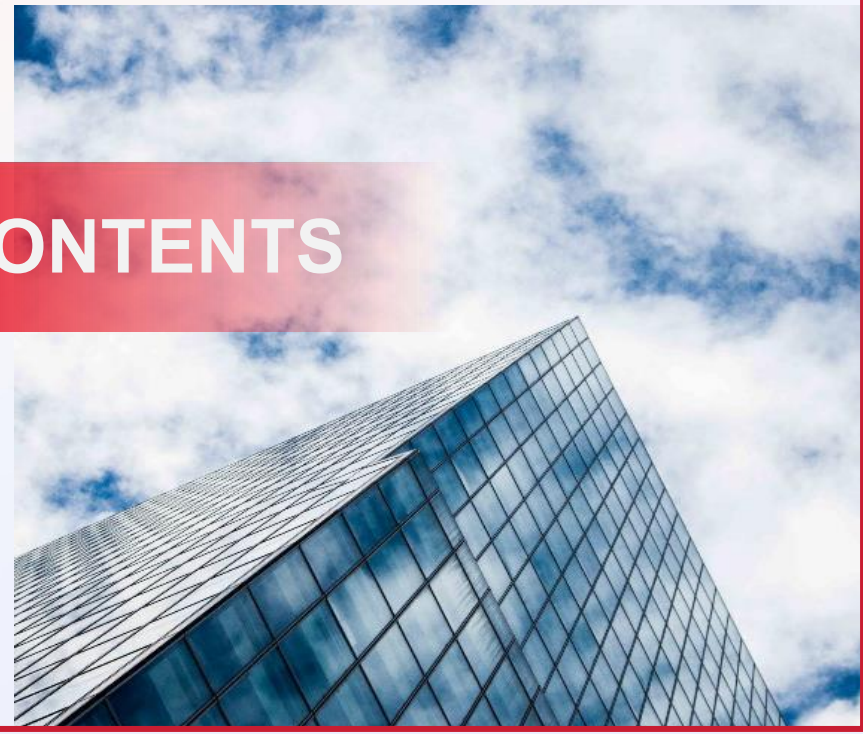
算力x联接，释放保险AIGC无限潜能

卢爱周 | 新华三金融事业部

目录

CONTENTS



AIGC在保险行业的应用场景

大模型相关金融的应用场景

前台 通用金融场景

智能客服、智能投顾、智慧营销、智能助手、客户获取、客户维护、客户跟进、智能话术、智能投研、金融数字人……

前台 专用金融场景

银行	保险	证券
智能征信	智能核保	智能投行
贷款催收	智能理赔	智能资讯
……	……	……

中台

智慧风控、反欺诈、安全服务、自助分析、智能知识库……

后台

智能运维、智能代码生成、智能运营、知识管理、人力管理、智能培训、文档处理……

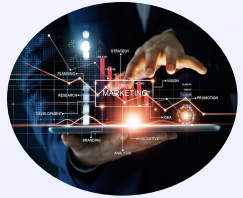
保险行业适合的场景

前台赋能

保险产品智能配置
保险客服
代理人销售支持

定损理赔

智能定损
智能理赔



风控减损

事前预防减损
防欺诈

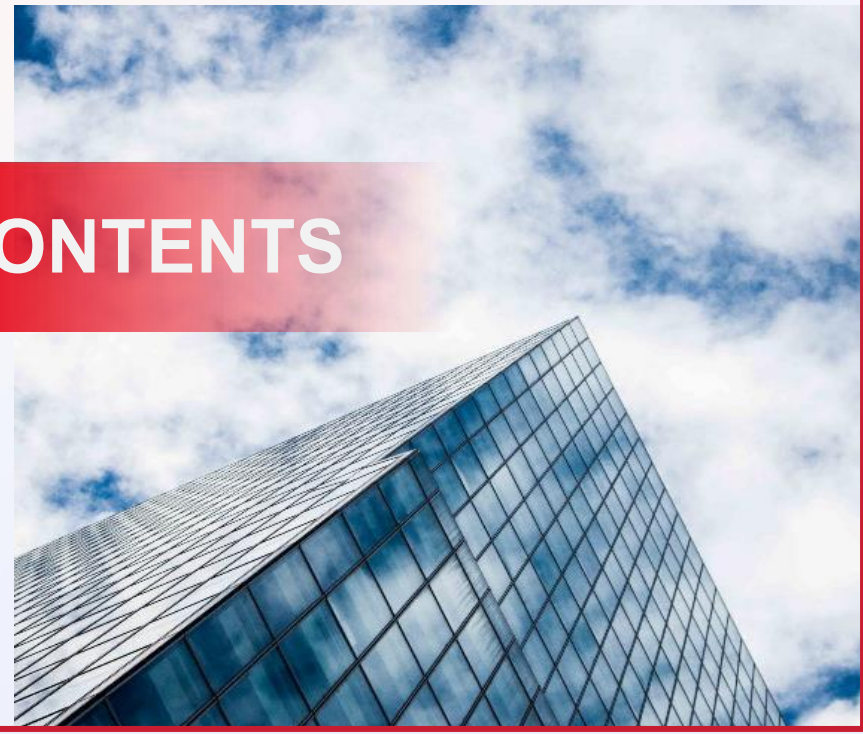
产品开发

车险定价
代码生成

参考财通证券《保险+AI 深度报告》

目录

CONTENTS



新华三多年深耕ICT领域

20年深耕积累

新华三具备AI全栈解决方案能力

新华三技术战略

AI in ALL

让新华三产品和解决方案更智能

AI for ALL

赋能百行百业，让客户的应用更智慧

算力

算力多元化

CPU、GPU多元化选择满足不同市场不同客户的需求同时提升供应链供给保障

×

联接

联接标准化

服务器内/外部联接标准归一化，实现异构智算集群，降低算力部署成本

软件生态标准化，便于算力资源和大模型互联互通，提升智算资源利用率

充分发挥算力和联接双基石的乘数效应，带来倍增的技术整合优势和智算效率

新华三保险AIGC解决方案

- **大模型全栈能力，**支撑从模型到基础设施一站式应用部署
- **统一的数据共享平台，**支持行业数据要素统一治理和管理
- 算力、存力、运力**协同感知**的高性能计算集群
- 符合国家“双碳战略”的**零碳绿色**算力中心

