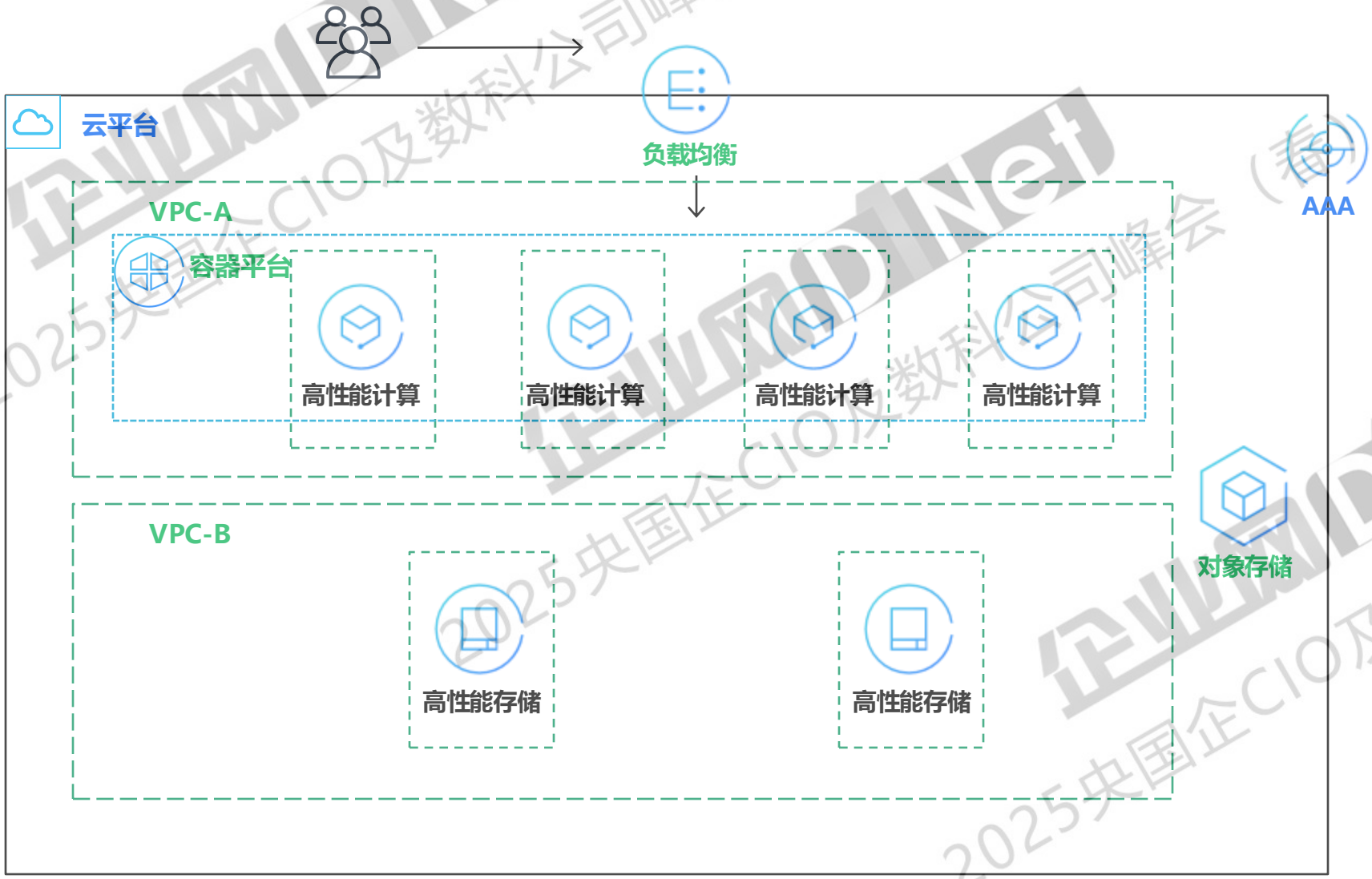
Deepseek开源模型
DeepSeek-R1/V3 671B原生模型、DeepSeek-R1-Distill-Qwen-1.5B/7B/14B/32B、DeepSeek R1-Distill-Llama-8B/70B

模型规模	推理硬件要求	**推理显存要求**	训练硬件需求	训练显存需求	模型大小（FP16）
1.5B	单卡GPU（如：RTX 3090或A10）	6-8G	单卡GPU（如A100 40GB）	10-16GB	3GB
7B	单卡GPU（如：A100 40GB或RTX 4090）	16-24GB	多卡GPU（如：4xA100 40GB）	32-48GB	14GB
8B	单卡GPU（如 A100 40GB或RTX4090）	20-28GB	多卡GPU（如8xA100 80GB）	64-128GB	16GB
14B	单卡GPU（如A100 80GB, H20 96GB）	32-48GB	多卡GPU（如8xA100 80G）	64-128G	28GB
32B	多卡GPU（如2xA100 80GB,4XH20 96GB）	64-96GB	多节点分布式训练（如：16XA100 80GB）	256-512GB	64GB
70B	多卡GPU（如4XA100 80GB，4XH20 96GB）	128-192GB	多节点分布式训练（如：32XA100 80GB）	512-1024GB	140GB
671B	多卡GPU（16XA100 80GB，8XH20 96GB，32X昇腾910B*）	1-1.5TB	大规模分布式训练（如128xA100 80GB）	5-10TB	1.342TB

