

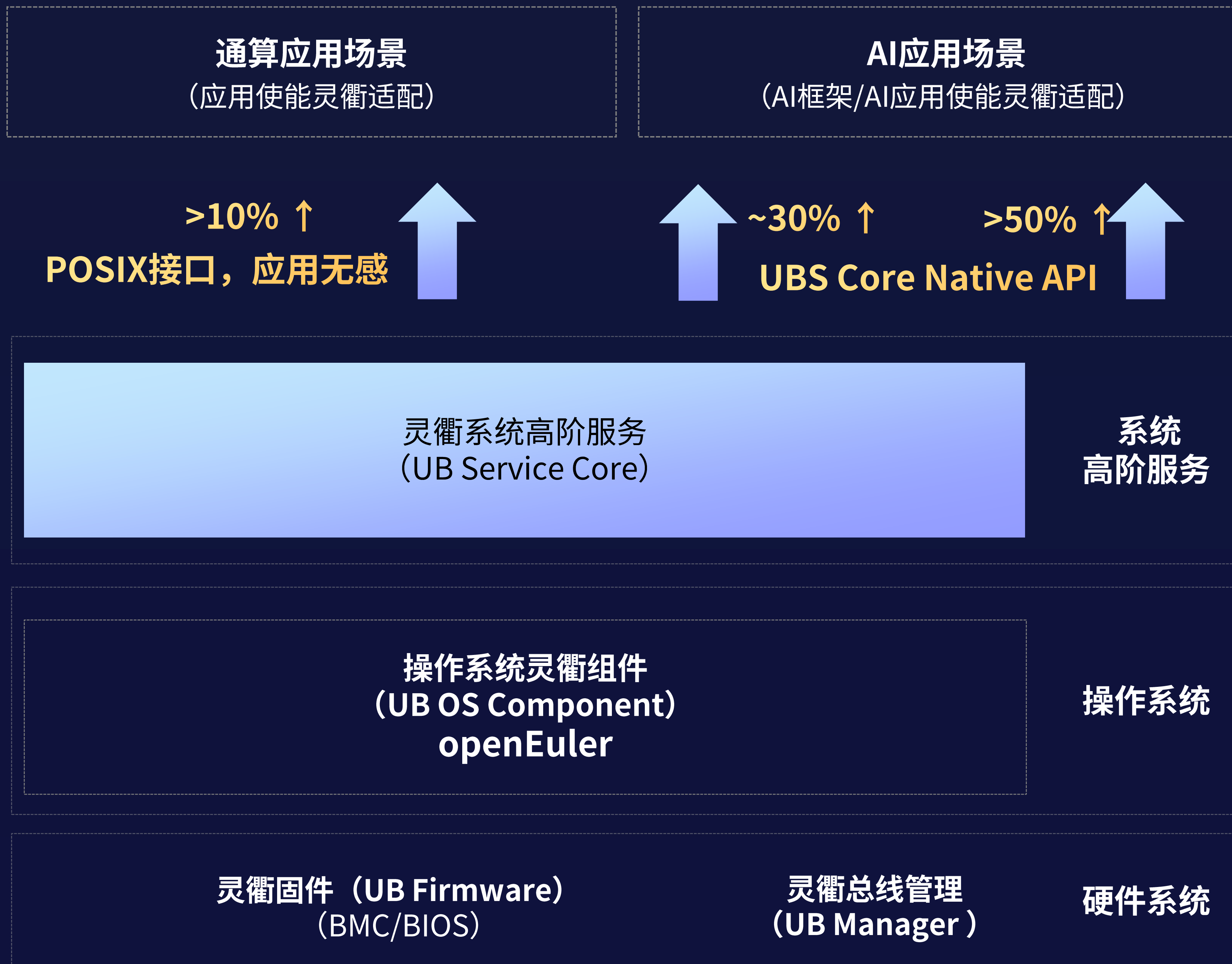
灵衢系统高阶服务关键技术和应用

姚弋宇

sig-UB-ServiceCore Maintainer



操作系统以及系统高阶服务使能灵衢系统，开源开放使能应用价值变现



关键竞争力与价值主张

1.支持平滑演进，易部署，支持可定义计算

- UBS Engine跟随openEuler发布、部署，无需额外硬件/软件安装
- UBS Engine支持去中心化，对等池化，实现集群内灵活组合编排，可定义计算管理

2.应用三种使用模式，实现Matrix集群高性能

- OS提供Posix接口，应用无需修改即可获得**10%**收益
- 应用集成UBS Core加速库/运行时，少量适配可获得**30%**性能收益
- 应用针对UB进行深度修改和架构优化，可获得**50%**以上性能增益

3.多样性算力集群，打造ScaleUp新能力新场景

- OS支持多虚1（NPU/DPU）、超级容器（热迁移）、VM确定性迁移等新能力
- 从单节点**3个9**到超节点**4个9**挑战**5个9**：外设Nway冗余，核故障检测与冗余替换、VM极速迁移实现计划内0中断，节点随机下电高可用

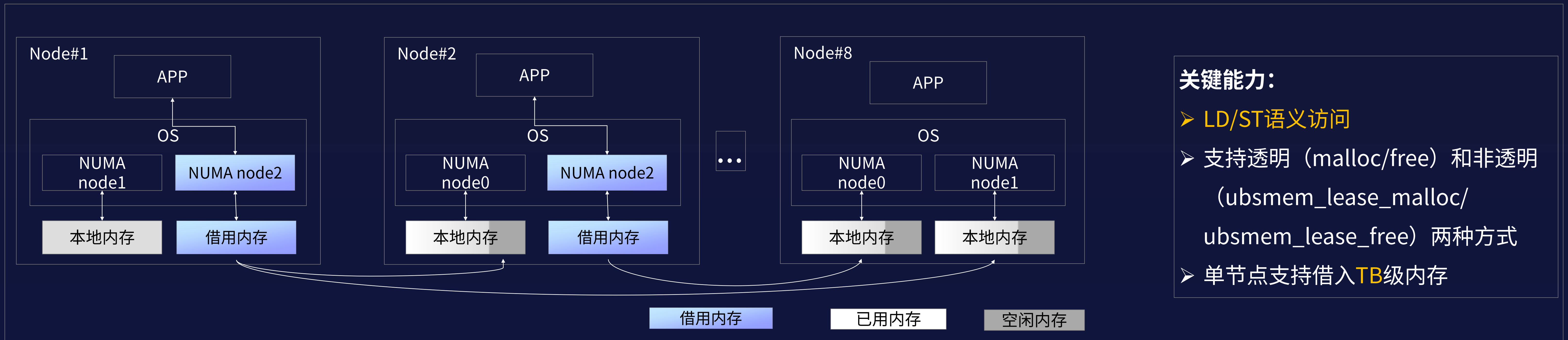
UB Service Core:提供内存池化、通信、IO等系统高阶服务，使能应用加速，充分释放超节点架构优势

