

超节点和关键应用场景

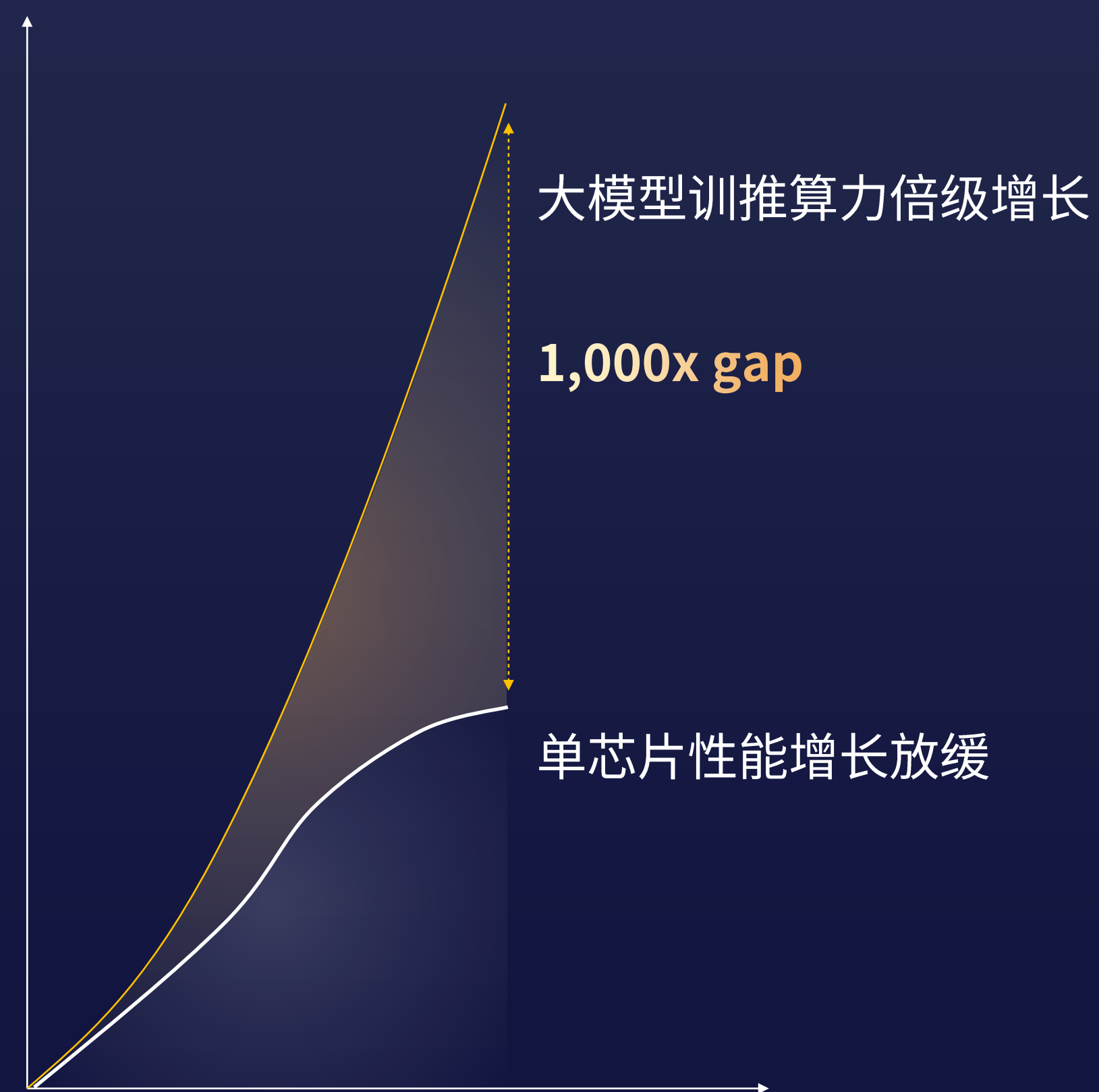
张健 openEuler社区Contributor



智算和通算技术与应用快速迭代演进，算力有效性和时延要求不断提升

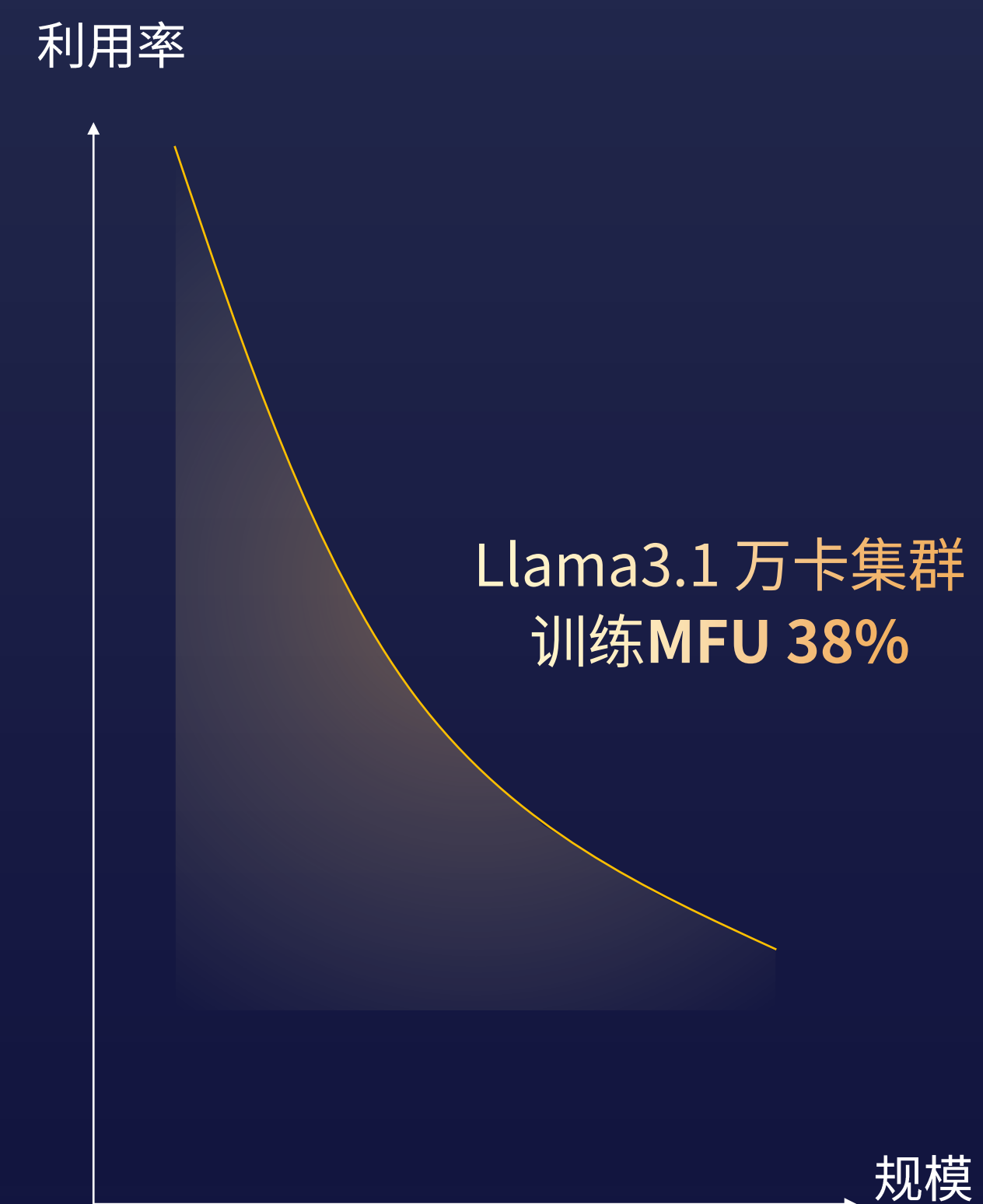
单芯片无法满足 日益增长的算力需求

受限于半导体工艺，单芯片性能和业务对算力需求差距大

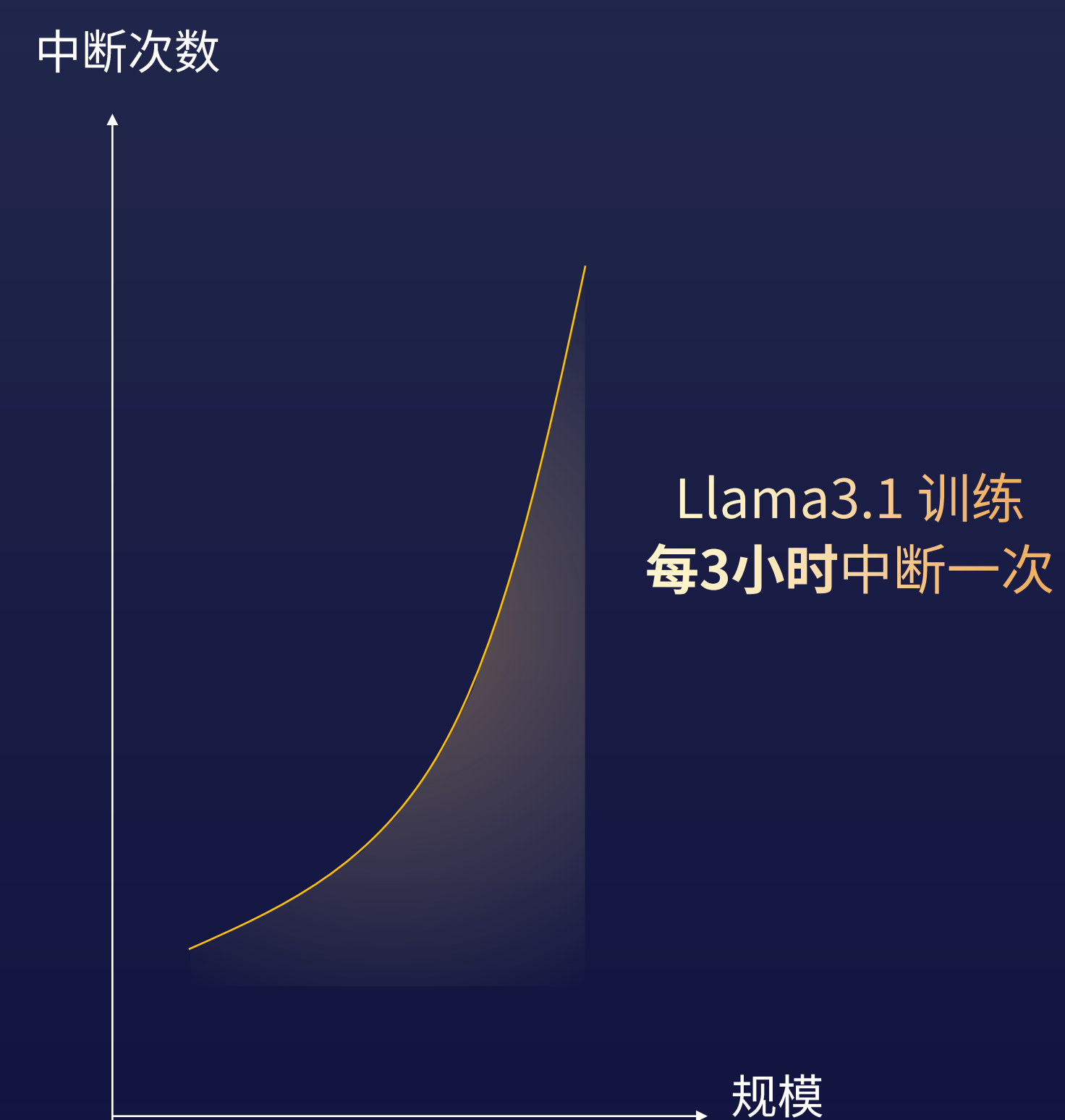


面向智算，服务器堆叠 不能满足规模的算力供给

规模越大，算力利用率越低



规模越大，训练中断越频繁



面向通算，资源碎片化等问题 无法有效解决

资源碎片化
利用率低

计算、存储等资源
隔离在不同节点

数据处理
无效搬移

20%算力用于数据搬运，
协议栈处理等低价值任务

跨节点
通信效率低

传统RPC、Socket 等机制跨节点
交互路径长

需要新的
计算架构

开创超节点计算架构，释放集群规模算力的潜能

超节点架构

物理上由多台机器组成，逻辑上以一台机器学习、思考、推理



资源池化

计算、内存、互联和存储等均可池化
全局组件平等直达支持任意多对多协同

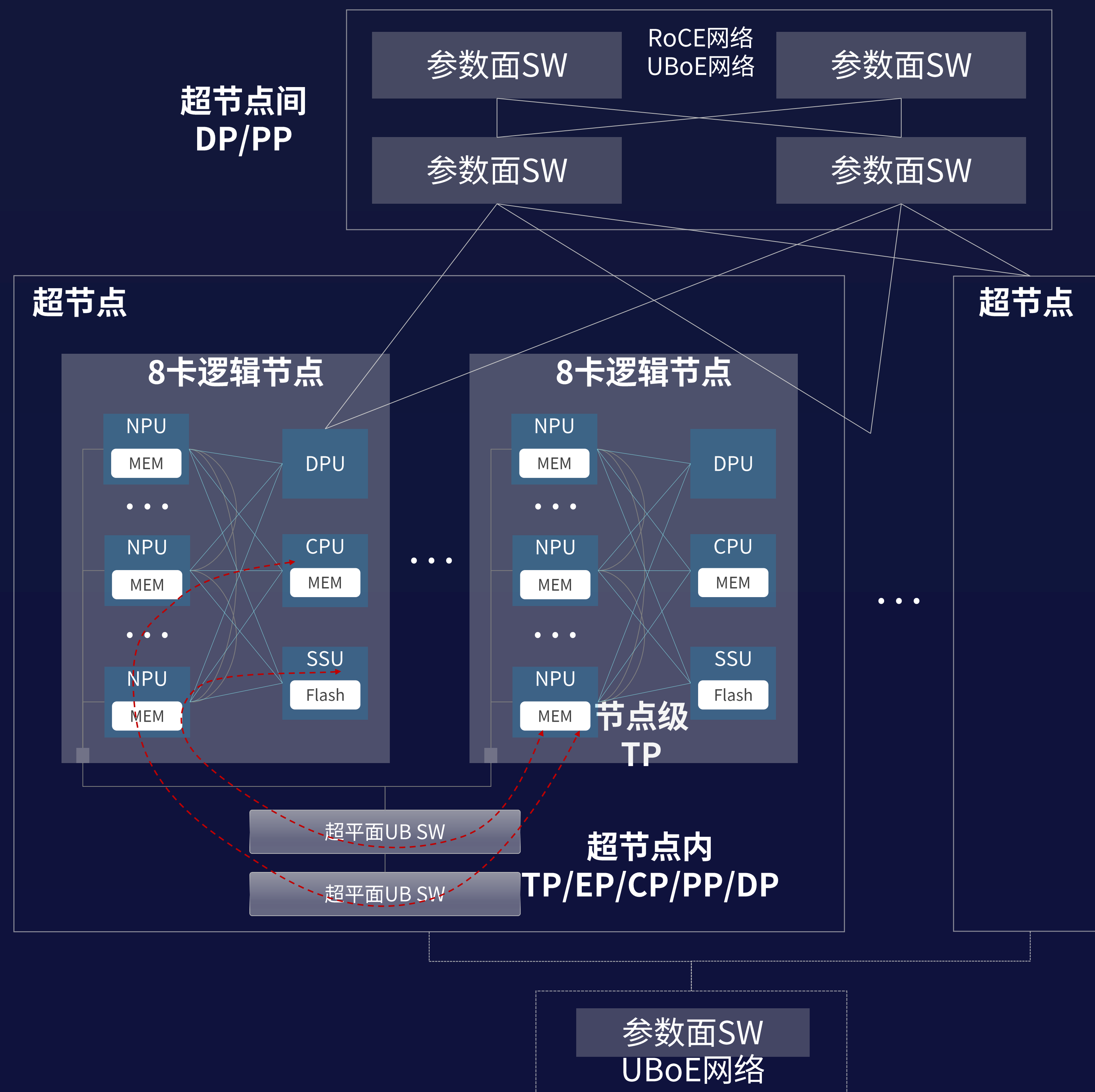
规模扩展

按需灵活扩展，高利用率
大范围、整系统相同编址机制
协议归一，无额外转换开销
控制静态、传输、动态时延

长稳可靠

10倍提升长稳运行时间
自动避障，自动恢复
易部署，机房环境适应性强
易扩展，从模组到柜级均可扩展

智算超节点系统示例：大规模、高性能、多样拓扑，支持多样算力资源池化



大模型训练/推理典型场景

- **大规模超节点**：通过UB交换、光电互连能力，超节点域可支持64~8k卡规模，可容纳各类并行流量的高性能通信；
- **负载亲和的拓扑**：支持nD-Fullmesh、Clos、OCS等不同拓扑融合部署，为不同并行流量提供高效通信；