国产AI加速卡市场情况

01 CPU+AI加速卡厂商

国产化服务器CPU厂商，同时也有国产AI芯片

HYGON 中科海光

海光+DCU　GPU

02 国产AI加速卡主流厂商

有大批量出货案例，客户基础相对扎实

昆仑芯 KUNLUNXIN

DSA

Cambricon 寒武纪

03 AI加速器初创厂商

创业公司为主，特定行业有一定客户基础

壁仞科技 BIREN TECHNOLOGY
摩尔线程 MOORE THREADS
天数智芯 Iluvatar CoreX
MEETAX 沐曦

GPU

Lynxi 灵汐科技
SOPHGO 算能
MOFFETT AI
Enflame 燧原科技

DSA

主流国产AI加速卡评价

1、性能

产品制程
训练卡
推理卡
大模型
小模型

2、通用性

CUDA兼容和迁移的难度
不同开发框架支持
新模型适配速度

3、生态

封闭
开放

4、功耗

单卡能耗
服务器能耗
单位功耗产生算力

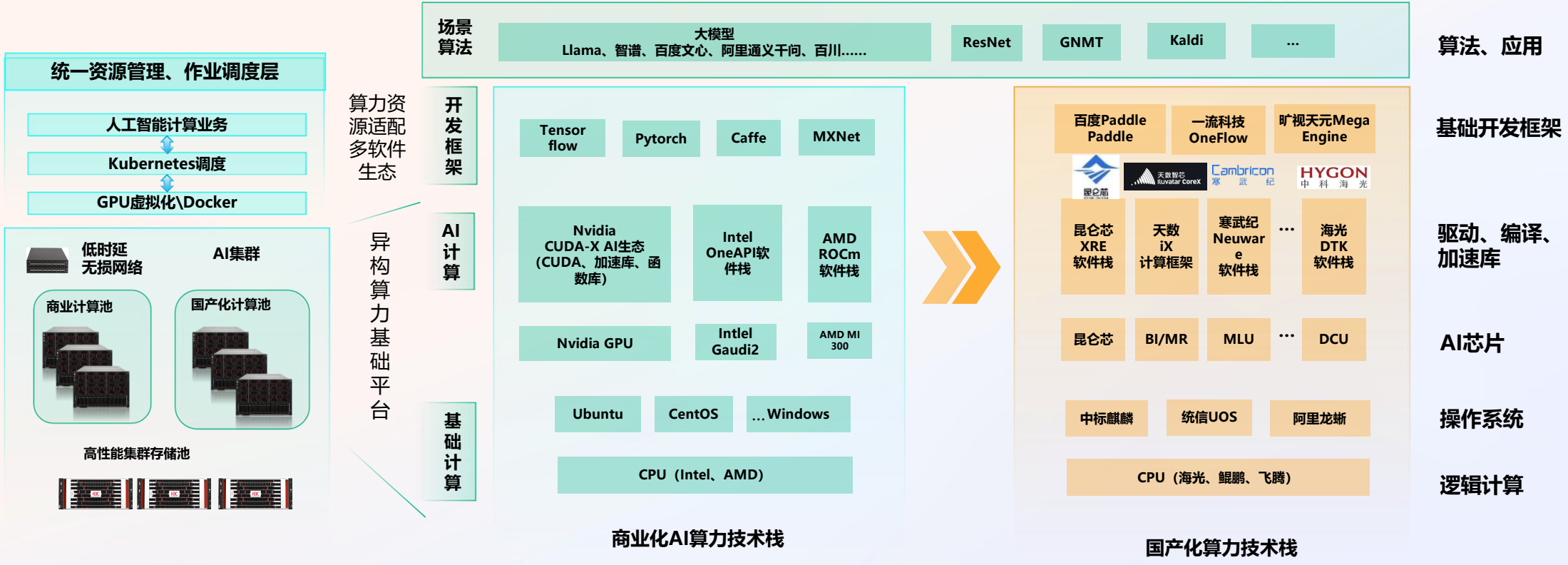
5、价格

算力卡价格
单机价格
单位成本产生算力

6、供应链

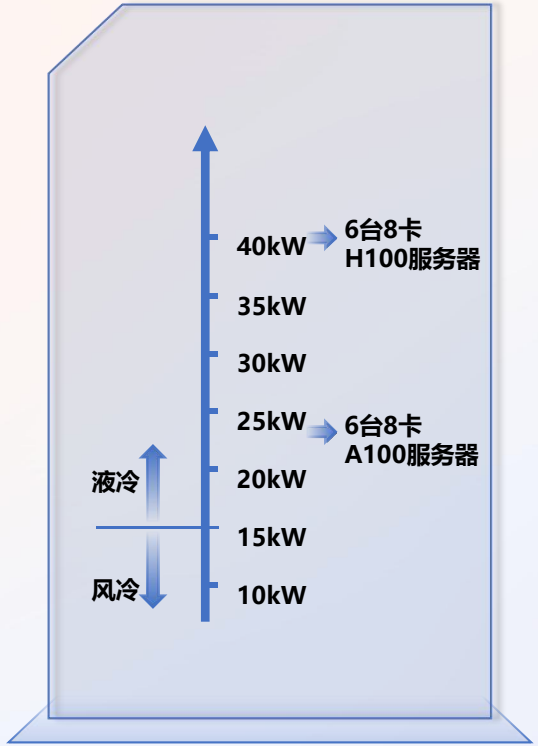
近年产品供应总量
是否被美国制裁
公司规模
产品路线图

构筑多元异构智算底座，支撑开放算力生态



数据中心能力更强大

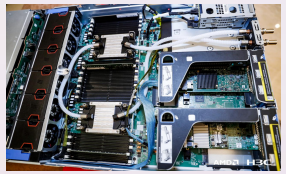
5亿 美元	460万 美元	19万 度电	85吨 碳排放
新建算力中心耗资	新建算力中心耗资	训练一次耗电	训练一次二氧化碳当量



