FY25Q1-CN-ZHENGYANG-RSMP-P
北京正阳恒卓科技有限公司

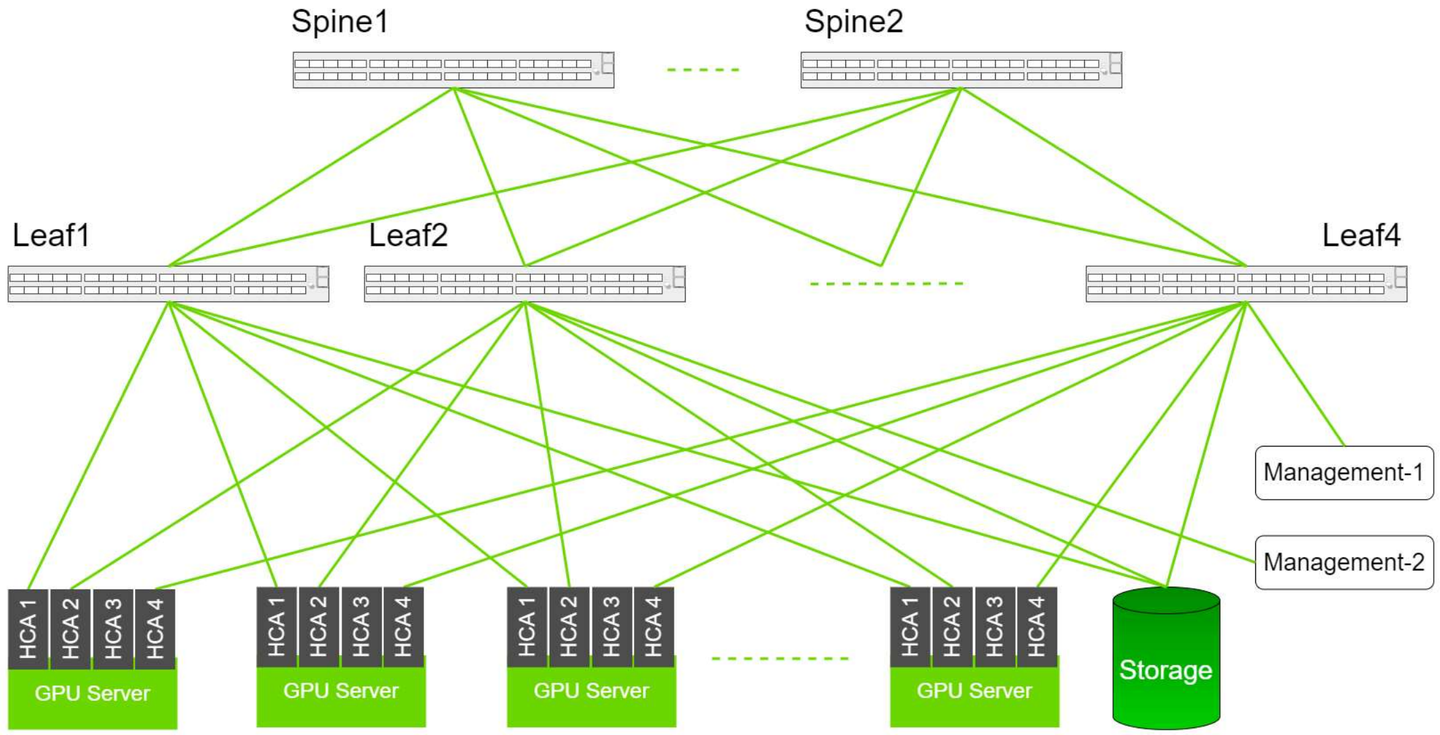
吴昌辉，2024年4月

01

案例01

高校实验室大模型训练平台解决方案

为构建高校小型大模型高速低延迟可扩展训练平台,采用Leaf-Spine架构设计实现了一个高效灵活的Infiniband高速互联解决方案



- 6 台 QM8790
- 12台GPU服务器，每台4张HDR网卡
- 4台存储服务器，每台2张HDR网卡
- Leaf-Spine架构高速互联网络
- 高速、低延迟、可扩展
- Infiniband 4xRails， SHARPv2支持

