三、分离式内存的发展趋势

讲者：

上海交通大学 计算机系 SAIL实验室

2023年2月

上海交通大学
SHANGHAI JIAO TONG UNIVERSITY

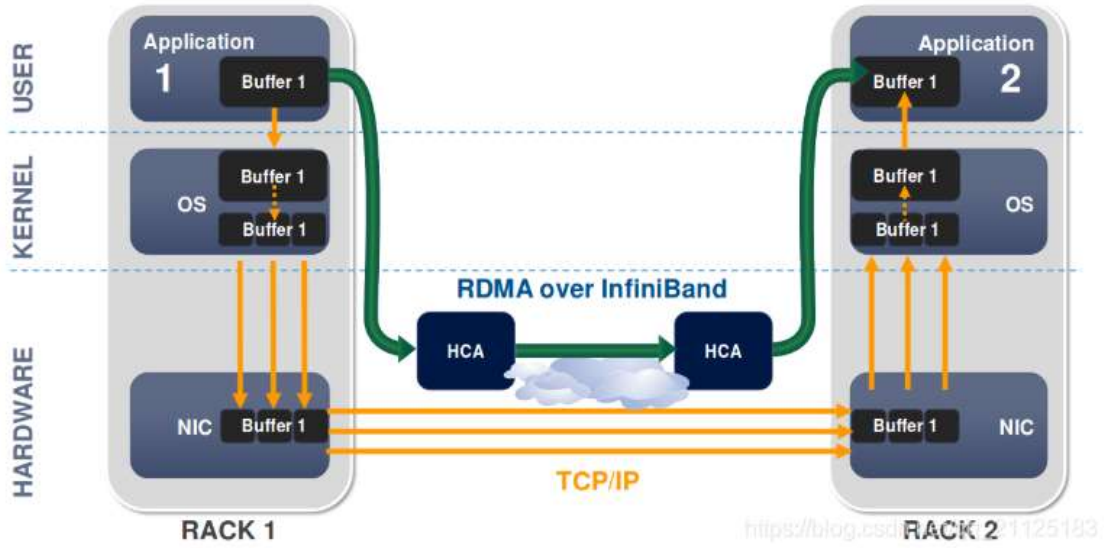
目录

1.网络技术：整体趋势

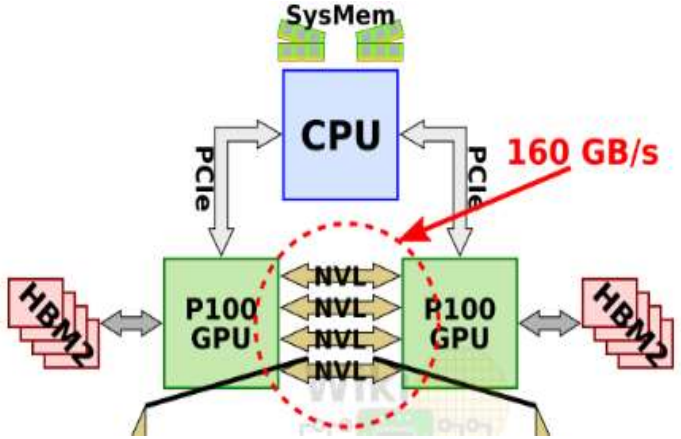
- 高速
  - 任务执行模式
  - 数据卸载的性能成本和潜力
- 安全
  - 分离式内存的市场空间
  - 分离式内存的通信成本
- 可扩展
  - 分离式内存的组织架构
  - 分离式内存的资源管理模式

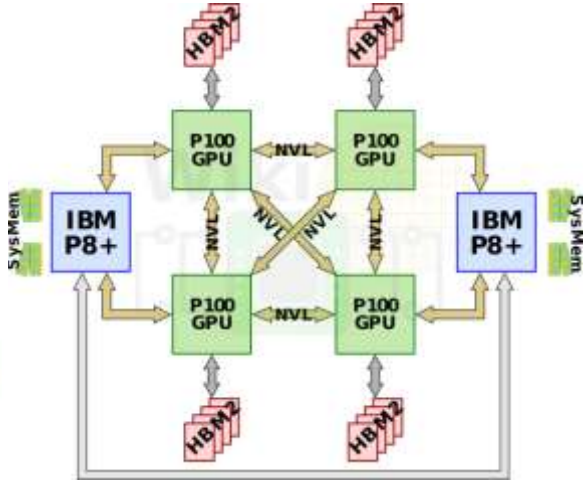
1.网络技术：先进网络技术及性能

RDMA



NVLink



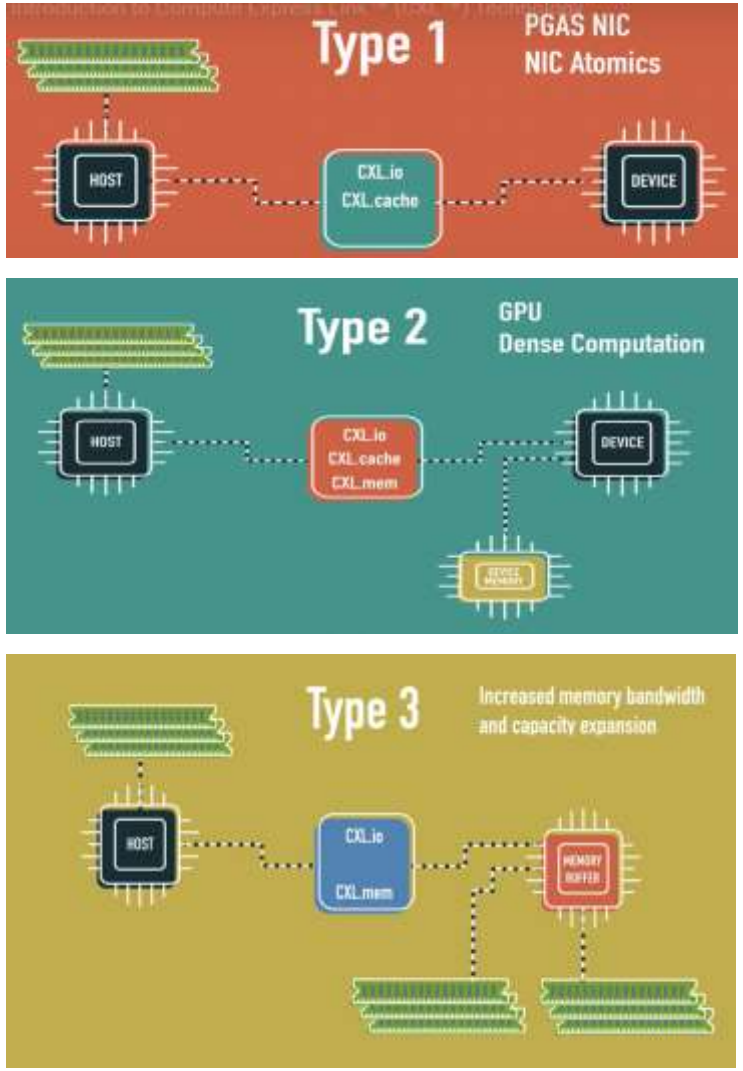


网卡版本	带宽	接口
ConnectX-5	100Gb/s	16x PCIe Gen3
ConnectX-6	50Gb/s	8x PCIe Gen4
ConnectX-5	200Gb/s	16x PCIe Gen4
ConnectX-7	400Gb/s	32x PCIe Gen5

1.网络技术：先进一致性通信协议

Compute Express Link (CXL)：

- **基本概念：**计算互联CXL 标准支持创建内存池和加速器池的互联，支持分离式内存和可组合式虚拟机的构建，从而更加高效地使用内存资源。
- **硬件依赖：**PCIe 5.0版本及相关主板（暂时没有商用）
- **发展状况：**CXL 3.0版本在 2022 年 8 月发布
- **种类：**
  - CXL.io (Host与加速器)
  - CXL.cache （加速器与Disaggregated memory ）
  - CXL.mem （Host与 Disaggregated memory）
- **优势：**
  - 延迟比RDMA低一个数量级，与DIMM相当（理论）
  - 内存访问机制路径短
  - 能够解决数据一致性问题

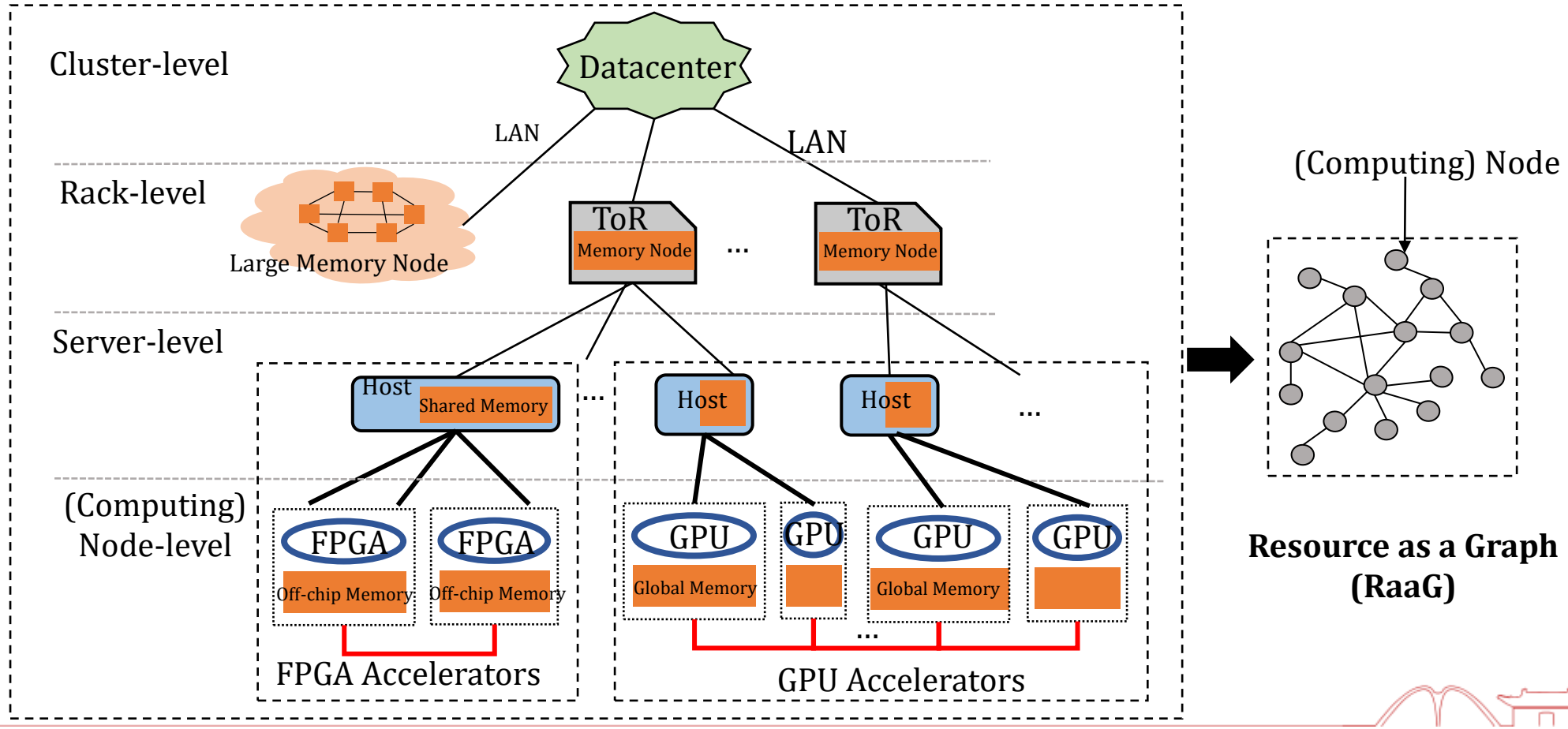


1.网络技术：对内存互联结构的影响

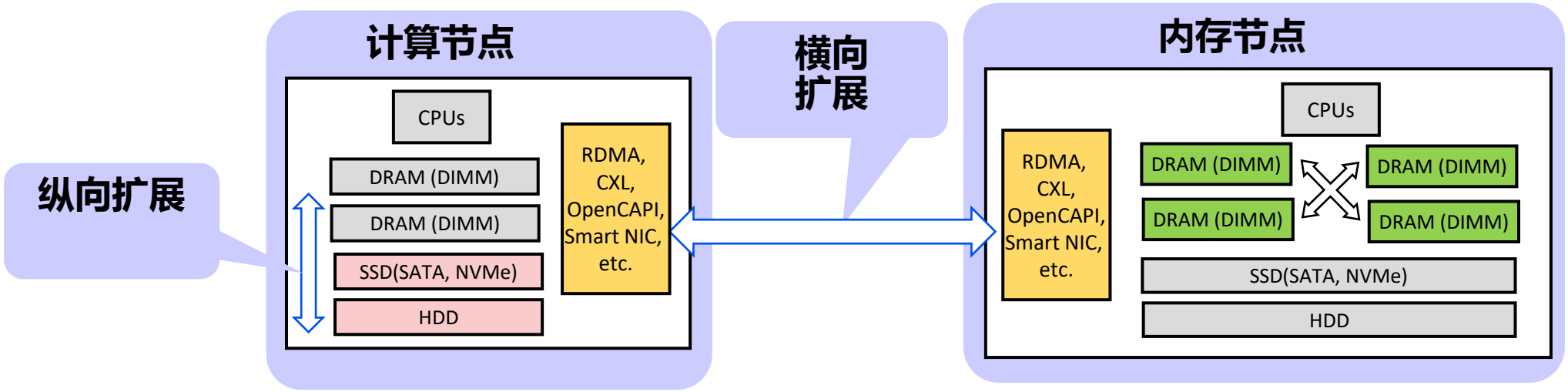
网络硬件设备、互联方式的更新对系统中内存互联结构有较大影响

可插拔内存设备->灵活的内存池构建

高速互联->多层级内存池构建



1.网络技术：对远内存访问模式的影响



（节点内）纵向远内存：

→ 更高的I/O 延时
→ 更小的数据传输带宽
→ 很大的非共享的内存容量
→ 被动交换出数据

（节点间）横向远内存：

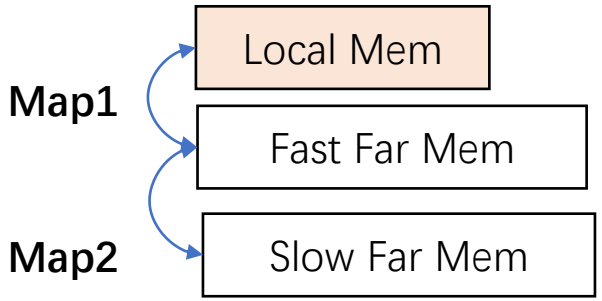
→ 更快的远内存访问
→ 更大的数据传输带宽
→ 灵活但有限的内存容量
→ 可以主动或者被动卸载数据