INT8/INT4量化版本整体推理资源需求可以减少30%-40%

注*：鲲鹏910B不支持FP8，需要转换为BF16，显存资源量倍增

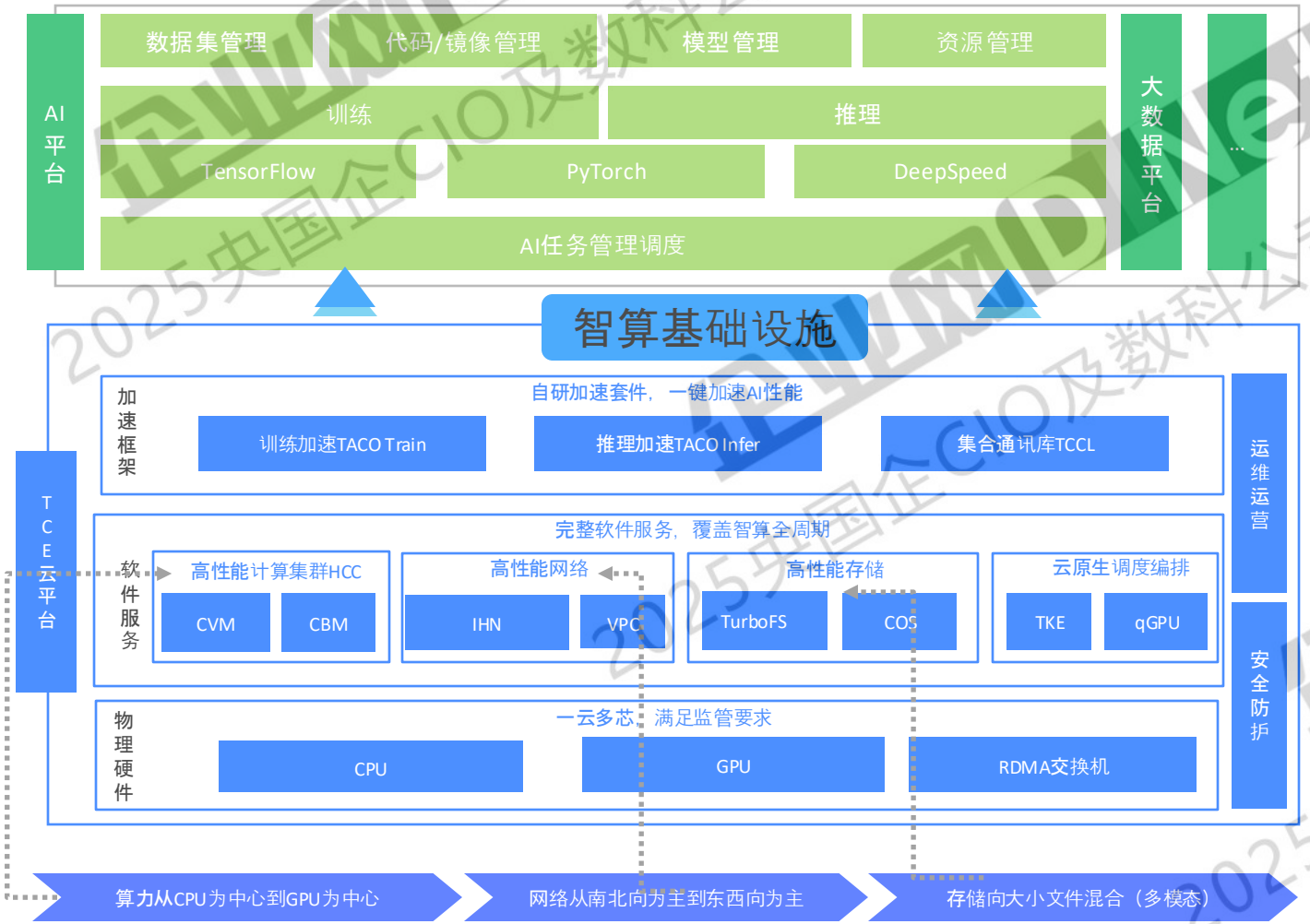
基于DeepSeek的应用架构



- **前端访问服务：**OpenWebUI的Web页面服务和DeepSeek R1模型的API服务都可以通过CLB ingress对外提供高并发高性能高可用的访问接入；
- **模型算法服务：**承载DeepSeek R1模型的vLLM推理框架，部署在TKE容器服务平台实现高可用高性能高扩展性；
- **算力资源服务：**HCC、IHN以及TurboFS等为TKE容器服务平台提供各种各样的计算、存储、网络资源。

DeepSeek满血版的坚实基座：腾讯专有云智算方案

- TCE智算方案定位于智算场景下的算力基础设施平台。包括了HCC、IHN、TurboFS等产品，提供完善的高性能计算、存储、网络能力，为上层应用平台提供基础算力



服务业务，拥抱AI

利用完整的软硬件智算生态，可快速建立企业的AI解决方案

开放兼容，安全合规

一云多芯，适配兼容多种GPU芯片和服务器硬件

多元算力

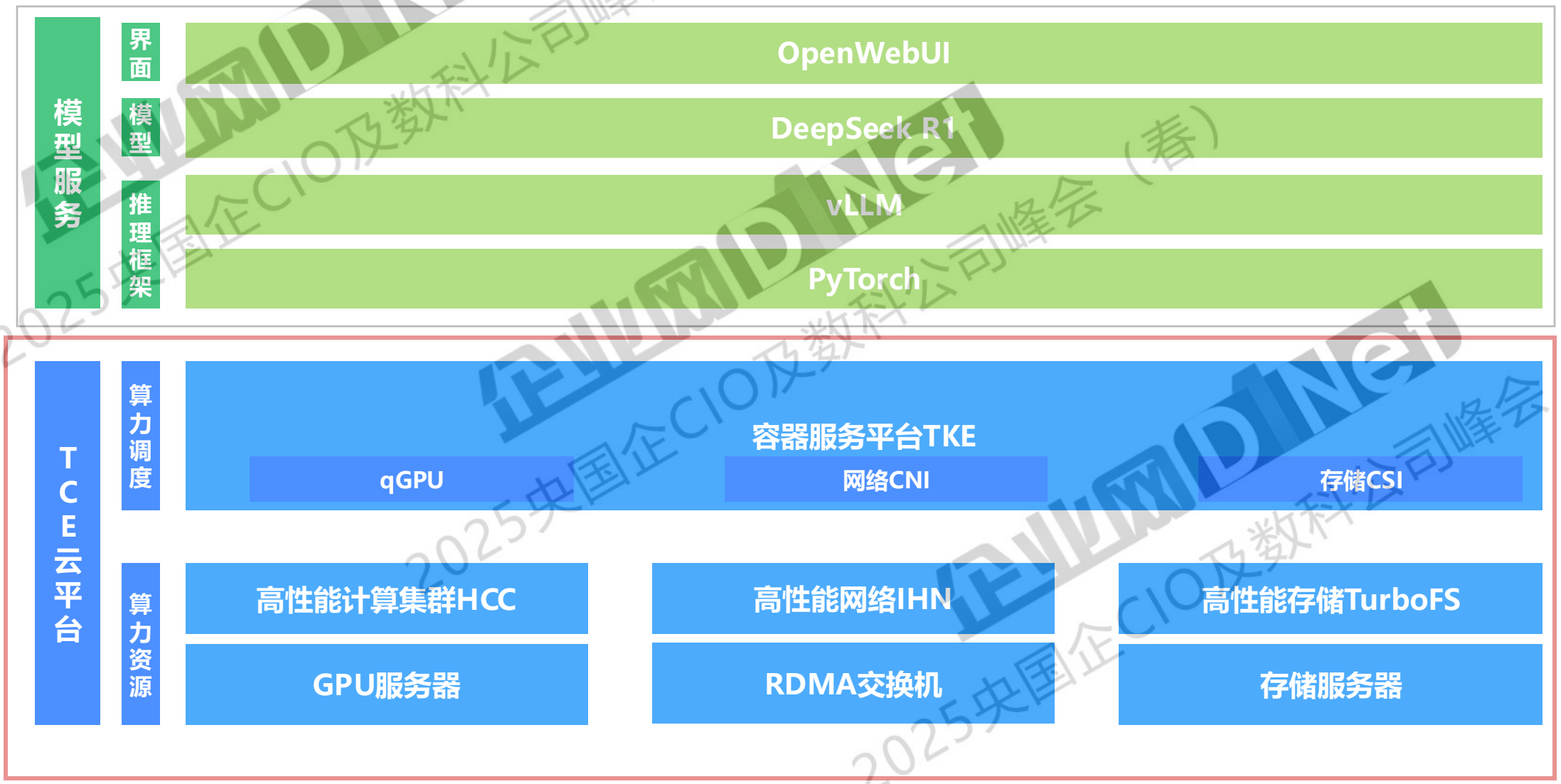
提供训练、推理等场景的多元算力，包括单机单卡、单机多卡、多机多卡，并可通过qGPU技术为智算算力进行精准划分和调度