- **总线级高性能**：支持同步ld/st低时延通信，以及异步read/write大带宽通信，TP/EP/CP等流量可使用最佳通信模式；
- **灵活参数面组网**：可以通过DPU支持RoCE组网，也可以基于UB交换支持融合的UBoE组网，有效提升性价比；

超长序列、Agentic AI、多算力协同等新场景

- **CPU池化互连**：CPU支持大带宽UB，实现紧耦合多样算力池，支持CPU和本地/远端NPU高效通信、CPU和NPU灵活配比；实现高性能多样算力部署、分级内存池；
- **SSU池化互连**：支持将Flash通过UB接入超节点，支持SSU和CPU/NPU高效对等通信，实现大容量、高带宽、高吞吐的分级KVcache、向量池化能力；

LLM训练场景挑战：模型规模增大，并行规模远超节点内高带宽域，不可掩盖通信占比高

负载：小专家成为主流，大并行域大量通信无法掩盖

集群规模（卡数） = DP * TP * PP * EP （并行域）

随着集群规模增大，DP域无法持续扩大

$GBS \text{ (固定值)} = MBS \text{ (固定值)} * Num_MB * DP\text{域}$

$Bubble = (PP-1) / Num_MB$

DP扩大会导致计算Bubble变大，模型收敛挑战

GBS: Global Batch Size; MBS: Micro Batch Size

小专家模型规模仍在增长，PP/EP扩大

目前EP和PP并行>8，并行范围已经超出了传统的8卡服务器的高带宽域

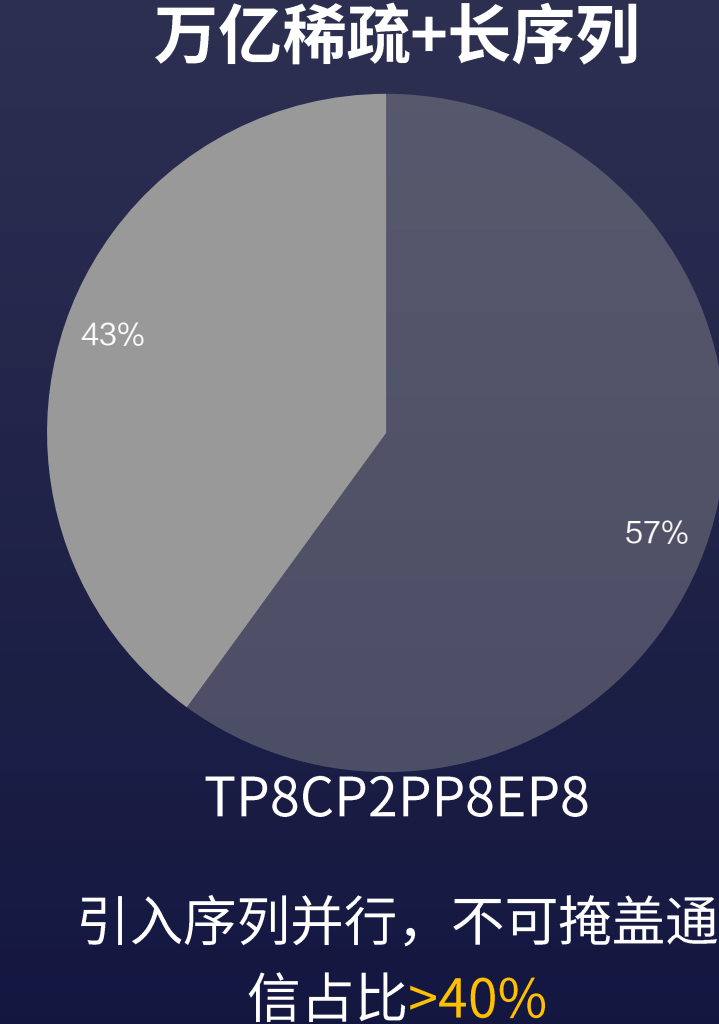
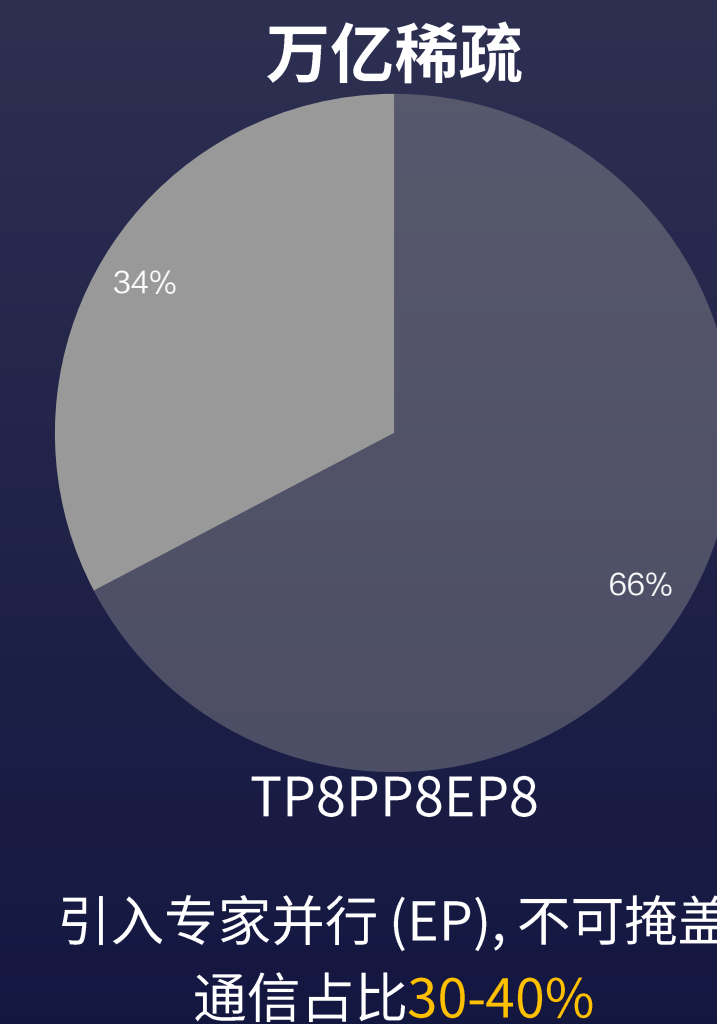
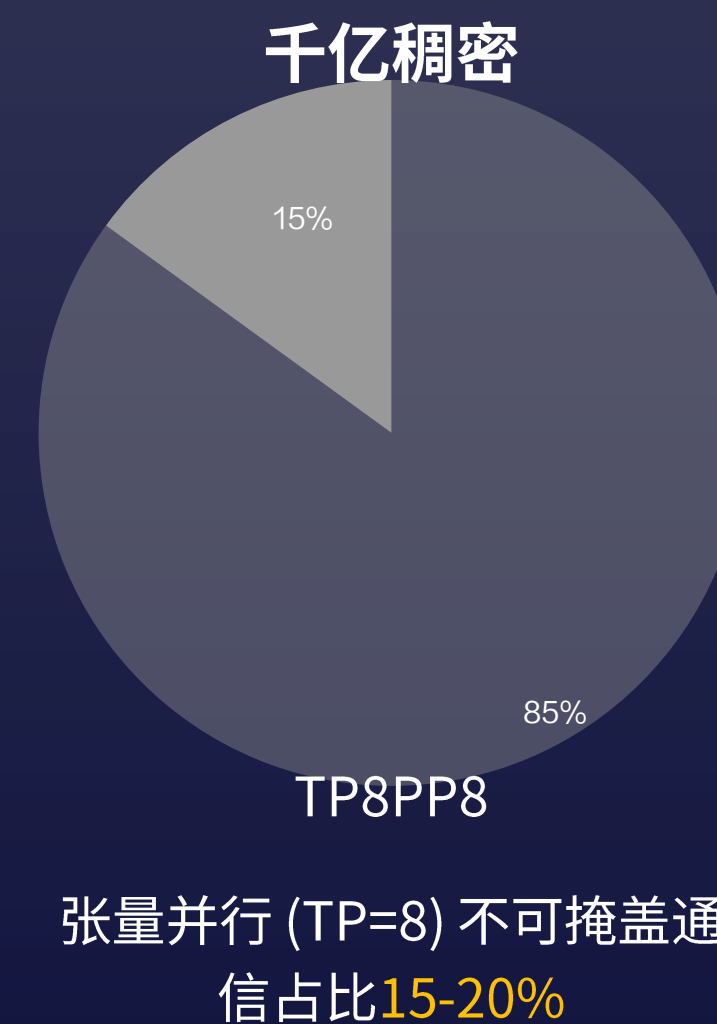
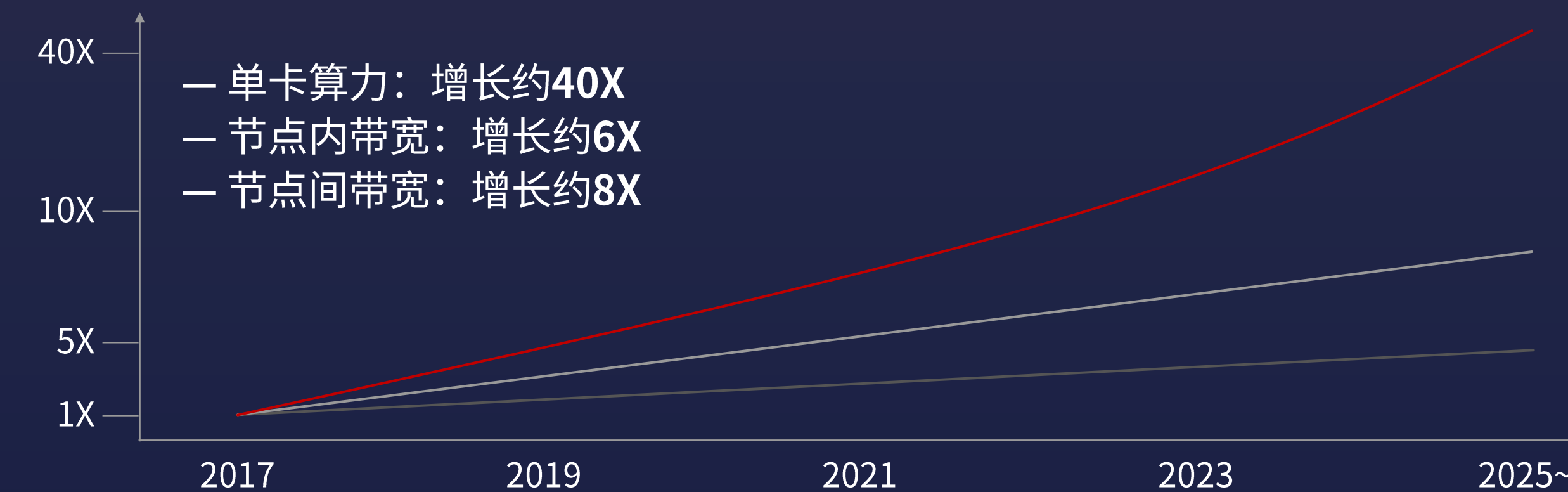
传统集群涉及大量跨节点网络通信

