

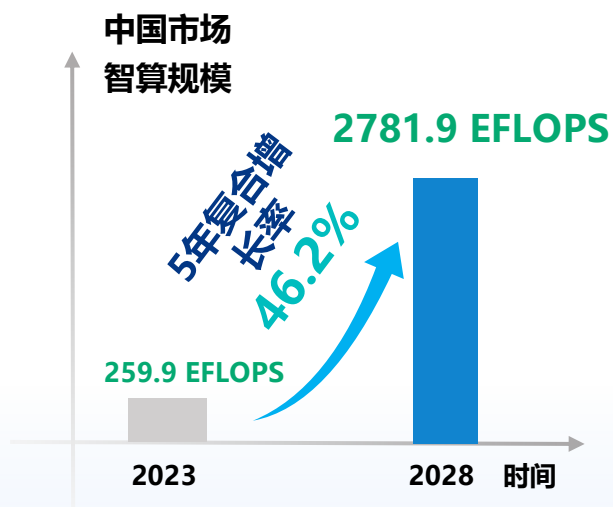
AI时代，IT基础设施如何 满足算力爆炸需求

超云数字技术集团有限公司

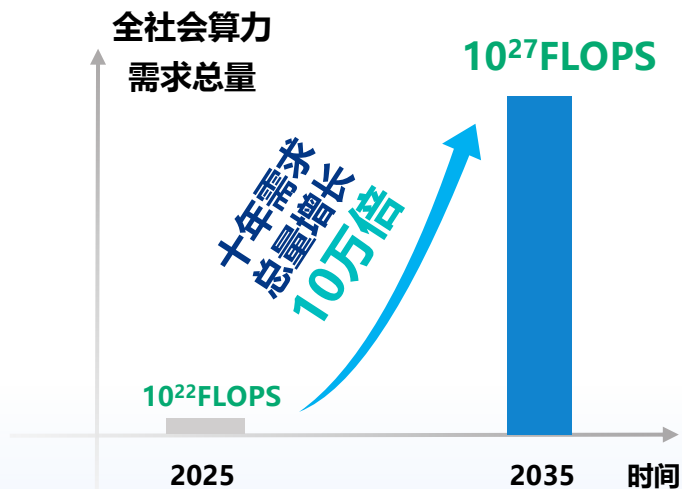
AI需求的爆发，推动算力需求爆炸增长

人工智能技术迅猛发展，随着智能化的普及和渗透，多元算力成为社会运行的基础设施

训练与推理需求爆发，带动AI算力需求持续高速增长



源引: IDC 《2025中国人工智能AI算力评估报》



源引: 华为 《智能世界2035》