结合横向远内存和纵向远内存的优势可以得到更高的内存效率和任务吞吐

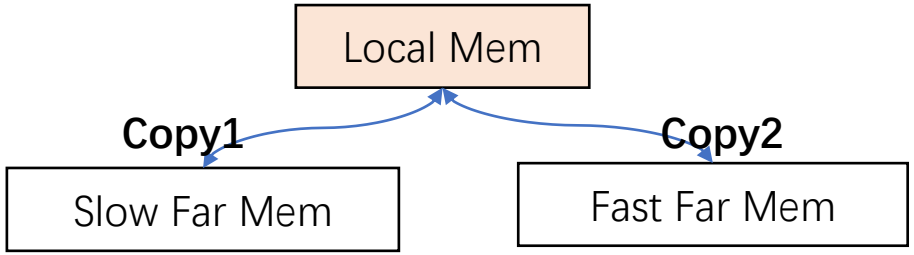
[1] Tmo: transparent memory offloading in datacenters, ASPLOS'22
[2] Can far memory improve job throughput?, EuroSys'20

1.网络技术：对内存层级的影响

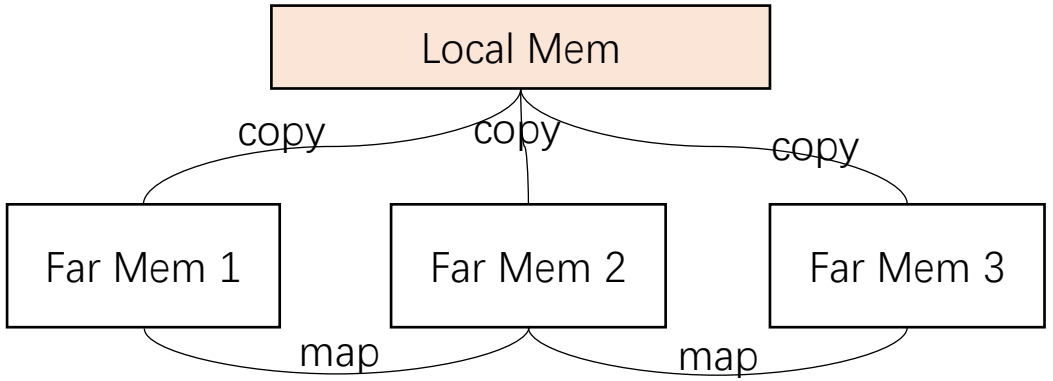
基于swap的传统层级映射式[1]



基于一致性的直接缓存式[2]



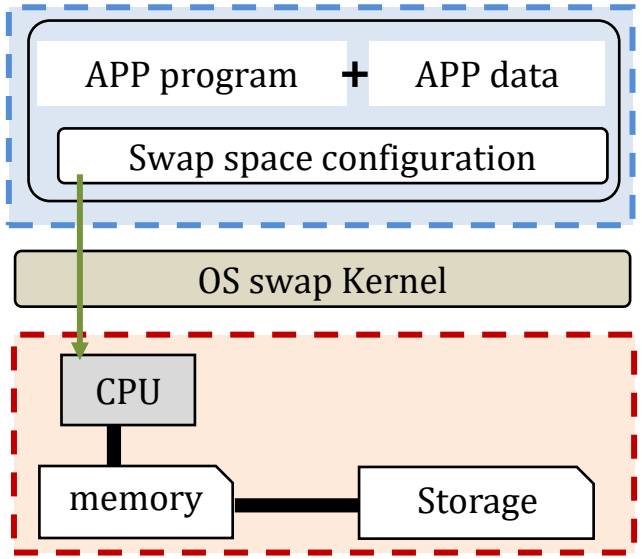
Cache与swap混合的新型层级



- Exclusive：
  - Swap-and-Map-based migration
  - 充分利用空间
  - 性能不如cache好
- Inclusive：
  - Copy-based caching
  - 性能好
  - 浪费空间
- Inclusive/Exclusive：
  - 性能好
  - 空间合理利用

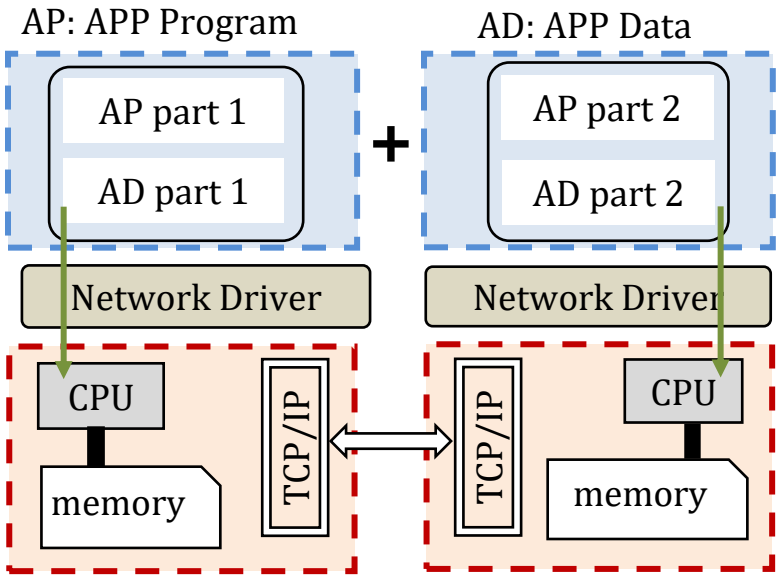
[1] Transparent and Lightweight Object Placement for Managed Workloads atop Hybrid Memories, VEE 22
[2] Panthera: Holistic Memory Management for Big Data Processing over Hybrid Memories, PLDI 19