全面开源至openEuler、分层解耦、简单易用、生态兼容



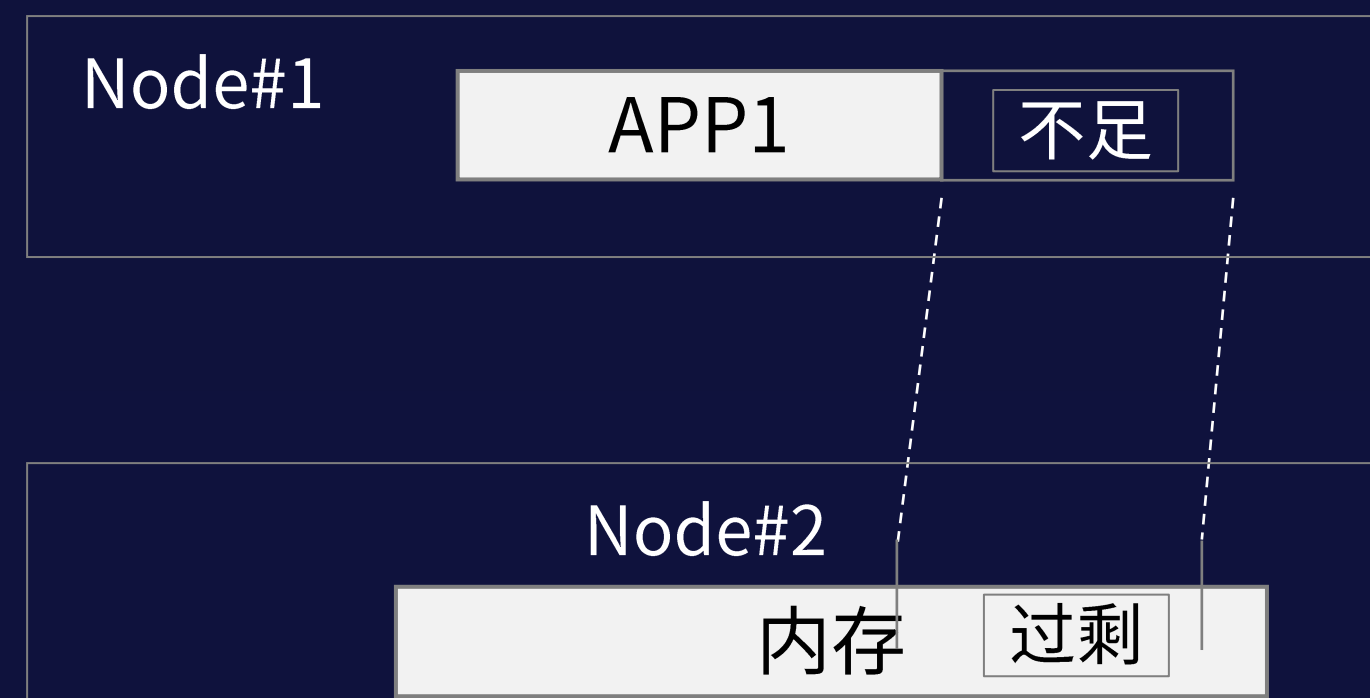


核心能力1--内存借用：将邻居节点空闲物理内存远程映射，实现负载弹性扩缩内存

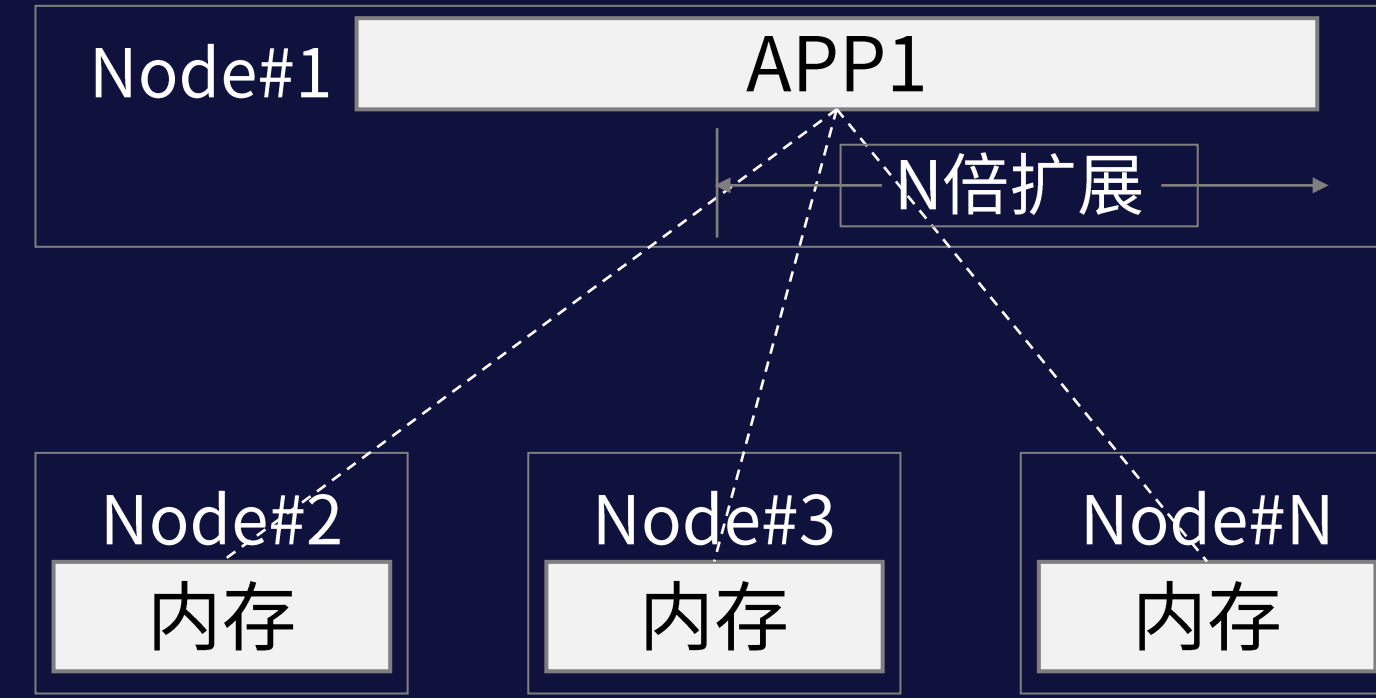


应用范式示例

模式一：内存临时借用
削峰填谷、pageCache加速IO

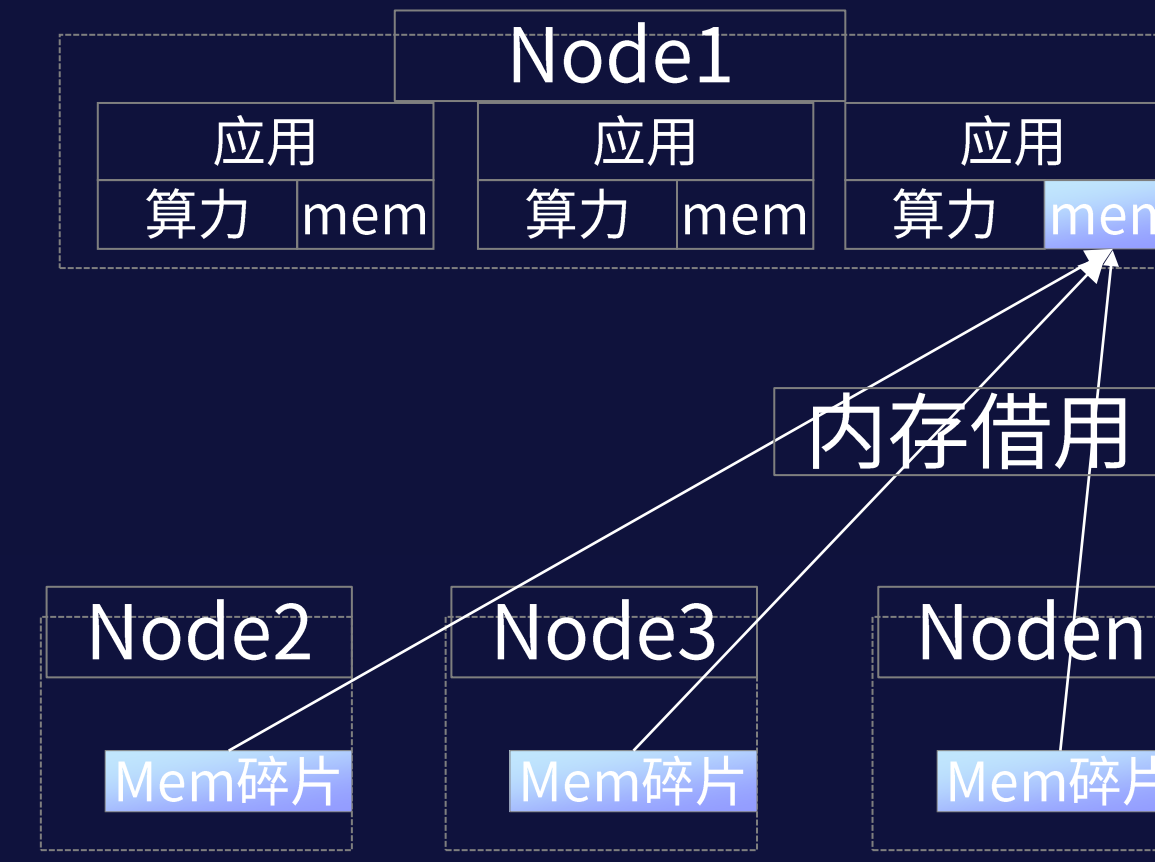


模式二：一借多
N倍扩展，超大内存

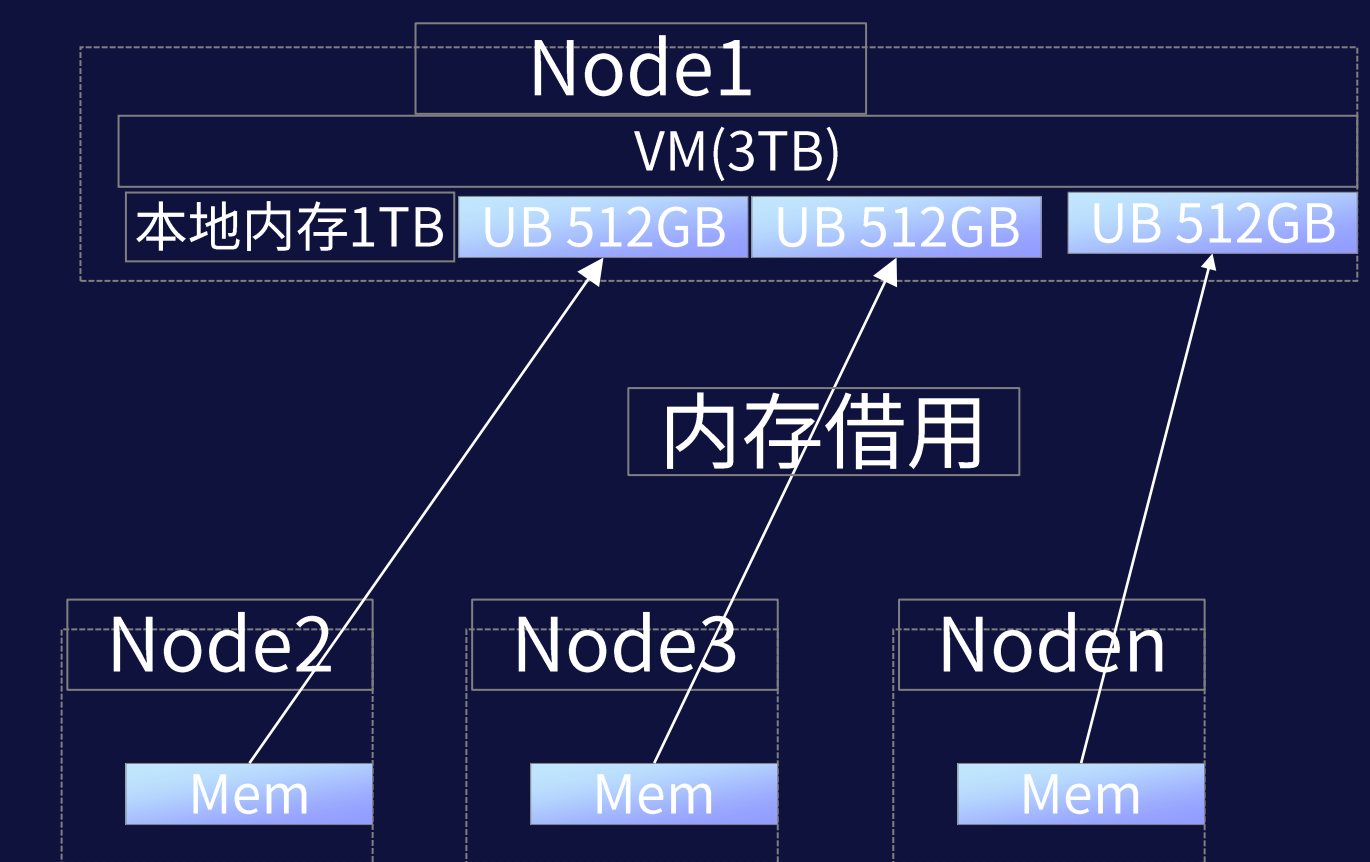


应用场景示例

CASE1:内存碎片



CASE2:3T超大规格虚机



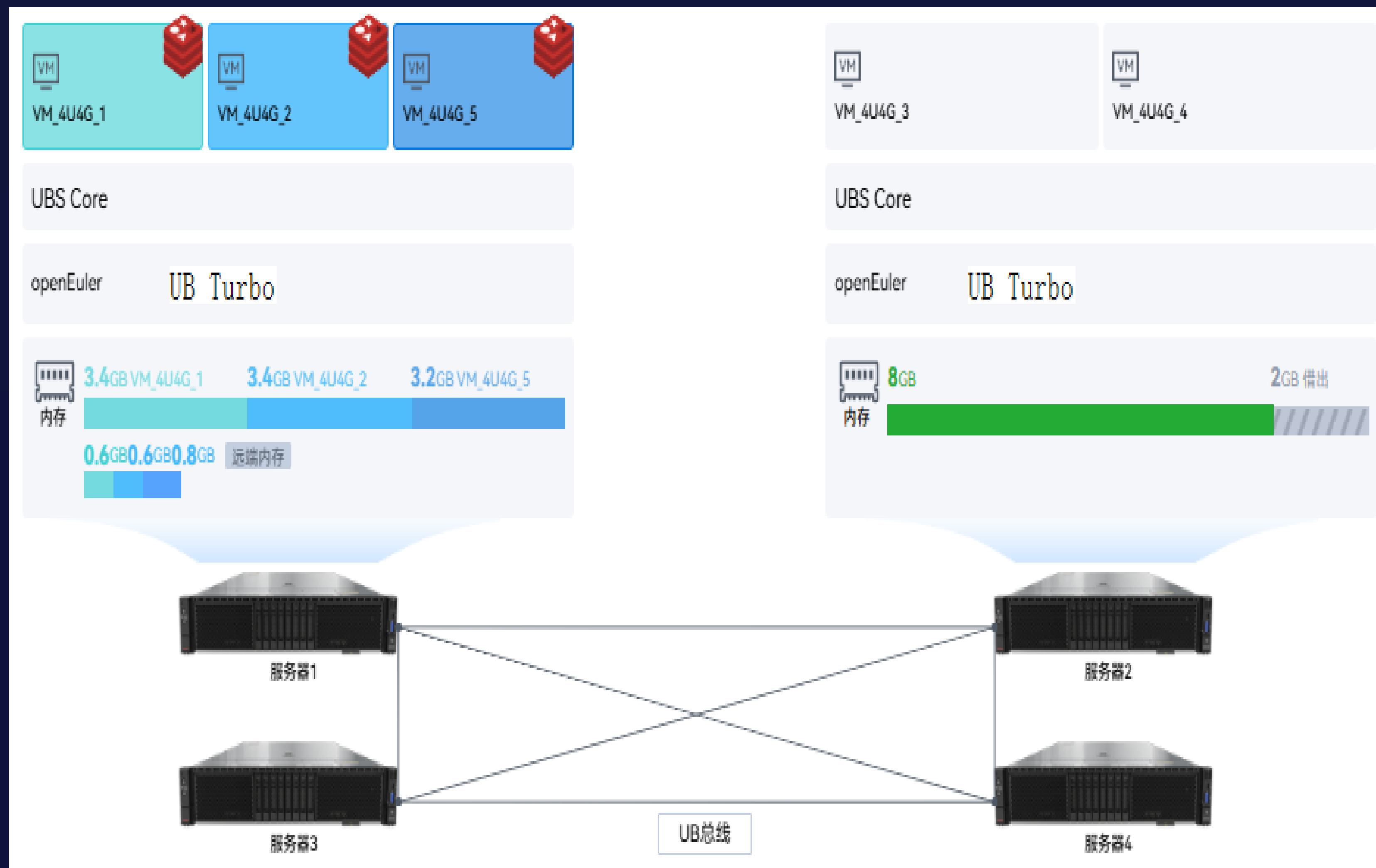
CASE1：集群资源碎片消减，提升利用率25%

（AsIs） 问题和挑战