TP、PP及EP涉及大量集合通信，且难以被完全掩盖

并行模式	单卡通信量	通信不可掩盖性程度
数据并行 DP	GB级	低，可大部分掩盖
流水并行 PP	GB级	中，可大部分掩盖
张量并行 TP	百GB级	高，通常不可掩盖
专家并行 EP	百GB级	高，通常不可掩盖

挑战：互联带宽的增长远不及算力，训练系统面临“木桶效应”

节点间通信带宽8年仅增长4倍，远低于算力增长40倍



■ 计算 ■ 通信（不可掩盖）

*数据来自服务器集群性能仿真

LLM训练场景解决方案：

大规模/大带宽/低时延超节点互联，降低通信占比，提供更大超参寻优空间

大带宽/低时延互联 → 降低通信占比
大规模超节点互联 → 更大超参寻优空间

64~1k卡超节点：可完整覆盖TP、EP、CP寻优空间
>1K卡超节点：应对模型不确定性、优化资源调度效率

超节点应用TP*CP*EP

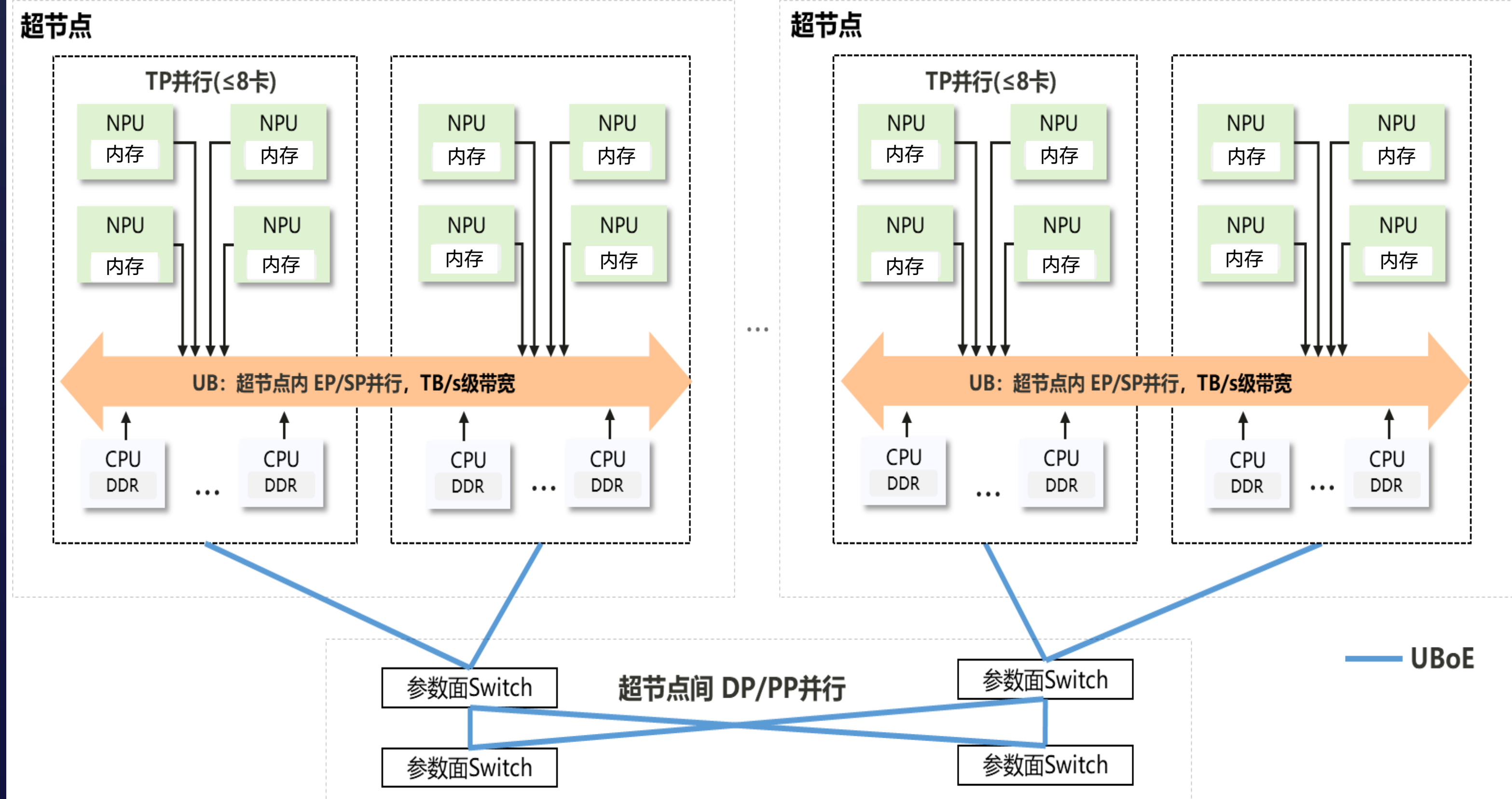
nD FM+超节点
提供大带宽TP

超节点内EP
大带宽、分层EP

超节点内CP
优化CP性能

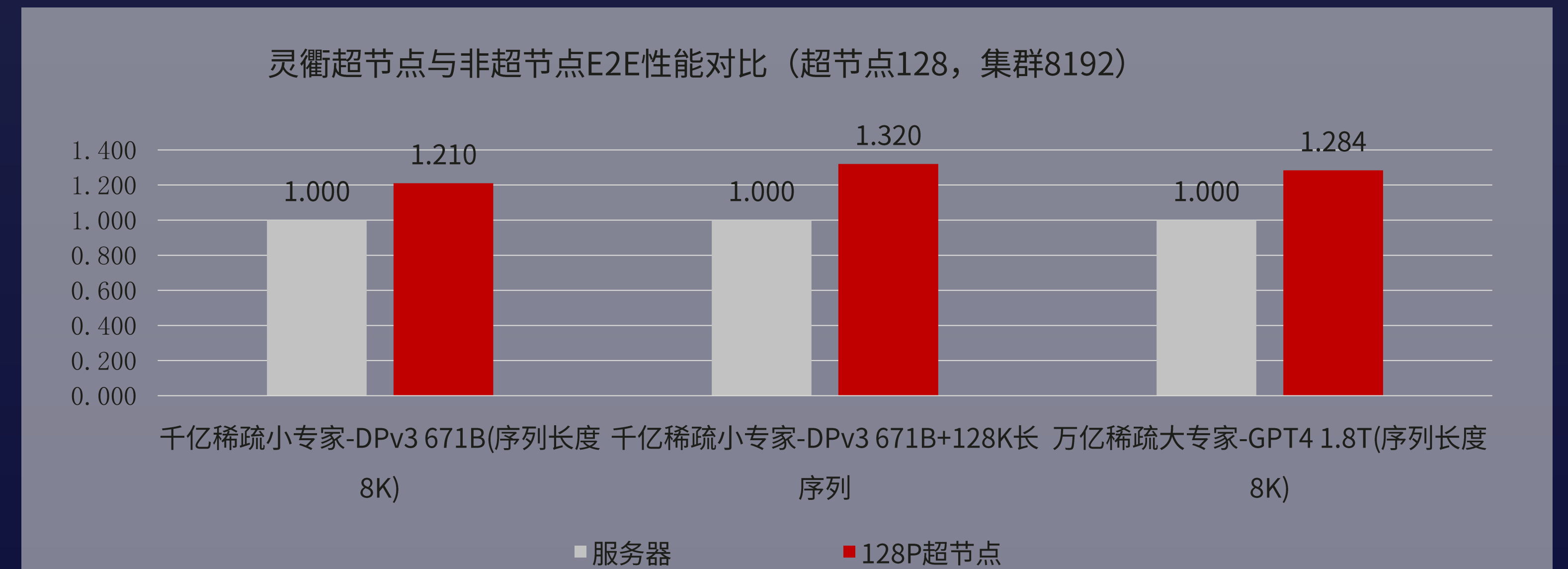
超节点分层DP
降低大网DP数据量

超节点增强TP/CP
降低PP域



端到端性能典型收益：20%+
参数规模、序列长度、专家域持续提升，收益越高

超节点与非超节点单步通信时间占比实测数据（超节点规模128，集群规模256）

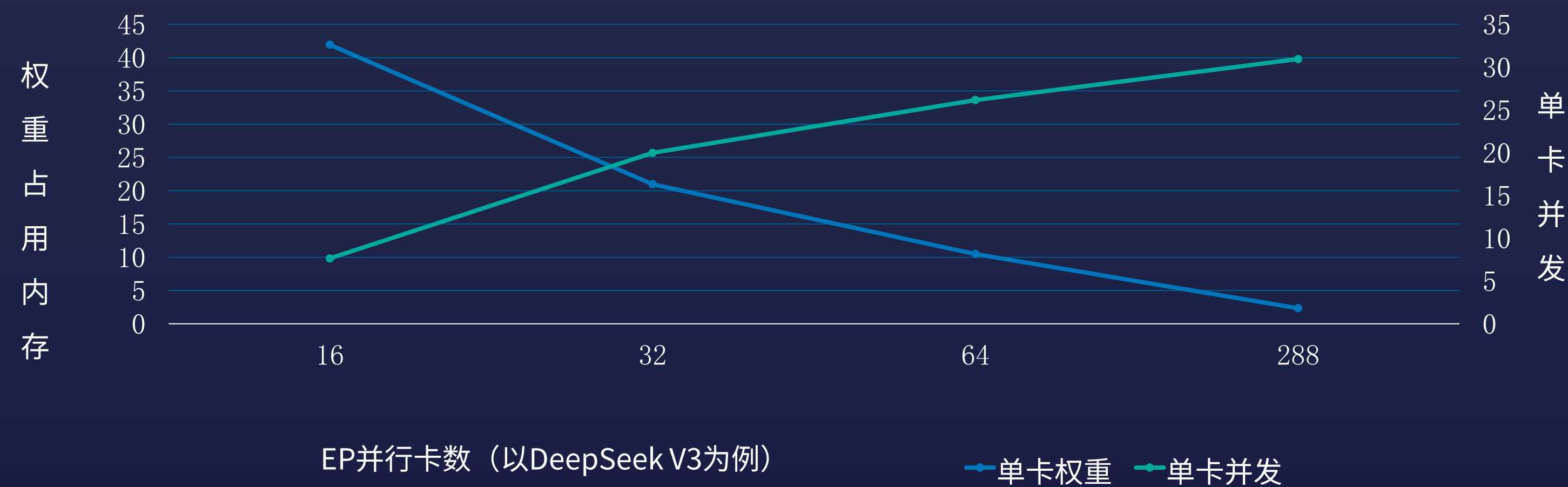