为了实现最高密度的算力 液冷服务器是最佳方案

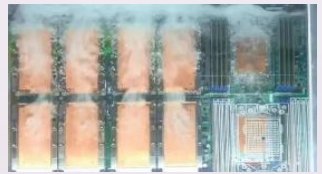
冷板式液冷

- 年平均PUE值1.2左右
- 单机柜达45KW



浸没式液冷

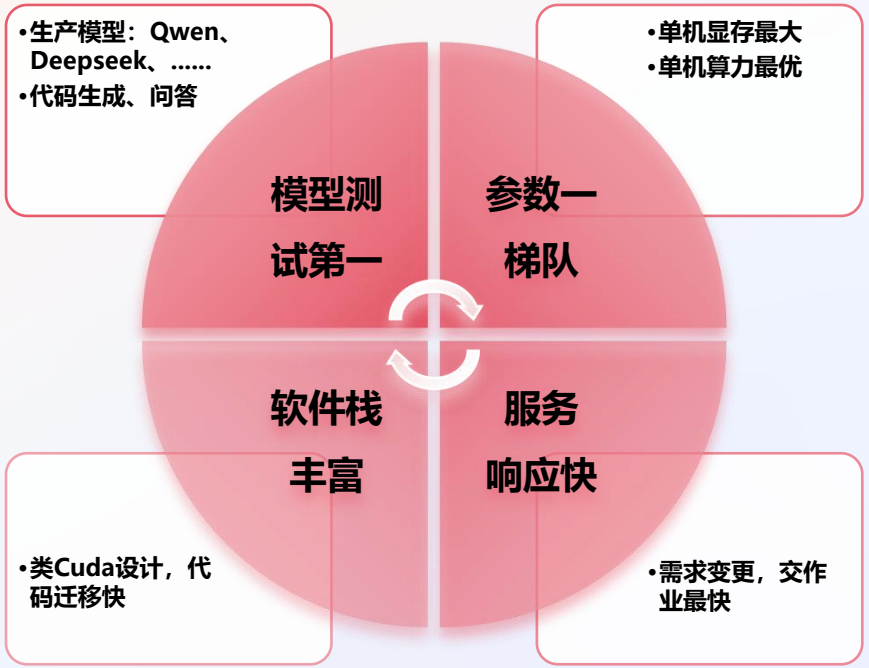
- 年平均PUE值1.13左右
- 单机柜功率基本无上限



案例 | 某头部银行构AI芯片资源采购项目第一份额

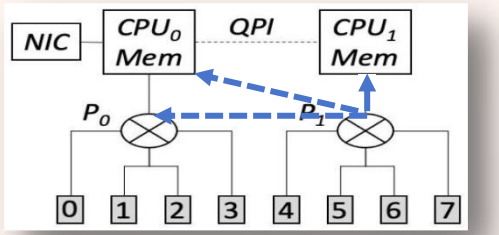
65台

H3C R5500 G7服务器
(昆仑芯P800 模组)