云基础设施虚拟化场景，选择资源创建虚机是装箱问题，虚机大小未知导致内存搁浅，成为碎片（客户反馈25%以上），资源利用率低

（Now） 效果和目标

基于UB内存借用基本能力，将搁浅内存资源重利用，提升资源利用率**25%**，创建更多虚机，Redis集群场景性能提升**10%+**



价值场景

①虚拟化

②大数据

③数据库

关键技术&价值

1. UBS Core识别集群中的**资源碎片**，将碎片内存动态借入到有空余算力的节点，创建更多虚机
2. UB Turbo将借入的内存在已有虚机间进行分配，更多的远端内存被分配给不敏感虚机（如空闲虚机）
3. UB Turbo基于鲲鹏芯片**冷热感知**能力迁移数据，保障分级内存场景业务性能

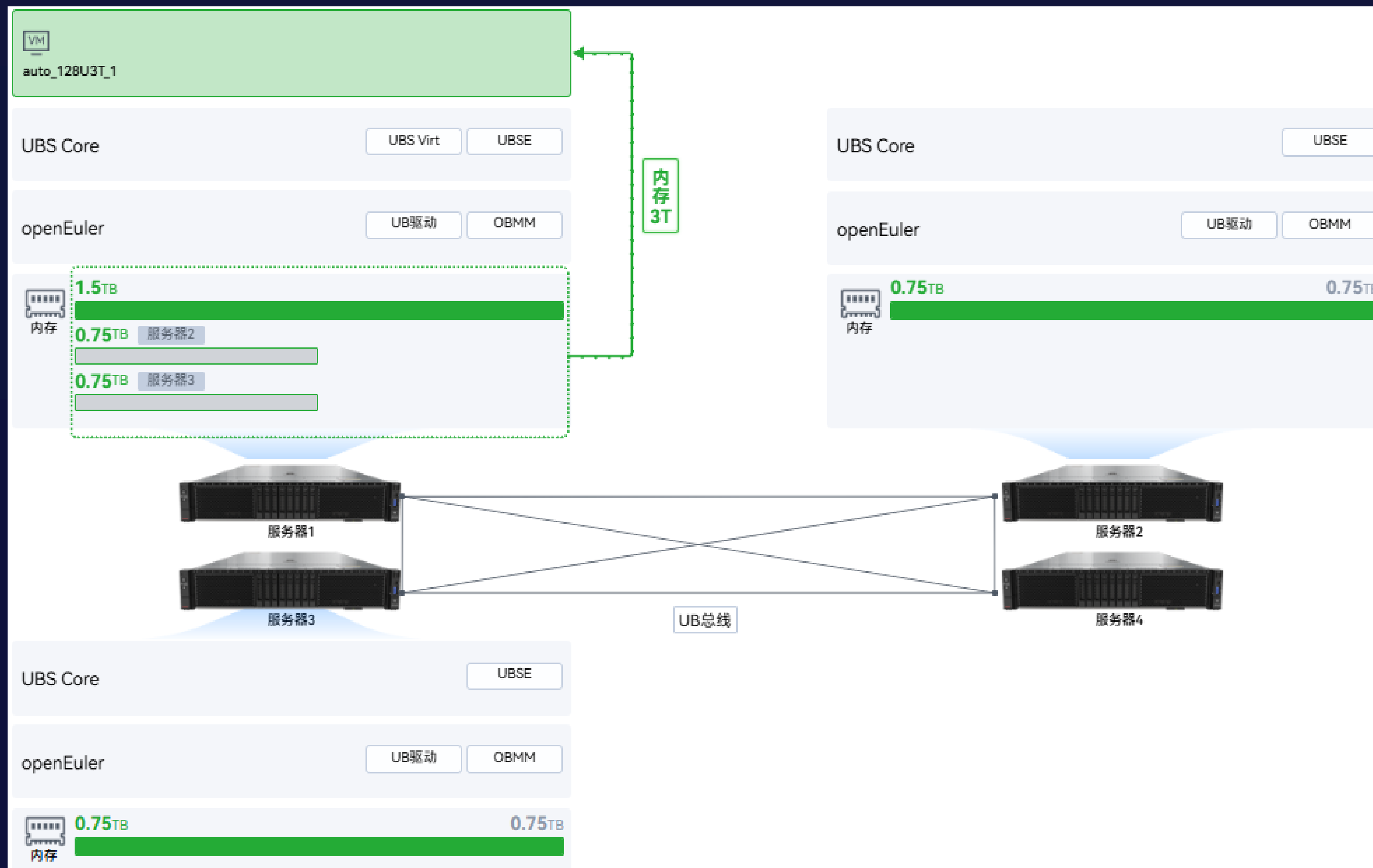
CASE2: 超大虚机 ($\geq 3T$) 通过UB系统提供大内存大算力聚合虚拟机, 降低成本