智算推理场景挑战：大模型稀疏化、长序列趋势明显，All2all通信、内存容量成为推理性能关键瓶颈

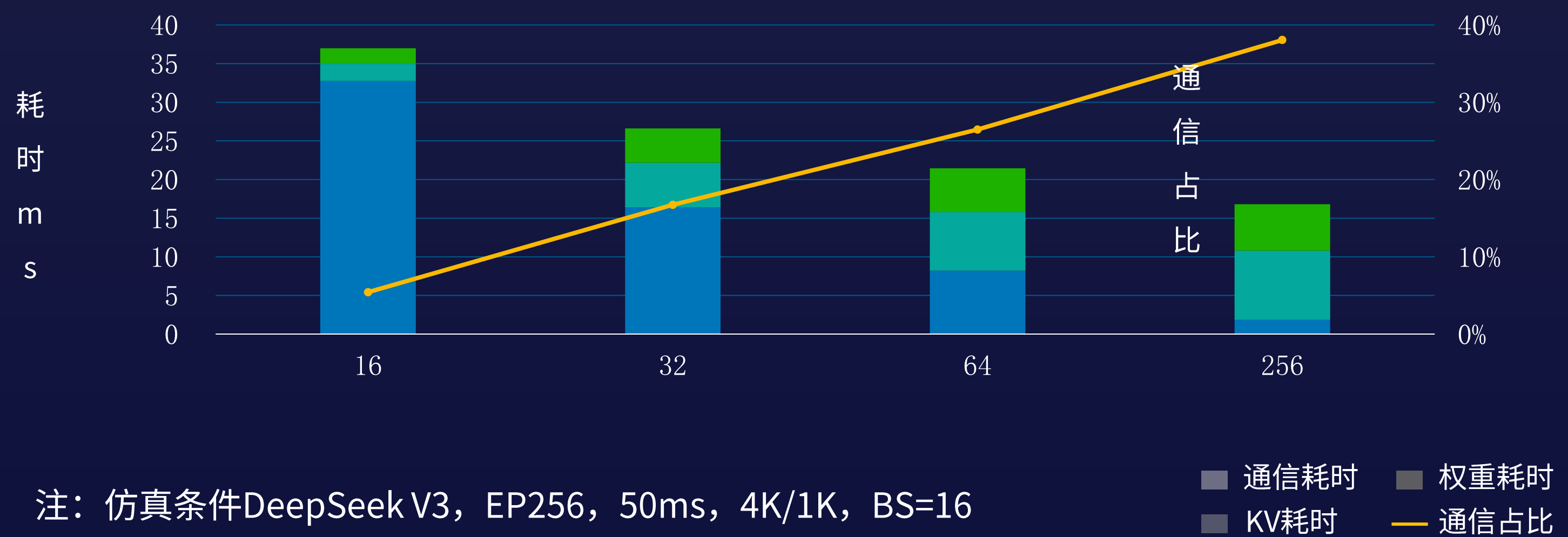
MoE稀疏化、大EP带来通信性能挑战

大EP并行成为MoE大模型推理加速关键技术

EP并行越大，单卡权重越少，单卡可支持并发越大



仿真测算，256卡EP并行，all2all通信耗时占比30%+

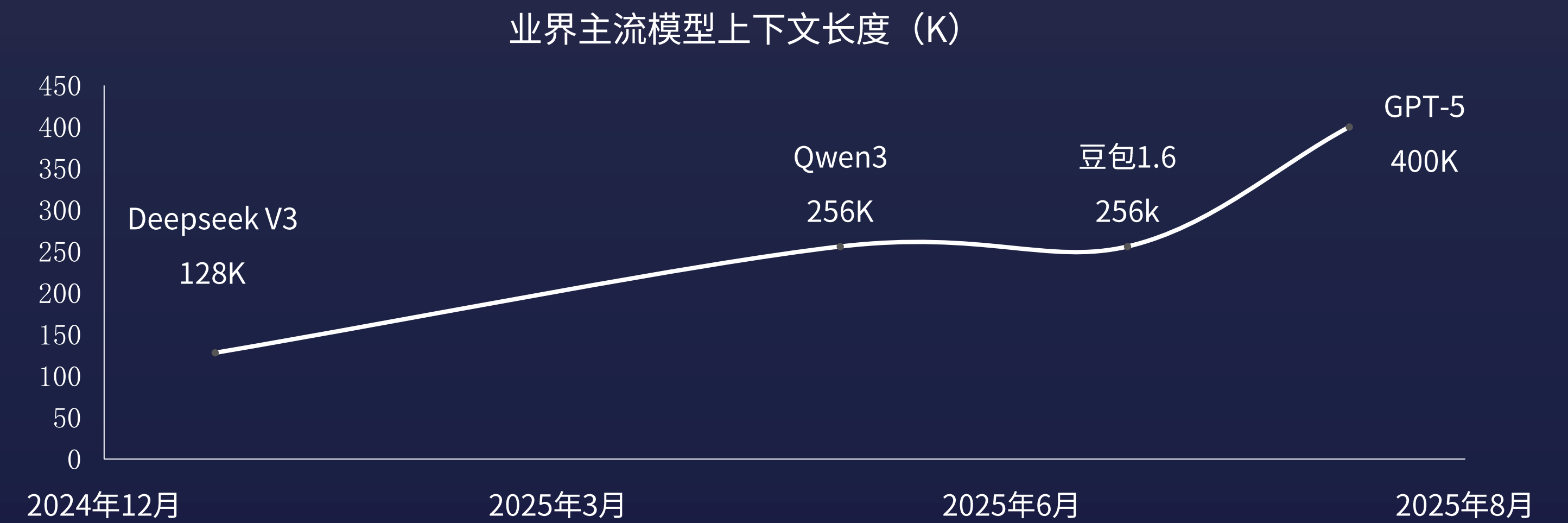


注：仿真条件DeepSeek V3, EP256, 50ms, 4K/1K, BS=16

■ 通信耗时 ■ 权重耗时
■ KV耗时 — 通信占比

长序列推理，带来高性能内存容量挑战

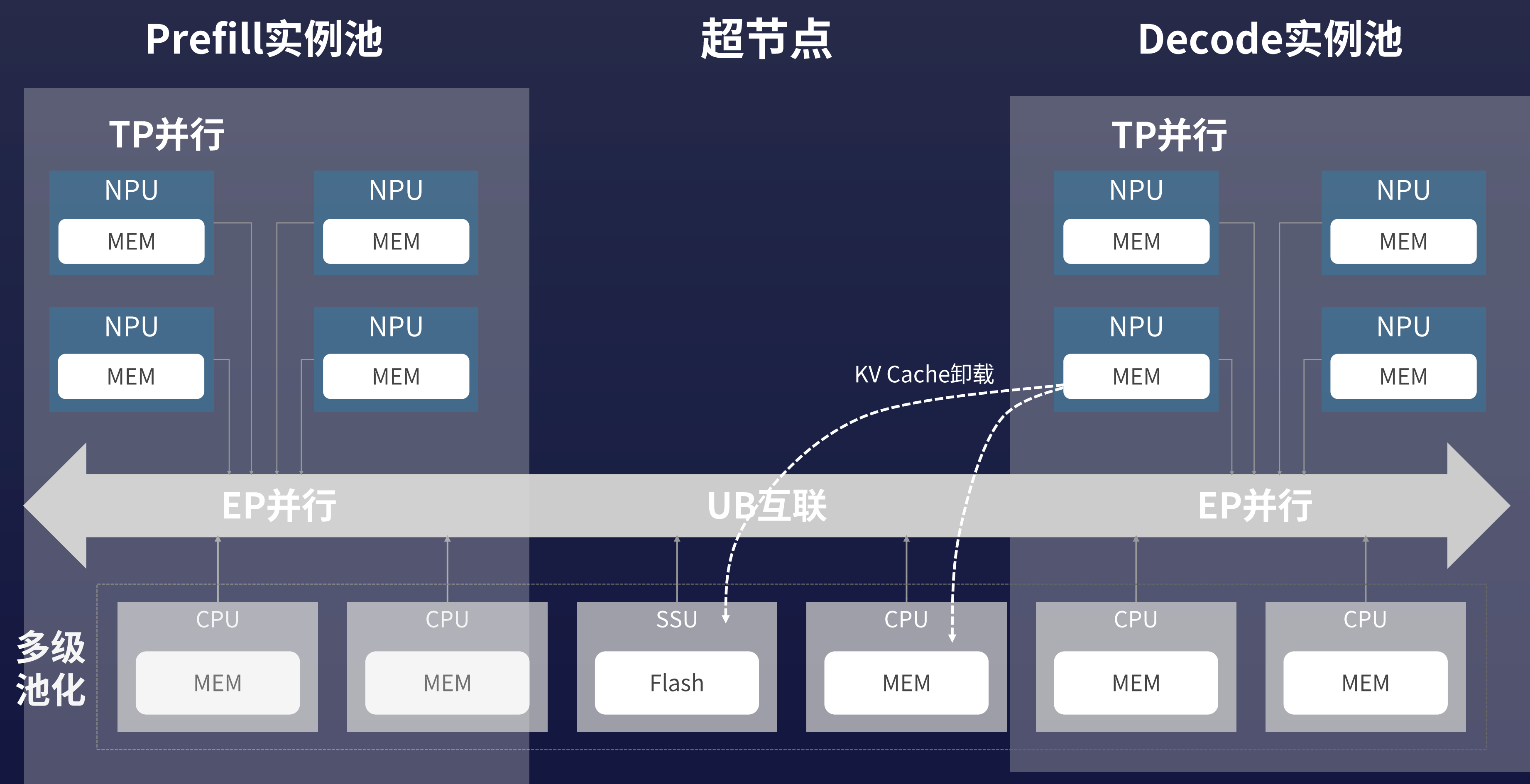
主流模型支持序列长度逐渐增加，长序列成关键场景



- **联网搜索+RAG**：输入序列长度大幅增加；根据KIMI开源数据分析，当前5%用户请求的序列长度超过49K；
- **行业生产系统应用**：多图、长文档、视频类分析，长序列是确定性需求，预计长度可达~M级；
- **稀疏化注意力**：业界开始采用稀疏化ATTN，大幅降低长序列的计算量，推动商用部署；但仍需要缓存全量KV cache；
- **片上内存/DDR容量瓶颈**：以DeepSeek V3 64~128K序列为例，每NPU需要~400GB容量+百GB/s带宽，节点内片上内存/DDR容量不足，需考虑扩展Memory或Flash；

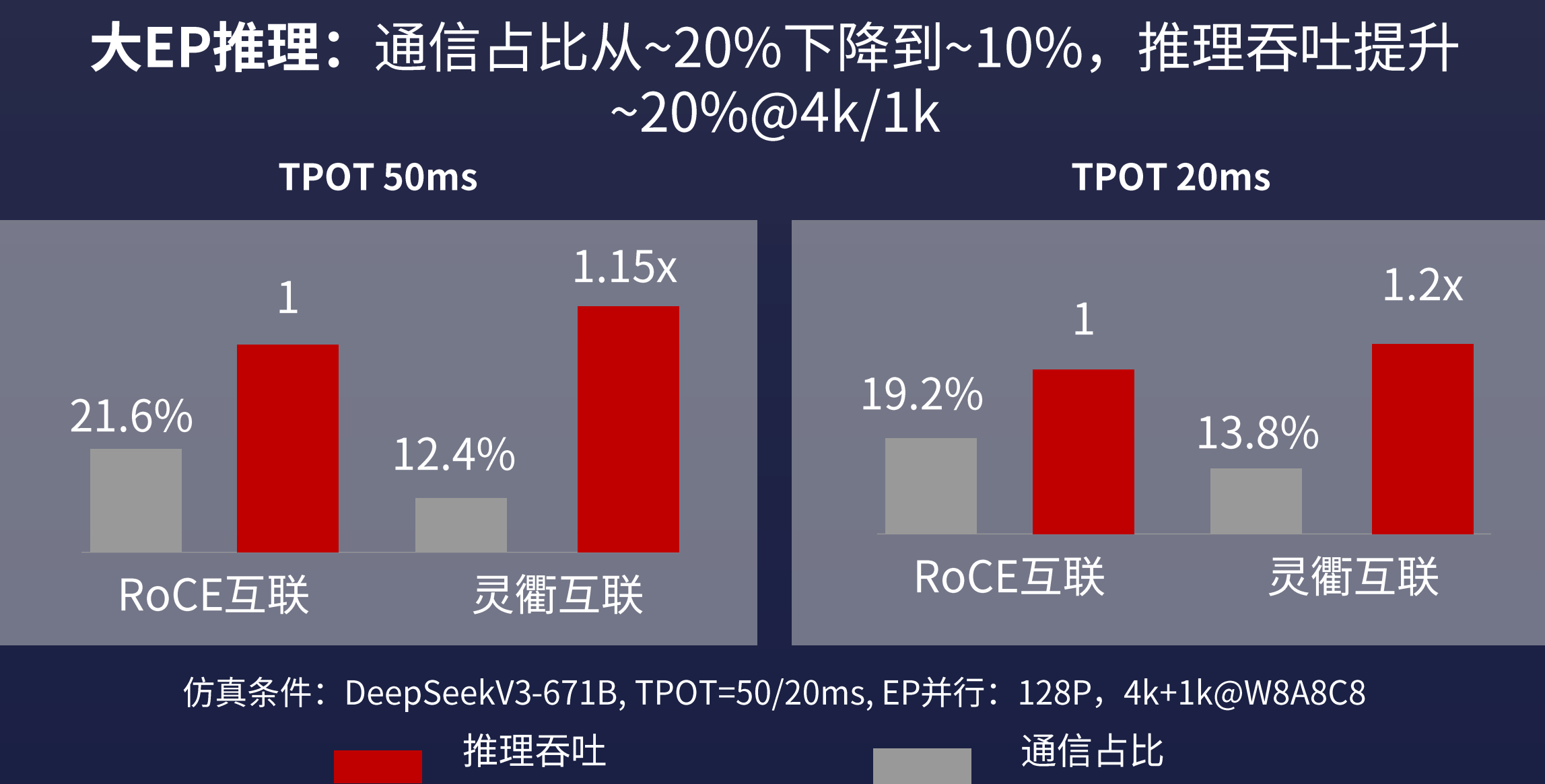