1.网络技术：对应用执行模式的影响



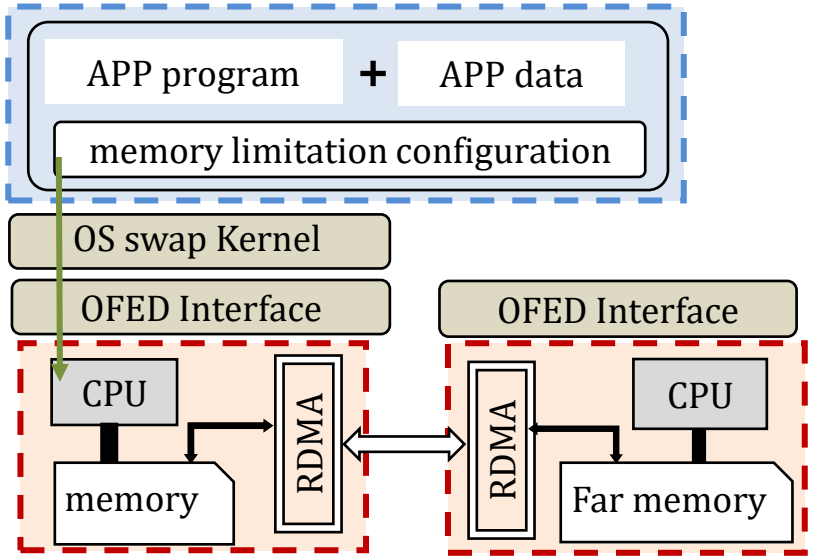
单节点处理

- 使用本地的存储空间
- 额外数据访问依赖I/O
- 应用透明



分布式处理

- 使用网络传输较小通讯信息
- 任务级别并行度增大
- 应用非透明



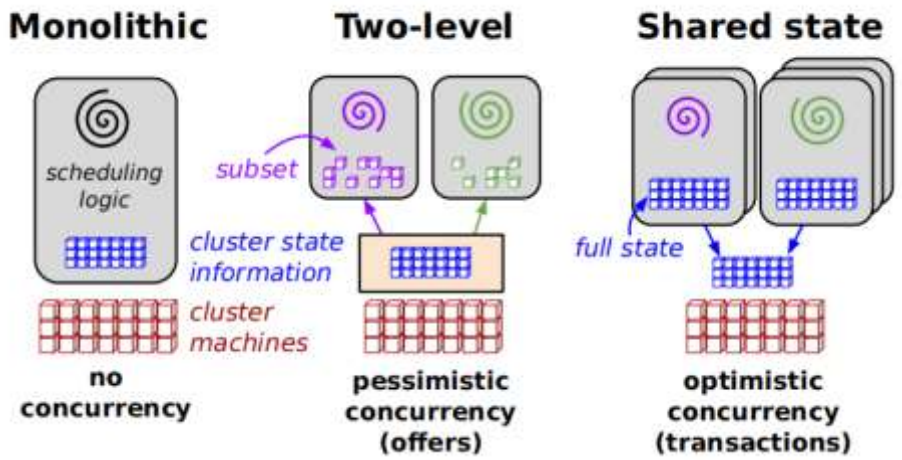
远内存处理

- 使用网络传输大块数据
- 任务和数据级别并行度增大
- 应用透明

1.网络技术：对资源调度策略的影响

网络部署规模、资源总量的动态变化对资源分配的策略有较大影响

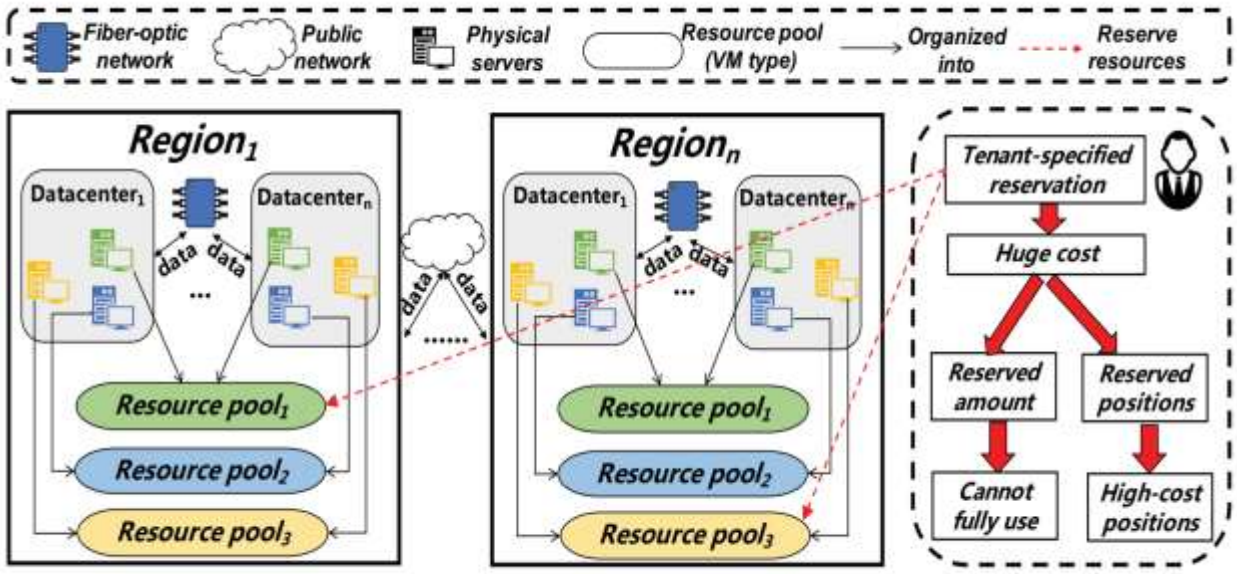
规模过大->分布式共享视图



- 建立共享视图资源架构，对于集群资源进行统一管理调度
- 使用智能预测模型，根据历史数据规划资源分配情况。

Omega'EuroSys13

资源受限->基于带宽的资源调度



- 建立通信开销在资源和性能视角下的数学模型。
- 将通信开销纳入整体调度算法考虑范畴，影响任务分配优先级。

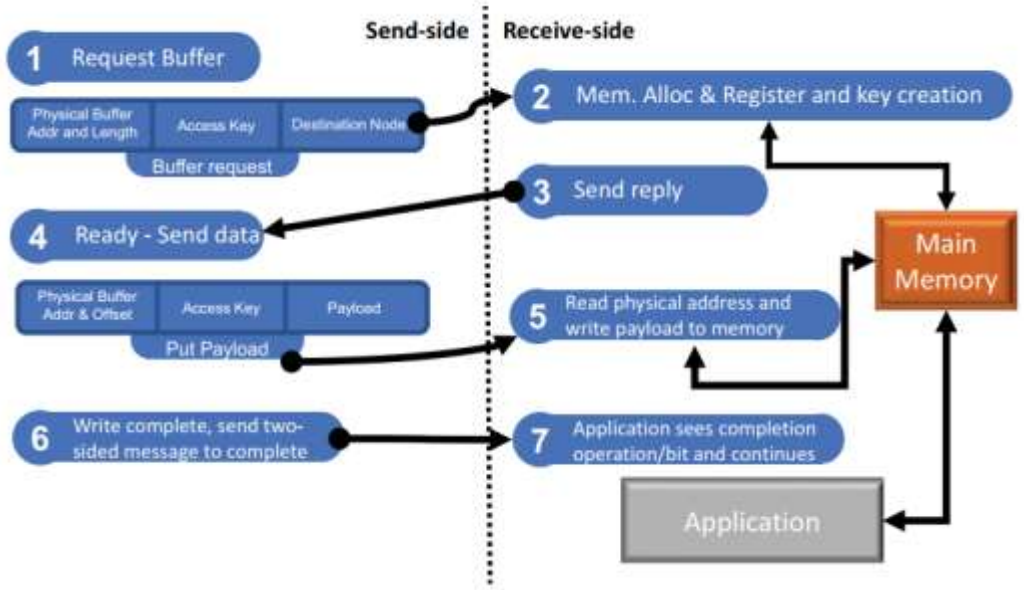
ROS'SoCC22

1.网络技术：对内存访问框架设计的影响

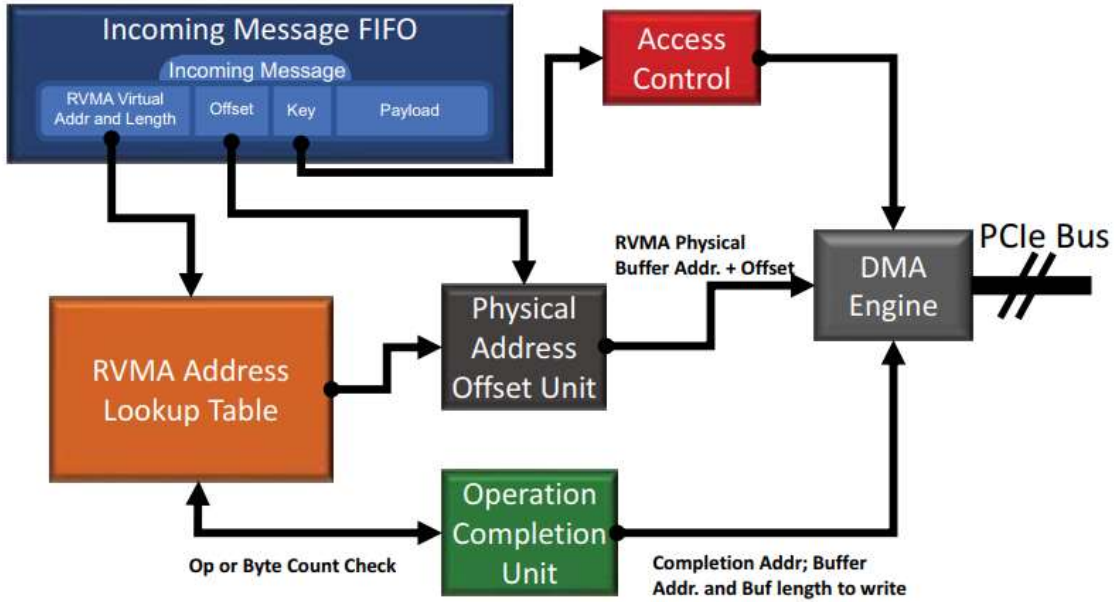
为了网络通道的安全性，人们设计了保证数据传输鲁棒性和安全性的远内存访问框架

- **Better usability**
- **Fault tolerance**
- **High performance**

传统RDMA 读写操作



为传输信息增加一层封装，使用目标物理内存地址的底层详细信息，只使用基本的虚拟邮箱地址和目标邮箱注册的缓冲区中的偏移量



[1] RVMA: Remote Virtual Memory Access, IPDPS'2021

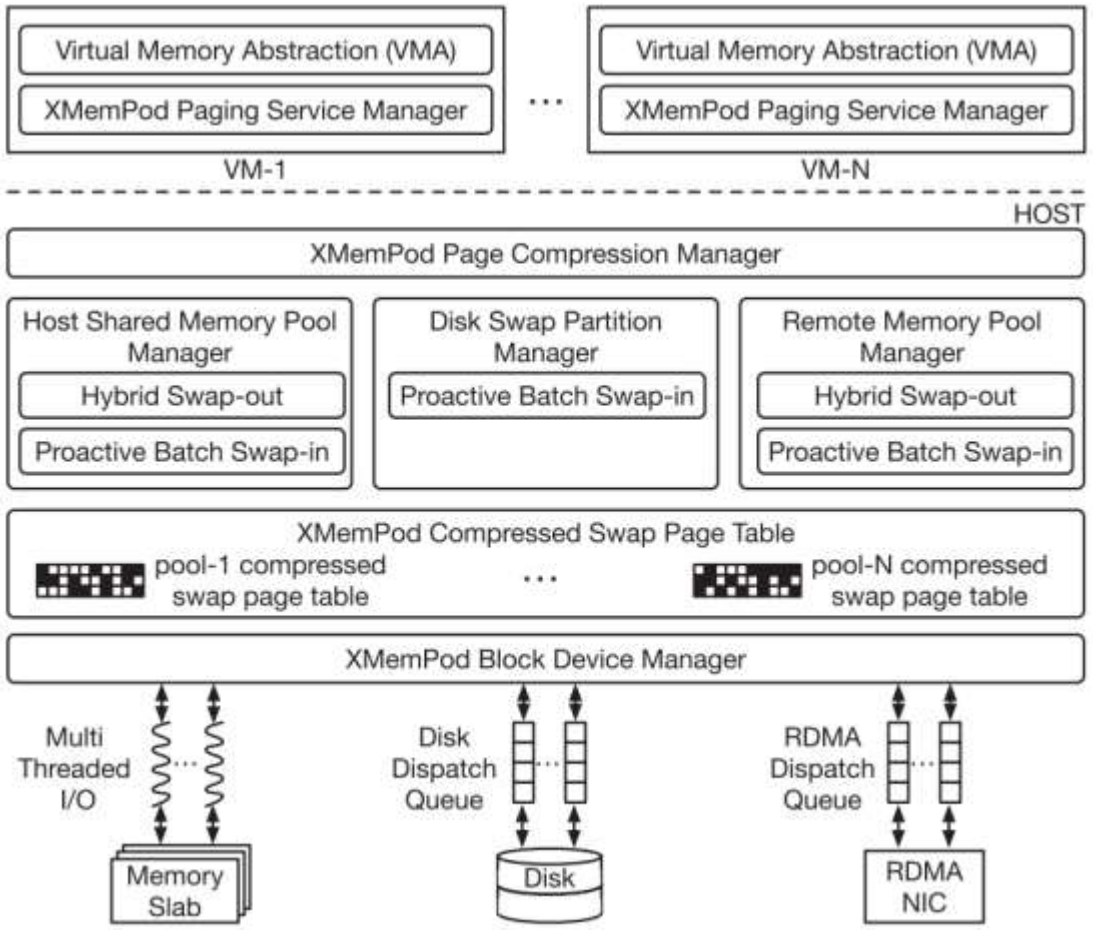
上海交通大学
SHANGHAI JIAO TONG UNIVERSITY

目录

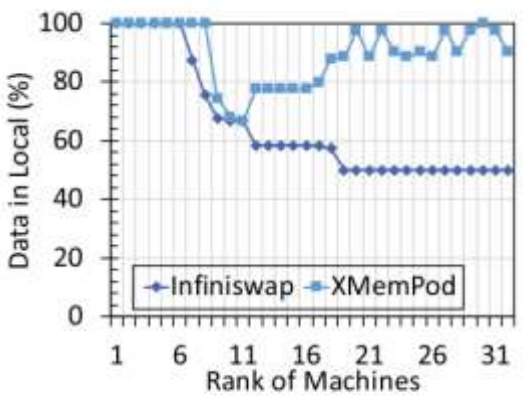
2.软件技术：整体趋势

- 高效率
  - 整体延时低
  - 并行度高
  - 利用率高
- 高易用
  - 应用透明
  - 编程简单
- 支持安全
  - 容错处理
  - 数据安全
- 支持异构
  - 性能更高
  - 资源分配合理

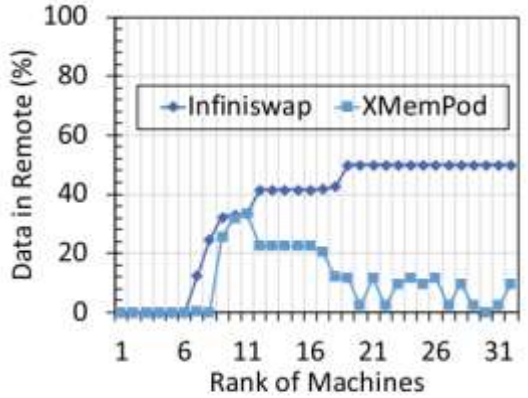
2.软件技术：硬件虚拟化部署->高并行度



- XMemPod提供高效、透明、动态的可用内存共享，这些可用内存在同一主机或集群中的不同虚拟机之间分解。
- XMemPod提供了一个分层内存扩展框架，允许虚拟机上内存密集型工作负载先扩展虚拟化主机内存，然后扩展远程内存，然后才求助于外部磁盘



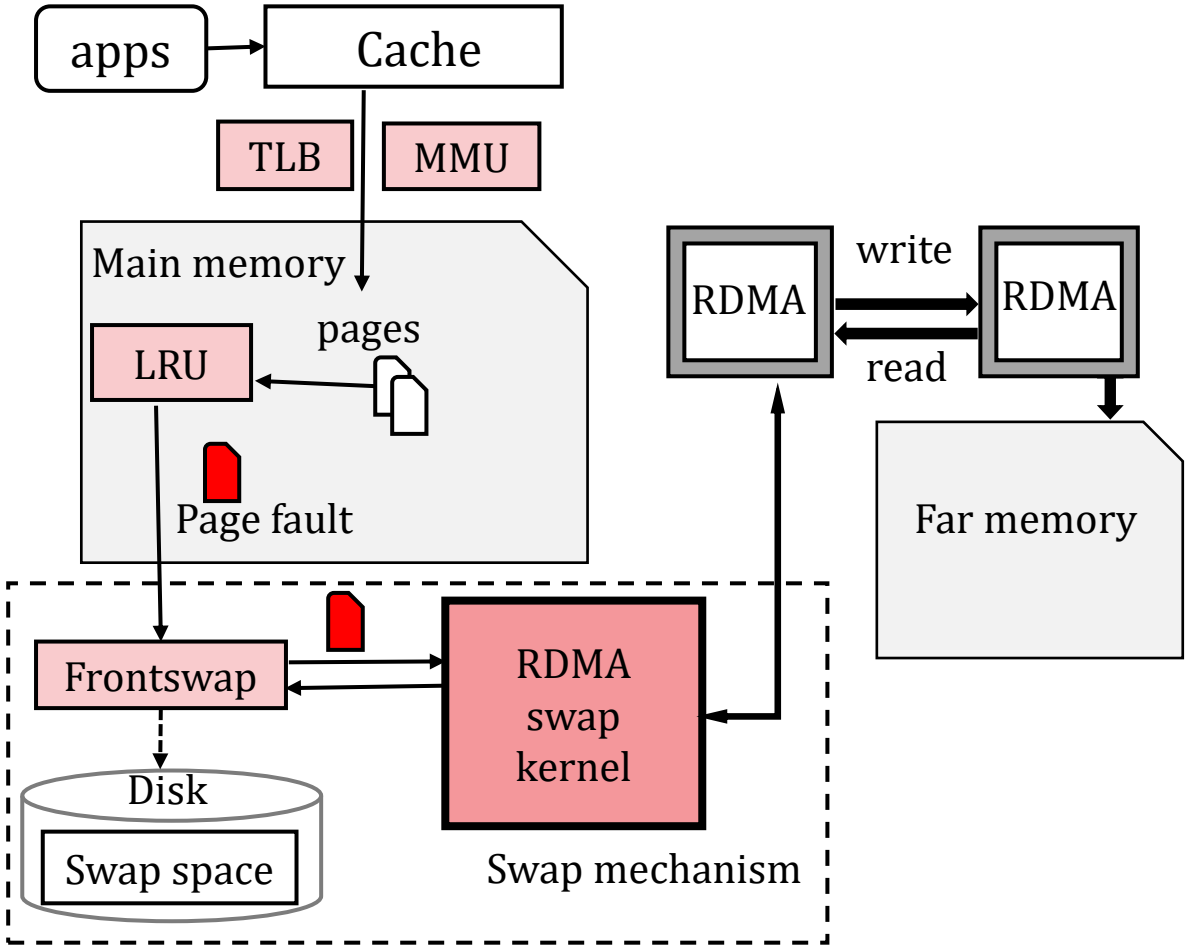
(a) Data Stored in Local (%)



(b) Data Stored in Remote (%)

[1] Hierarchical Orchestration of Disaggregated Memory, TC'2020

2.软件技术：系统软件部署->应用透明



当前的优势:

- 应用透明
- 比本地存储性能要好

当前的问题:

- 内核开销大，不能利用RDMA不实用内核的优势，有明显的上下文切换开销
- 基于swap被动卸载数据，无法灵活配置后端卸载
- 基于固定的页面大小的数据卸载浪费带宽

2.软件技术：应用框架部署->编程简单

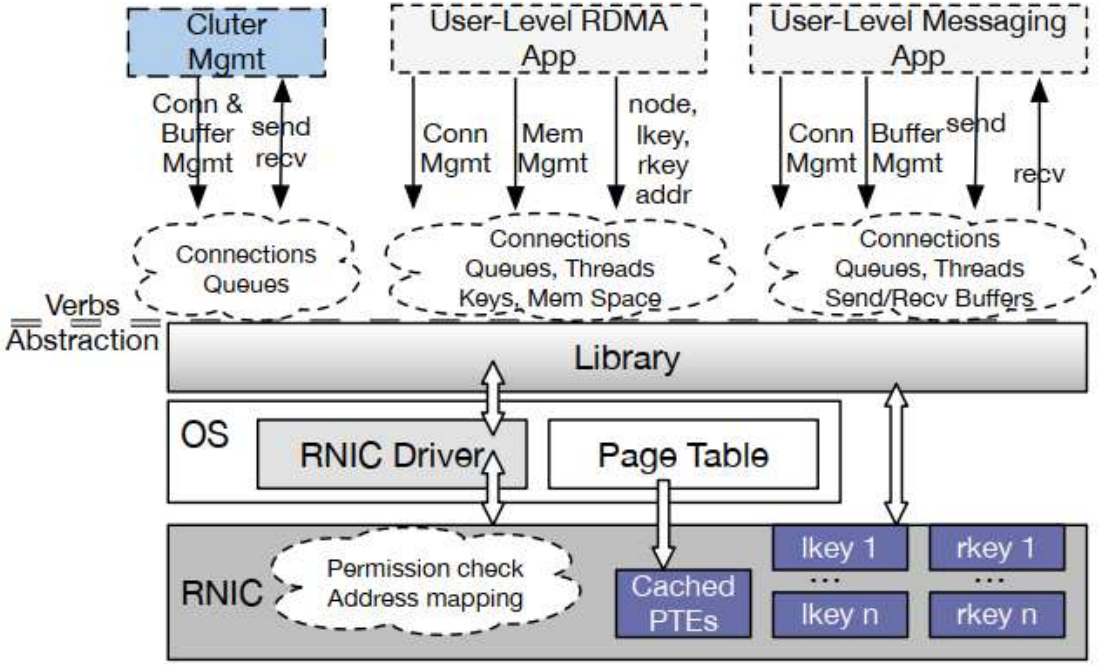


Figure 1: Traditional RDMA Stack.