卓越性能

自研软硬一体高性能网络IHN：3.2T带宽接入，多轨道聚合流量架构，自适应通信优化。

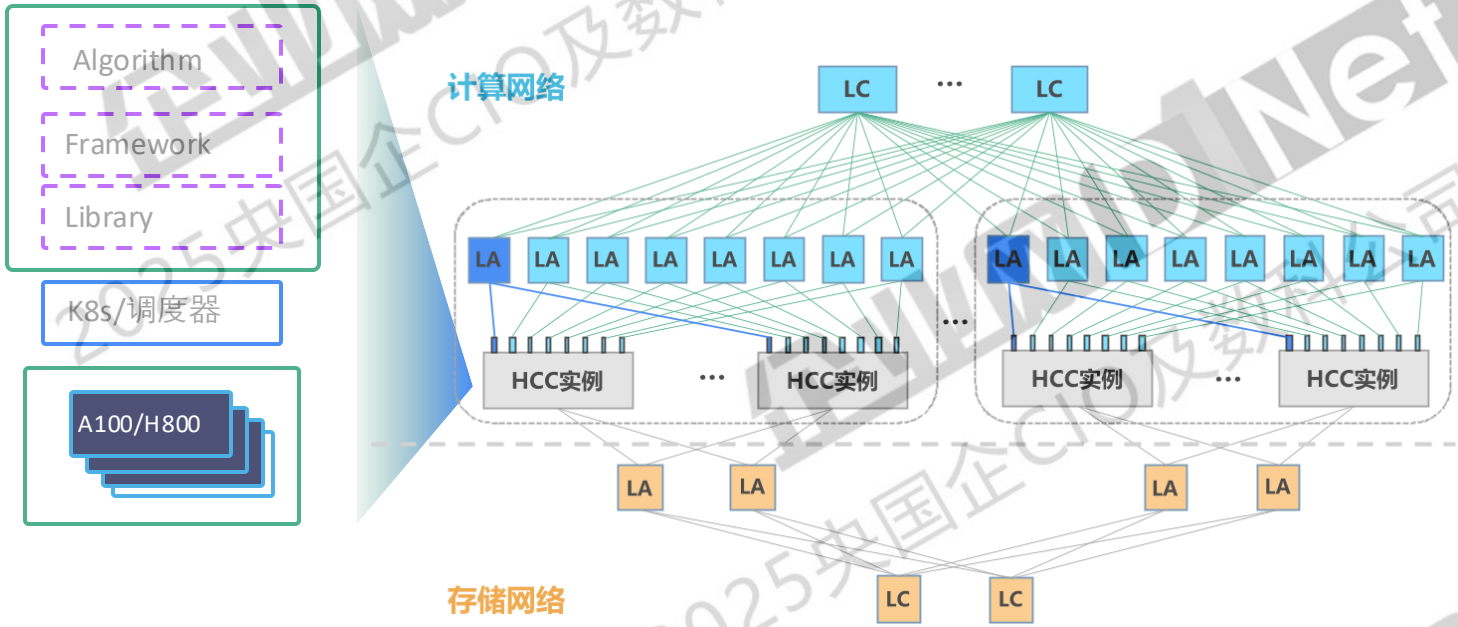
分布式训练/推理加速框架TACO

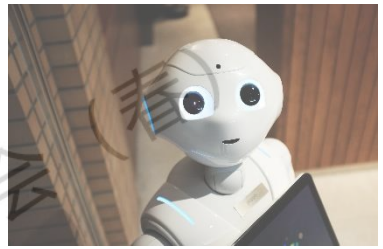
DeepSeek应用在TCE上的部署架构



高性能计算HCC：异构算力池调度器

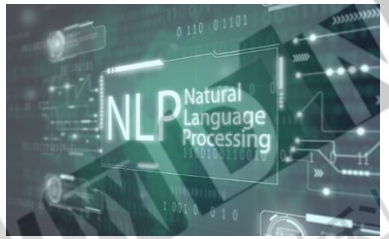
高性能计算集群 (HCC)，面向大规模AI及高性能计算场景，广泛适用于大型语言模型、自动驾驶、商业推荐系统、语音识别、图像识别、AI制药等人工智能模型训练场景。