智算推理场景解决方案：降低大EP推理通信时延，提升长序列批次大小，整体提升20%+

方案：UB互联提升EP间通信性能，分级KV Cache支持长序列

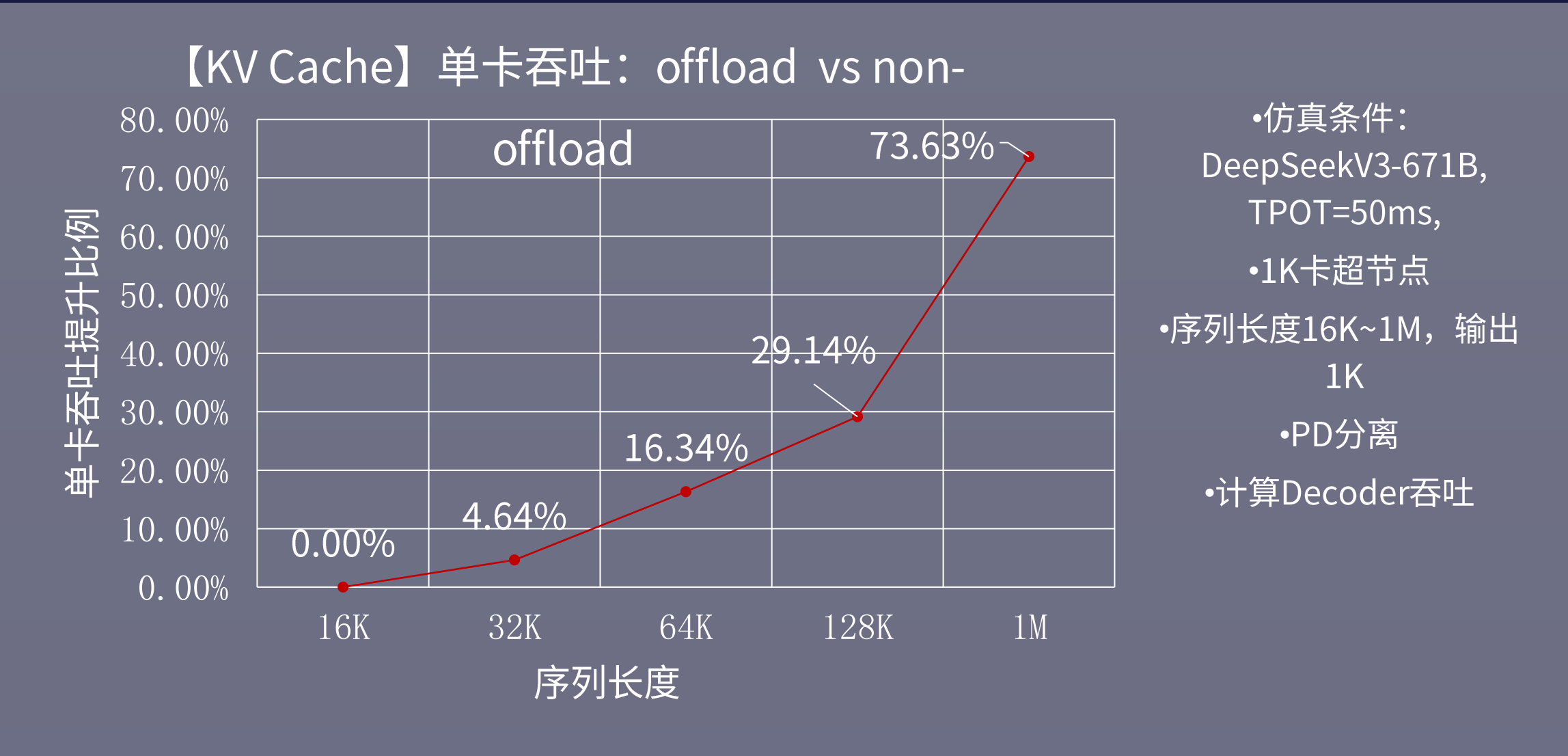


- 通信优化**：利用UB互联NPU间大带宽、低时延、统一内存语义实现高性能Dispatch和Combine通信，大幅降低多专家通信时延；
- 存储优化**：基于全局资源池化，提供多层次KV Cache卸载能力，纵向多层次介质扩容量，横向NPU和介质通信加速，以查代算提升整体系统性价比；

收益：吞吐提升20%~30%



长序列推理：吞吐提升~30%@128k序列；序列越长，收益越大



智算场景超节点关键能力 → 灵衢互联带来的智算业务价值汇总

灵衢使智算场景业务价值

LLM训练	大并行域、大数据量 TP/CP/EP并行通信性能压力	更优并行超参空间 E2E性能收益20%+
大EP推理	大EP并行域、低时延TPOT EP并行通信性能占比20%+	更优EP通信性能 E2E性能收益20%+
长序列推理	搜索+RAG、行业应用 ~M级序列，片上内存容量问题	片上内存/DDR/SSU池化 128k序列优化30%+
强化学习	训推集群间反复迭代交互 互连带宽和时延成为瓶颈	256+卡训推共卡部署 参数实时更新，20%性能提升
生成式推荐	高精度与低时延业务诉求 Prefix KV Cache容量带宽	片上内存/DDR/SSU池化 低时延场景吞吐提升50%
Agent	多算力、长记忆、多轮次 10X+缓存/异构算力高并发	CPU/NPU/各类介质池化 吞吐性能提升50%+