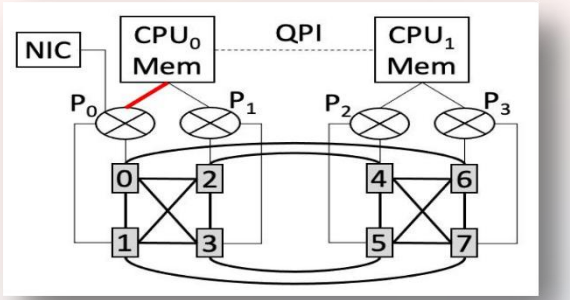


AIGC网络形态

GPU服务器节点内通信

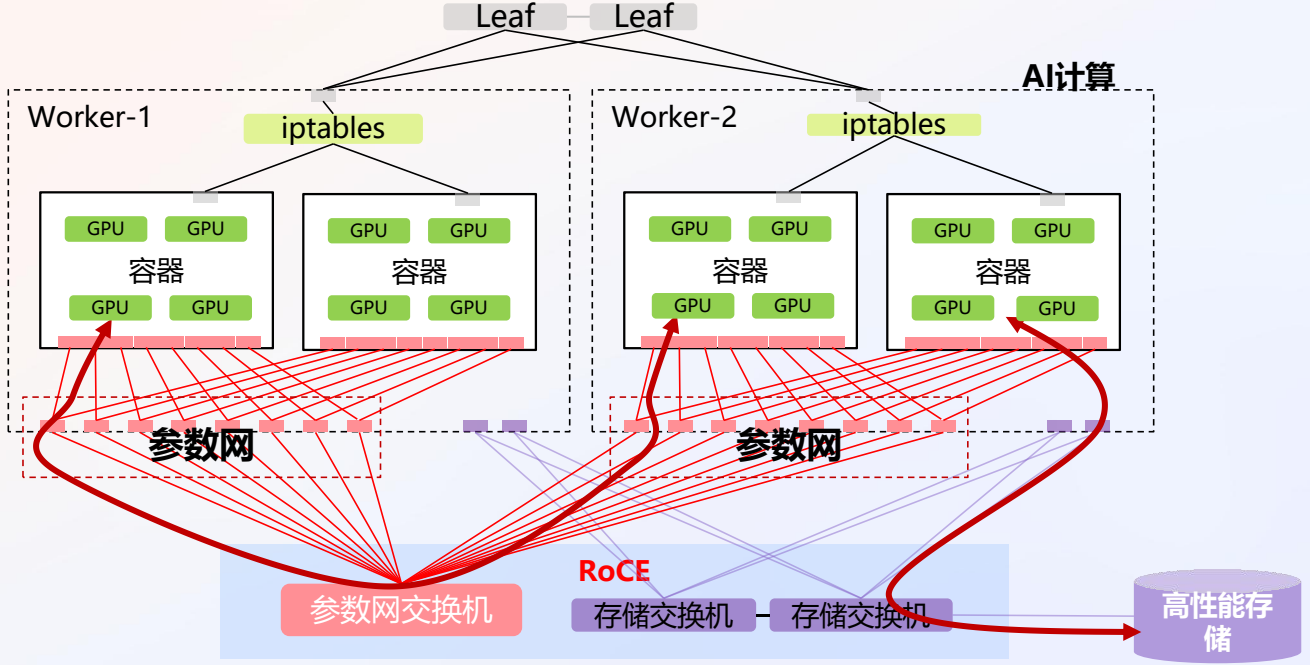


PCIe-only 型拓扑



NVLink-based 拓扑

GPU服务器节点间需要高速网络来实现模型训练参数交换和存取



通过**400/200Gbps**的RDMA网络，实现训练任务的GPU间高速网络互通

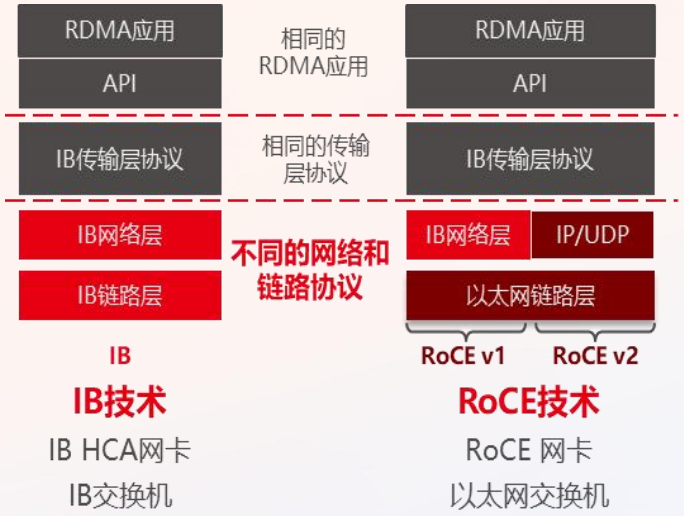
AI算力资源池建设网络规划

适用场景

适用：需要进行**多服务器**分布式训练/微调/推理(极少)场景，进行参数交互或数据读写
可不用：**单卡或单机**训练的场景，大模型出现之前，AI训练基本没有RoCE要求

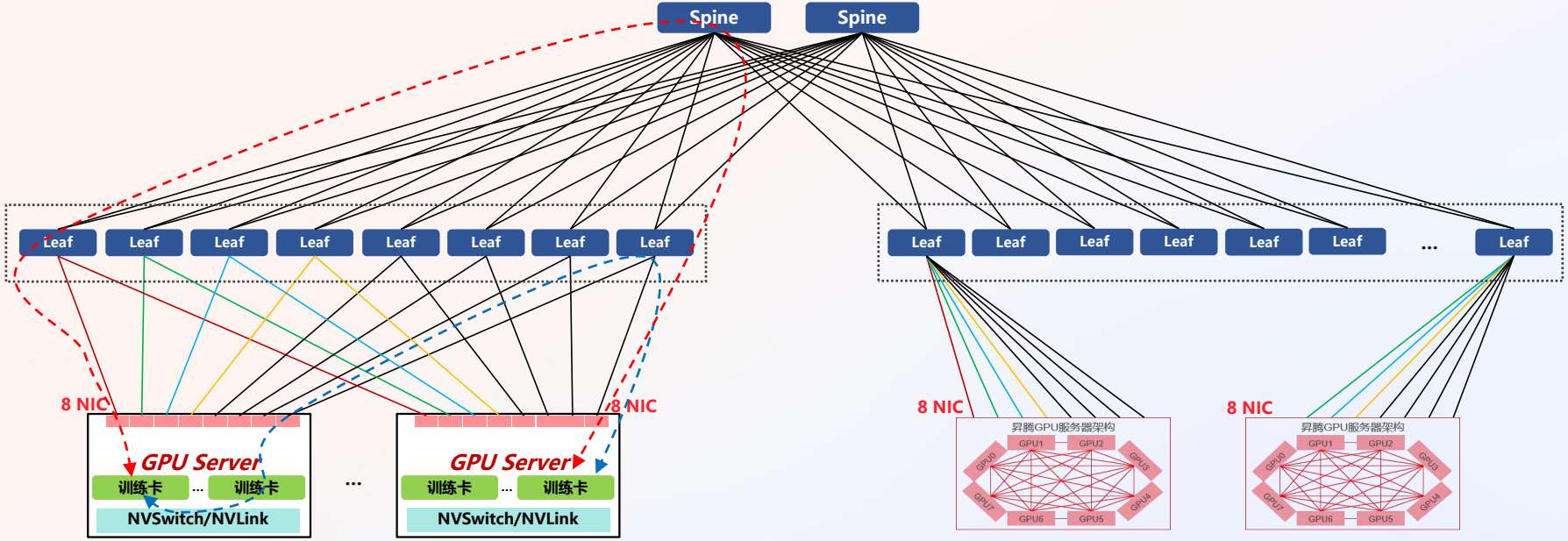
带宽需求

参数网性能需求：PCIE4.0，主推100G/200G；**PCIE5.0主推200G/400G**
存储网性能需求：主推100G



技术	IB	RoCE
开放性	私有技术，专用专网	**开放**以太，融合网络
国产化	被美国NVIDIA收购	基于**开放**标准，自主可控
带宽	演进较慢（400G刚发布）	**领先**（400G已商用，800GE标准制定中）
时延	静态时**延低**，端到端时延总体持平	静态时延略高，端到端时延总体持平
流控	无损-信用值	无损-PFC
管控面	集中式，大型网络部署规模受限	分布式
开局部署	**即插即用**	较复杂，可自动化简化部署
云化部署	差（不支持VXLAN）	**好**，支持租户配置动态部署
建网成本	高	**低**
运维	难，封闭网络，专网专人运维，需要原厂服务	**易**，开放网络
产业生态	一家独大，发展受限；不相容IP；运维支持能力有限	规模**大**，发展**迅速**；**相容**IP；运维**支持好**