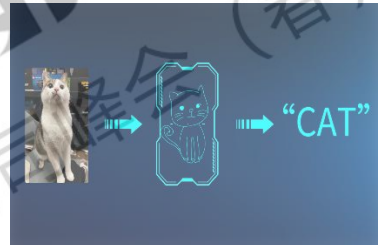


AIGC大模型生成

自动驾驶



自然语言处理



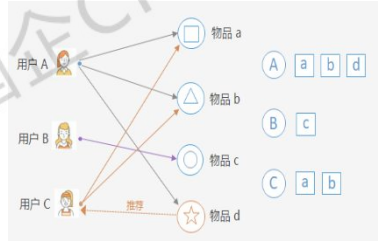
图像识别

产品关键特性

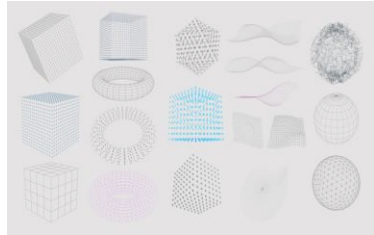
搭配高性能GPU：产品支持A800及H800 NvLink GPU，提供强大算力

低延时RDMA网络：节点互联网络低至2us，带宽支持800G-3.2Tbps

GpuDirect RDMA：GPU计算数据无需绕行，跨机点对点直连

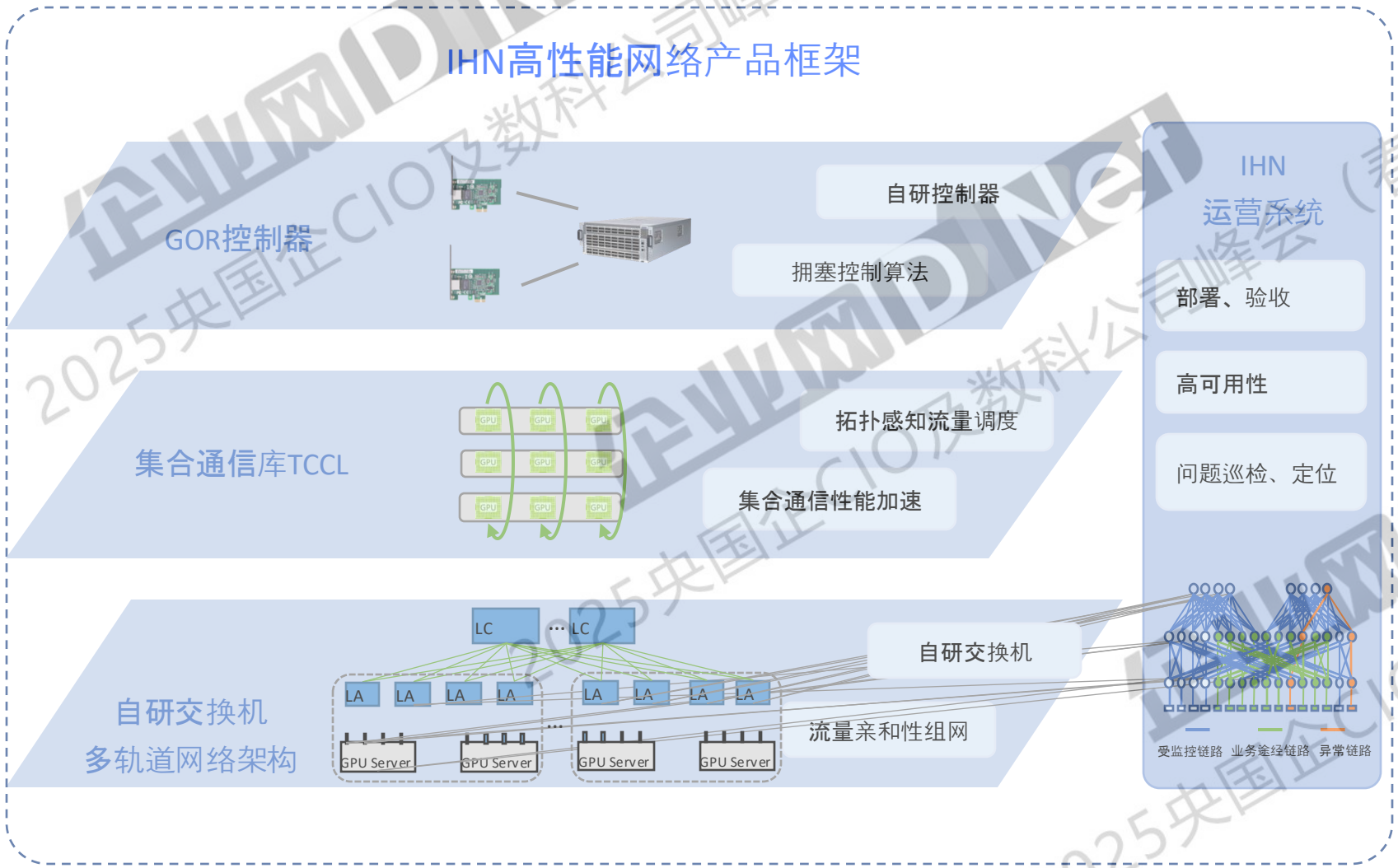


推荐系统



材料仿真

高性能网络IHN：提供训练及满血版推理依赖的RDMA网络



■ 产品能力

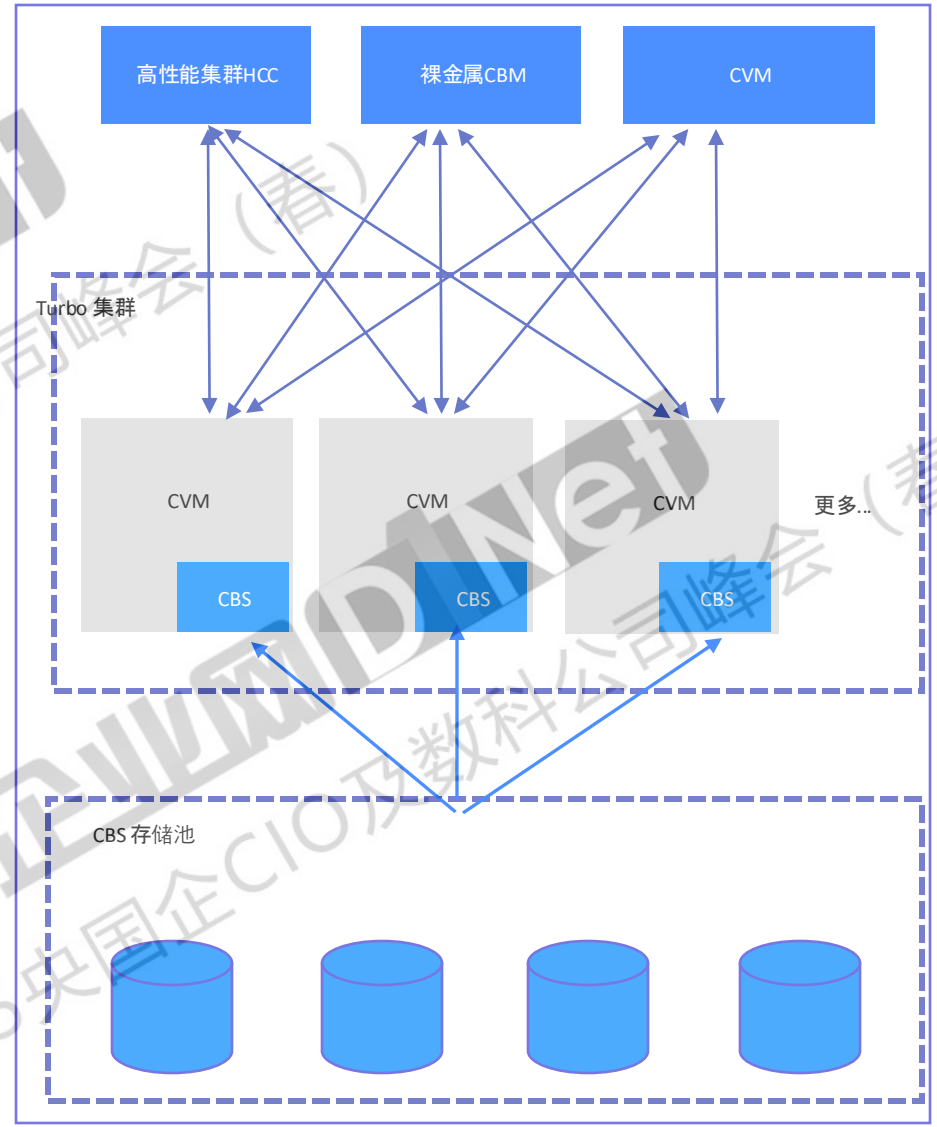
- 低成本、大规模多轨道网络架构：
  - 全栈自研交换机，支持接口100G/200G/400G；
  - 方案与AI流量亲和，路径时延降低40%，支持高冗余bonding组网；
  - 超大规模：通过多轨道组网，最大弹性支持100K+GPU卡规模
- 高性能无损网络：
  - TCCL，感知拓扑进行流量亲和性调度，实现AllReduce的负载率达到90%以上。