基于灵衢的智算超节点关键能力

超高性能：百GB~TB/s高带宽、us级低时延

超大规模：支持32~8k XPU的超节点规模

多样算力：CPU、NPU、SSU的超节点高性能池化

丰富语义：内存统一编址，支持同步+异步访存、
URPC

通算超节点系统架构示例：基于UB总线实现内存池化、资源池化、大带宽超节点



内存借用、内存共享场景

- **极致性能直连拓扑：**UB支持H2H内存语义互访，CPU之间通过多端口UB实现8节点Fullmesh直接电缆互连，是通用风冷机柜最佳内存池化拓扑方案；可通过2D Fullmesh继续扩大直连域规模；
- **内存池化和共享能力：**节点之间支持内存借用，借入内存支持本机Cache一致性；节点之间支持内存共享，多方读写同一空间；

资源组件池化场景

- **DPU池化互连：**DPU支持UB总线接入，并支持多节点共享、多节点聚合、动态即插即用等能力；实现高性能DPU资源池；
- **SSU池化互连：**SSU支持UB总线接入超节点，支持SSU和CPU/NPU高效对等通信，支持大带宽高吞吐Flash、分布式EC、自主数据读写、动态即插即用能力；实现高性能SSU资源池；

CPU间高性能互连场景

- **CPU间高性能通信能力：**CPU支持大带宽UB互连，支持CPU之间通过同步ld/st和异步read/write对等通信；实现低时延、超大带宽的H2H通信，支持极速热迁移、高性能通信库等能力；