02 案例02

对外翻译

智能翻译平台需要一款支持60多语种、达到专业翻译质量、可大规模分布式训练的自定义机器翻译系统,以满足不同场景应用和灵活部署的需求。

- 多语种机器翻译能力
- 支持60多种语言,覆盖各行业领域
- 要达到专业翻译质量水平

- 强大的自然语言处理能力
- 分析理解不同语言的语法和语义
- 生成流畅自然的翻译结果

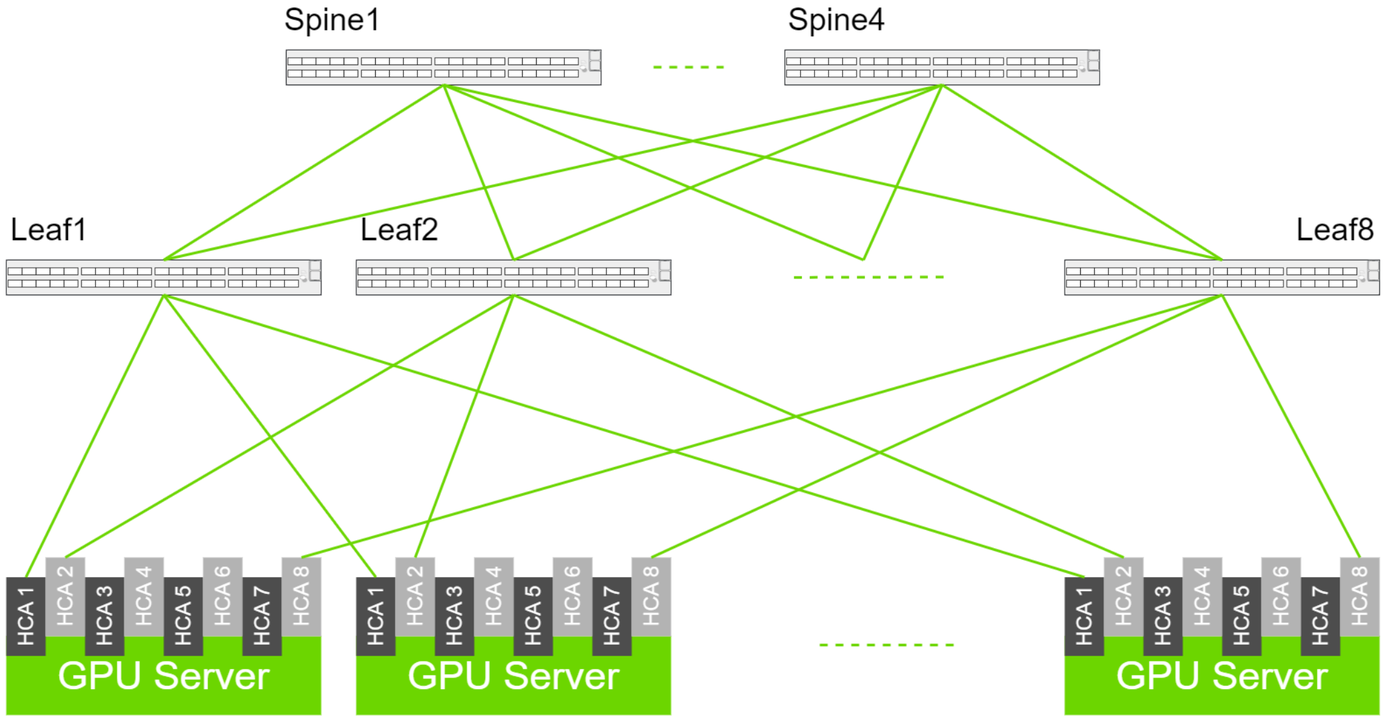
- 多场景应用和部署
- 支持API、云平台及边缘部署
- 整合到各类系统和流程