  - 自研协议TITA&GOR控制器，全局业务流精准的监控、选择、决策和调度，3分钟内完成拥塞消除。
- ms级时延度量，分钟级自愈运营系统：
  - GOM：全QP精细监控，可快速定位网络、GPU故障节点，实现集群自愈
  - 网络故障1分钟发现、3分钟定位、5分钟自愈

并行文件存储TurboFS：100GB+ 吞吐

- TurboFS可与CVM、TKE、CBM等云产品搭配使用，实现数据的高效共享访问
- 百 GB 级吞吐，千万级 IOPS，应对最苛刻的HPC业务大小文件混合的性能挑战

产品名称	TurboFS并行文件存储
产品定位	高吞吐、大容量：大规模吞吐业务 高IOPS、低时延：高负载下延迟敏感型业务
适用场景	大规模高性能计算、AI 训练
容量上限	10TiB-100PiB
性能规格	带宽：100GB/s IOPS：1000万
延迟	4K 单流读：0.2ms 4K 单流写：1.5ms
支持协议	POSIX/MPI
支持操作系统	Linux系统

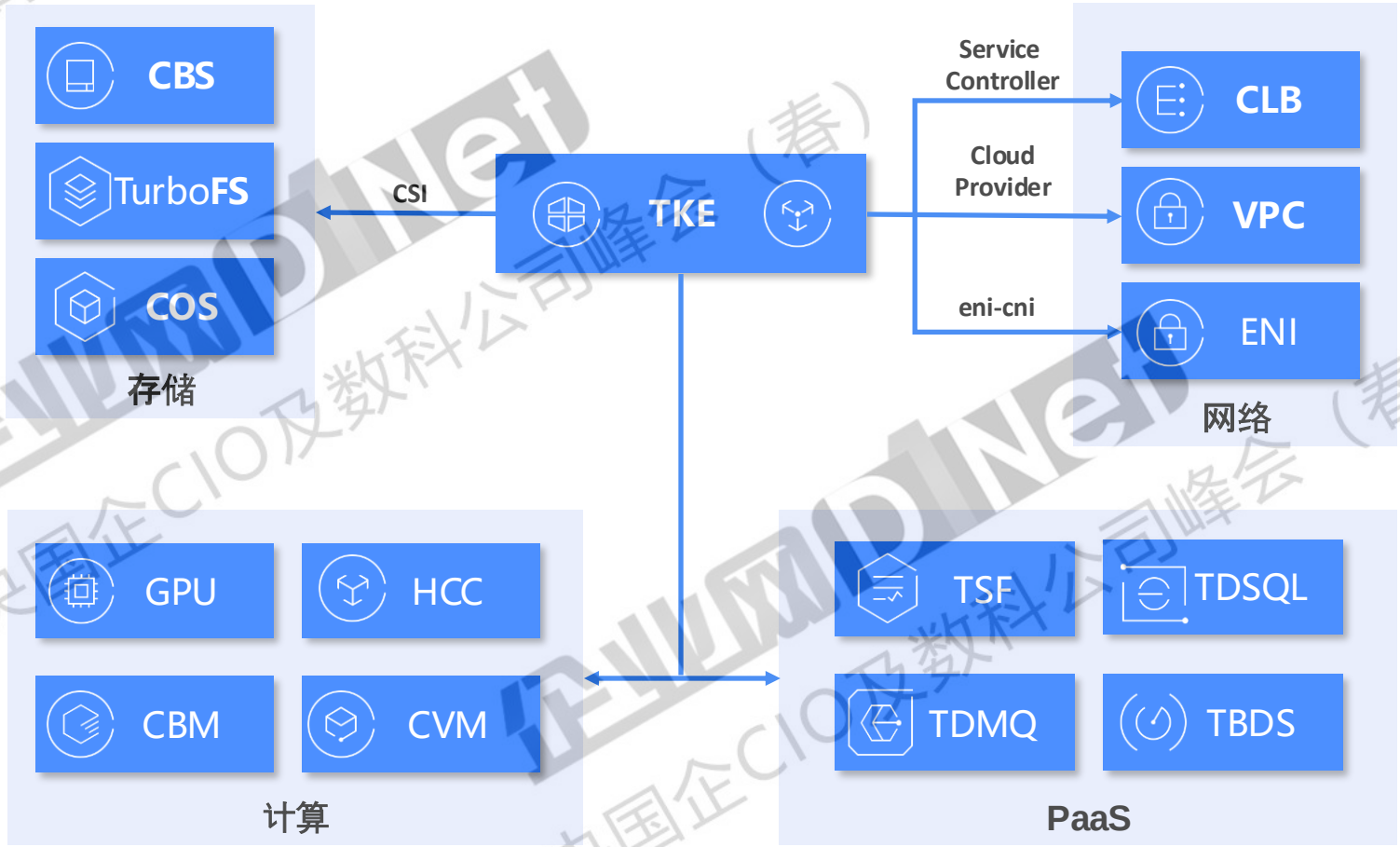
- 采用非对称架构，数据节点和元数据节点独立部署
- 资源在底层进行隔离，保障存储集群独享
- 三副本强一致架构能力，CVM位于不同机架的独立物理服务器严格反亲和
- 提供接入机热迁移的机制，保障数据的可靠性和服务的高可用