释放极致算力，引领通算场景能力持续创新



大数据

超大容量 免数据落盘

池化算子 | Shuffle聚合加速

Spark 1TB 数据处理时长
3,580s

Spark 1TB 数据处理时长
2,557s, 30% ↓



数据库

超低时延 内存池高效交换

全局Buffer Pool

一写多读，一主两备
1.85 mn tpmC

多写多读，三节点多写
5.4 mn tpmC



虚拟化/容器云

超高弹性 进程错峰借用

内存借用扩带宽

在离线混部资源利用率50%

资源利用率 70%



搜推广

超大带宽 池化数据读写

异步内存访问通路 | 微秒级元数据同步

在线推荐请求处理时延
~350ms

时延降低40%+



证券极速交易

超短路径 直通远端L3 Cache

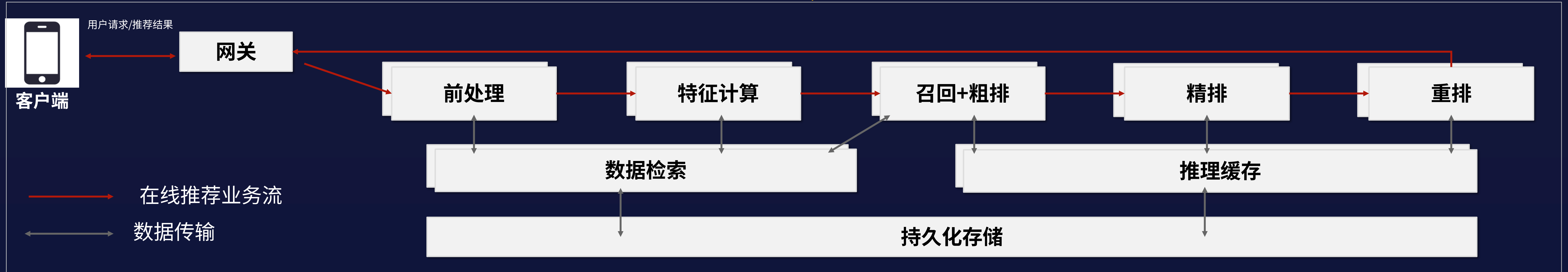
内存语义百ns 远端访问

256Byte 典型交易报文
1200ns

256Byte 典型交易报文
降低50%

搜推广挑战与方案：灵衢超节点加速RPC&KVC/Emb读写，应对在线推荐百ms极致低时延诉求

业务痛点：在线推荐E2E时延需满足300ms硬约束，更低时延带来更优用户体验 & 广告曝光效果



挑战1：多服务RPC通信效率低

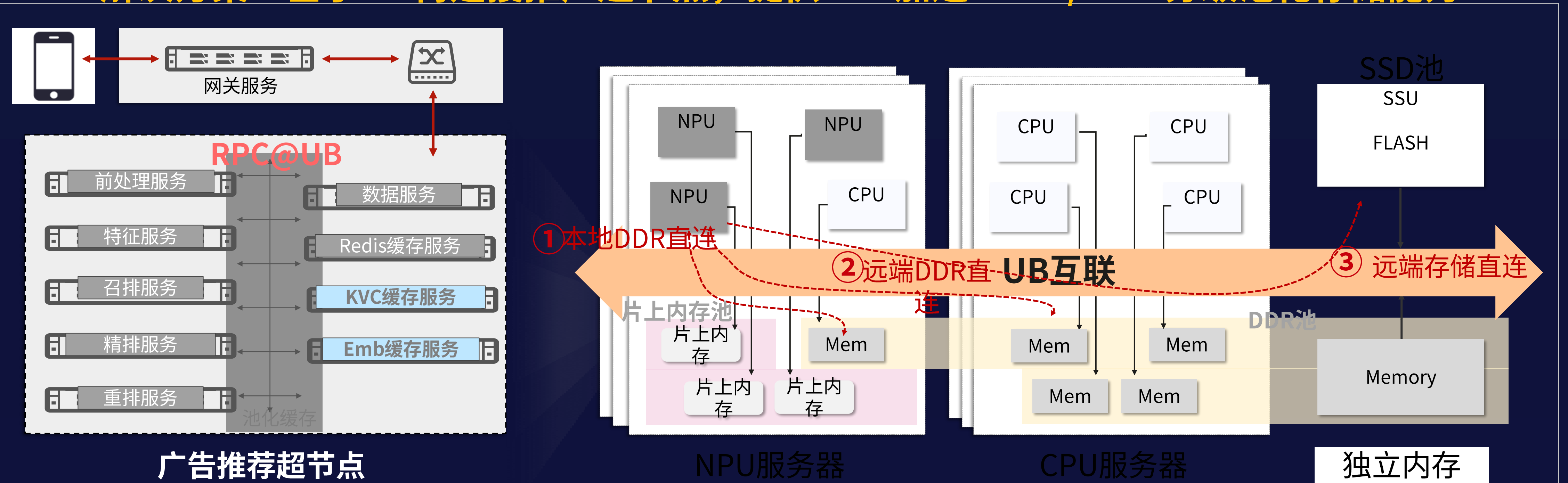
单请求10+服务间RPC通信，MB级数据通信量，RPC耗时占比可高达30%+

挑战2：片上内存/Mem/SSD单一介质无法同时满足KVC/Emb容量&带宽需求

➢ Emb：用户/商品Emb随业务扩张，未来可达百TB级

➢ KVCache：GR推荐场景KVC大小随模型规模 & 用户数增长，未来可达PB级，需求TB/s级加载带宽

解决方案：基于UB构建搜推广超节点，提供RPC加速&KVC/Emb分级池化存储能力

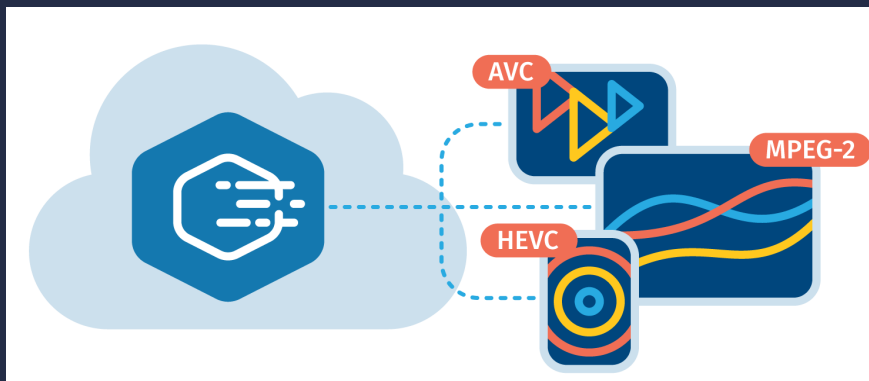


➢ RPC通信加速：基于灵衢URMA & Load/Store，实现TCP Socket能力，透明支持主流RPC框架

➢ KVC&Emb池化：构建PB级片上内存-DDR-SSD分级池化存储，提供TB/s高速KVC&Emb数据读写能力，算力/容量/带宽可按需独立扩展

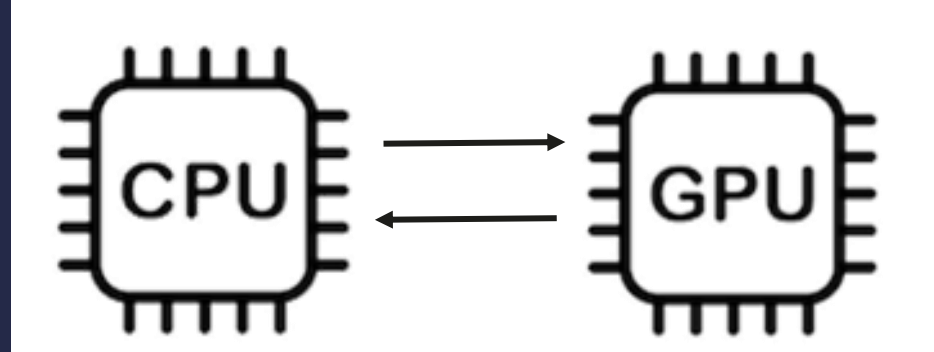
视频分布式转码解决方案构想：灵衢共享内存池，以传代算，提升吞吐

场景1：视频分布式转码



- 跨节点共享内存，免共享存储
- 以传代算，减少解码次数

场景2：视频增强/特征处理



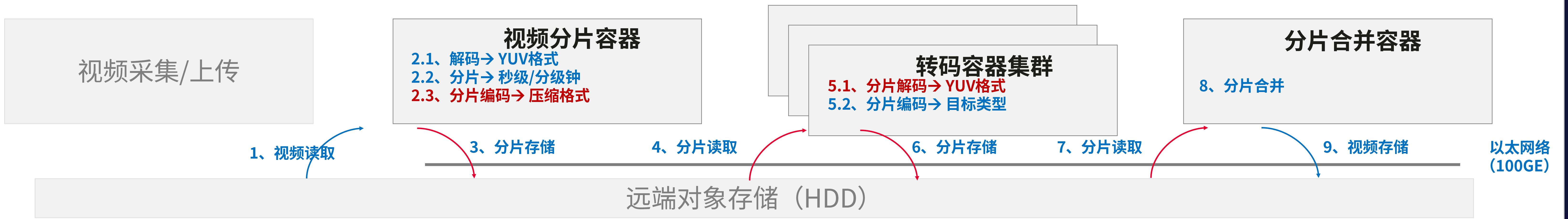
- CPU-GPU间高速通信
- 提升数据搬运效率

场景3：视频（协作）直播



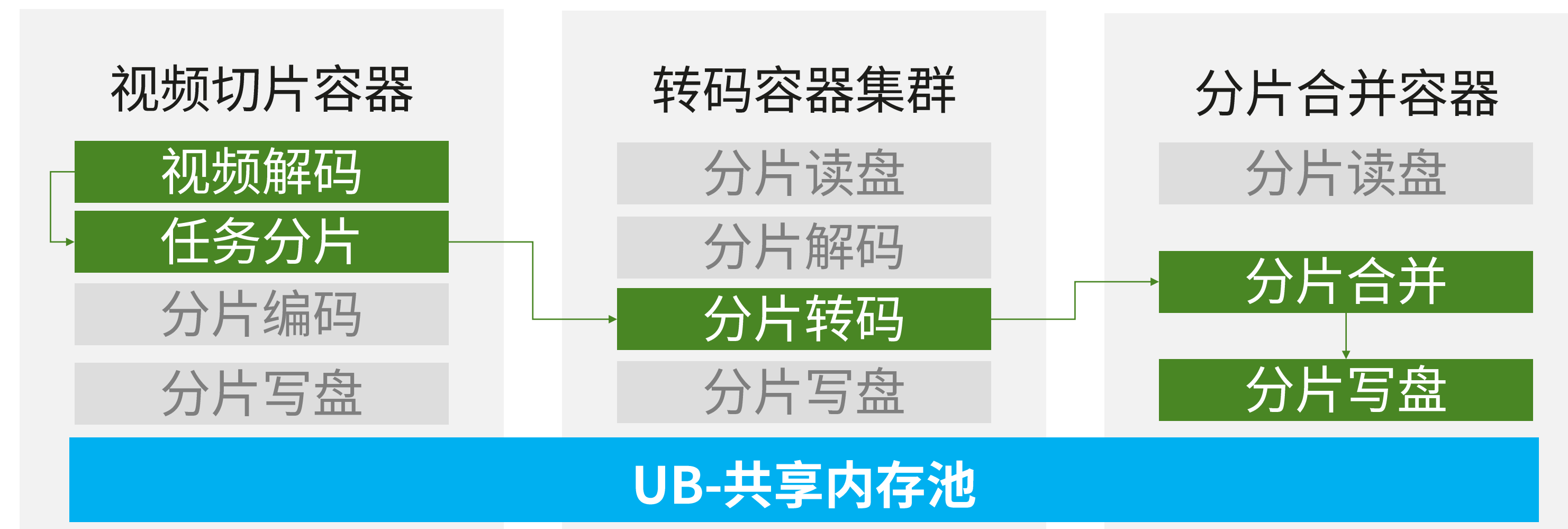
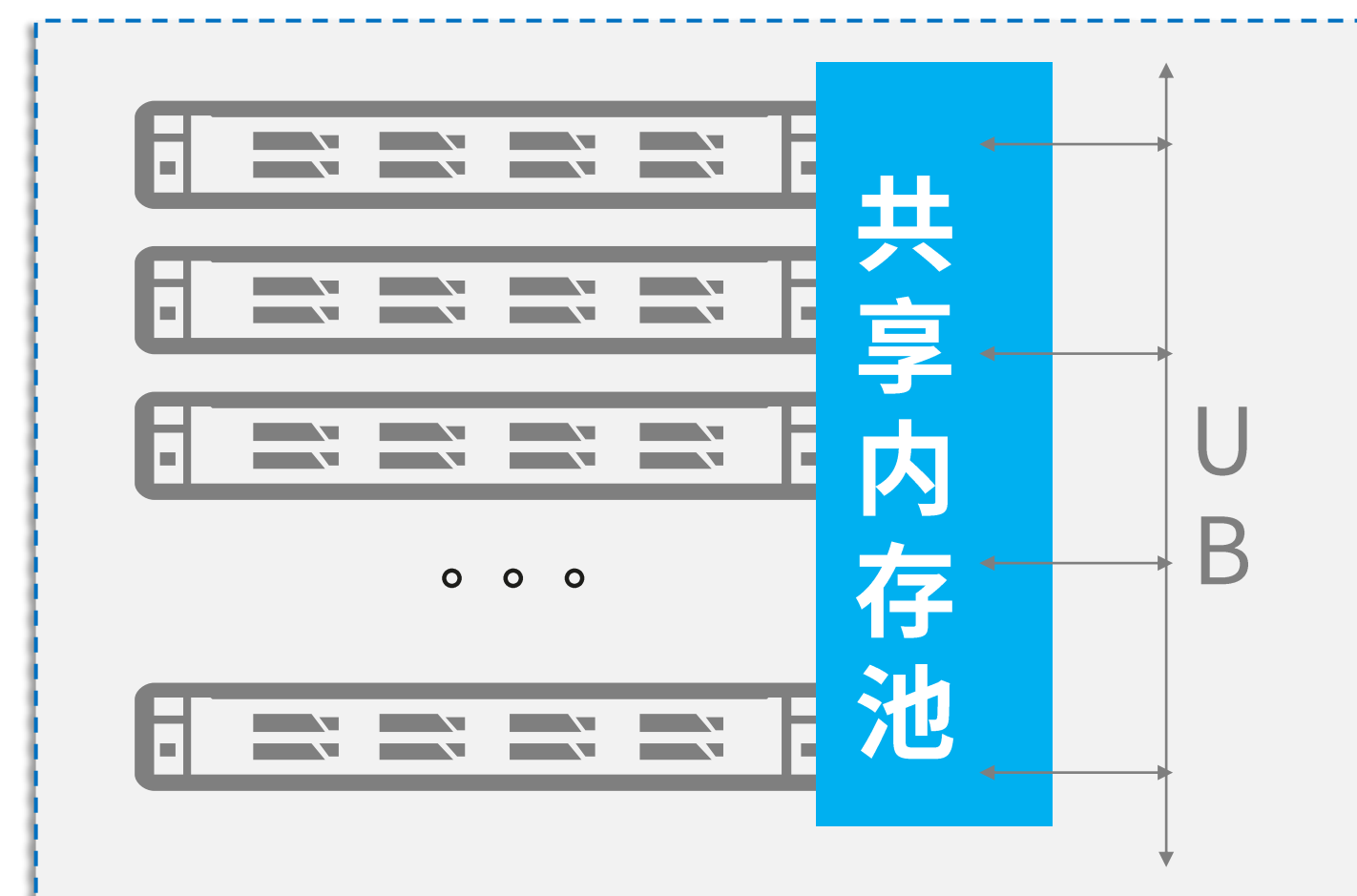
- GPU资源池化
-