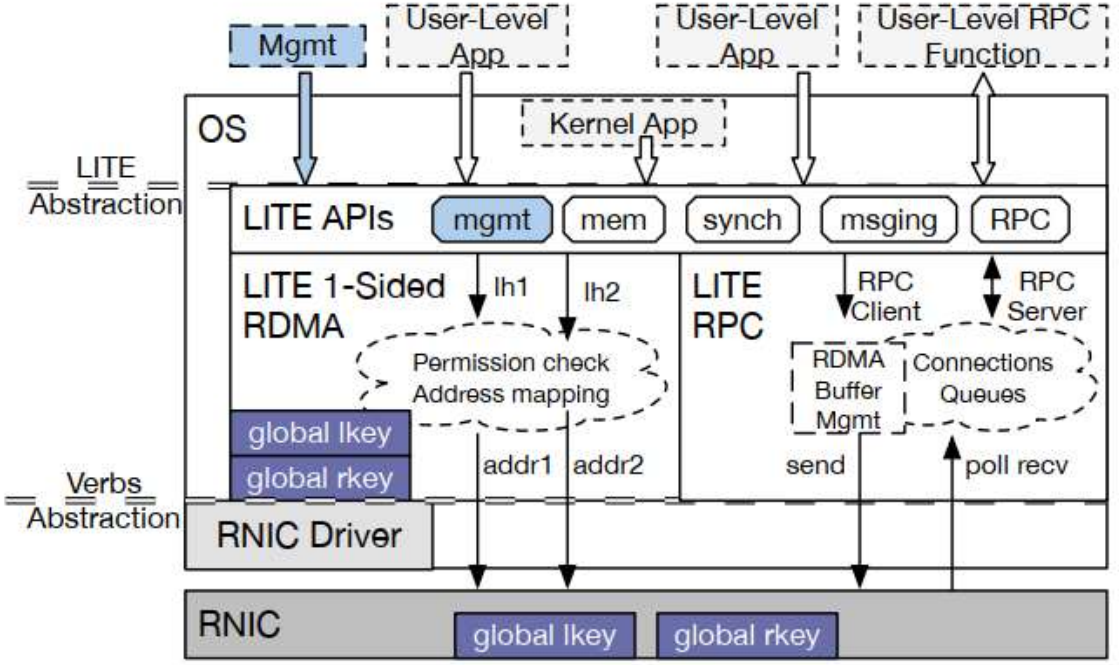
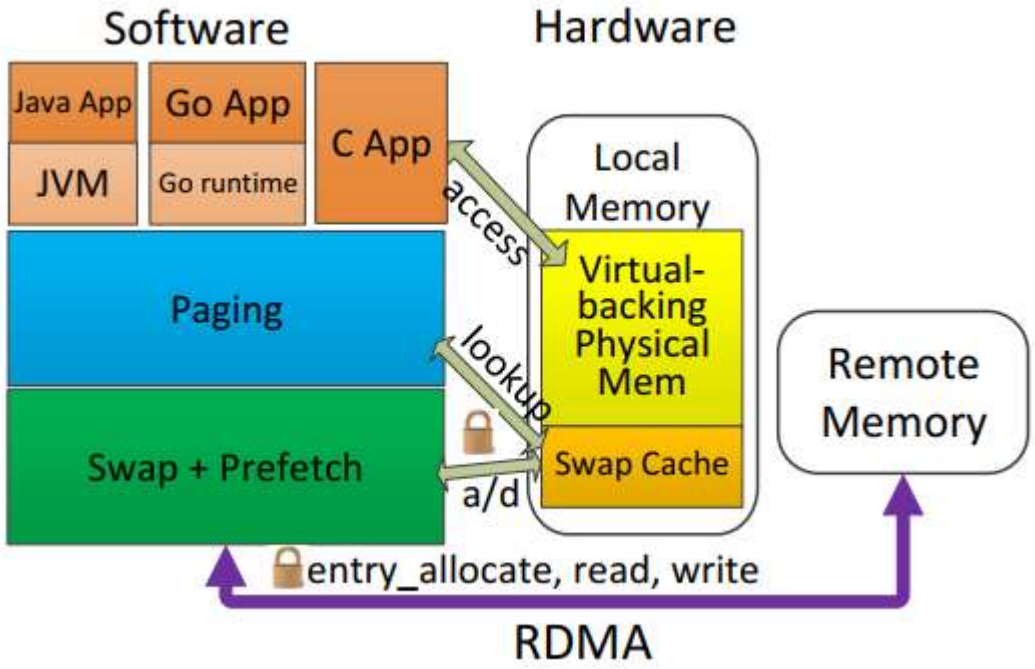
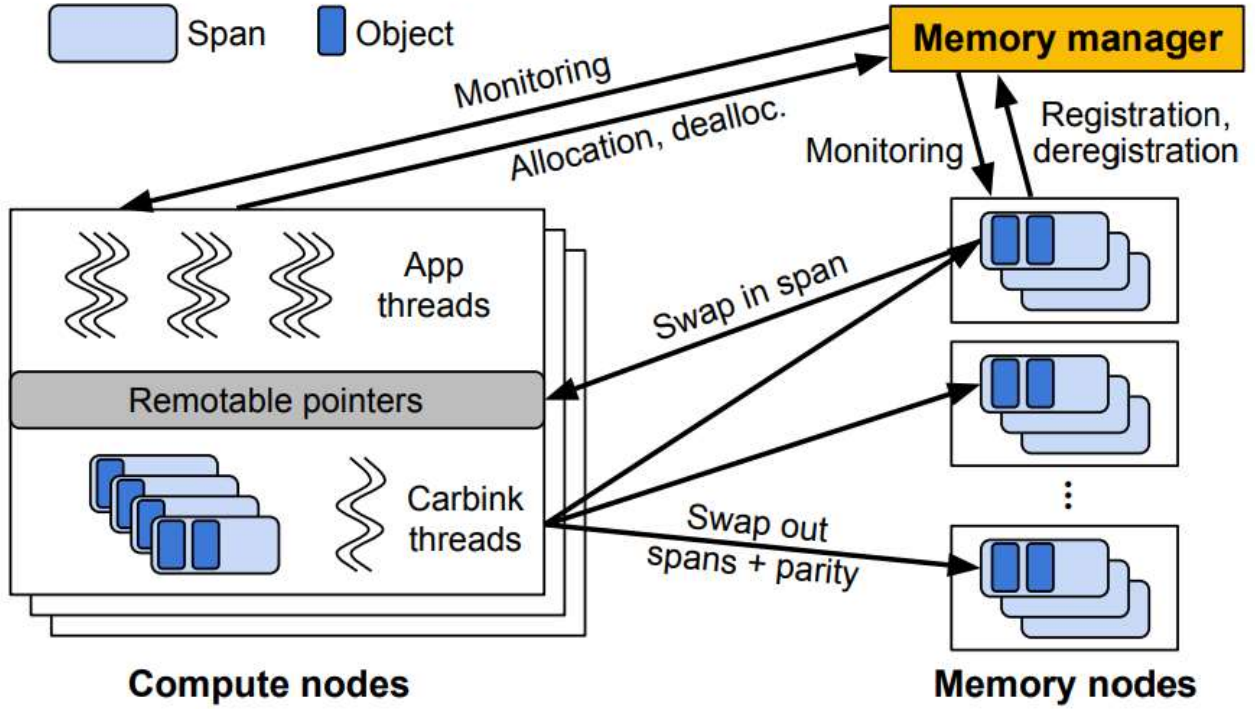


Figure 2: LITE Architecture.

- 利用RDMA原语设计应用接口，通用内核级间接虚拟化RDMA，最小化性能开销。
- 应用程序可以轻松地使用LITE执行低延迟的网络通信和分布式操作。
- 可以通过LITE来管理和保护其资源，从而降低其硬件复杂性和rnic上的内存。