容器服务平台TKE，实现高性能基础设施插件化支持

特性

- 对 K8S 无入侵，支持原生 K8S API，通过插件机制和高性能基础设置进行集成。



qGPU实现的算力共享与精准调度，实现GPU利用率倍增

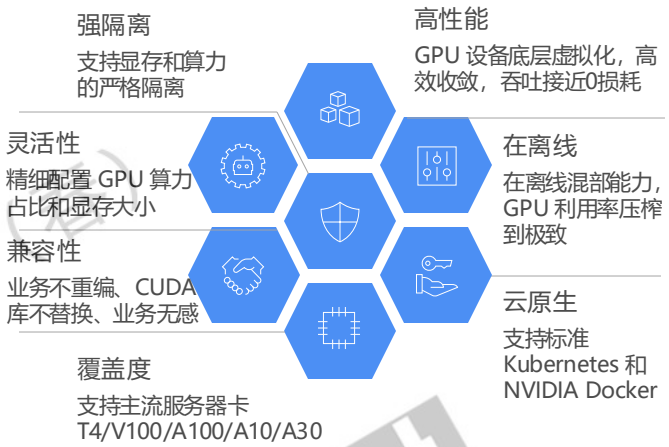
GPU算力&显存利用率低的问题

- GPU资源价格昂贵、利用率低
- 资源共享后使用不便、安全性问题

资源共享带来QoS问题

- 显存 / 算力隔离不准确，资源抢占干扰，QoS无法保证，损坏故障隔离性
- 资源切割不灵活

极致GPU利用率 + 强隔离QoS保障 → GPU算力共享技术



客户痛点

- 在线业务独占使用 GPU，利用率大多在 40% 以下
- 线下 IDC 很难满足业务需求增长，线下采购周期长
- 线下 IDC 故障隔离性差，运维成本较高

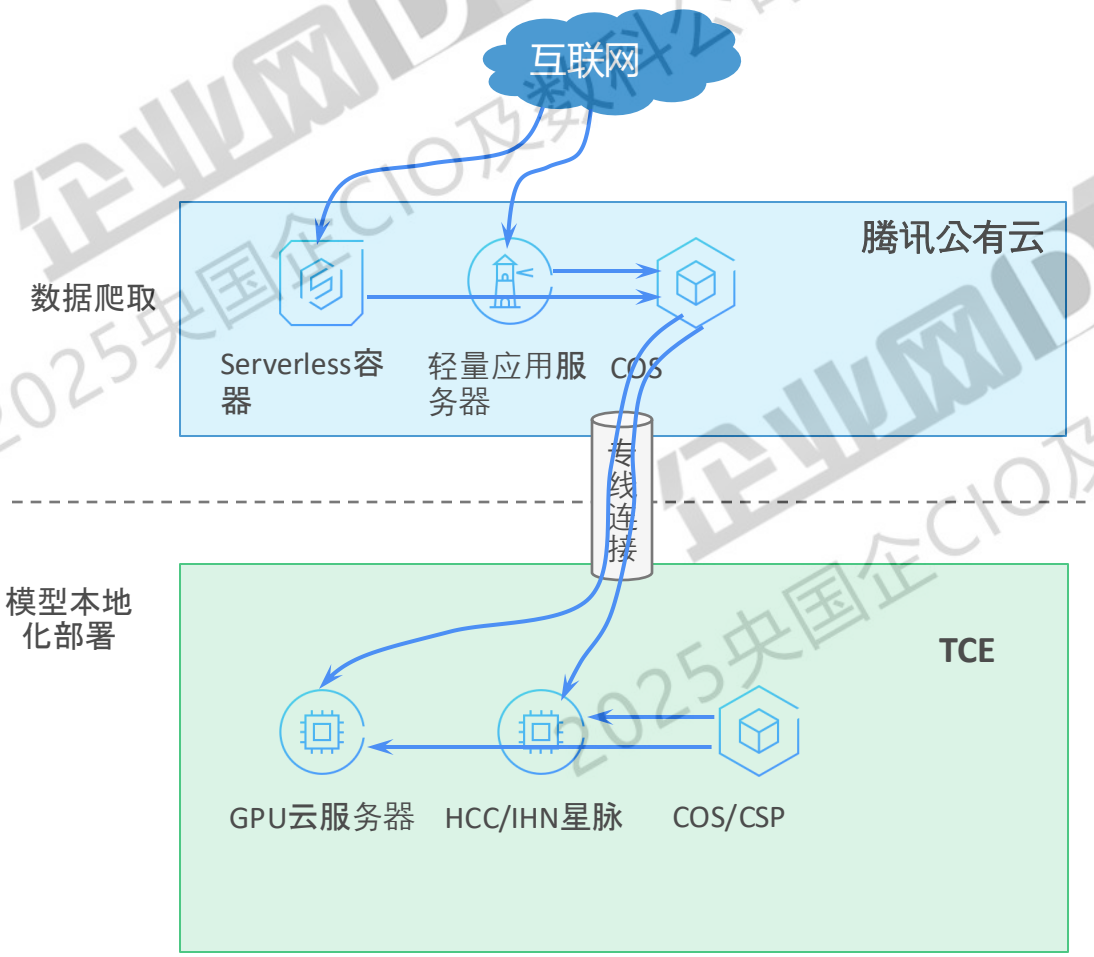
解决方案

qGPU 是腾讯自研的新一代容器GPU虚拟化方案，实现算力隔离等能力的同时，从根源上解决特殊场景下的GPU共享的干扰问题

云上收益

- 节省一次性投资成本，随用随取，减少资源闲置
- CVM 弹性扩容优势，涵盖空间、时间、大小和数量，可根据业务快速动态扩容