参数训练网络多轨组网 vs 单轨组网



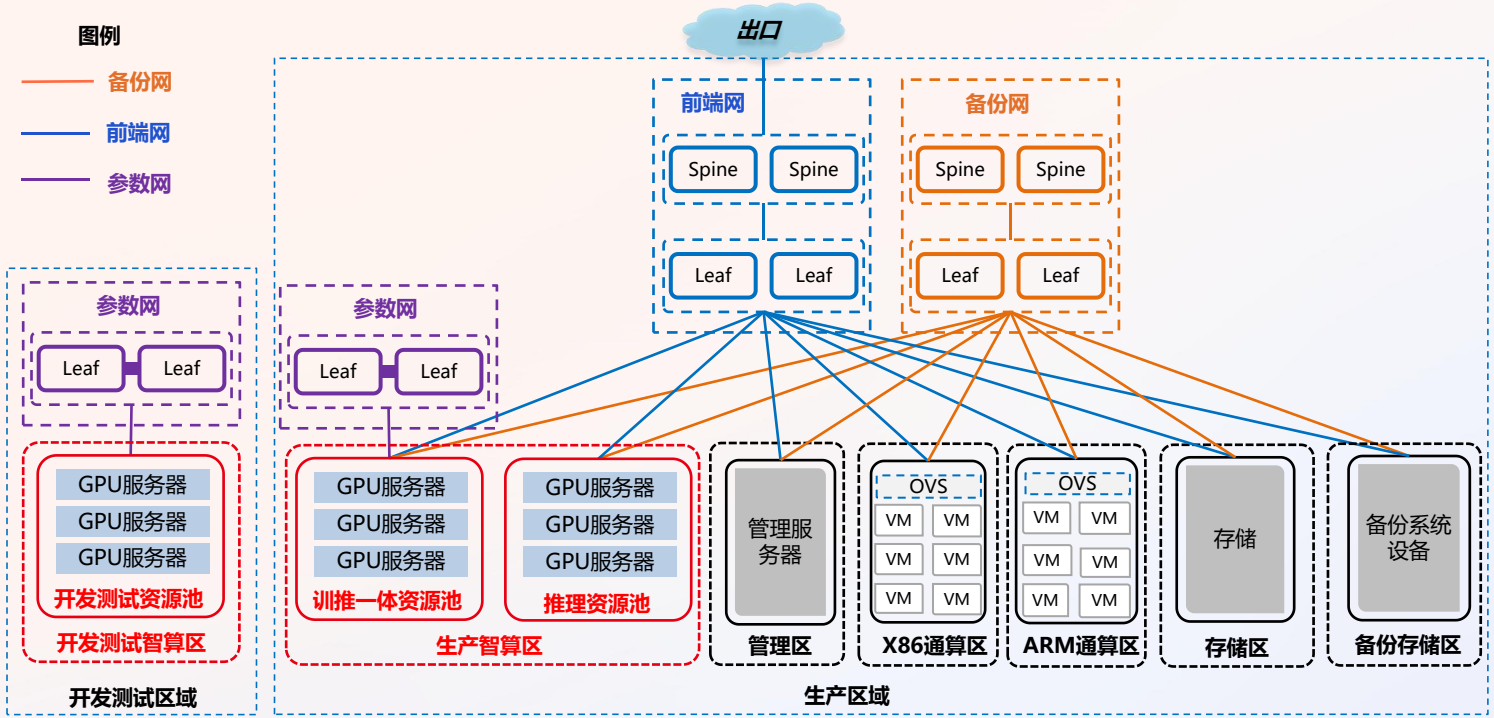
集群通讯效率最高，大部分流量经一级Leaf传输（+机内GPU代理）

Leaf故障影响GPU Server数量多

可用性高，Leaf故障影响GPU Server数量少

集群通讯效率偏低

金融智算系统部署示意图



智算系统总体架构

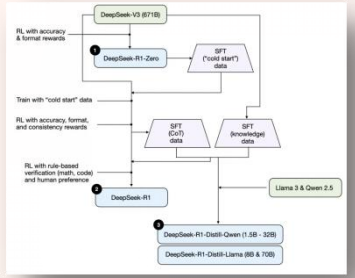
- 包括开发测试和生产两部分
- **开发测试资源池**：主要用于大模型训练和模型验证
- **推理资源池**：主要用于生产场景的大模型推理应用
- **训推一体资源池**：主要用于在生产区域大模型推理应用，也可进行大模型训练