(AsIs) 问题和挑战

某客户基础设施团队提出需要运行大规模应用, 如果采用一致性大系统 (4P, 8P), 成本很高, 希望华为提供聚合虚拟化解决方案

(Now) 效果和目标

阶段一: 通过超节点加**UB总线实现大内存虚机**
阶段二: **跨节点算力扩展**, 实现内存+算力大规模虚机



价值场景

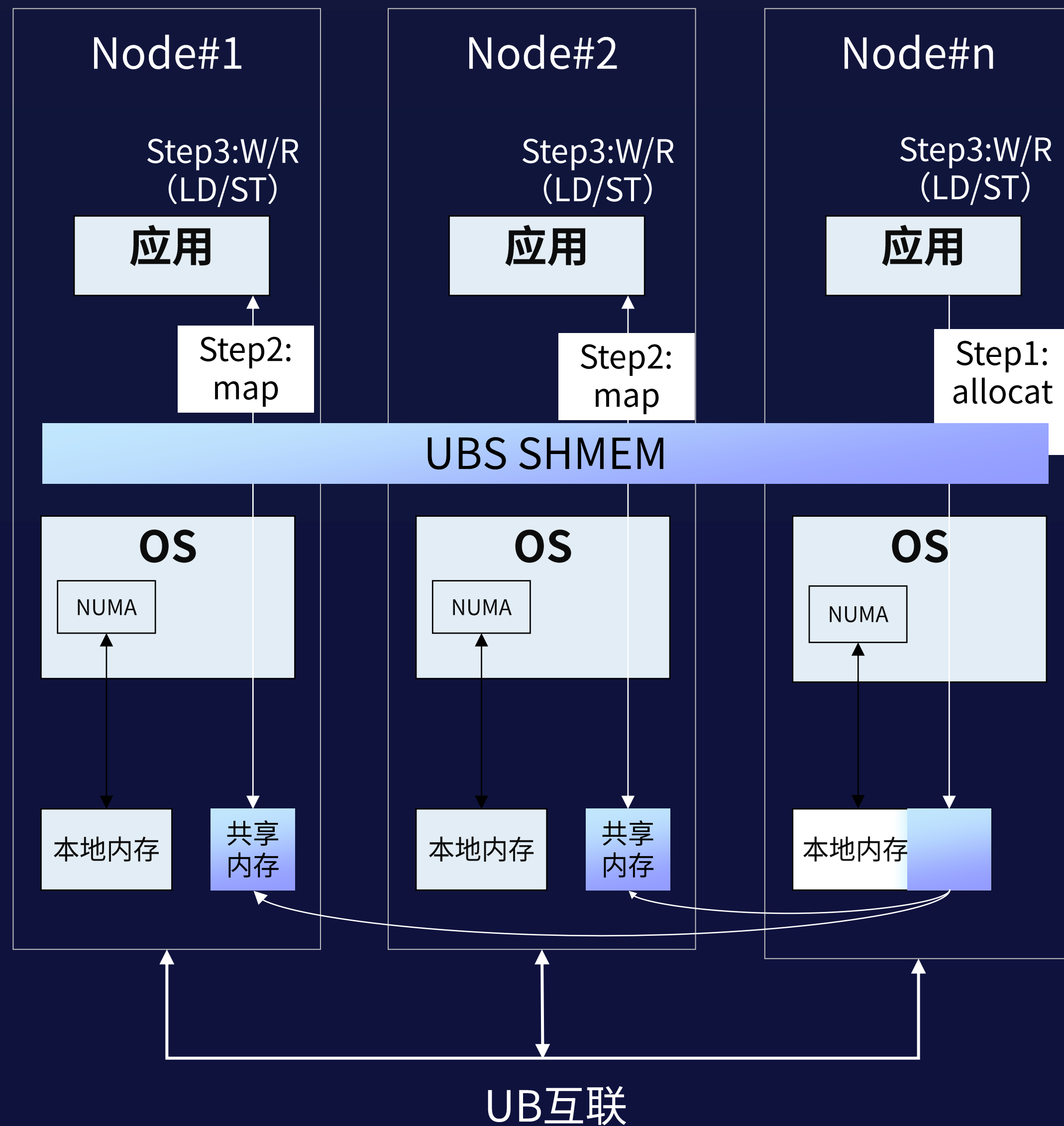
① 虚拟化

关键技术&价值

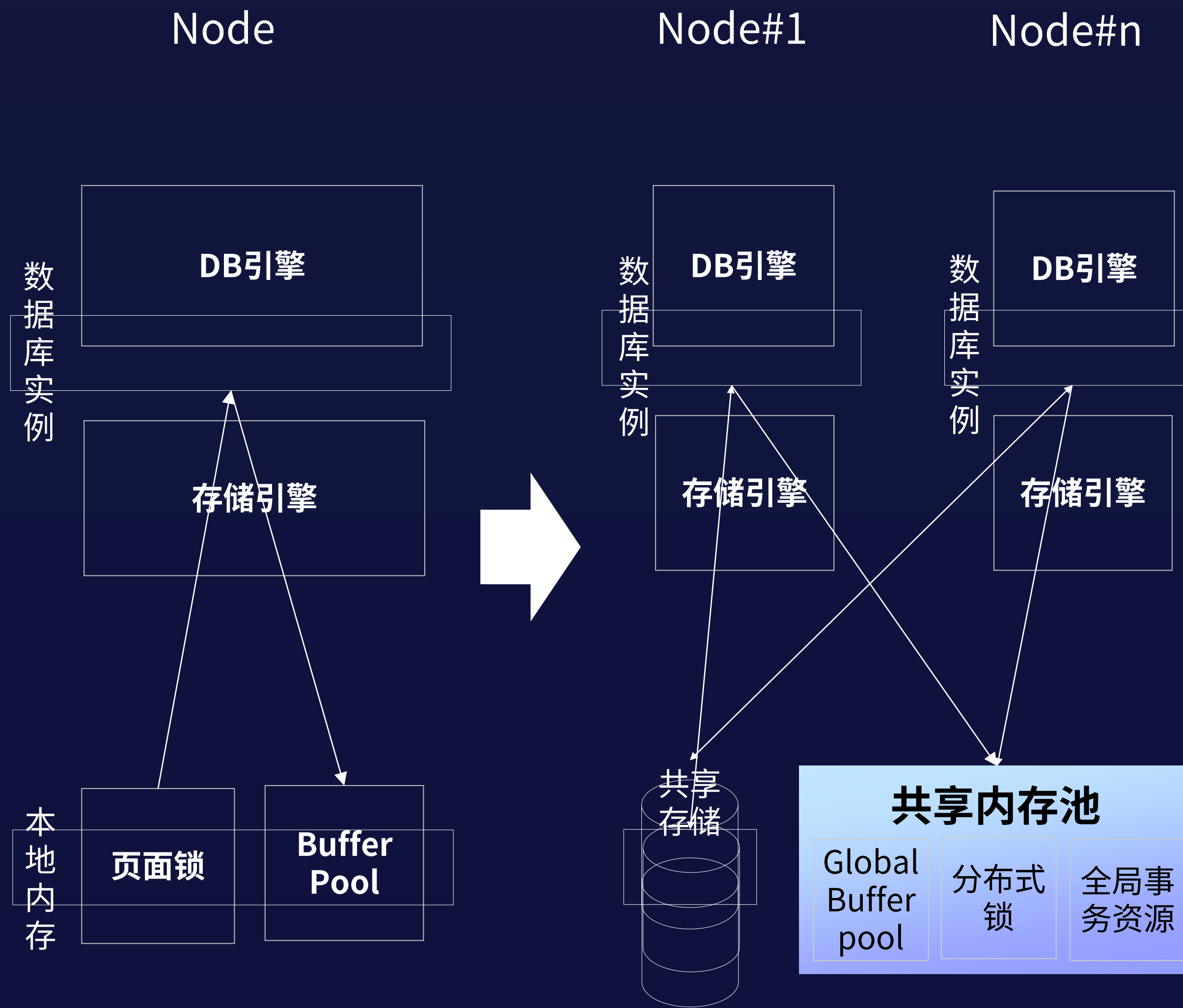
1. 通过**内存借用**, 打破单服务器物理内存的限制, 实现大规格 VM, 满足大规模应用诉求, 降低高P服务器采购成本。
2. UB互联总线支持**内存语义**, CPU可直接访问邻居节点内存
3. UB驱动, UB相关的芯片配置需要对应的驱动使能
4. OS使能, 融入现有OS框架, 包括内存借用, 上线, 下线
5. UBService Core, 集群服务, 含UBSE集群资源管理