1、大模型持续朝万亿、多模态和长序列演进

- 深度上，从千亿到超万亿参数
- 广度上，从单模态向多模态
- 长度上，从短序列到长序列

2、大模型训练对智能算力提出新的需求

主流路线是通过“**大参数+大算力+大数据**”探索模型能力上限；

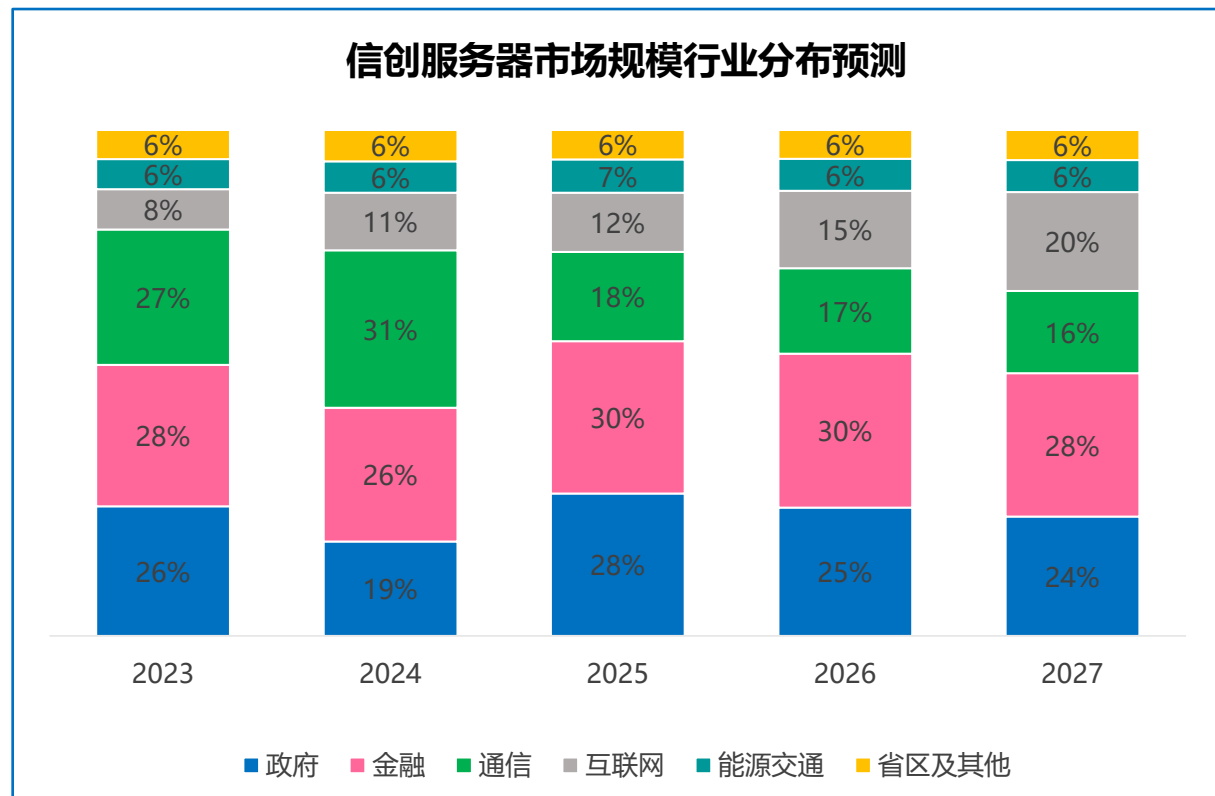
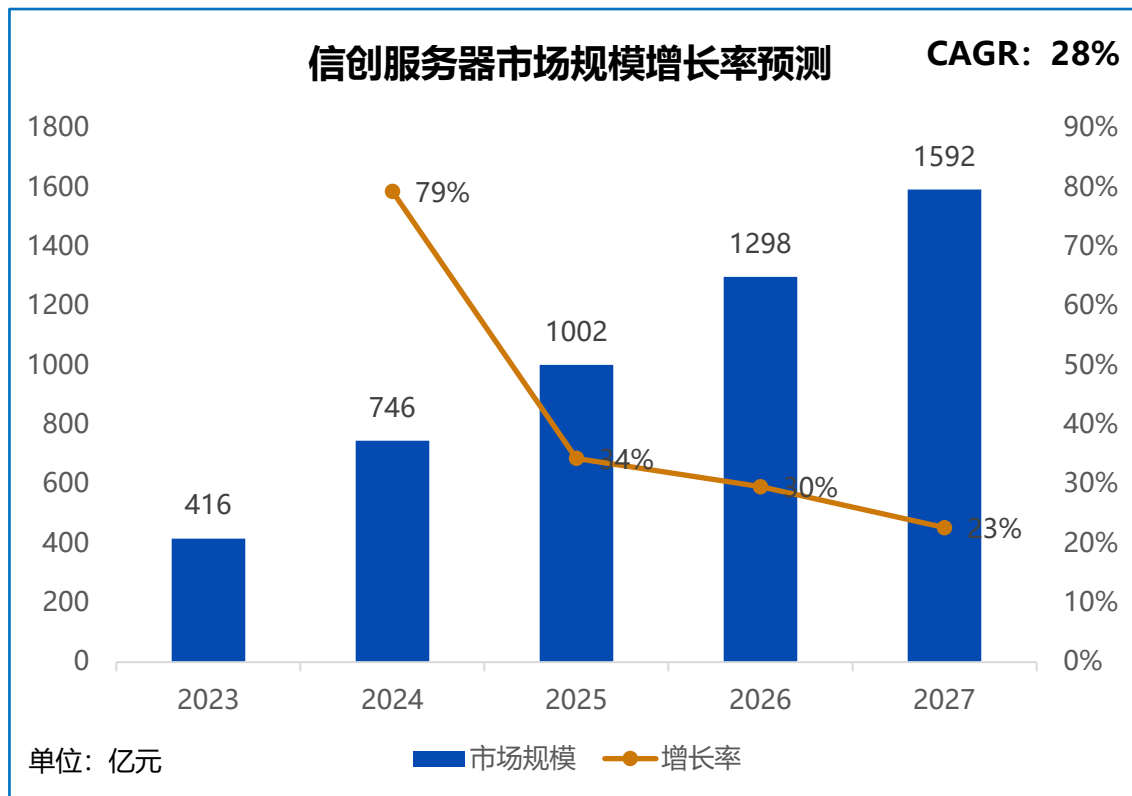
新兴路线是通过“**模型算法+算网设施**”的联合优化探索模型成本下限，两条路径并行探索