分布式转码跨机同步效率低下：数据同步存在4次存储IO栈，网络传输存在1次额外分片编码



基于灵衢400GB超大带宽内存池：免共享存储IO传输和分片编码

转码超节点（8~16）



穿刺增益测试评估

原视频格式	H.264 1080p
目标格式	H.265 / H.264
目标分辨率	1080P / 720P / 480P
视频分片	60s

以传代算，减少冗余解码次数，**提高转码效率25%**

新行业：低时延解决方案@证券极速交易

业务挑战

百ns级时延竞速

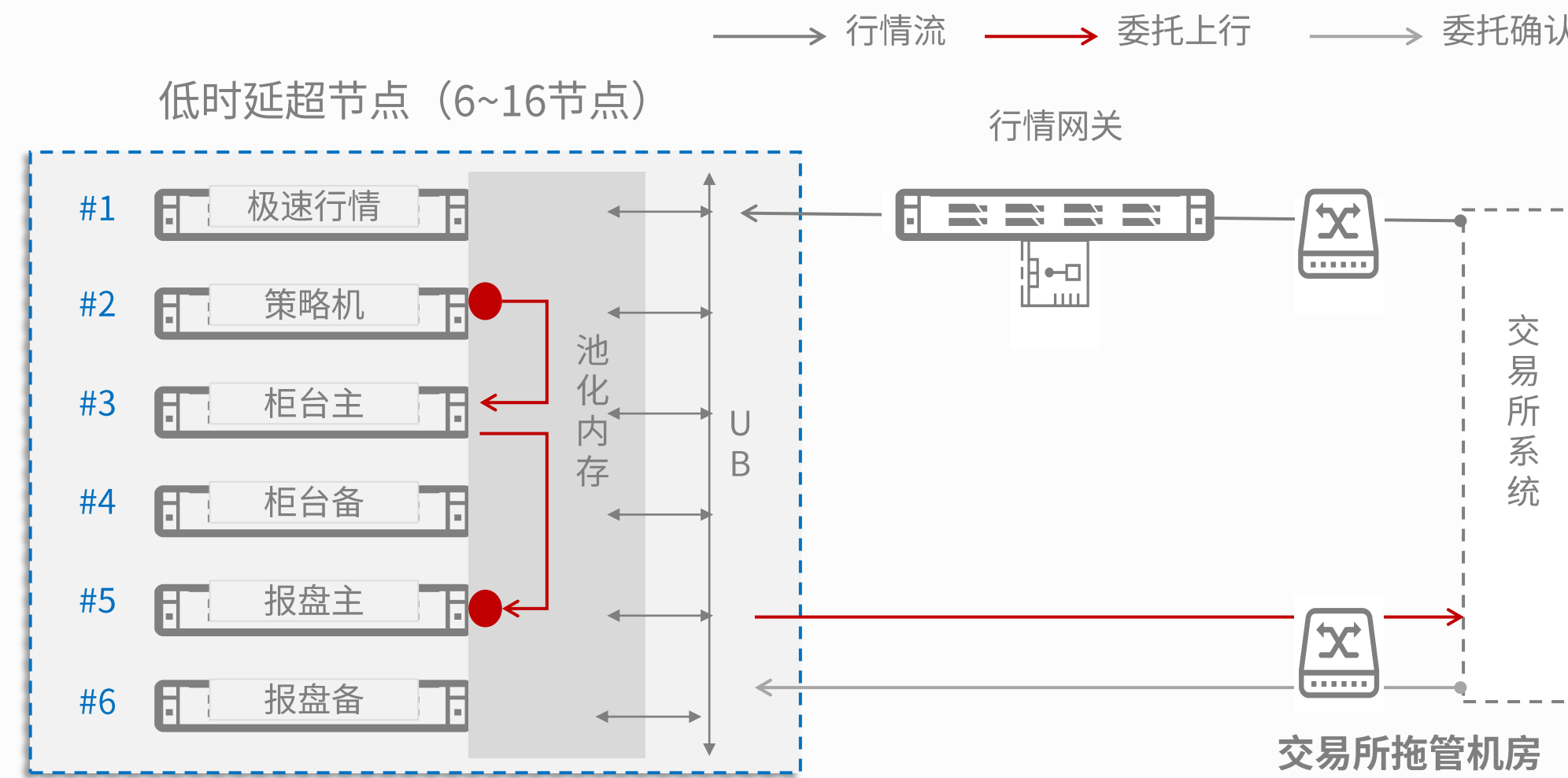
- 先到先得：证券采用竞价交易模式。

主备复制导致架构臃肿

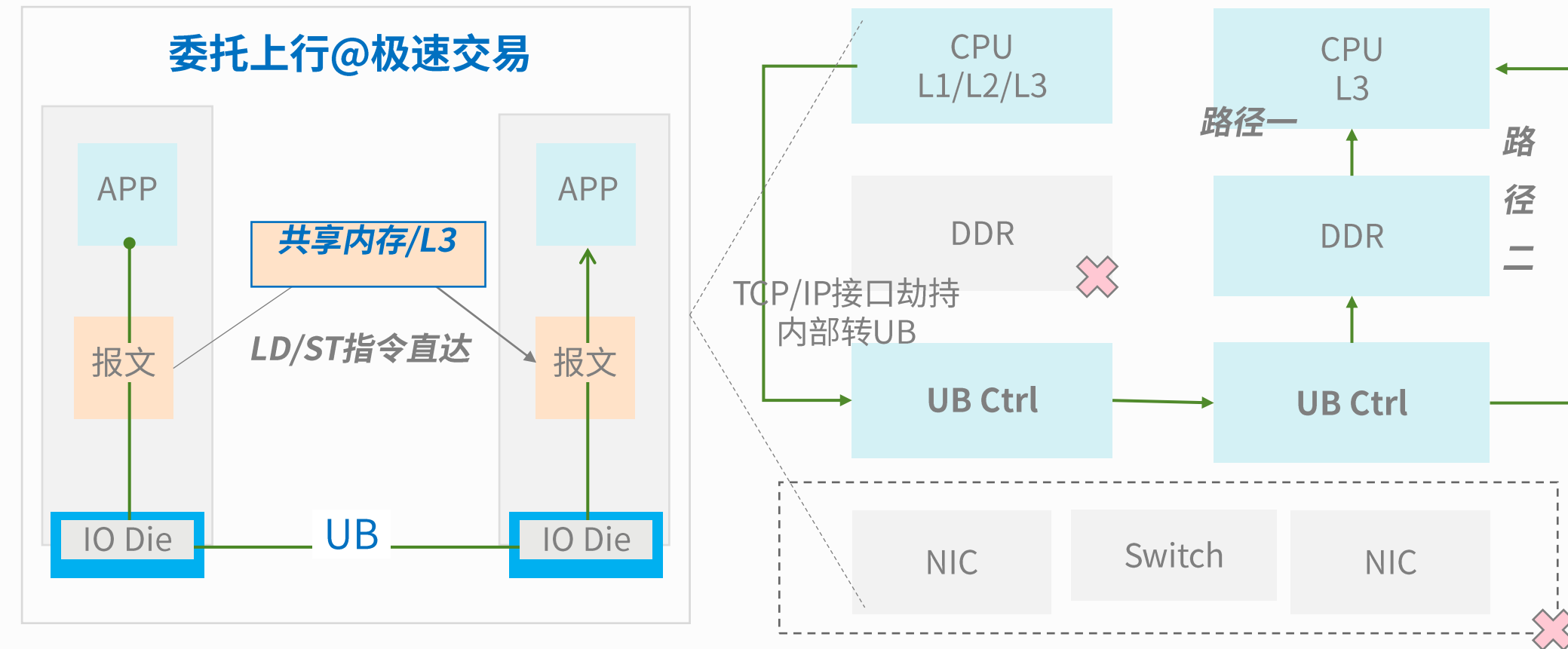
- 可演进性差：依靠应用实现同步，难扩展
- 性能差：依靠网卡，单事务50us。

解决方案：基于UB-M构建低时延超节点，提供亚us机通信和高可用内存池

关键能力一：LD/ST内存语义通信，替代传统网络，重新定义数据通信方式

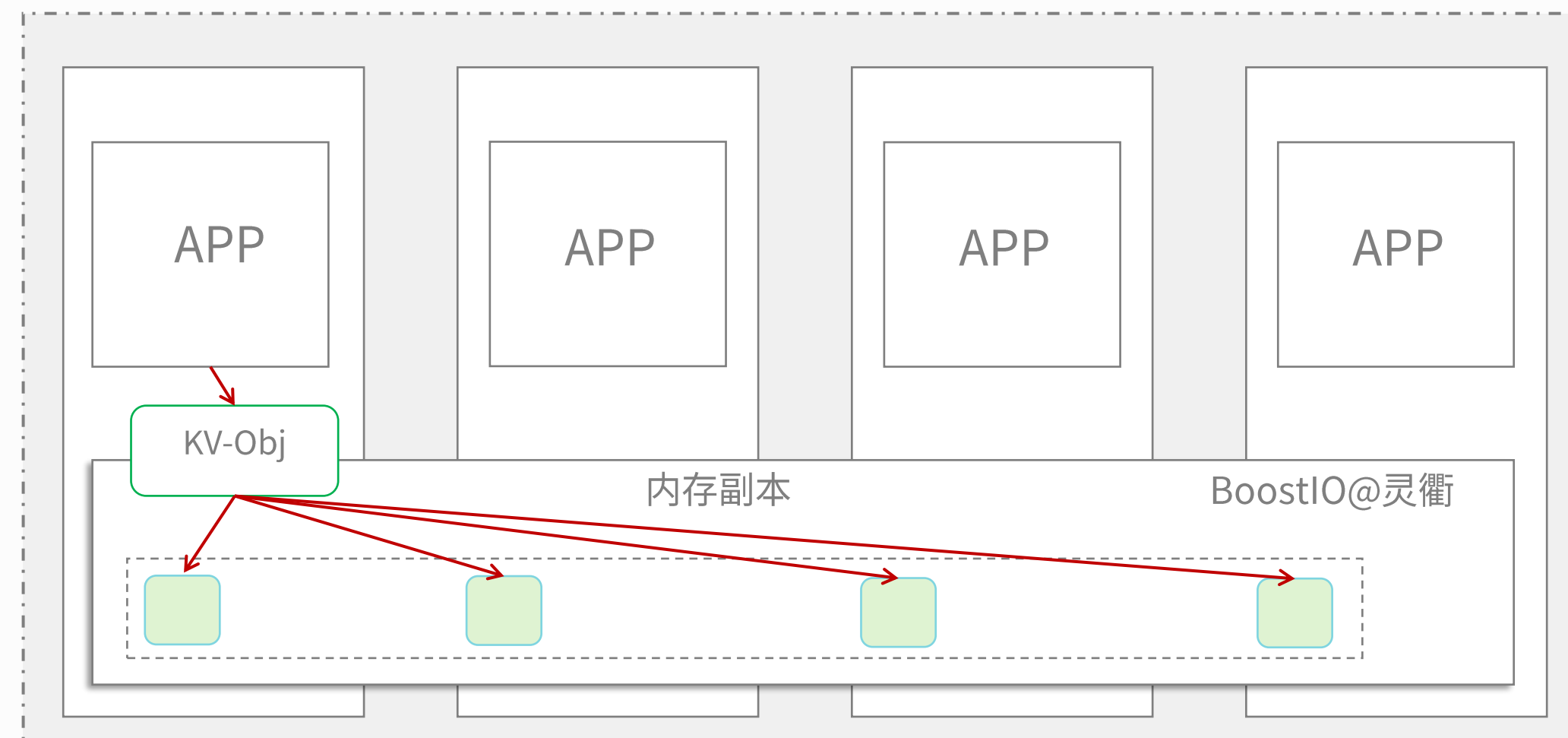


新形态：核心交易一体机

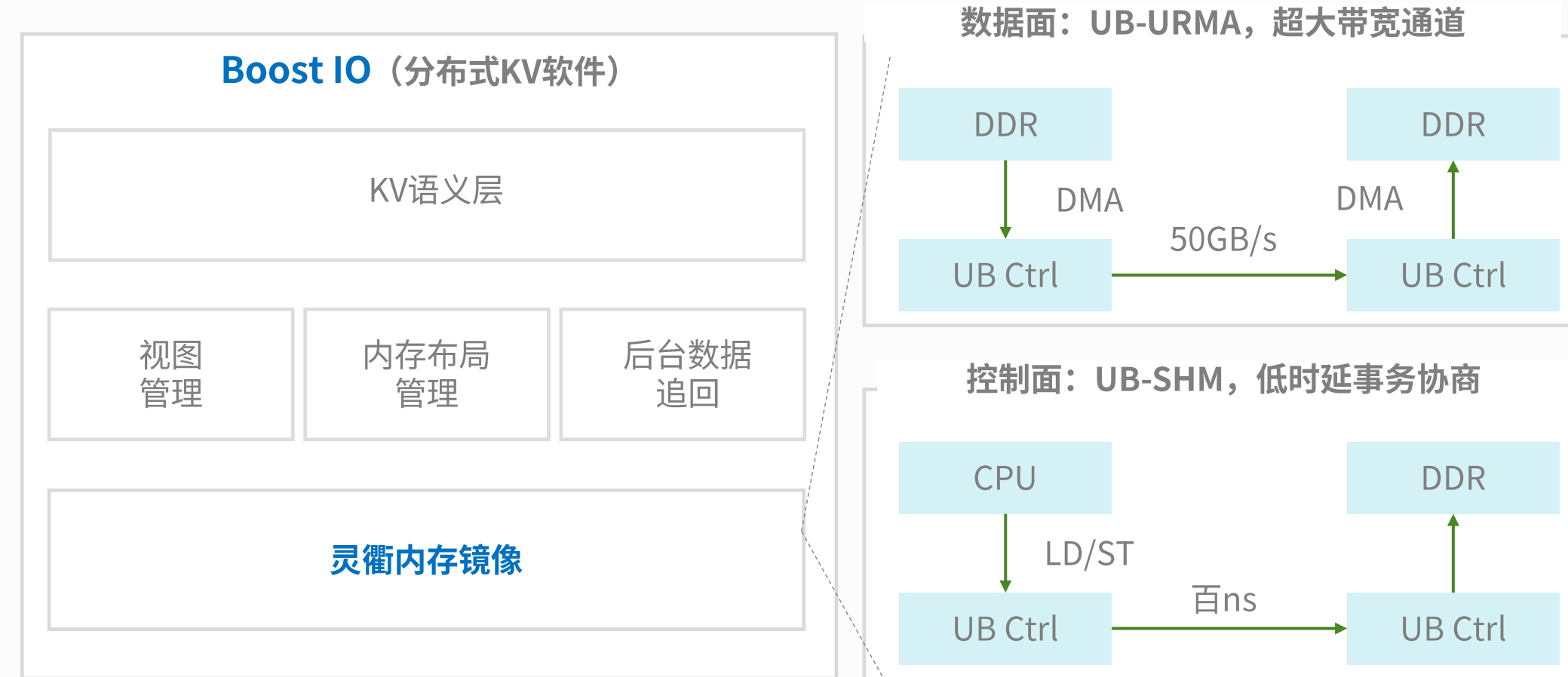


关键技术：共享内存替代网络通信

关键能力二：内存池化，构建分布式一致性&事务能力，提供高可用数据底座



新形态：共享内存池底座



关键技术：灵衢结合BoostIO，提供分布式KV内存池

业务价值

极速通信

- 极致性能：共享L3 Cache，通信时延下降一倍。
- 全链路使能：行情、策略、柜台、报盘。
- 高可靠：池化故障抑制。

独立内存池

- 高可靠：最大支持8副本主备同步。
- 架构解耦，使能应用敏捷：主备逻辑下沉，业务无状态化。
- 高性能：基于UB加速数据面和控制面。

通算场景超节点关键能力 → 灵衢互联带来的通算业务价值汇总

灵衢使能通算场景业务价值

虚拟化	多种业务负载不均、资源争抢、通信开销10%+ CPU利用率<50%，内存利用率<30%	内存超分25%，50ms虚拟迁移 资源池化，利用率提升20%+
数据库	数据同步、分布式事务等网络/IO开销大 分布式池化架构，故障快恢挑战大	内存、数据、事务池化，多写多读一致性， TPCC提升20%，RTO<2s
大数据	CPU和内存利用率低，平均30% 数据预处理算子CPU-bound	Spark资源利用率30%→60% NPU加速算子性能50%
分布式存储	服务器主从架构约束SSD性能； 传统网络性能不足，5倍带宽收敛	SSU+内存对等池化，百GB带宽存储池； 盘即存储，免服务器TCO优化15%
搜推广	RPC通信和KVCache/Embedding查询时延高	内存语义，RPC时延下降50%； 内存池带宽4倍提升，KVC/Emb时延下降60%
证券低时延	百ns级时延竞速，先到先得 传统网络栈时延高	内存语义通信，时延优化一倍

基于灵衢的通算超节点关键能力

超高性能：百GB~TB/s高带宽、百ns级低时延

灵活规模：直连拓扑8~16节点，交换实现更大规模

高性能池化：内存、DPU、SSU的高性能池化

丰富语义：内存统一编址，支持同步+异步访存、URPC

Thank You



Q&A

?