核心能力2--内存共享：同一内存多节点共享，实现基于内存语义访问的数据共享

超节点共享内存 @ UB



数据库应用示例@UB SHM



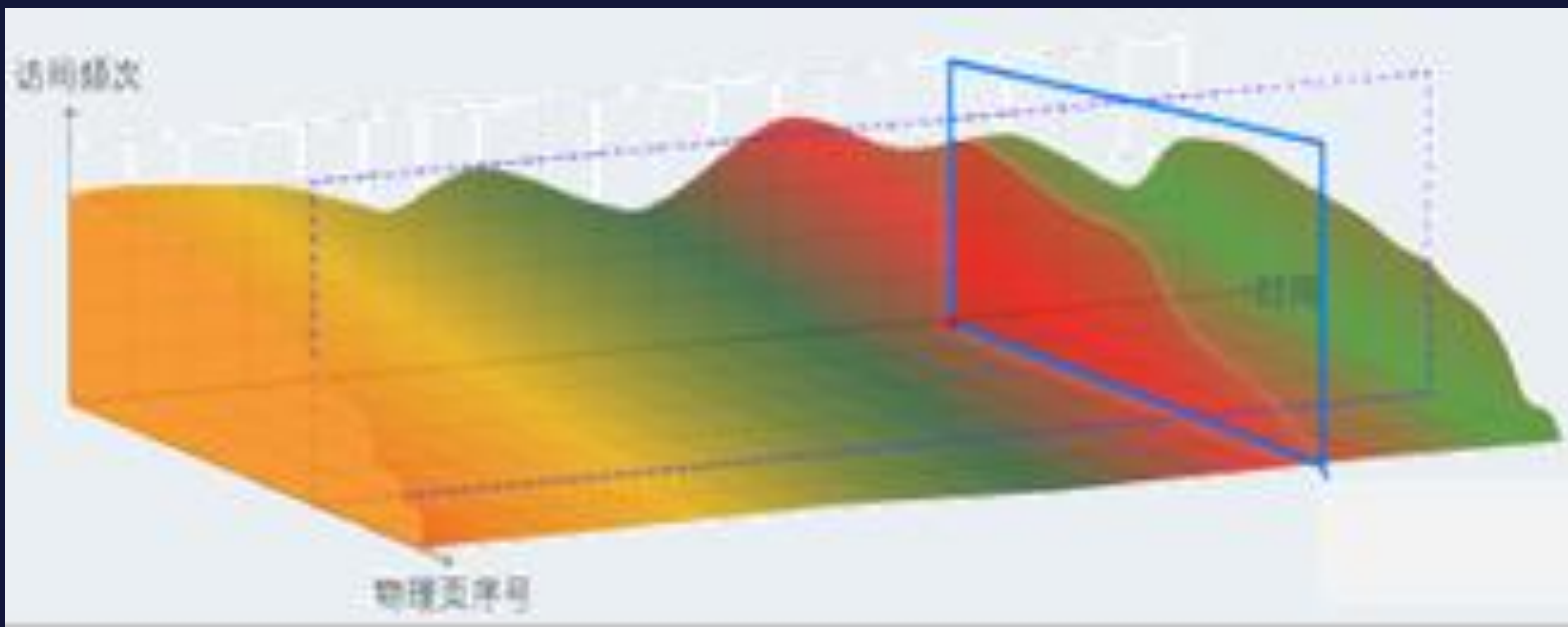
基于UB SHM的数据库实现

DB架构创新
主从->多主

收益
性能提升20%，内存节省30%，线性度大于0.75

- 关键技术**
- ① Global Buffer Pool内存语义直写数据，消除IO读写放大
 - ② 原子语义实现分布式锁和全局事务号等，简化和加速Remote Page Metadata处理
 - ③ 基于SHM的超低时延（百ns）RPC加速跨节点消息通信

核心能力3--UB Turbo软硬协同冷热判断，数据在本地与远端合理布局，应用性能无损



业务访存特征：业务访存行为不确定，存在热区形状不规则，变化剧烈等现象

① SMAP冷热统计阵列

芯片在内存访问路径上集成冷热统计阵列，减少软件实现的扫描开销，实现高精度判热

② 数据迁移策略

基于SMAP统计信息，对比近远端冷热数据，进行数据迁移，适应虚拟化，大数据，数据库等场景

节点

openEuler OS

UB Turbo

远近内存比例动态调整

② 数据迁移策略

冷热扫描

热区识别

自适应速率页面迁移

内核

SMAP冷热统计驱动

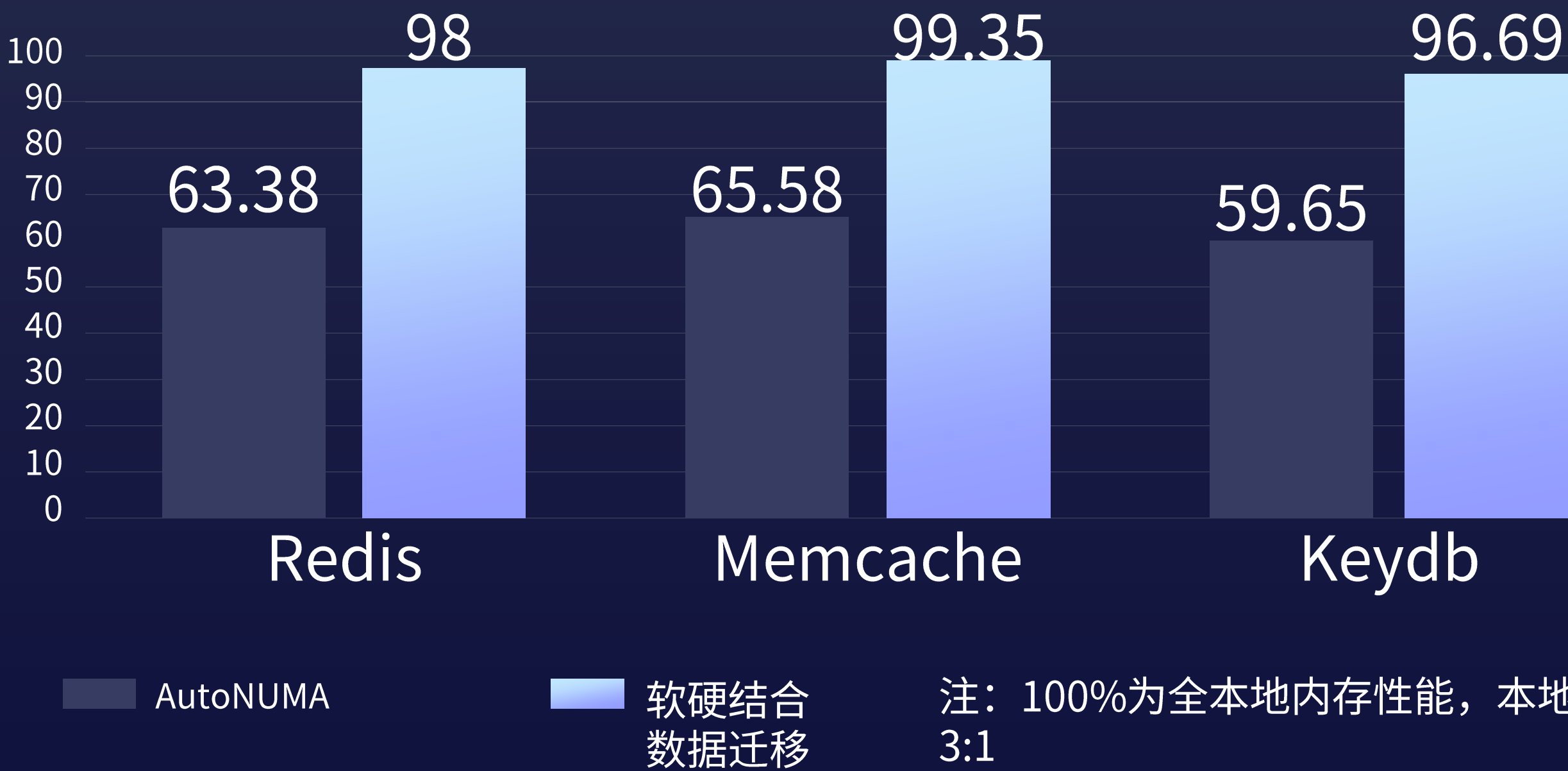
CPU

IO Die

① SMAP冷热统计阵列

客户价值

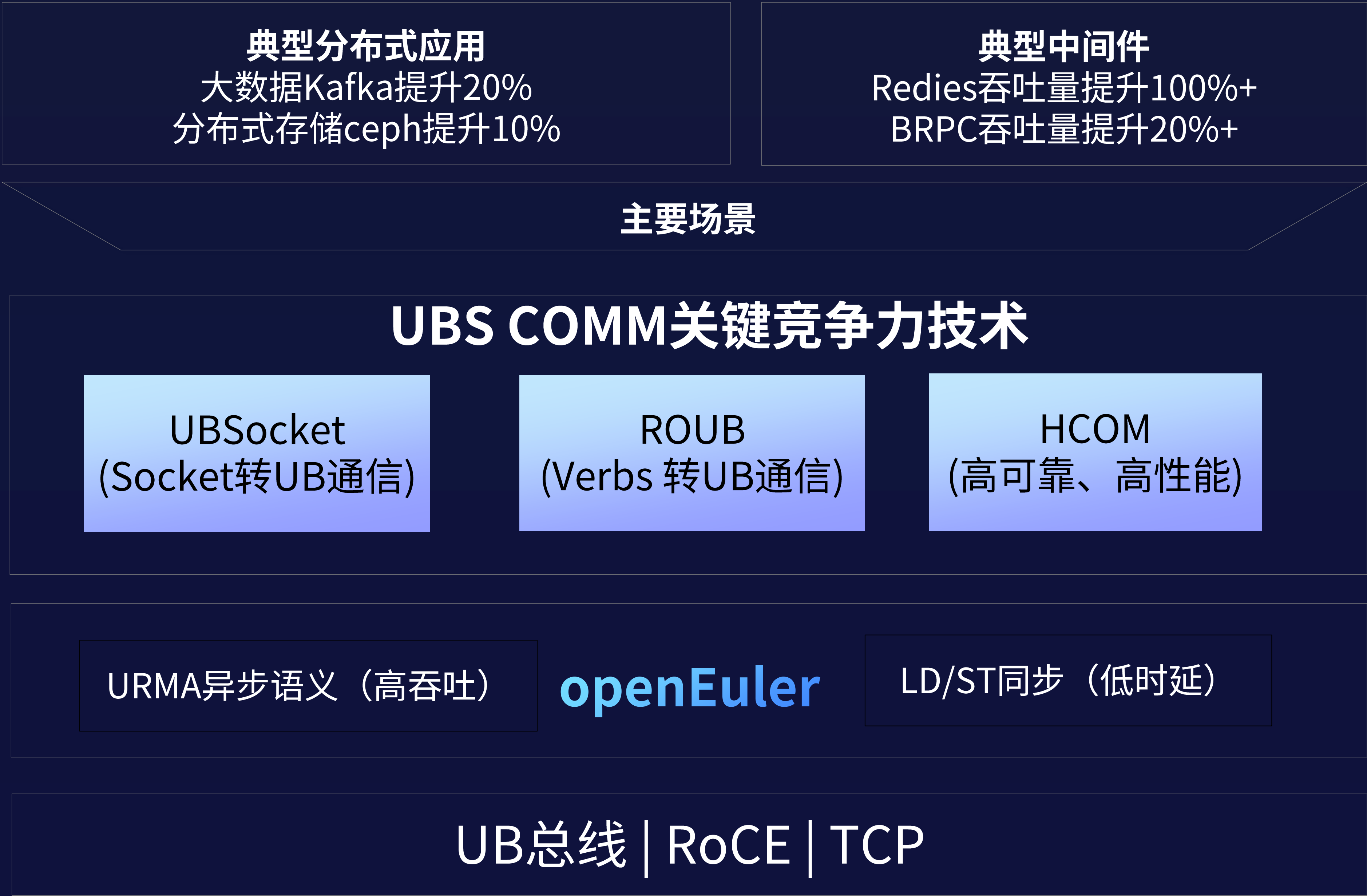
分级内存场景性能数据对比：与Linux原有AutoNUMA方案相比，软硬结合数据迁移方案在分级内存场景性能提升明显