3、推理业务持续爆发，增加端侧算力需求

推理网络将呈现“集中+广分布”特点，**P、D分离**，大小模型**云边协同**等新架构不断涌现

金融领域市场预测

- 2025年信创服务器整体市场规模已达1000亿，信创服务器市场复合增长率达28%，到2027年市场规模达到1500亿以上。
- 市场聚焦从党政扩展到行业市场，金融、通信、能源交通等行业。



AIDC智算中心面临五大挑战

芯片墙

制程和性能
供应和产能

功耗墙

单芯片功耗
数据中心算力密度

存储墙

万亿模型储存需求
传输低于计算速度

网络墙

节点通讯影响集群效率
并行计算无损要求

生态墙

生态垄断
兼容性适配



多样化算力布局，突破芯片墙禁锢

加速

训推一体化全栈覆盖，多芯片适配，AI 开箱即用

私有云

软硬件全栈自研，覆盖存储硬件，超融合，云管平台等，一站式交付

通用

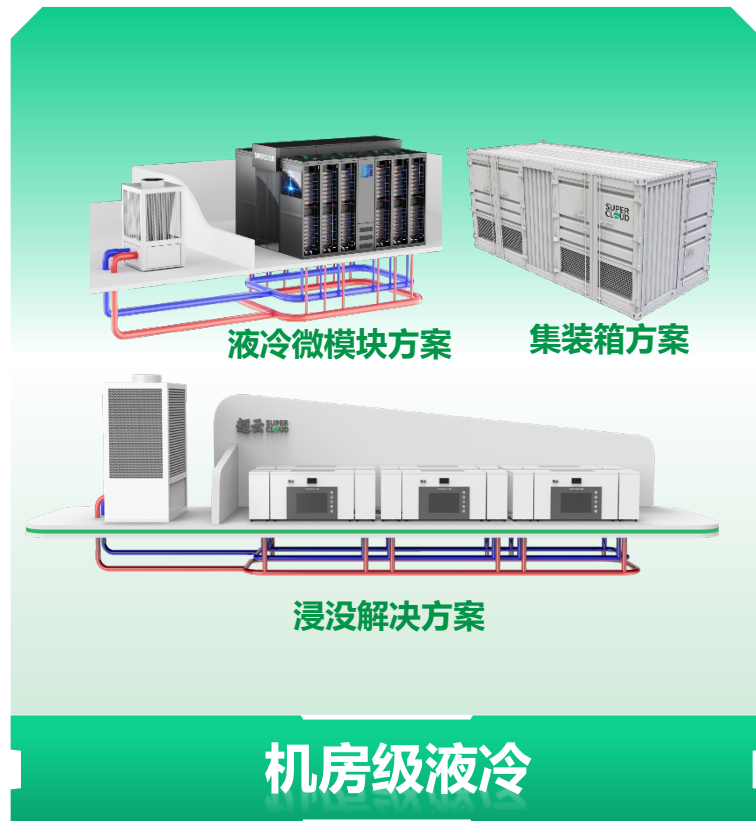
20+机型全场景覆盖，成熟稳定，运维成本低

信创

全信创路线覆盖，高国产化率，全栈合规适配

方案支撑体系：提供从方案咨询、规划设计到交付运维，全生命周期一站式算力服务

全链条液冷产品矩阵



性能

提升
20%

故障率

降低
15%

噪声

降低
47%

机柜密度

提升
13倍

节约低碳

节省
25%

AI存力分层，计算与传输的极致平衡



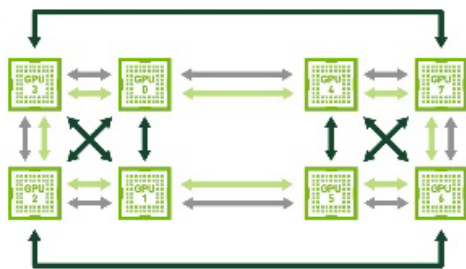
NVIDIA Spectrum-X Technology for AI Storage
— SuperCloud
November 2025