网络部署

网络系统包括参数网、前端网、备份网、带外管理网（未画出）等

- 参数网：多机训练或多机推理联接高速网络
- 前端网：前端包括业务网、存储网、管理网等功能。

DeepSeek技术解析



DeepSeek是系统工程学的典范

模型层

新型MOE框架 | 小模型distill | 强化学习

Infra优化

FP8 混合精度 | PD分离 | 通算融合 | 避开TP并行

底层优化

PTX重构 | Dual Pipe | GPU core通信 | GPU控制IB

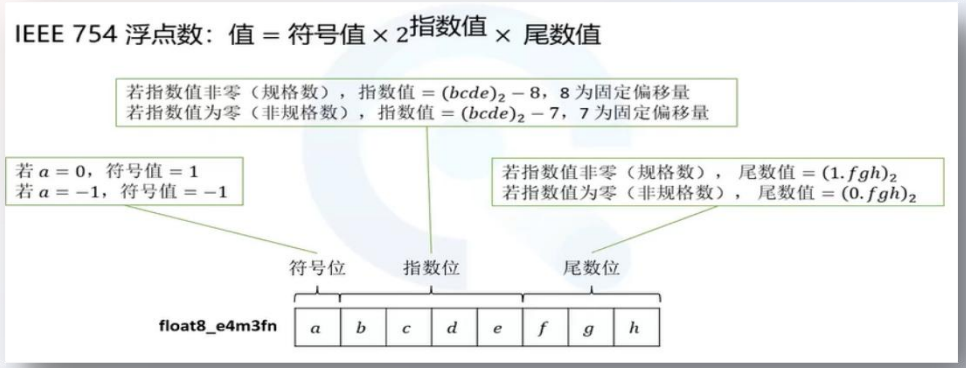
模型名称	标题创作	文章背景	全文概览	读者收获	开放性提问	关键词选择	总分	梯队	分析
DeepSeek-R1	5	5	4.5	4	4	4	26.5	第一梯队	在标题创作和文章背景维度表现完美，全文概览接近满分，展现强大的语义理解和结构化输出能力
DeepSeek-V3	4	4	3	4	4	4	23	第二梯队	适合需要行业洞察的场景
DeepSeek-R1-Distill-Qwen-32B	3	3	4	4	3	4	21		蒸馏模型中表现最佳，全文概览和读者收获优势明显
DeepSeek-R1-Distill-Qwen-14B	3	3	3	4	4	2	19	可用梯队	读者收获和开放性提问较强，但关键词选择较弱
DeepSeek-R1-Distill-Llama-70B	3	4	3	3	4	2	19		文章背景和开放性提问较强，但关键词选择较弱
Qwen/Qwen2.5-7B-Instruct	3	3	3	3	4	2	18		在开放性提问维度表现较好，适合发散性任务
gpt-4o-mini	3	3	2	3	3	4	18		在开放性提问维度表现较好，适合发散性任务
DeepSeek-R1-Distill-Llama-8B	3	0	2	3	2	2	12	待改进梯队	小参数模型普遍表现不佳，完全无法理解任务需求
Meta-Llama-3.1-8B-Instruct	0	0	0	1	1	0	2		
DeepSeek-R1-Distill-Qwen-1.5B	1	0	0	0	0	0	1		
DeepSeek-R1-Distill-Qwen-7B	0	0	0	0	0	0	0		

DeepSeek参数精度推荐

DeepSeek-R1 671B模型在训练阶段采用了 FP8（E4M3）格式

	sign	范围指数（E）								精度尾数（M）																							
		1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
FP32	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
FP16	X				X	X	X	X	X	X	X	X	X	X	X	X	X	X	X														
BF16	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X																
FP8（E5M2）	X				X	X	X	X	X	X	X																						
FP8（E4M3）	X					X	X	X	X	X	X	X																					
INT8	X		X	X	X	X	X	X	X																								
INT4	X						X	X	X																								

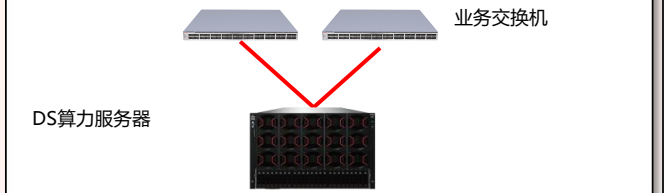
- **FP8（E4M3），最优选择**
  - ✓DeepSeek原生精度
- **FP16/BF16，国产卡，显存和算力需求比FP8大幅度增加**
  - ✓需要把FP8精度转换成FP16/BF16精度，推理精准度不降低，参数显存需要接近1.4TB
- **INT8，国产卡，显存优选**
  - ✓量化成INT8精度后，精准度略有降低
- **INT4、动态量化，国产卡，显存占用最小，牺牲精准度**
  - ✓兼容主流软硬件生态，AI领域全场景覆盖