- 高性能分布式计算
- 处理大规模(数百GB)数据集
- GPU内存256GB起步,需要多机协同

- 定制化解决方案
- 根据行业需求开发自定义模型
- 提供私有化部署和本地化服务

02 Solution Highlight

通过InfiniBand无阻塞Leaf-Spine拓扑、SHARP v2通信库、以及NCCL集合通信库实现高效的多机分布式训练,动态扩展支持超大规模语言模型。

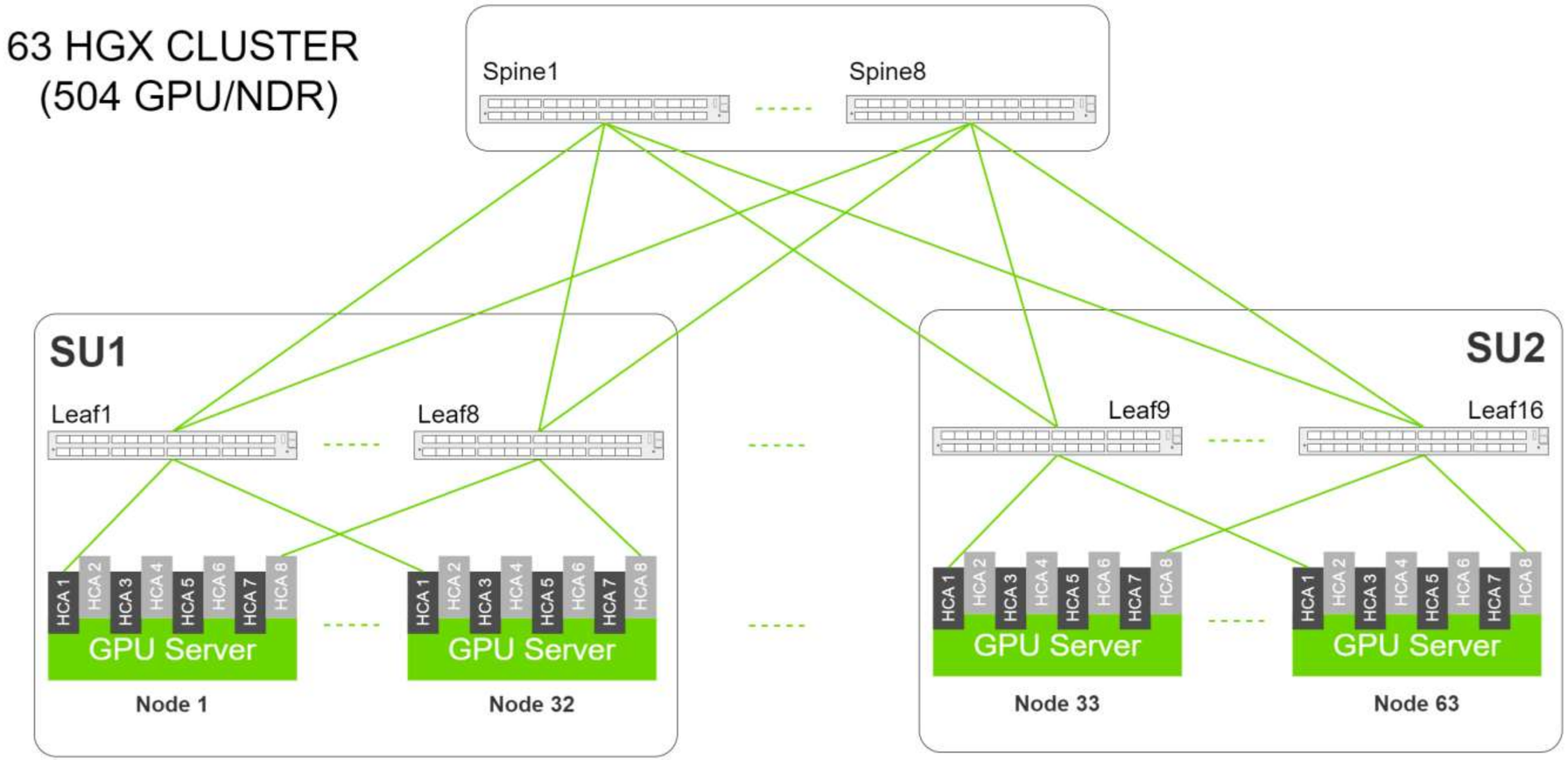


- 2台QM8700 + 10 台 QM8790
- 20 台 GPU服务器，每台8张HDR网卡
- Leaf-Spine架构高速互联网络
- 高速、低延迟、可扩展
- Infiniband 8xRails， SHARPv2支持