Next-Generation Distributed Storage Architecture:
Low-Latency and Network Optimization for AI Workloads

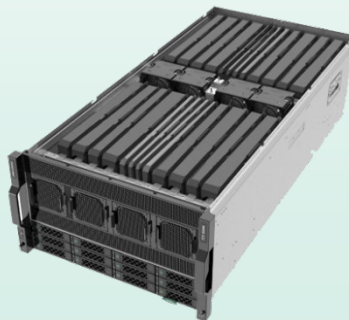
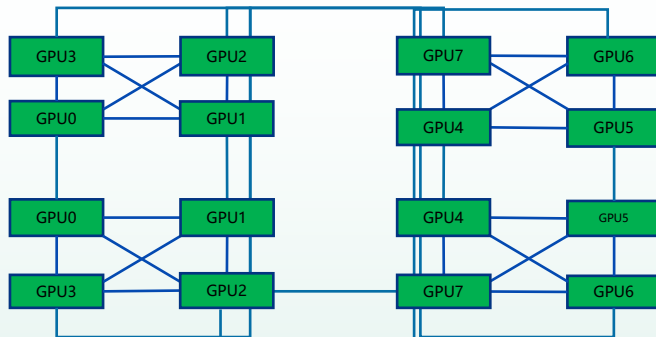
G1	G2	G3	G3.5	G4
GPU HBM	System Memory	Local Storage	NVIDIA CMX Context Memory	High-Performance Storage
Active KV Cache Only	Spillover Tier	Warm KV Reuse	Warm KV Reuse	KV Reuse
Extremely Fast	Very Fast	Limited capacity	Power Efficiency	Massive shared capacity at lowest cost
Limited capacity	Limited capacity	Difficult to share	Scalable capacity	Optimized for durability
Difficult to share	Difficult to share	Endurance limits	Optimized sharing at scale	Bandwidth and capacity scaling increases cost and power
		Serviceability	Ease of serviceability	Suboptimal for large-scale inference
			High throughput	
			Lower TCO	
			No data durability	
Compute Node - Local			SU Scaled	Data Center Shared