2.软件技术：数据安全->隔离与容错



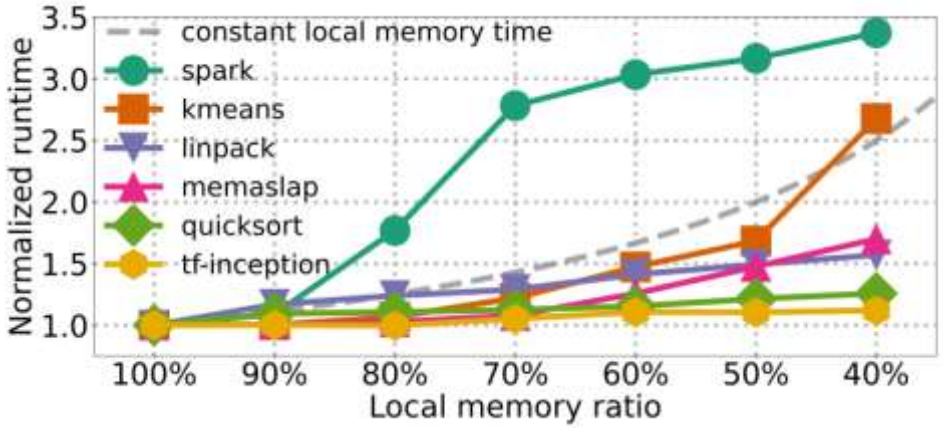


数据隔离

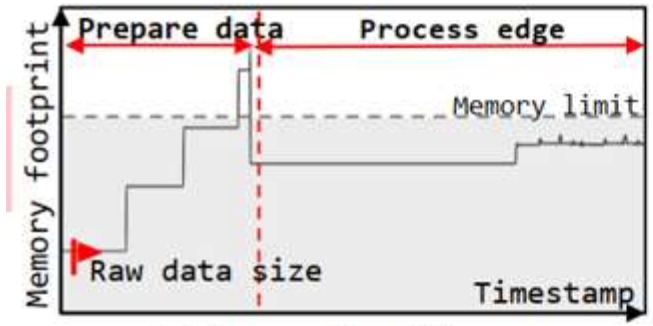
容错处理-提供冗余

需要一种安全的方式来更新远程内存中保存的数据，为数据提供空间、时间和引用上的安全性

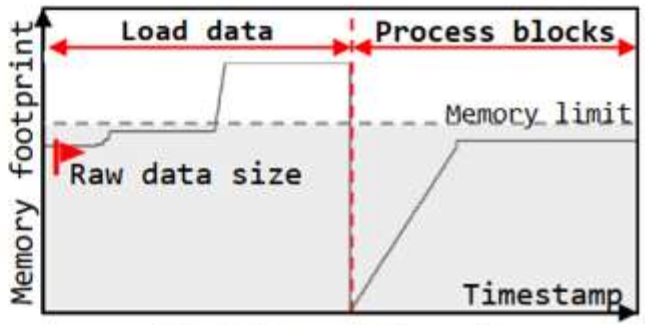
2.软件技术：资源管理->高利用率



任务在远内存系统上表现有差异

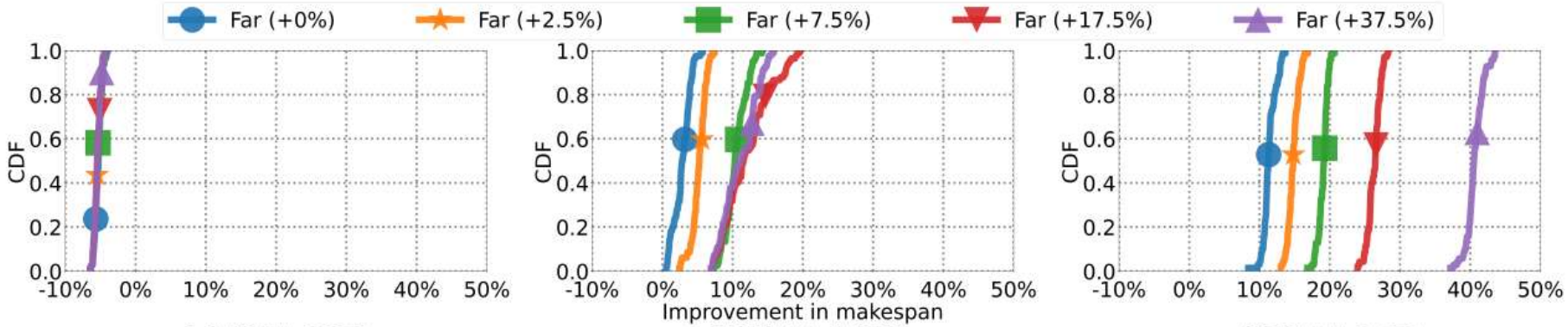


(a) Pagerank on Ligra



(b) BFS on Gridgraph

任务在远内存系统上的差异具有动态性



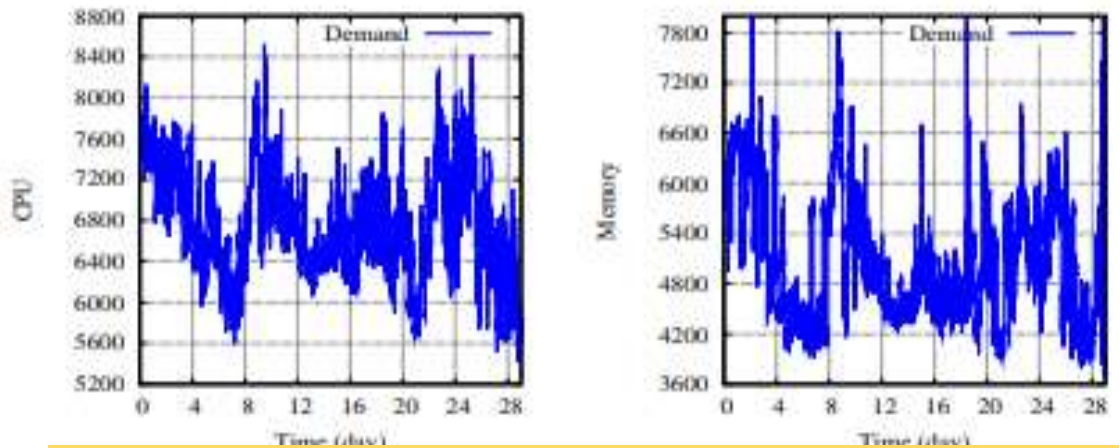
(a) [0.75, 0.85]

(b) [1.15, 1.25]

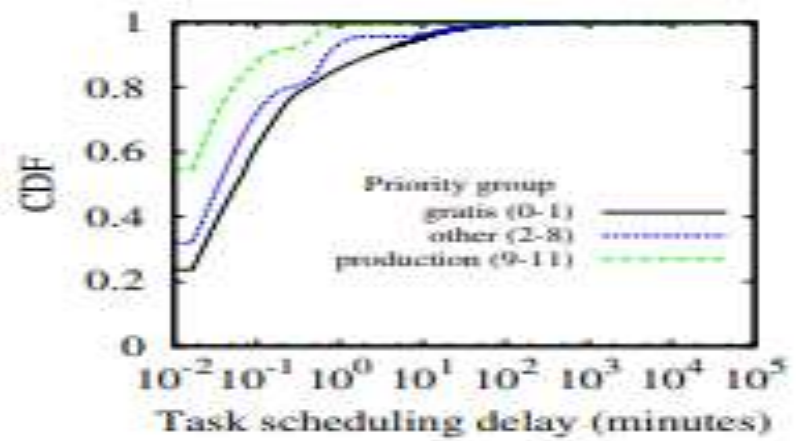
(c) [1.55, 1.65]