- TKE qGPU 容器增加 1-3倍 业务部署密度，实现 GPU 多业务共享，算力厘米级，显存 MB 级隔离，大幅降低用卡成本
- 节省运维成本，腾讯云上提供了 TKE、qGPU、COS 等各类产品组合使用
- 年 TCO 成本节约 50%+，利用率提升 100%

与公有云构建混合云，兼顾安全性与知识库时效性



业务需求

本地化部署DeepSeek等生成式AI模型，保证信息安全；

通过爬虫技术等方式从互联网获取数据，扩充知识库；

企业私有知识库部署在本地；

解决方案

公有云侧使用serverless容器或轻量化应用服务器部署爬虫，并将公有知识库存放在COS；

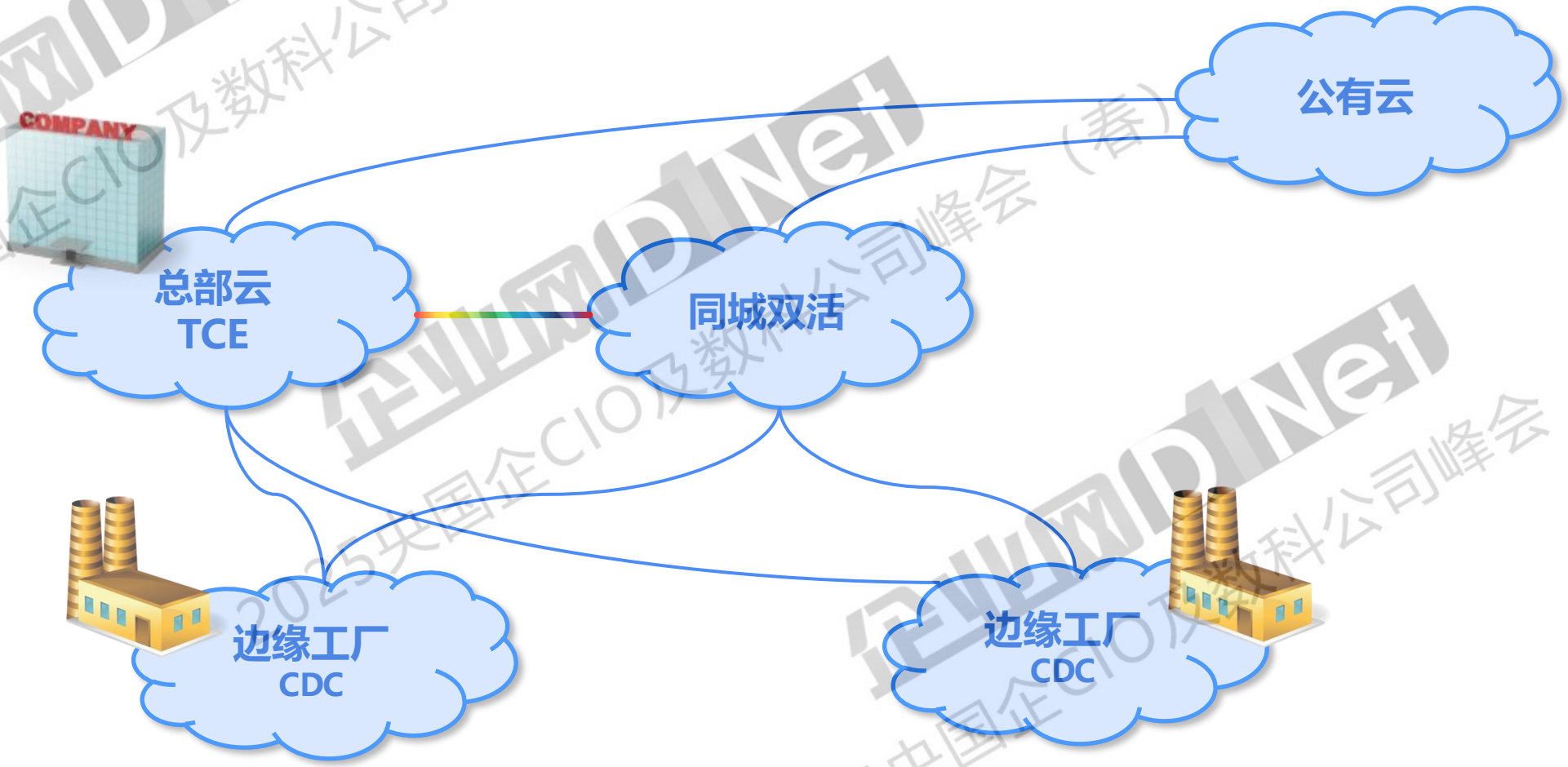
TCE侧使用GPU云服务器/HCC部署AI推理，本地存放私有知识库；推理时同时从公有知识库和私有知识库生成回答；

客户价值

本地部署AI模型，保证企业私有知识库安全性，推理过程不泄漏企业内部数据；

利用公有云的低成本，高效地从互联网爬取数据更新知识库；

云与边缘协同，实现算力下沉到边缘节点



云与边缘计算的融合协同，实现总部统一管理各边缘节点算力资源

生态共赢：广泛兼容国产芯片

GPU/异构计算

昇腾：昇腾910B

海光：Z100L，K100

其他：燧原S60、昆仑芯、沐曦...