03 案例03

人工智能算力中心

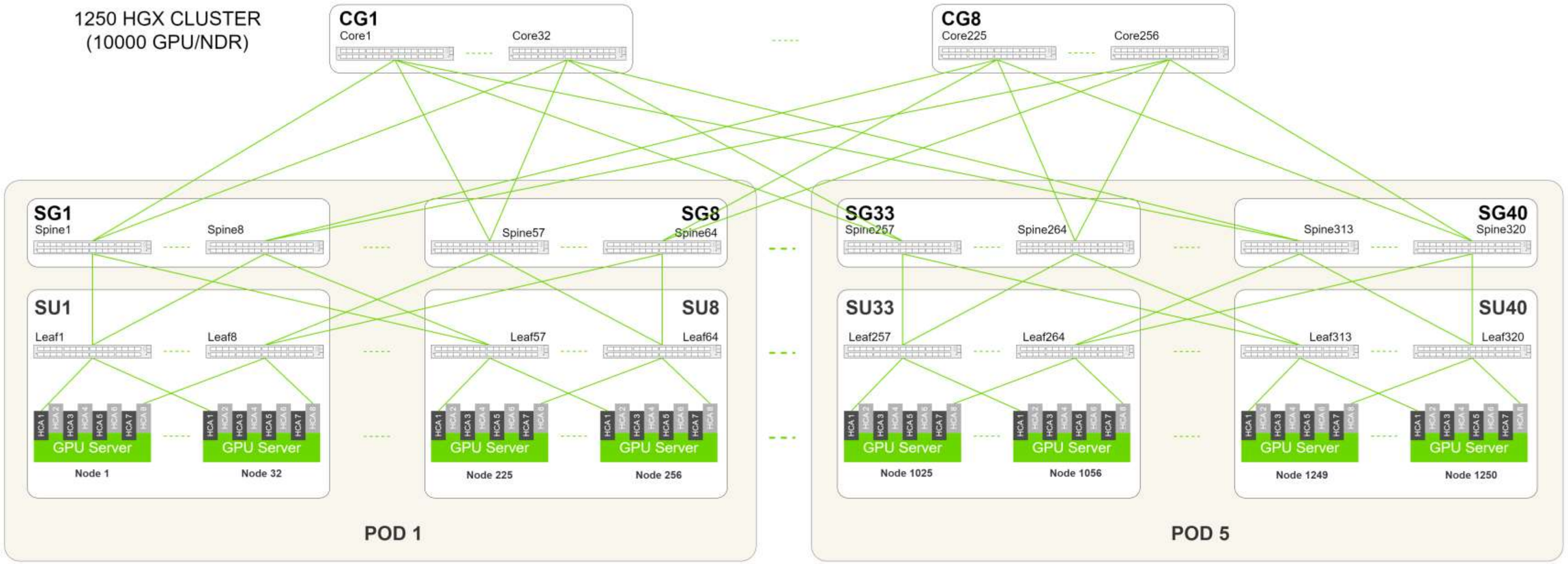
通过部署InfiniBand网络，我们建立了一个具有1000PFlops@FP16计算能力的智算中心，实现了高吞吐量与低延迟的数据处理，为AI研发提供了强大的硬件支持和高效的管理服务。



04 Solution Highlight

超大规模大模型训练平台

构建基于400Gbps Infiniband高速互联网络和集成GPU服务器集群,实现EB级的分布式AI模型训练能力。



Thanks!