根据任务表现限制其内存并触发远内存访问，可以提升利用率和任务吞吐量

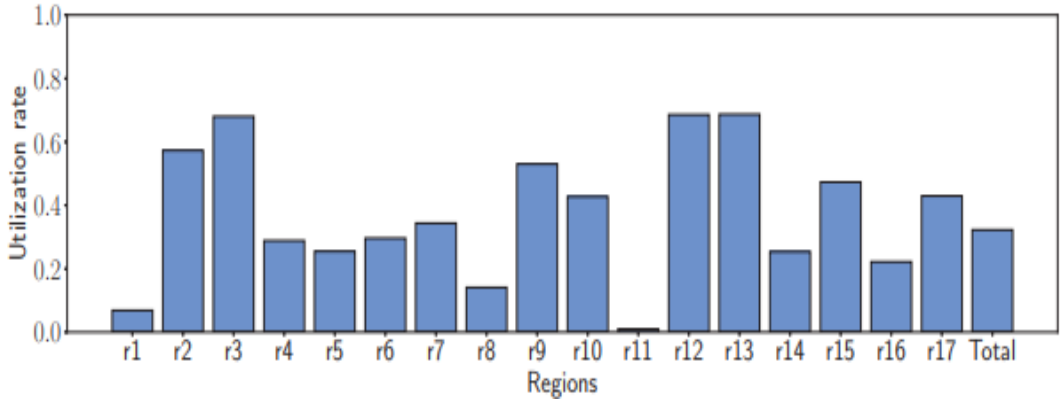
2.软件技术：异构资源调度



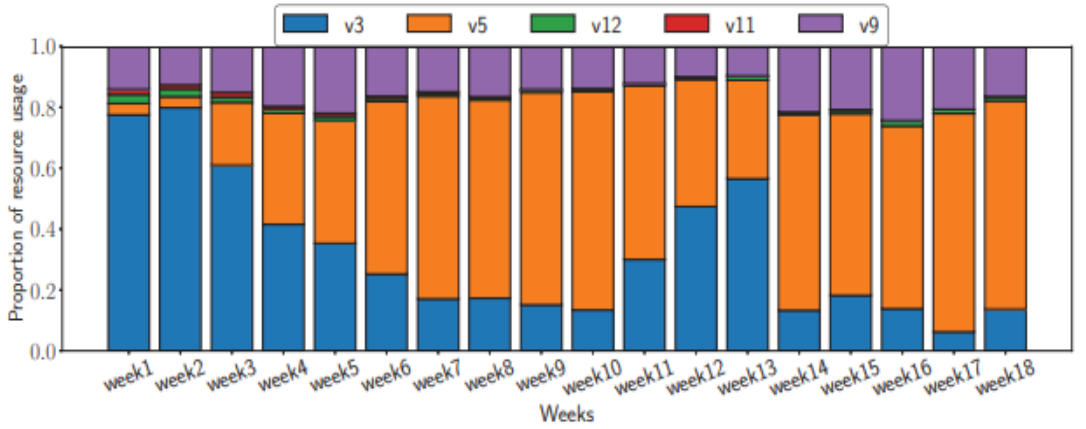
任务对于异构资源需求差异



异构资源调度决策开销巨大



分布式数据中心资源利用不均

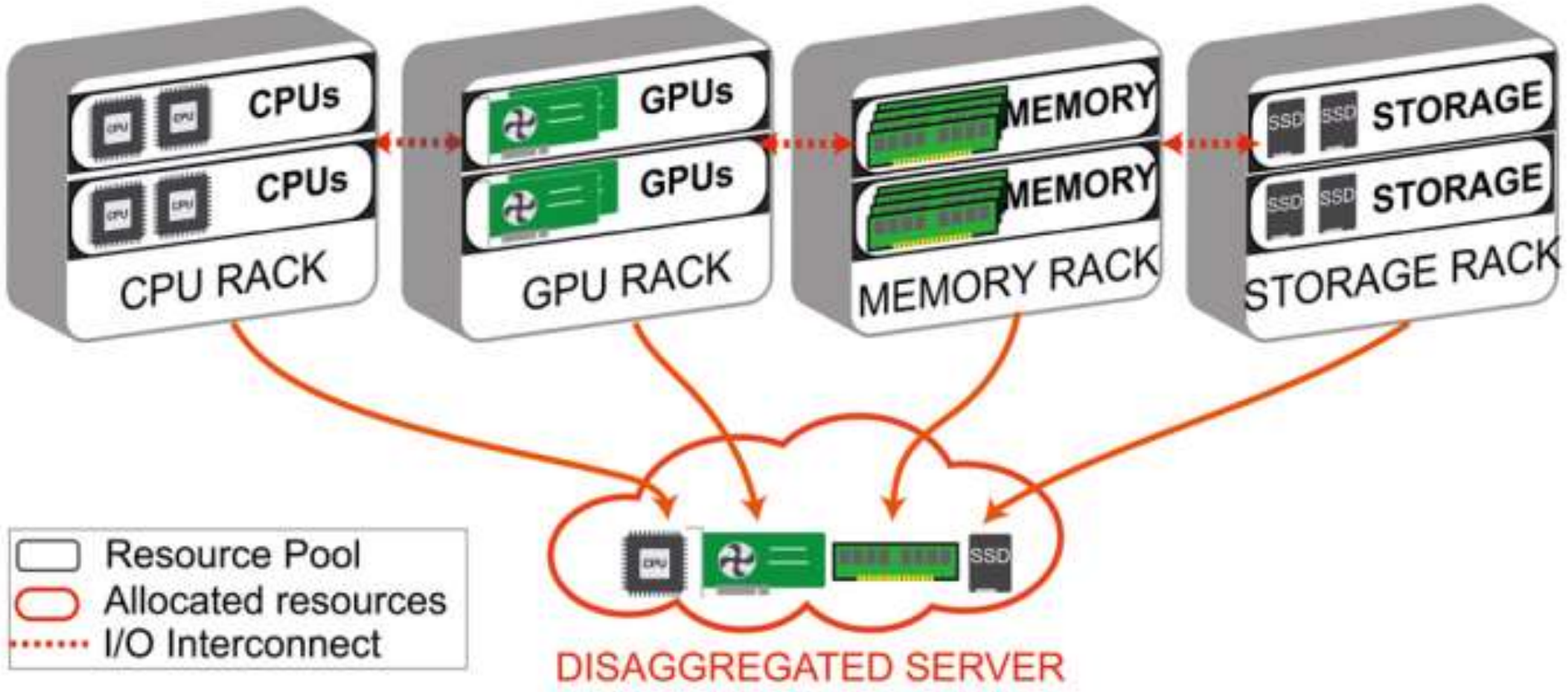


分布式数据中心资源决策不均

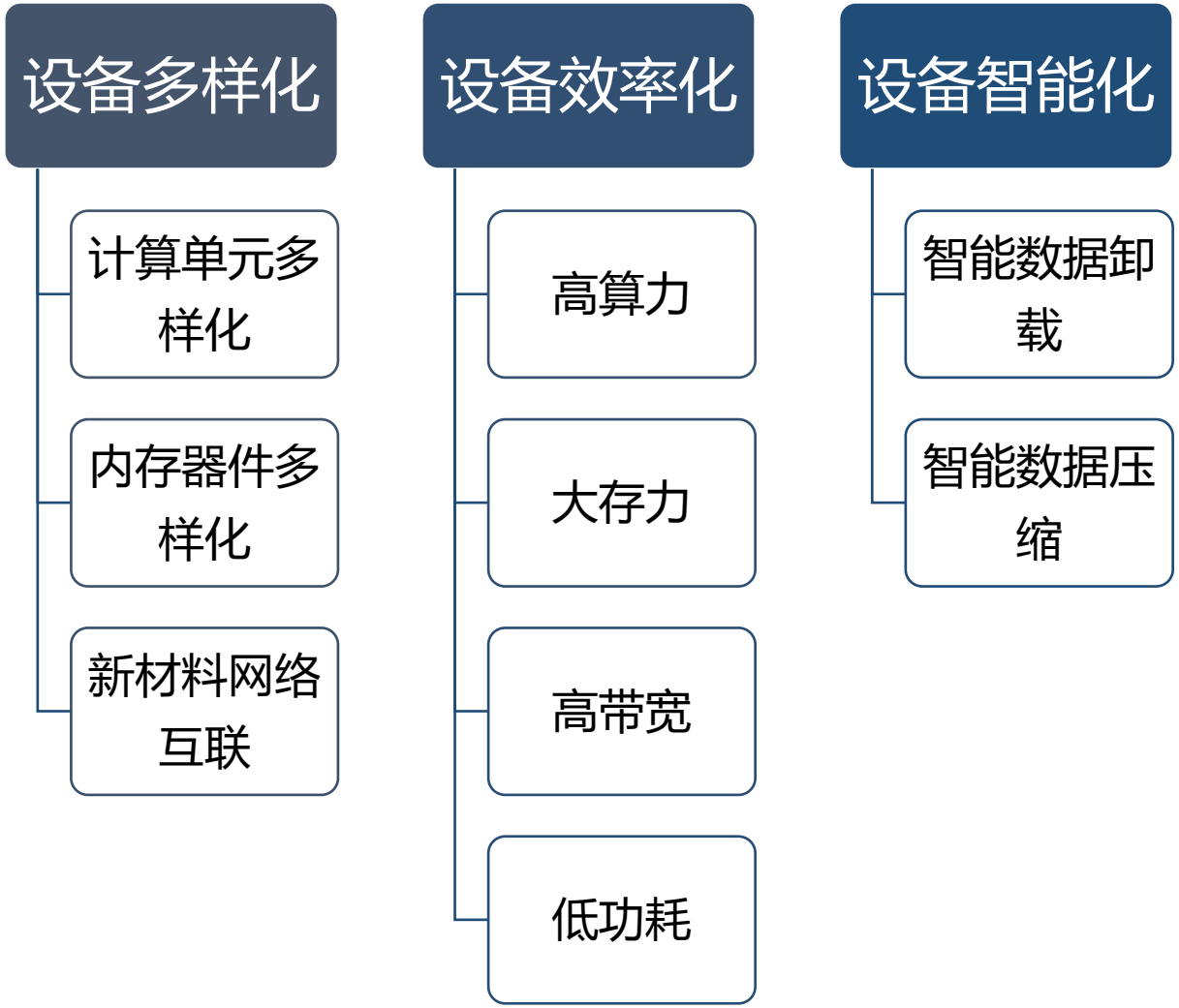
上海交通大学
SHANGHAI JIAO TONG UNIVERSITY

目录

3.硬件技术：分离式架构主要部件



3.硬件技术：整体趋势


设备多样化
计算单元多样化
内存器件多样化
新材料网络互联
设备效率化
高算力
大存力
高带宽
低功耗
设备智能化
智能数据卸载
智能数据压缩

3.硬件技术：设备多样化

计算单元多样化

- CPU擅长处理串行的逻辑推理算法，以及异构设备的调度控制
- 其内存为计算单元共同使用
- 功耗较高。

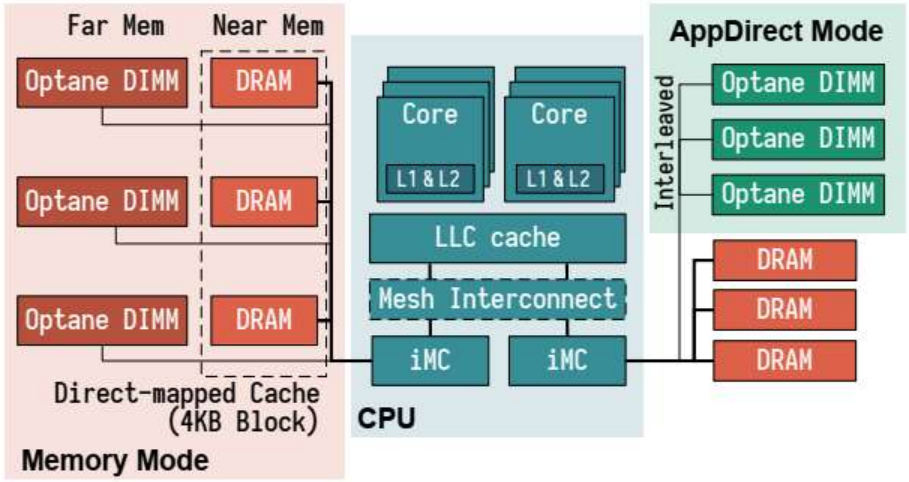
- GPU具有很多线程，其运行依赖于GPU的调控，可以显著加速并行度高的应用
- 其拥有一定的片上寄存器和片上内存，但通常也需要使用CPU内存
- 功耗很高

- FPGA/ASIC芯片采用硬件编程，可以设计超高并行度，也可以设计复杂逻辑，其设计过程依赖CPU
- 拥有大量片上寄存器，也拥有片上内存、存储等器件
- 但当前芯片编程困难，编译速度慢，开发周期长。
- 功耗很低

3.硬件技术：设备多样化

内存设备多样化

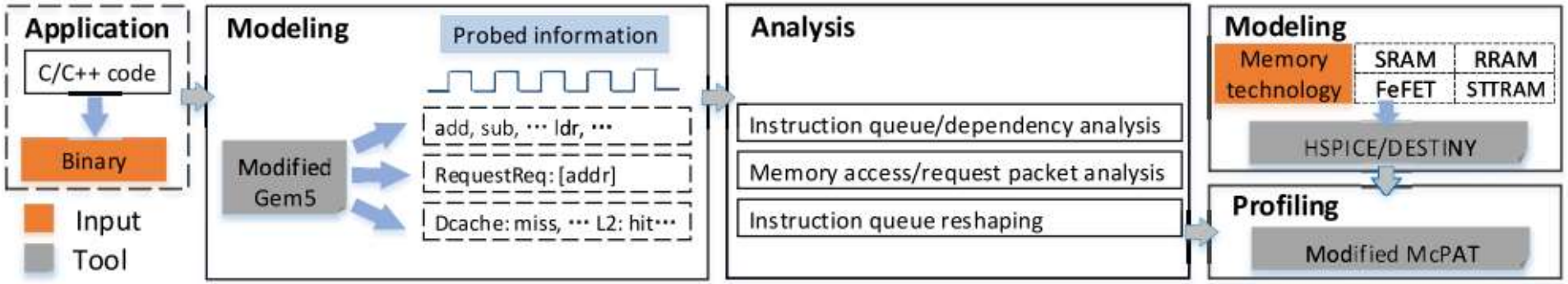
持久式内存



(a) Optane Platform Modes (Memory and AppDirect)

- 基于专用原语编程
- 基于仿真设计

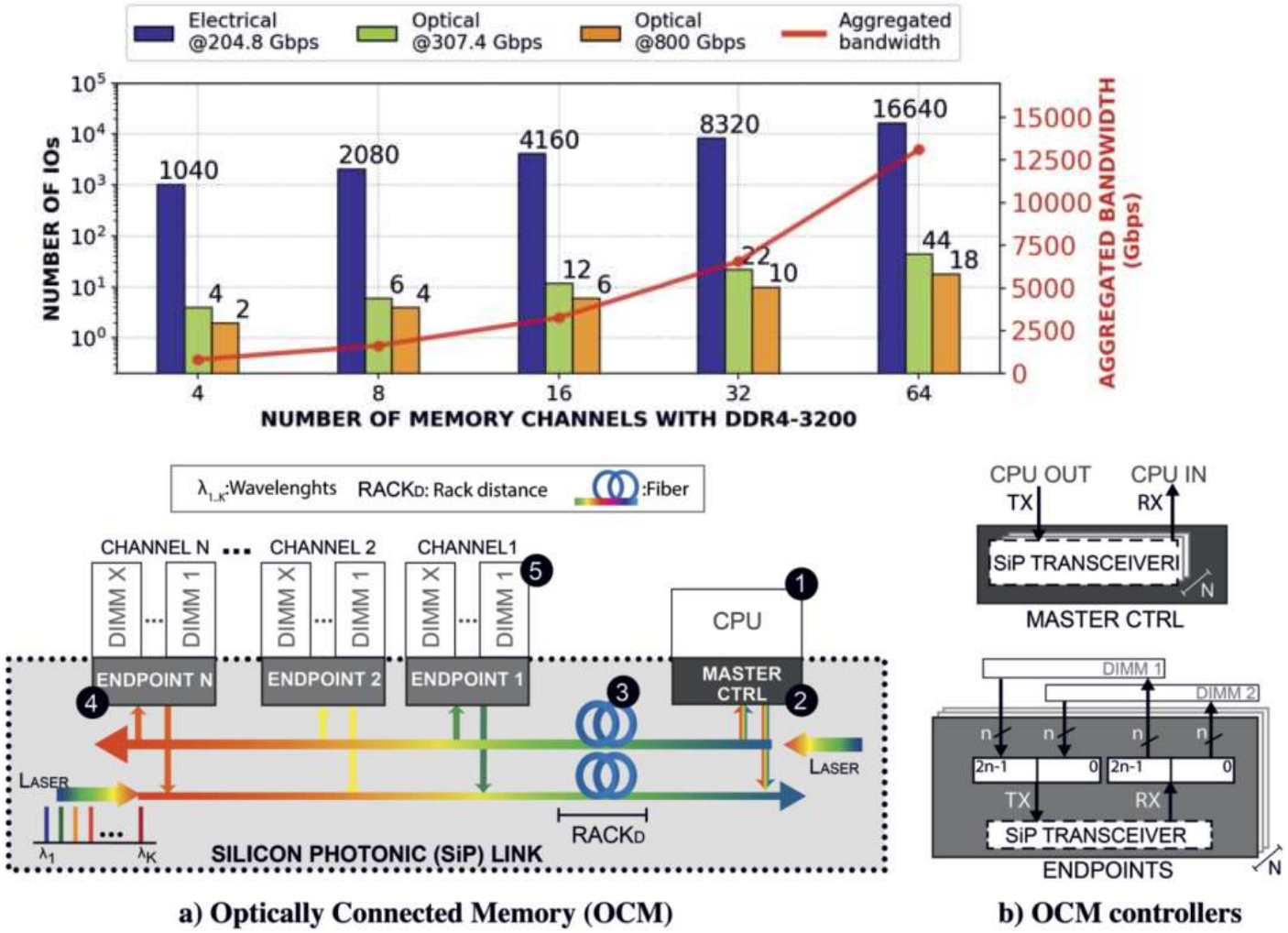
存内计算



- 基于硬件语言编程
- 基于仿真设计

3.硬件技术：设备多样化

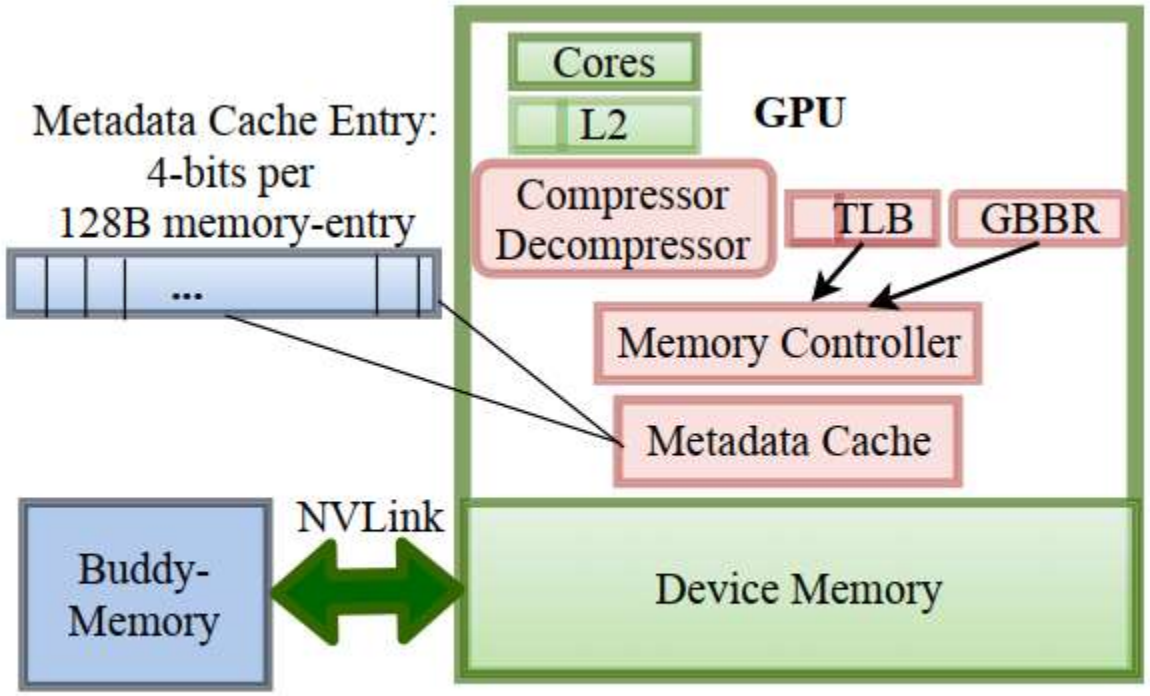
新材料网络互联



- Optically connected memory（OCM）
- 采用光学互联内存的新材料
- 超高带宽
- 超低功耗
- 执行速度比基于nic的分解内存快5倍