注：100%为全本地内存性能，本地远端容量比例3:1

核心能力4-- UBS COMM: 基于灵衢集群的高性能、高可靠以及生态兼容的通信协议

UBS COMM架构图

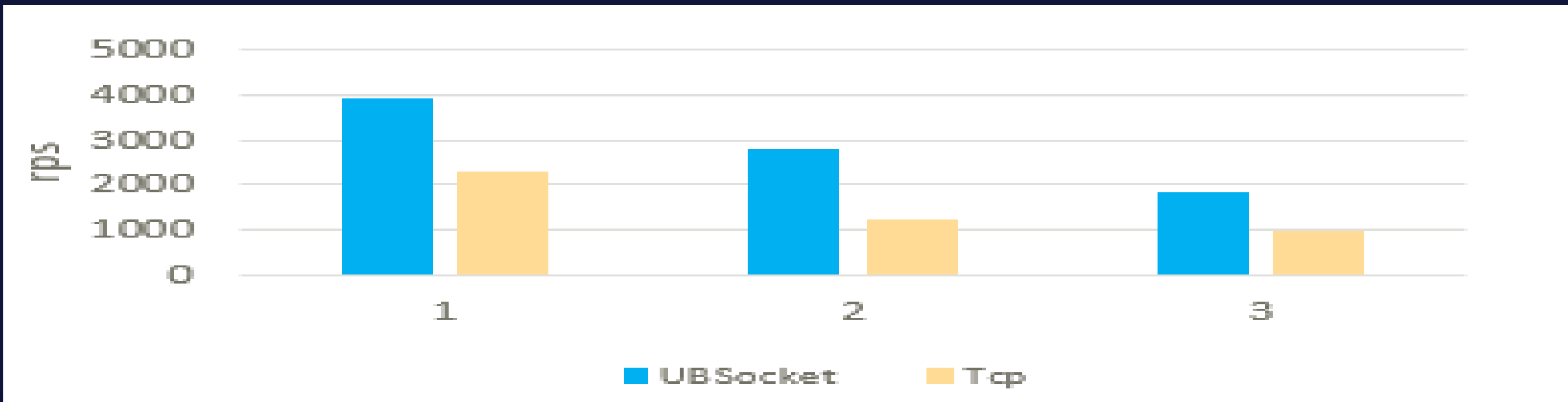


关键技术&价值

一、生态兼容，应用无感，性能提升30%+

➤关键技术1-UBSocket

北向兼容Socket应用， Socket转UB高性能内存语义转发，性能提升30%以上（如下是Redis的测试数据）



➤关键技术2-ROUB

北向兼容Verbs应用， Verbs转UB高性能内存语义转发

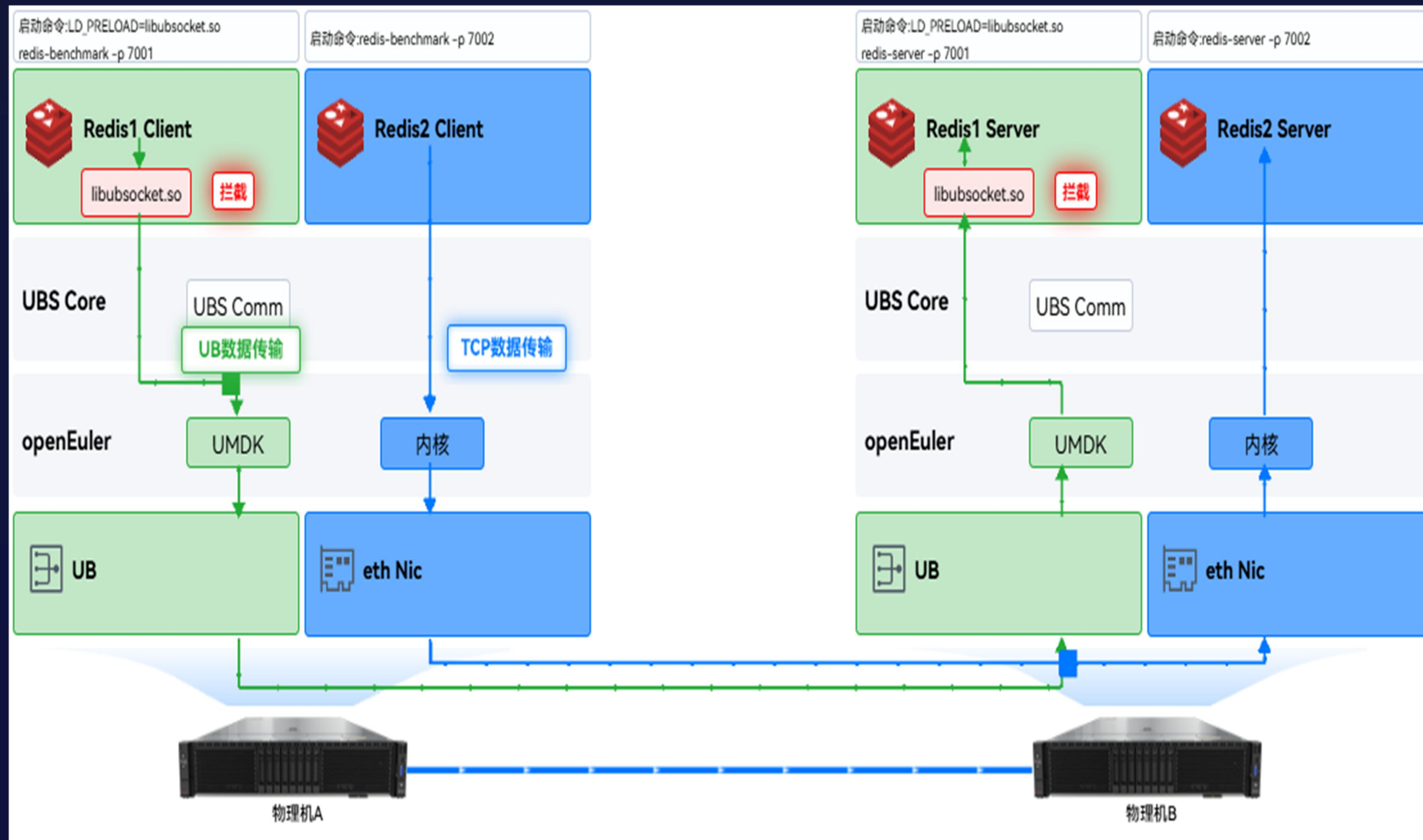
二、极致性能，极致可靠，性能提升50%+

➤关键技术3-HCOM

北向提供统一私有API，南向不同协议的基础通信库，简化编程难度，实现高可靠、易用性、高性能的访问

CASE3: UBSocket应用“零修改”，使能UB高性能网络，通信性能提升

方案说明



方案描述：在双机的集群环境，互联带宽超100Gbps，启动两个Redis实例，其中一个加载UBSocket特性，北向接口维持兼容，南向处理上通过bypass TCP/IP协议走高性能UMDK内存语义提升性能，对比两个实例的性能；

价值场景

① 虚拟化

② 大数据

③ 数据库

关键技术&价值

- 北向接口兼容，开箱即用：**支持POSIX标准SOCKET接口，通过LD_PRELOAD接管原有接口，业务软件无感知；
- 南向发挥灵衢高性能：**bypass内核态TCP/IP协议栈，拦截转换为UB高速通信；
- 典型应用性能提升：**如下Demo验证所示，绿色实例相比蓝色性能提升明显。