CPU

鲲鹏：鲲鹏920，鲲鹏920B

海光：海光1/2/3/4代

飞腾：FT 2000/FT S2500/FT S5000-C

其他：合芯国产Power

成功案例 SH科技SunClouds实现DeepSeek服务WX用户

尚云AI弹性算力云平台SunClouds：https://www.sunclouds.com/index

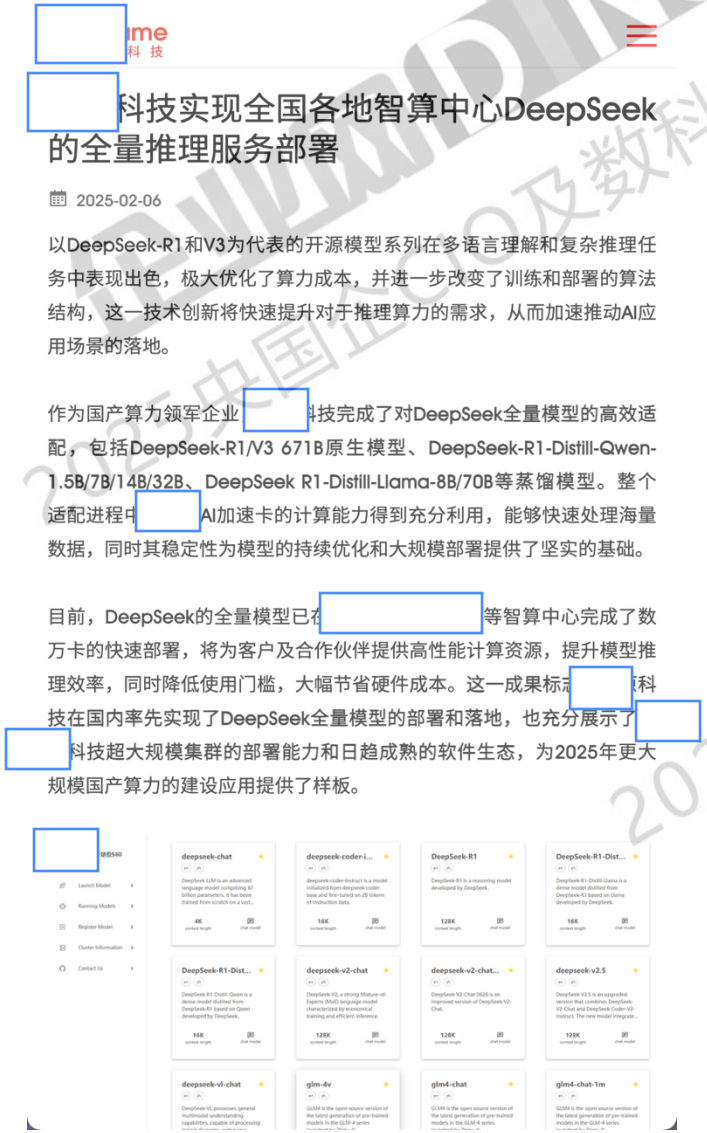
SunClouds采用腾讯专有云TCE作为算力云平台的核心基座，承载多种异构计算、通用计算、分布式存储、云原生及安全服务，为SH 云终端用户提供一站式自助用云、弹性用云和敏捷部署大规模并行计算及联机大模型训练/推理服务。

关键服务类型：

- 多元算力：CPU：Intel、AMD；GPU：Nvidia A800/A10/T4
- 高性能网络：NVlink及200G Infiniband
- 高性能存储：分布式并行存储TurboFS，分布式全闪块存储

https://mp.weixin.qq.com/s/2pa7mzrop9BPa9hPuQjo-g

成功案例：SY亿算云国内率先实现DeepSeek全量模型的落地



科技实现全国各地智算中心DeepSeek的全量推理服务部署

2025-02-06

以DeepSeek-R1和V3为代表的开源模型系列在多语言理解和复杂推理任务中表现出色，极大优化了算力成本，并进一步改变了训练和部署的算法结构，这一技术创新将快速提升对于推理算力的需求，从而加速推动AI应用场景的落地。

作为国产算力领军企业 科技完成了对DeepSeek全量模型的高效适配，包括DeepSeek-R1/V3 671B原生模型、DeepSeek-R1-Distill-Qwen-1.5B/7B/14B/32B、DeepSeek R1-Distill-Llama-8B/70B等蒸馏模型。整个适配进程中 AI加速卡的计算能力得到充分利用，能够快速处理海量数据，同时其稳定性为模型的持续优化和大规模部署提供了坚实的基础。

目前，DeepSeek的全量模型已在 等智算中心完成了数万卡的快速部署，将为客户及合作伙伴提供高性能计算资源，提升模型推理效率，同时降低使用门槛，大幅节省硬件成本。这一成果标志 科技在国内率先实现了DeepSeek全量模型的部署和落地，也充分展示了 科技超大规模集群的部署能力和日趋成熟的软件生态，为2025年更大规模国产算力的建设应用提供了样板。

https://www.enflame-tech.com/news/328

燧原科技国产万卡算力集群点亮仪式

腾讯助力“东数西算”工程，依托专有云TCE建设支撑QY市人工智能产业蓬勃发展的重要云“底座”。结合国产化SY S60 GPU，构建全新一代的国产化万卡GPU智算数据中心，项目的建设标志着国产算力集群取得了阶段性的成果。未来将面向泛互联网和大模型应用行业客户提供推理算力资源和大模型应用服务。

关键服务类型：

- 超大AI算力集群：91台通用服务器+1252台SY S60*8服务器，万卡集群
- 高性能存储：分布式并行存储TurboFS，分布式全闪块存储
- 国密测评

感谢倾听