[1] Optically connected memory for disaggregated data centers, JPDC'2022

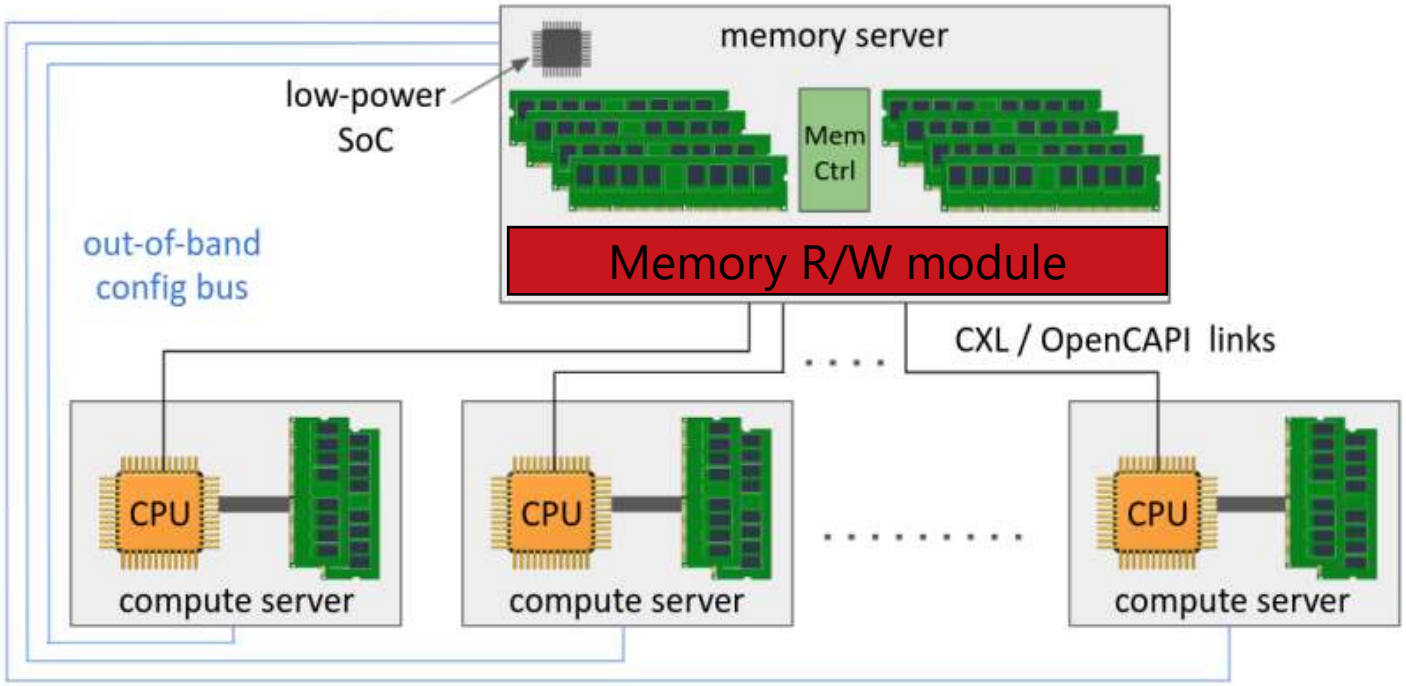
3.硬件技术：设备智能化-智能压缩



(a) Architectural overview

- 将冷数据压缩后换出
- 访问数据时解压缩
- 可以节省内存空间
- 可以提升任务吞吐

3.硬件技术：设备智能化-智能卸载



- 内存节点使用较少计算资源
- 将卸载模块实现在内存节点上
- 卸载模块需要数据访问和读取的控制

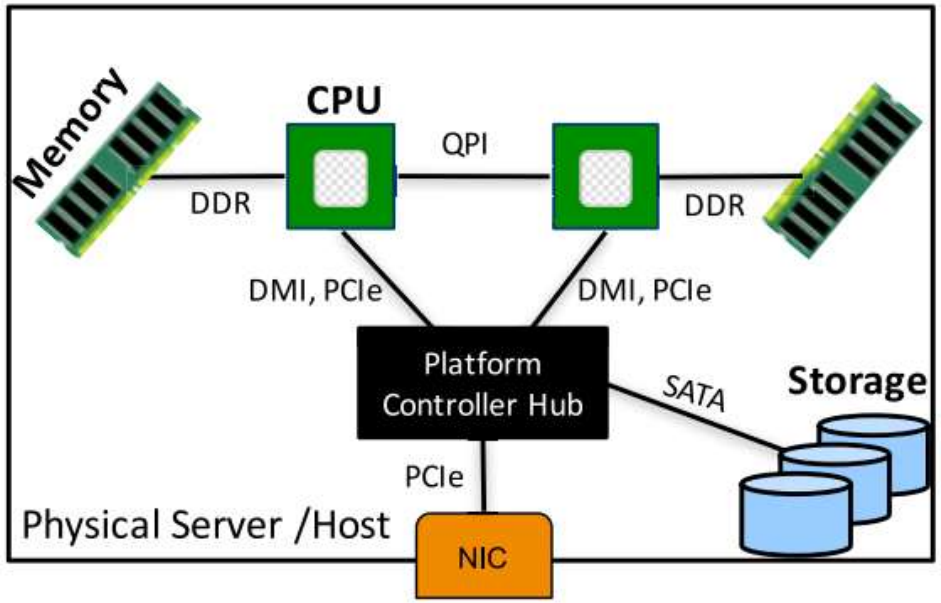
上海交通大学
SHANGHAI JIAO TONG UNIVERSITY

目录

4. 超融合基础设施：软件定义硬件 (SDHI)

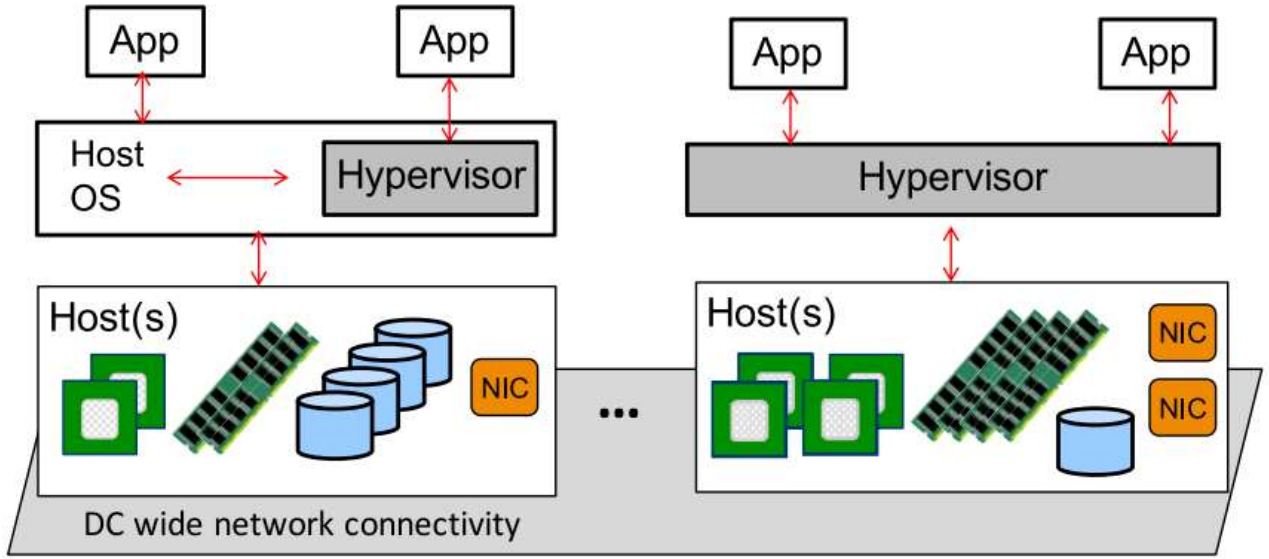
上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

Hardware infrastructures



实际硬件结构

Software defined infrastructures (SDI)

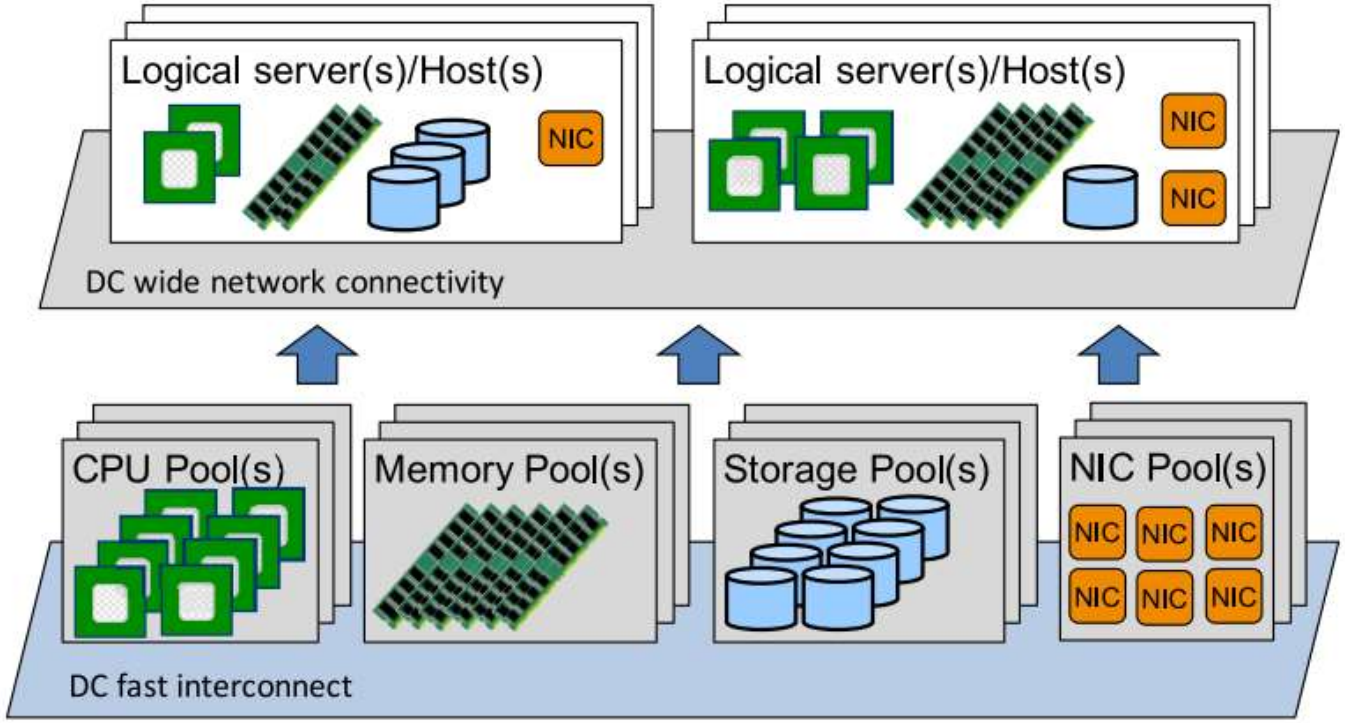


Server oriented SDI

软件定义的架构

4. 超融合基础设施：软件定义硬件 (SDHI)

Software defined hardware infrastructures （SDI）



SDI based upon a disaggregated architecture

软件定义的硬件架构（分离式架构）

4. 超融合基础设施：软件定义硬件 (SDHI)

(SDI)

SDN [17] Software-Defined Networking	Separates the network control planes from data planes and physical network entities to improve programmability, efficiency, and extensibility of network.
SDS [21] Software-Defined Storage	Separates the control planes from the data plane of a storage system enabling heterogeneous storage to respond dynamically to changing workload demands.
SDC [11] Software-Defined Computing	Originated from the computing environment in which the computing functions are virtualized and managed as virtual machines through a central interface as one element.

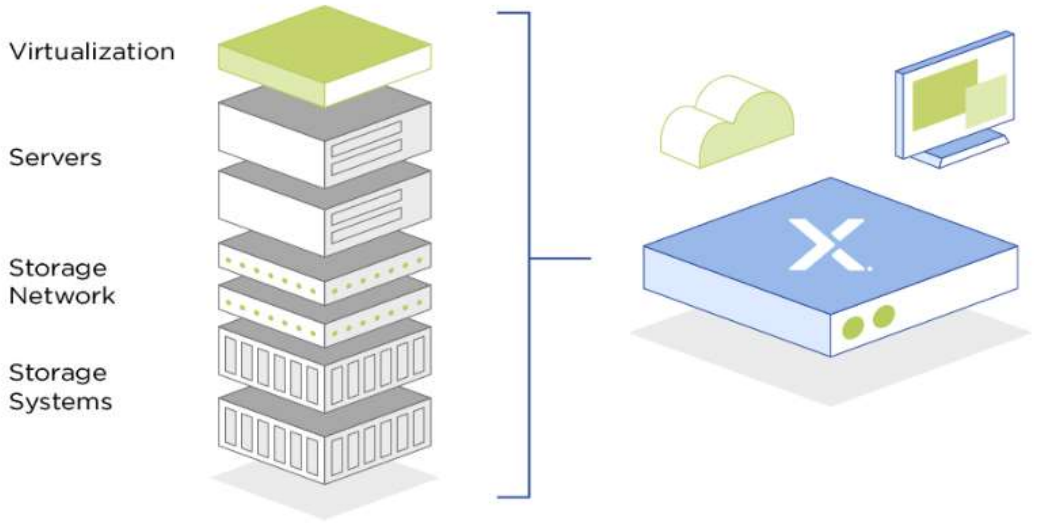
[1] Software-Defined “Hardware” Infrastructures: A Survey on Enabling Technologies and Open Research Directions, IEEE COMMUNICATIONS SURVEYS & TUTORIALS, 2018

4. 超融合基础设施： HCI简介

上海交通大学 SHANGHAI JIAO TONG UNIVERSITY

超融合基础架构（hyper-converged infrastructure，简称HCI）：

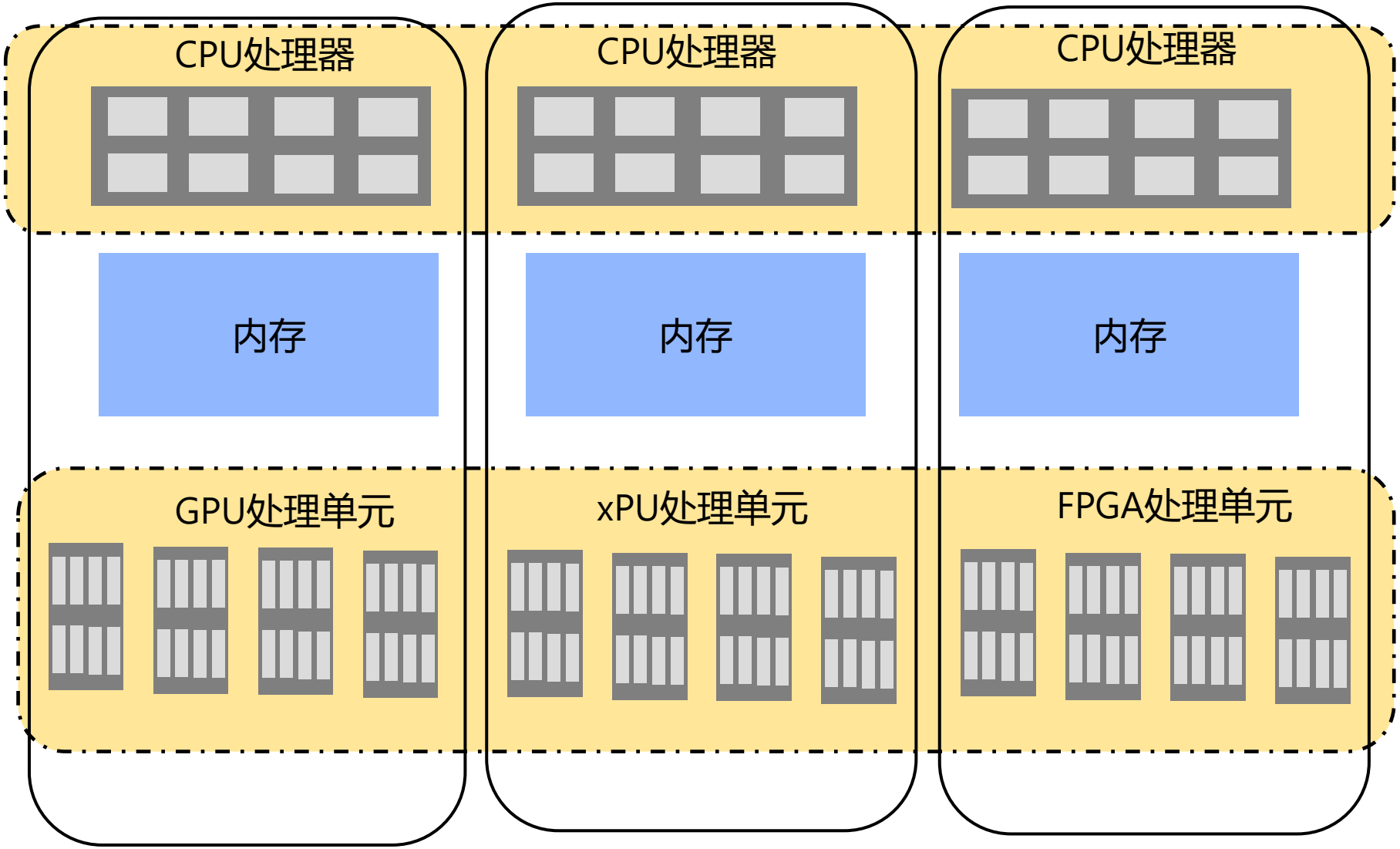
- 超融合基础架构是一个软件定义的 IT 基础架构，它可虚拟化常见“硬件定义”系统的所有元素。
- HCI 包含的最小集合是：虚拟化计算（hypervisor），虚拟存储（SDS）和虚拟网络。
- HCI将计算、存储和虚拟化资源紧密地集成在单个系统中，这些资源可以通过基于x86的设备交付，也可以作为可以安装在现有硬件上的软件交付。