全栈网络互联方案，高效并行计算

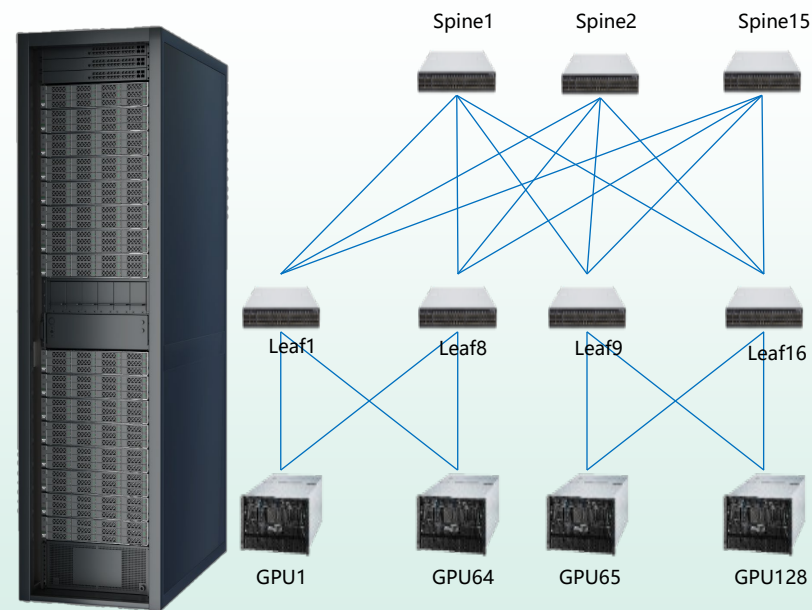
片间互联



卡间互联



集群互联



芯片--数据中心互联

极致的网络要求

无损网络、并行计算

超云生态，全面落地



海光信息

行业优势：深耕金融、政务领域，生态适配成熟
产品优势：性能矩阵完善，支持全算力场景覆盖
核心产品：BW1100 系列, BW1500 系列



摩尔线程
MOORE THREADS

摩尔线程

行业优势：北京政企行业，生态建设全面
产品优势：全栈支持推演渲染，视频处理能力领先
核心产品：S5000 高性能计算模组



天数智芯

行业优势：聚焦二线互联网及特种行业应用
产品优势：性能领先，支持超大规模训练
核心产品：天垓150S 加速卡, 智凯100 推理卡



昆仑芯
KUNLUNXIN

昆仑芯

行业优势：背靠百度生态，互联网应用数据优势
产品优势：软硬一体优化，大模型推理性能优异
核心产品：P800 模组 & PCIE 独立加速卡



沐曦

行业优势：专注特种行业，通用GPU架构
产品优势：高能效高性价比，降本增效
核心产品：CS500 模组, C500 PCIE 独立卡



XIANGDIXIAN
COMPUTING TECHNOLOGY
象帝先计算技术

象帝先

行业优势：专注图形渲染、视频解码与处理
产品优势：国内图形性能领先，生态适配完善
核心产品：高性能视频渲染类加速卡系列

超云海光产品组合

海光4号：

通用计算

74-4U4S



R3426

74-2U2S



R3226

54-2U2S



R3225

34-2U1S



R3221

34-单路工作站



HS222

存储服务器

74-4U2S



R3425

54-4U60盘



R3425

加速计算

74-2U2卡



R3226

74-4U8卡



R8428

74-8U80AM



R8868

海光3号：

通用计算

73-2U2S



R3216

53-2U2S



R3215

边缘服务器

53-2U2S



R3215 (边缘)

单路工作站



HS212

存储服务器

73-4U2S



R3416

加速计算

73-2U2卡



R3216 (GP236)

海光加速计算产品

2U8卡推理服务器



R3216

2U8单宽GPU

2U2双宽GPU

推理密集型

4U8卡训练服务器



R8428

4U8双宽GPU

支持PCIE 5.0

训练密集型

8U8 OAM训练服务器



R8868

8U8 OAM模组

支持国产模组

训练密集型

塔式渲染工作站



HS311

塔式2卡

支持2张双宽卡

图形渲染

海光通用计算产品

7400-2U2S



R3226

5400-2U2S



R3225

7300-2U2S



R3216

5300-2U2S



R3215

3490-2U1S



R3221

全面自研海光通用计算，从高端到入门，总有一款适合你

海光4号

7000

高性能高扩展

海光4号

5400

高性价比

海光3号、2号

7000&5000

明星旗舰机型

海光3号、2号

7000&5000

高性价比

海光3490*1

高扩展性

入门级

金融行业AI建设：三点核心体会

01

算力底座：AI创新与自主可控并重

金融算力底座要兼顾AI创新和自主可控，不能为追求AI性能而放弃国产化路线。

02

存算网冷：一体化整体规划设计

不能只盯着GPU算力，存储、网络、散热必须整体规划，存算网冷一体化设计，否则昂贵的算力硬件难以达到预期效果。

03

稳定优先：适配与全生命周期服务

金融业务对稳定性要求极高，软硬件提前适配、全生命周期服务，比单纯硬件参数更加重要。

成立于2010年，是由中国电子（CEC）、云基地（CLOUD VALLEY）联合注资成立的国家高新技术企业，国家级专精特新“小巨人”企业。超云作为专注于数据中心产品及服务的本土制造商，始终致力于构建自主可控的算力基础设施体系。

- 中国服务器市场第一梯队
- 连续三年服务器市场增速排名第一

NO.1

连续三年
X86服务器增速

TOP5

液冷服务器
总体排名

TOP7

国内服务器品牌
总体排名 

* 数据来源IDC市场研究报告

TOP5

信创整机
总体排名

NO.1

连续三年
推理业务总体增速

超云 SUPER
CLOUD
让数据中心更简单

让AI更简单

超云 SUPER
CLOUD
让数据中心更简单