测试情况(Redis)吞吐量提升**100%**



CASE4: HCOM(Hyper Comm)打造极致性能、极致可靠的通信库

方案说明

金融极致时延场景

HCOM

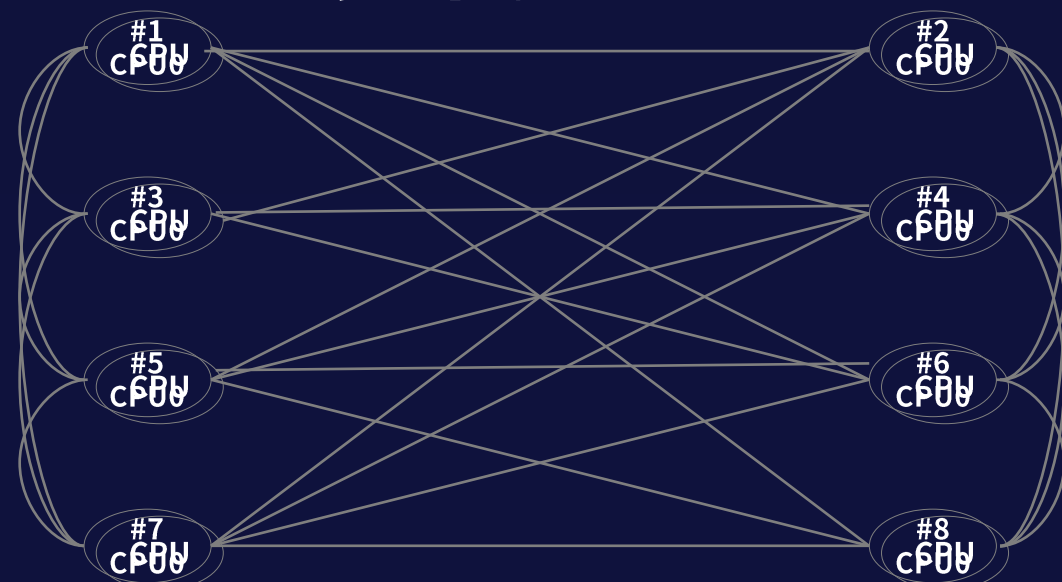
服务层统一API（易用）

南向自适应传输（大包URMA/小包内存语义）

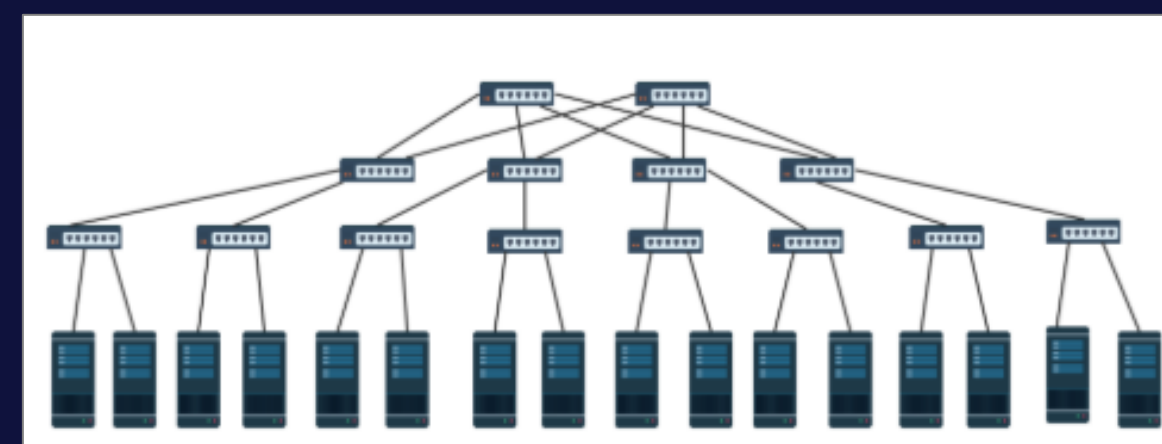
URMA异步语义

LD/ST同步语义

灵衢超节点



传统以太网



价值场景

①金融高性能、高可靠

关键技术&价值

1、统一接口：南向对接不同协议，北向提供统一接口

- 可以从TCP/RDMA演进到灵衢超节点，北向API保持一致

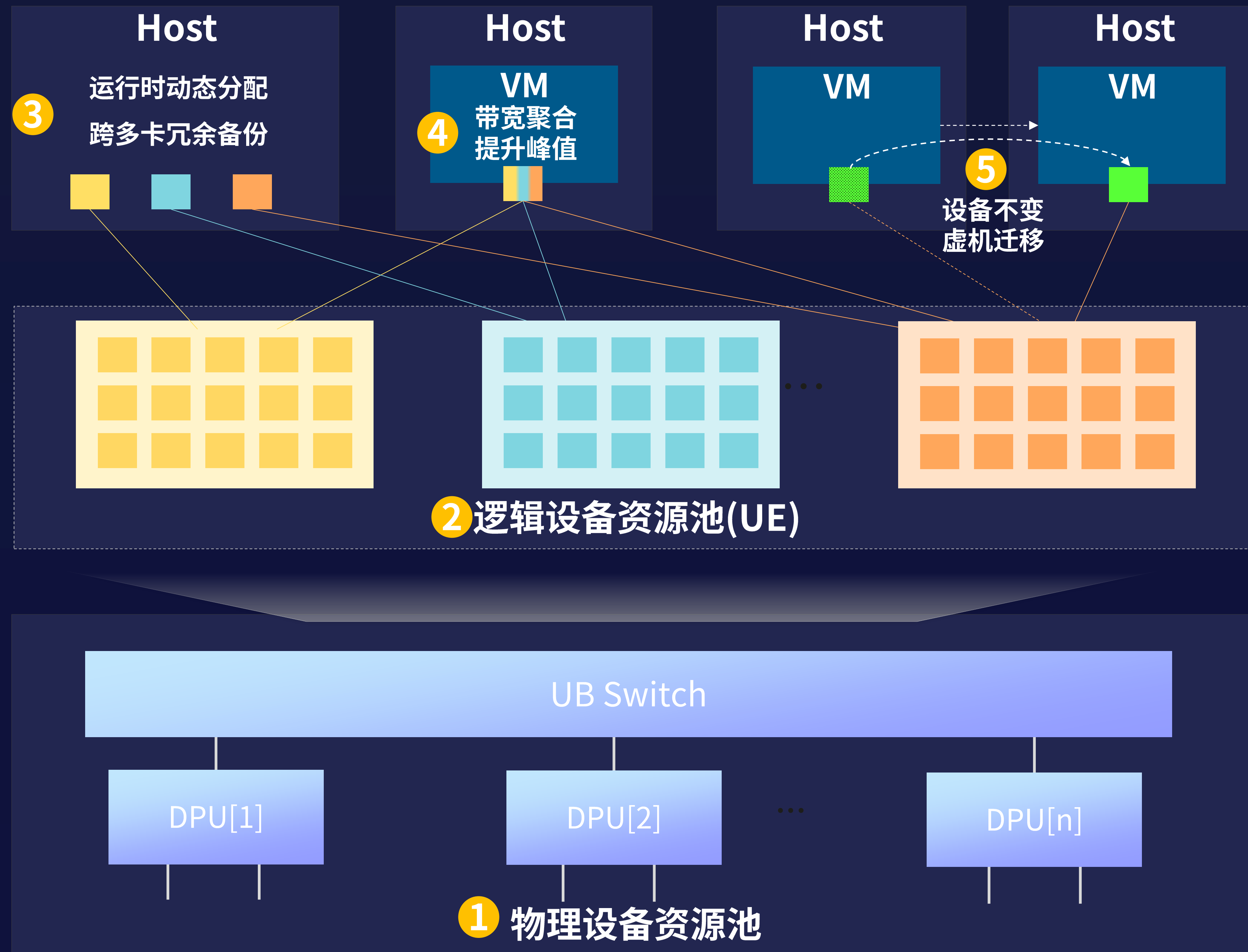
2、极致性能：TB级高带宽，1us级低时延

- 短包场景（比如<8K）采用ld/st同步语义降低小包时延
- 长包场景(>8K) 采用异步传输语义，发挥带宽优势达成TB级带宽

3、极致可靠：多路径多平面毫秒级切换

- 感知物理拓扑，出现故障路径后毫秒级完成链路切换，确保业务可靠性

核心能力5--资源池化 (DPU池化为例): 池化共享动态分配, 提升带宽利用率和峰值, 节约网络硬件成本



性能

- 5 虚拟机迁移, 设备不变, 缩短迁移时间, 提升迁移效率
- 4 多卡带宽聚合, 提升单点峰值带宽, 消减网络瓶颈

可靠性

- 3 主机物理网卡从1:1到1:N, 增加冗余, 提升可靠性

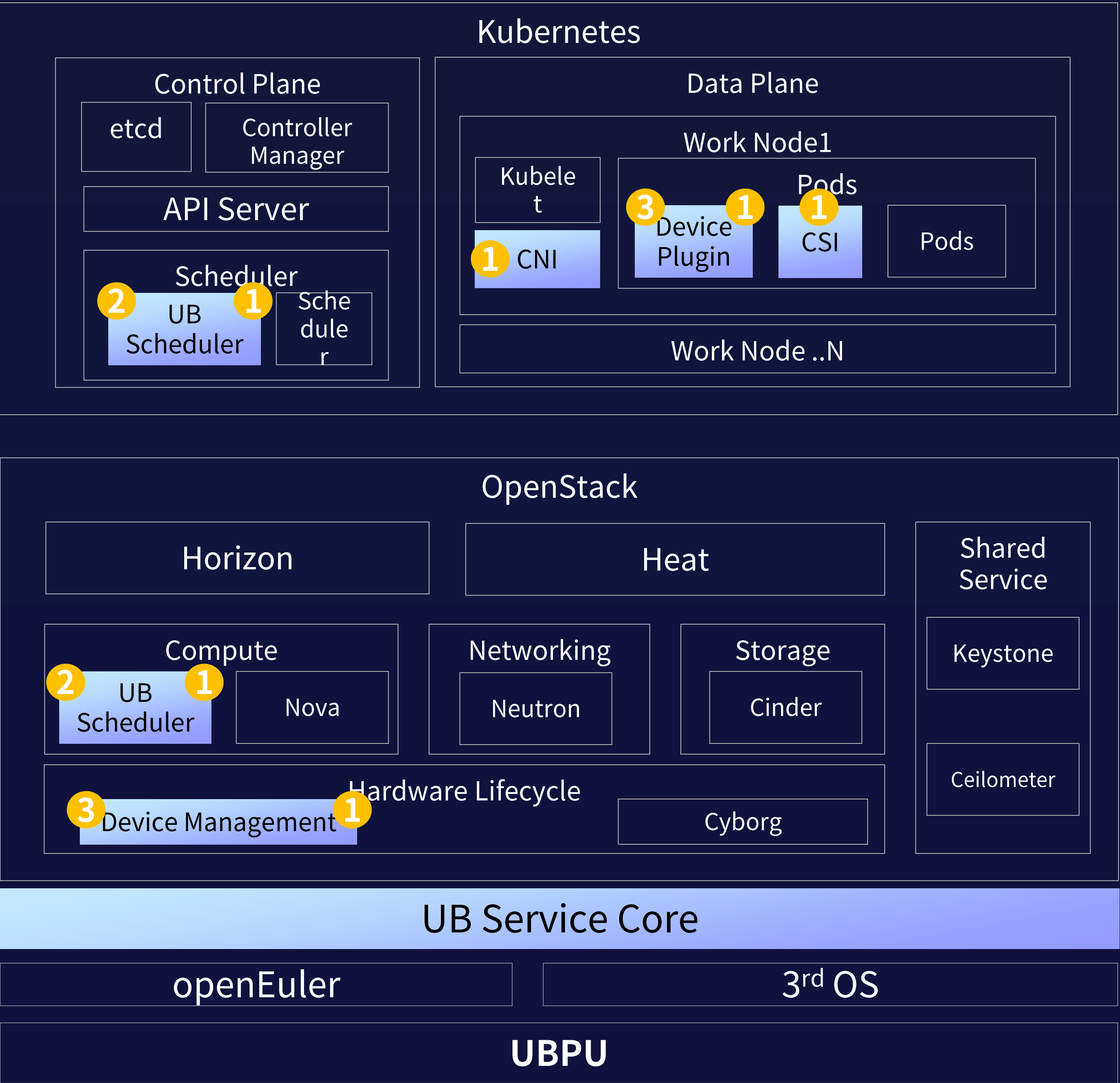
运维

- 2 逻辑设备资源与主机运行时动态绑定, 灵活组合, 无需硬件物理变更调整主机网络端口资源

成本

- 1 带宽资源池化, 运行时动态分配和调度, 提升DPU/网卡利用率, 节约主机网络硬件成本

核心能力6--UB-CloudEnabler：应用平滑接入容器/虚拟机生态，使能超节点池化资源管理调度



应用平滑迁移，提供参考实现，使能超节点调度

1 平滑迁移	- 遵循Kubernetes/OpenStack 扩展机制： K8S: 基于Scheduler机制扩展 OpenStack: 扩展Nova Scheduler能力 - 应用无需修改可直接运行