DeepSeek满血版算力部署最小配置

deepseek 满血版8卡服务器最小配置方案

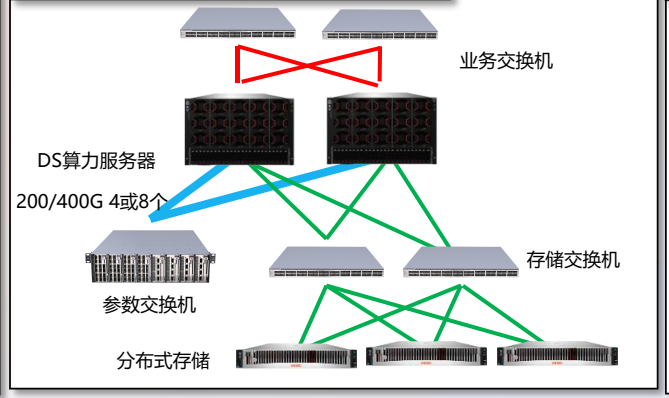
单机推理部署最小系统



方案一 单机8卡配置

1. NV HXX 8卡141G显存
2. KLX P800 8卡96G显存
3. 不需要参数交换机和共享存储系统

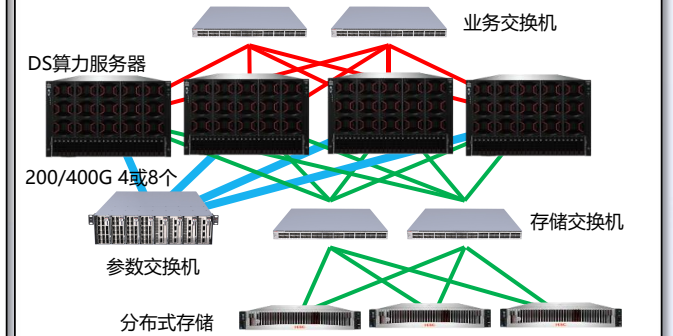
双机推理部署最小系统



方案二 双机16卡配置

1. 16卡64G显存，INT8精度
2. 需要参数交换机和共享存储系统

四机推理部署最小系统

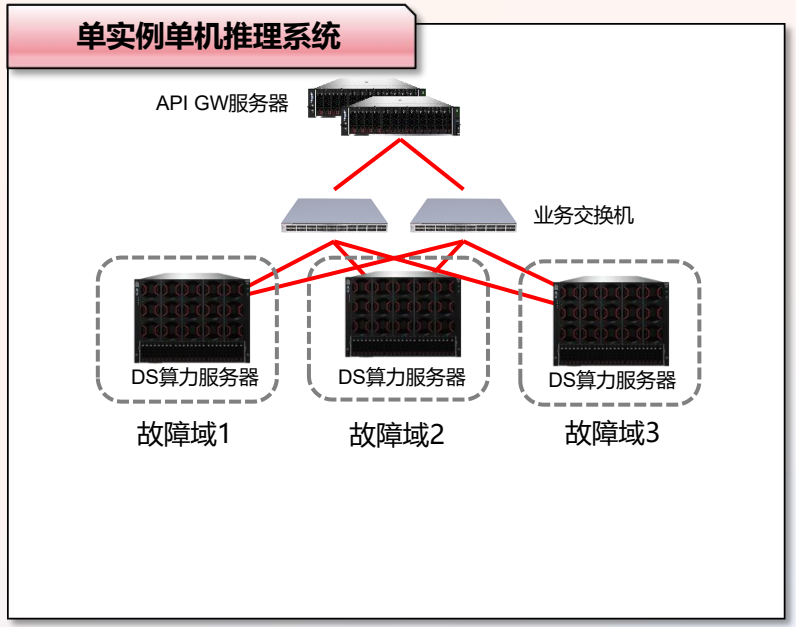


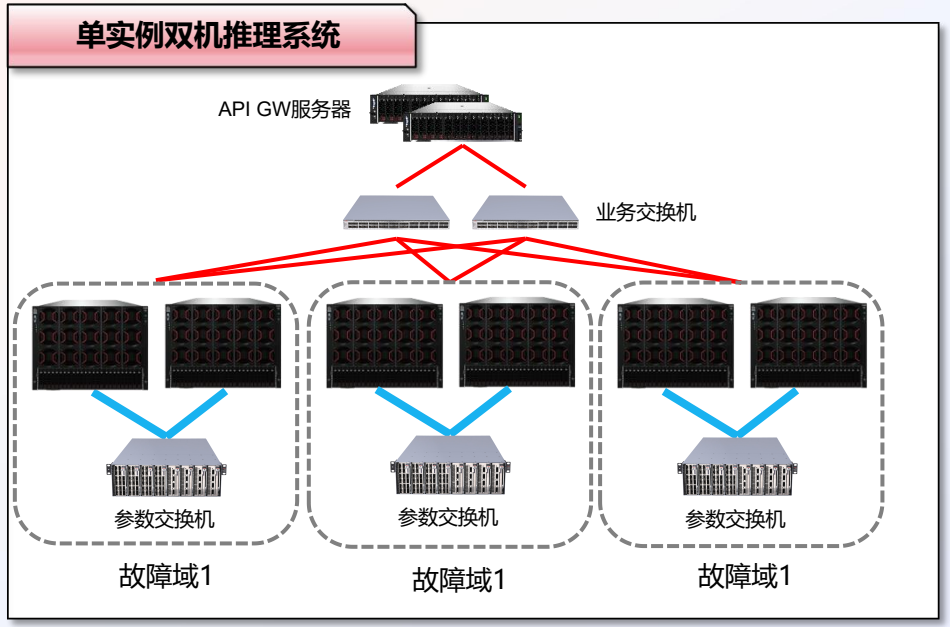
方案三 四机32卡配置

1. 32卡64G显存，FP16精度
2. 需要参数交换机和共享存储系统

DeepSeek 满血版单机VS双多机系统扩展和故障域划分

可能导致系统失效的组件包括：任一服务器的任一CPU、内存、主板、RAID卡、GPU、参数网卡，参数交换机，……





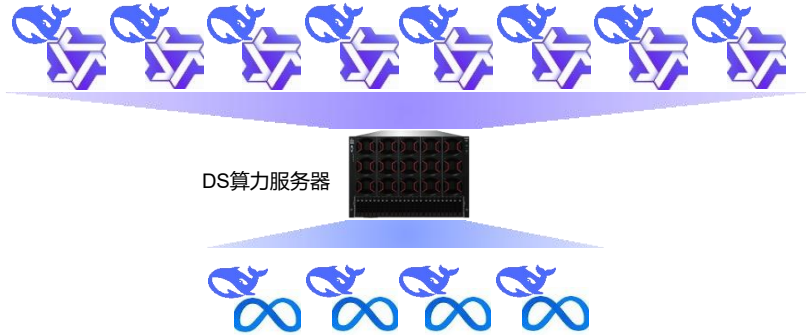
DeepSeek 蒸馏版本部署分析

部署32B参数版本需最少要1张96GB显存GPU或2张64GB显存GPU，运行BF16或FP16精度

部署70B参数版本需最少要2张96GB显存GPU或4张64GB显存GPU，运行BF16或FP16精度

96GB显存8卡服务器

支持在一台服务器上部署8个32B DeepSeek蒸馏版本实例

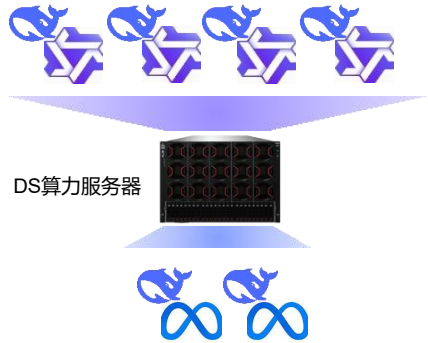


支持在一台服务器上部署4个70B DeepSeek蒸馏版本实例

单台服务器显存达到96GB的GPU，如英伟达、昆仑芯等

64GB显存8卡服务器

支持在一台服务器上部署4个32B DeepSeek蒸馏版本实例



支持在一台服务器上部署2个70B DeepSeek蒸馏版本实例

单台服务器显存达到64GB的GPU，如昇腾、沐曦、海光、天数等

96GB显存GPU服务器可以单机比64GB显存GPU服务器多承载1倍的蒸馏版本实例，供更多租户同时使用

DeepSeek将彻底改变金融行业大模型和AI算力需求格局

开放生态势不可挡

- 开源成为主流，驱动更多大模型开源，如百度文心
- 如不开源将会面临被淘汰的风险
- 围绕大模型的定制服务会成为刚需

大模型系统软硬解耦

大模型软件和GPU/网络实现解耦，金融客户面临更多的选择、不易被硬件厂商绑架

GPU选型侧重点变化

- DeepSeek R1会成为金融客户首选，但对GPU显存容量有很高的要求，带有大容量显存的GPU会受到欢迎
- 大模型测试周期及复杂度会不断缩短

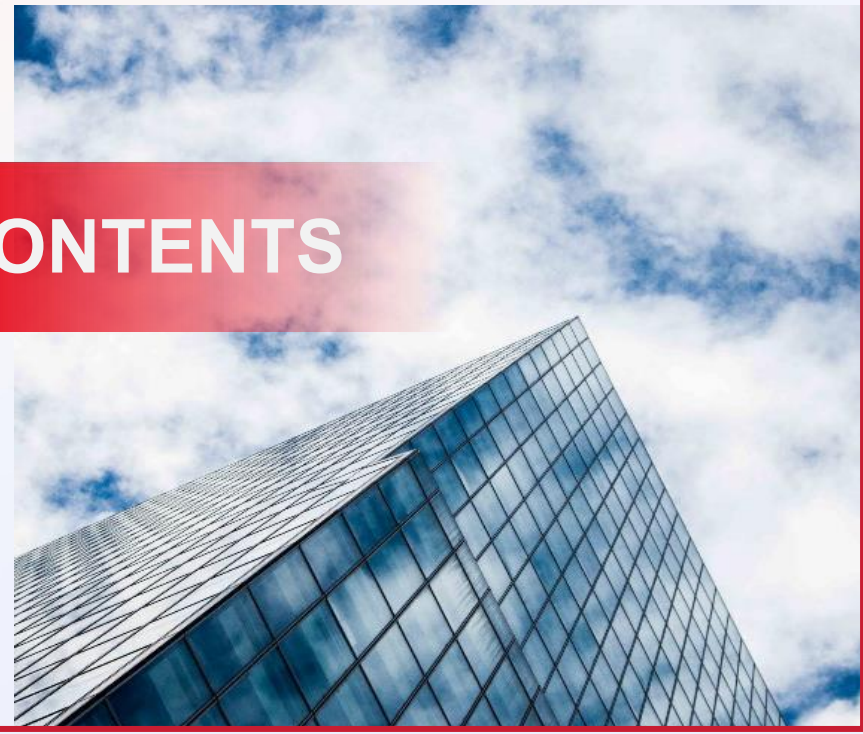
需要高速联机推理网

金融大模型最主要算力消耗是在推理场景，以前（阿里Qwen-72B、智谱GLM-130B等）的大模型只需要一台GPU服务器即可运行，DeepSeek R1 满血版往往需要多台国产GPU服务器才能运行，需以太网交换机把多台国产GPU服务器联网