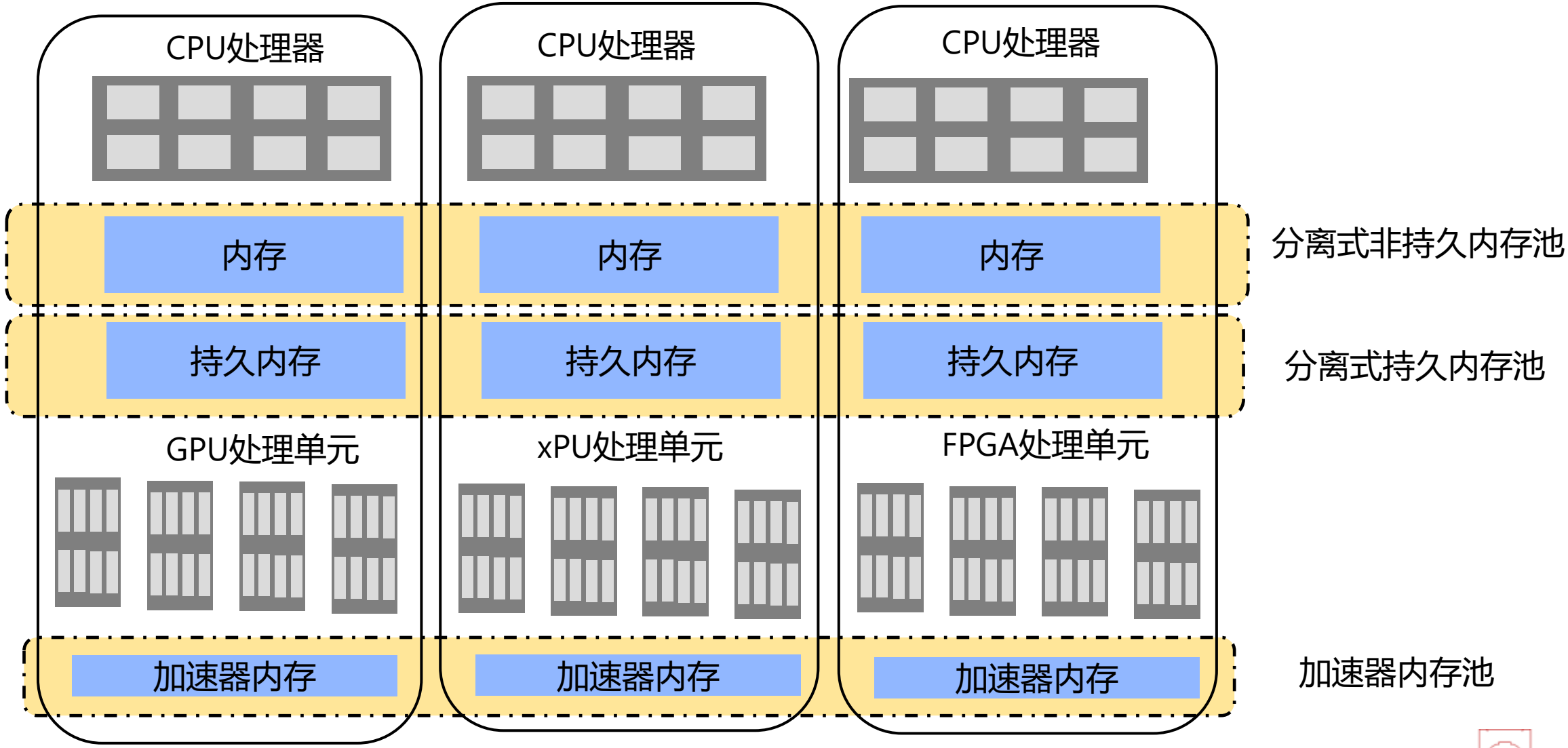
HCI 由两个主要组件组成：**分布式平面**和**管理平面**。

- **分布式平面**运行于节点集群之上，为虚拟机或基于容器的应用等客户应用提供存储、虚拟化和网络服务。
- **管理平面**支持从单一视图轻松管理所有 HCI 资源，无需为服务器、存储网络、存储和虚拟化单独制定管理解决方案。

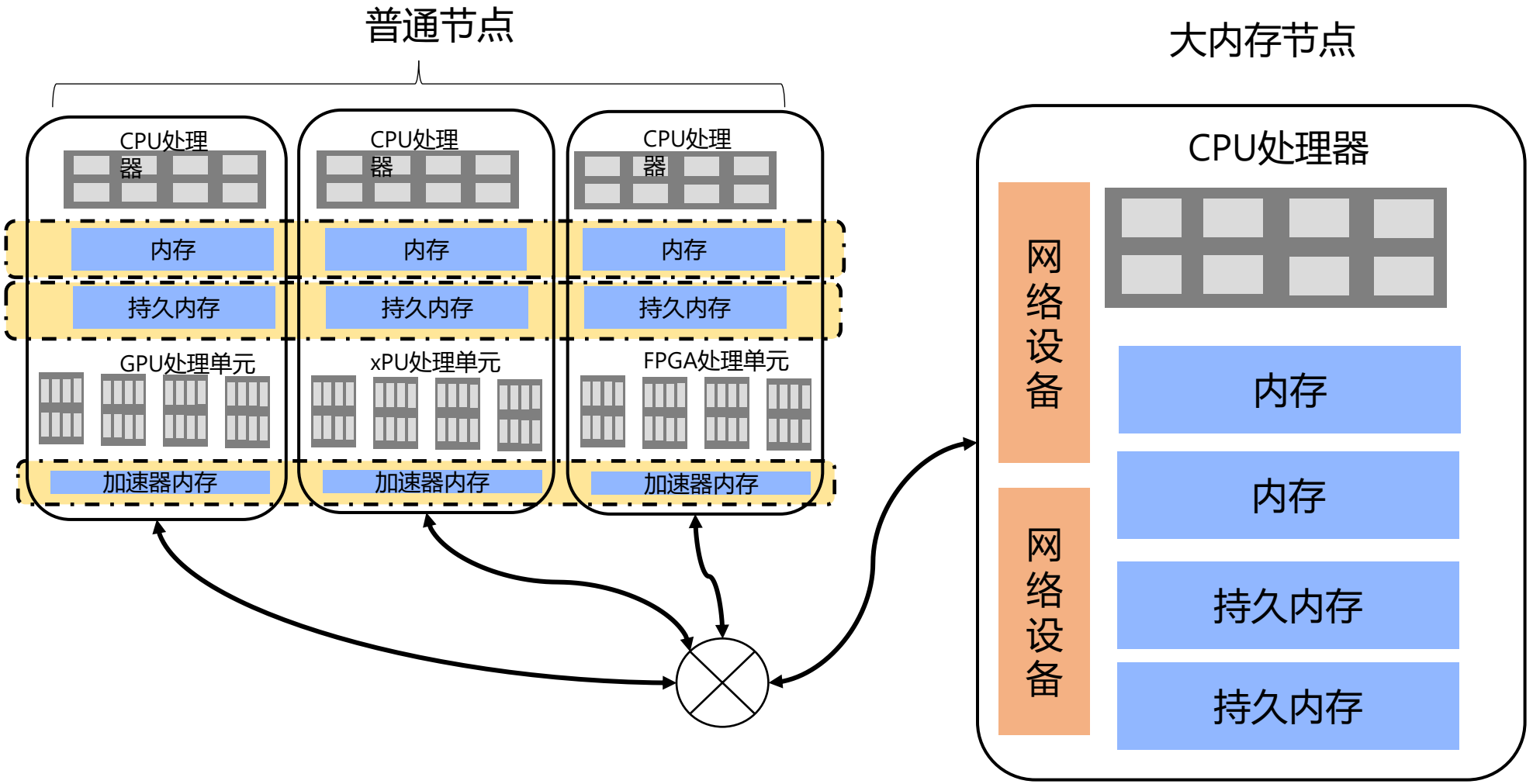
4. 超融合基础设施：异构计算架构



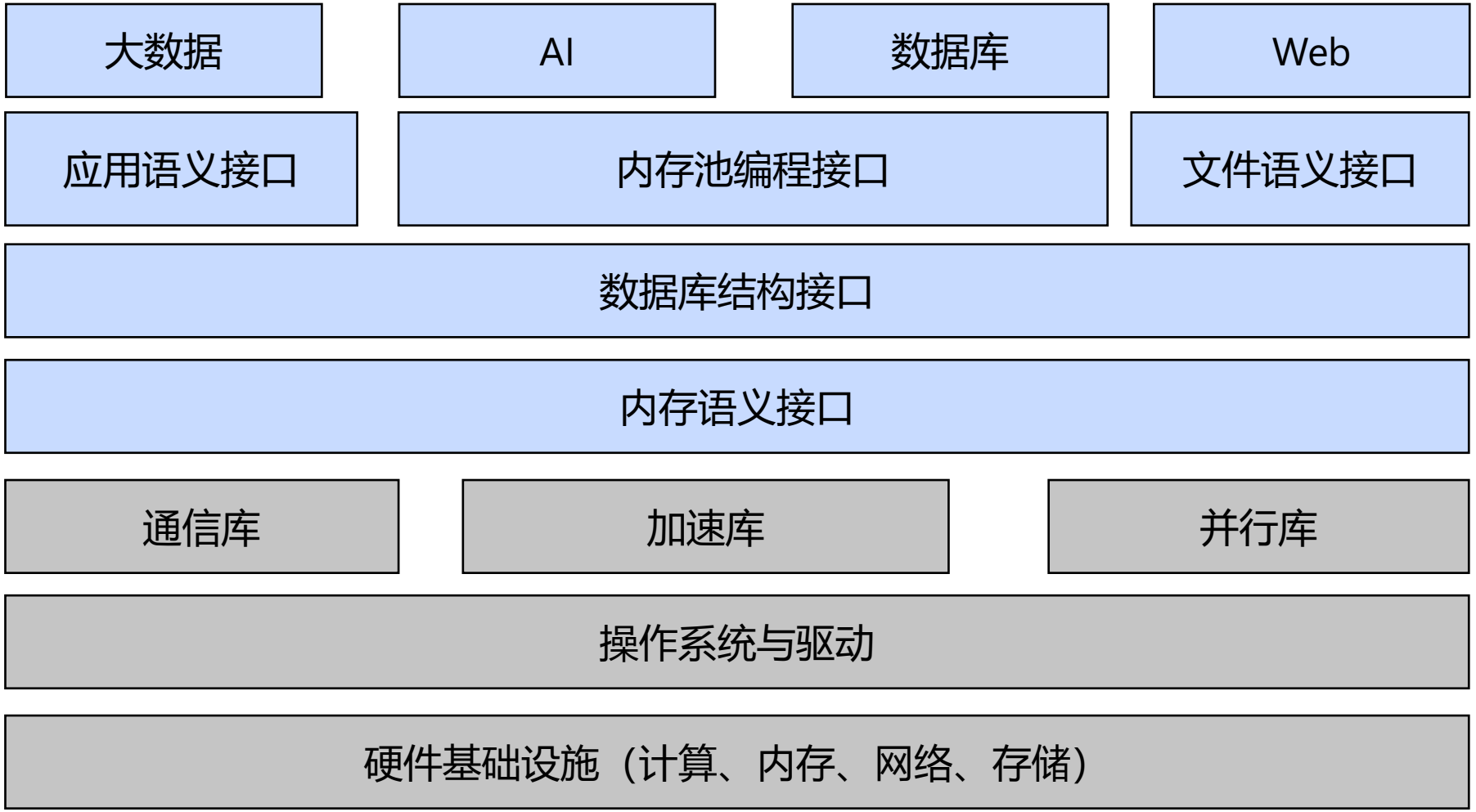
4. 超融合基础设施：内存层次



4. 超融合基础设施：网络互联资源池



4. 超融合基础设施：应用使用接口



上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Thank You