2 最小修改使能应用	- 部署模板 (yaml/XML) 少量适配使能超节点池化资源管理调度 资源无感借用: 感知超节点池化资源拓扑，提供容器、虚机的资源超分与调度 大规模虚拟机: 基于灵衢的内存借用和迁移功能，支撑创建大规格虚拟机，突破单节点物理限制
3 参考实现	超节点调度: 客户可基于当前Scheduler参考实现定制增强 设备管理抽象: 扩展新增UB设备类型，支持设备状态管理（如发现、重启、网络配置、绑定解绑、多卡直通等）

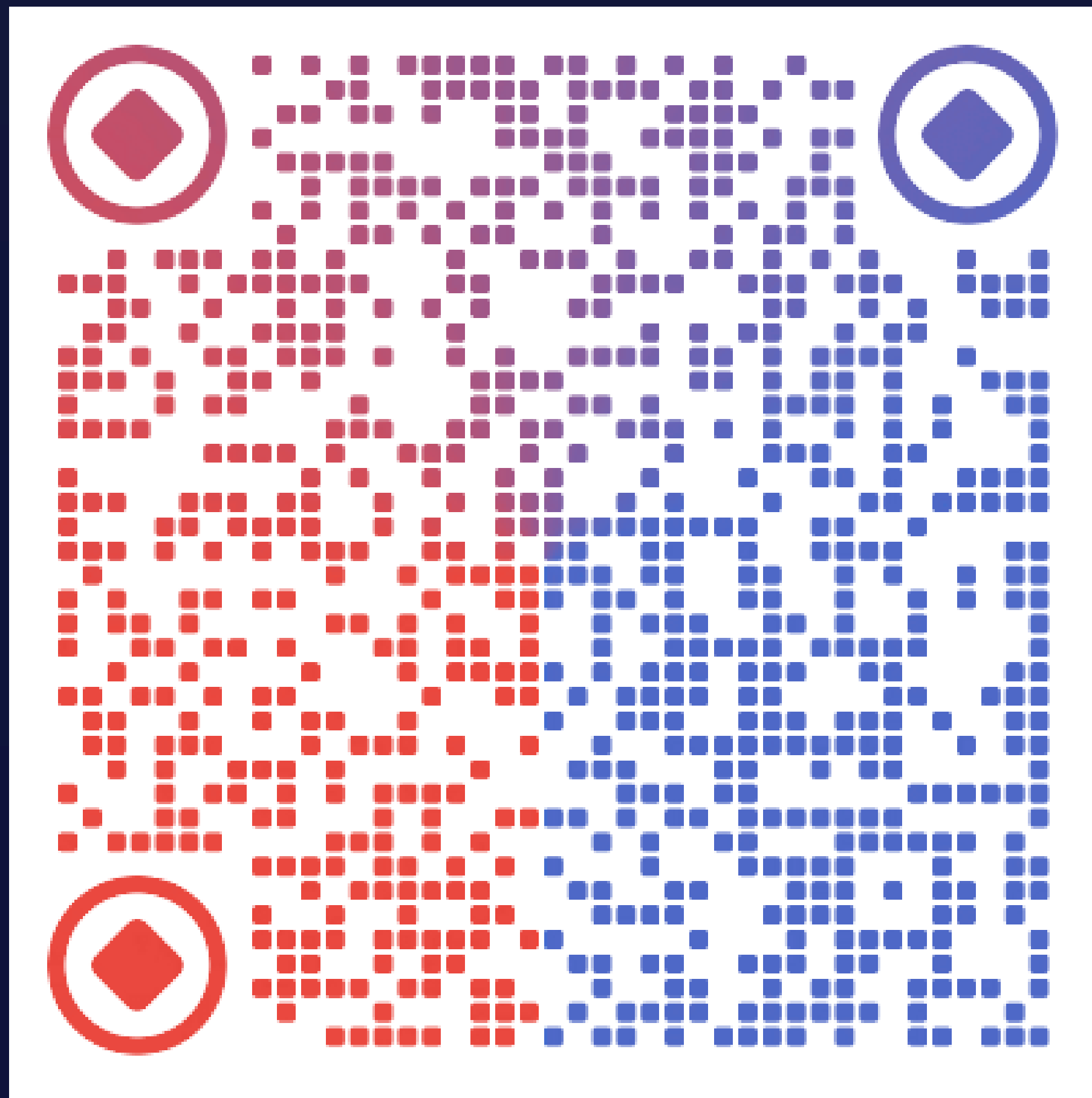
基于灵衢超节点7大核心能力构建十九大系统功能，实现灵衢

超节点在企业业务收益

场景	场景	系统功能	核心能力	类型
7生态使能 UB核心能力 1内存借用 2UB大带宽 3资源池化 4内存共享 5内存语义通信 UBS Core openEuler 软硬协同 灵衢计算集群 6UB Turbo	公共	系统功能一：灵衢总线语义及高可用技术	156	NEW
		系统功能二：集合通信	25	↑
		系统功能三：免拷贝和免序列化的高效通信	5	↑
		系统功能四：容器快起	4	NEW
		系统功能五：AI编译器并行调优	5	↑
		系统功能六：通用编译器 – 需要整合内容	25	↑
	智算	系统功能七：对标NVSHMEM构建aclSHMEM	257	↑
		系统功能八：智算KVCache	23	↑
		系统功能九：MoE专家高效负载均衡	5	↑
		系统功能十：基于内存语义的通算融合MoE加速	5	NEW
		系统功能十一：SSU直驱 – DPU直驱内容尽量整合	2	NEW
		系统功能十二：基于SIMT和UVM的多层级EmbTable	237	↑
		系统功能十三：推理高可用	25	↑
		系统功能十四：训练高可用	25	↑
	通算	系统功能十五：虚拟机热迁移	233	NEW
		系统功能十六：容器热迁移	234	NEW
		系统功能十七：节点/内存故障高可用	13	↑
		系统功能十八：集群资源碎片利用	1	↑
		系统功能十九：UBSocket使能UB高性能网络	27	NEW

UB Service Core 社区信息与相关交流平台

UB Service Core SIG



UB Service Core SIG WeChat



白皮书: <https://www.openeuler.org/zh/projects/ub-service-core/>