随着国产AI服务器算力的大幅度提升，400G网络接入、DDR5内存、PCIe5、128线程CPU等技术成为新一代服务器的标配

目录

CONTENTS



SDService-NSD 网络自服务解决方案

“一张工单走到底”
运维视角，网络变更全流程线上自动化

- 降本：统一变更管理，减少人力
- 增效：实现网络变更自动化，提升运维效率
- 提质：加强变更风险管控能力，减少变更引发事件数，提升运维质量

“一张工单走到底”
业务视角，端到端全流程线上自动化

把运维组从防火墙策略开通的繁琐工作中释放

- 根据安全需求自动定位防火墙、自动生成脚本、自动下发
- 根据安全管理要求做开通前合规校验
- 支持已经存在安全策略查询、治理

“网络数字地图”
可视化运维，助力运维人员提升网络应急效率

整合运维数据打破数据孤岛

- 动态拓扑：全网动态拓扑，逐层导航
- 搜素定位：支持任意设备的定位和关联拓扑展示
- 端到端路径导航：输入源目的IP/端口输出完整路径

网络变更自服务

全流程自动化

网络变更自服务	执行命令自动生成	自动执行
变更场景选择	回退命令自动生成	一键回退
变更前预检查	高危命令检测	变更报告

变更过程监控

- 配置备份
- 监控巡检
- 设备告警信息

变更场景管理

- 编排场景
- 场景启用停用
- 场景审核

变更套餐管理

- 套餐编排
- 套餐启用、停用
- 套餐审核

变更风险分析

- 基于规则风险打分
- 高危命令检测
- 变更关联分析

AI对接

- 自动生成脚本
- RAG知识库
- 防火墙策略分析

访问关系（防火墙）自服务

访问关系（防火墙）全流程自动化

访问关系自服务	脚本自动生成
合规性检测	自动化下发
防火墙自动定位	自动化验证

现网安全策略管理

安全策略采集	冗余策略检查
安全策略解析	隐藏策略检查
安全策略查询	未命中策略

安全策略需求智能分析

禁止开通需求	新建策略
冗余需求	需移动策略
重复需求	修改策略

访问控制管理

- 网络区域管理
- 访问规则管理
- 控制点管理

NAT策略管理

- NAT采集
- NAT查询
- NAT策略下发

端到端路径探测

端到端路径探测

- 内网互访
- 互联出入访
- 外联出入访
- 网络拨测
- 集成资产信息
- 集成变更
- 集成巡检
- 集成应急处置

自动化　网络驱动　采控　N-CMDB　流程　智能分析

1. 网络变更能力：加强变更风险安全管控能力
2. 网络服务化能力：保障网络服务有效性及时效性
3. 网络应急能力：构建网络运维“数字地图“，提升网络数据查询和分析能力

AIGC赋能SDService-NSD 网络自服务解决方案

通过百业灵犀AI助手提问：
OA系统无法访问，需分析其网络系统是否存在问题

通过网络自服务实现网络变更

SDService-NSD 网络自服务解决方案，以业务为视角，以网络服务为中心，参照智能运维国标框架指引，以数字化、自动化及智能化三大核心能力域建设为核心，提供面向业务的全流程、自助式网络服务，主推网络变更自服务和防火墙策略自服务两大场景，来源于金融数据中心6年的智能运维项目实践。

算力 x 联接
释放保险AIGC无限潜能

感 谢 观 看