



星火计划

中|国|电|信|转|型|专|业|领|军|人|才|培|养|项|目



分离式内存： 背景概念、关键技术、未来趋势

星火计划

中国电信转型专业领军人才培养项目



一、分离式内存概论

讲者：李超

上海交通大学 计算机系 SAIL实验室

2023年2月

星火计划

中国 电信 转型 专业 领军 人才 培养 项目



目 录

1

**内存背景及需求：
大容量高密度内存简介**

2

现代扩展后的内存层级

3

**处理器和加速器间的
统一内存访问机制**

4

现代服务器的存算分离架构

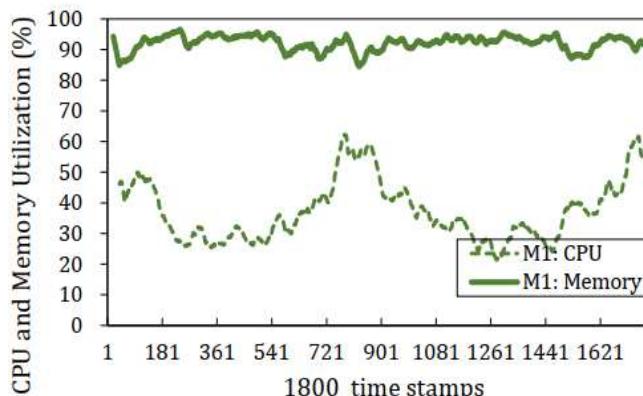
星火计划

中国 电信 转型 专业 领军 人才 培养 项目

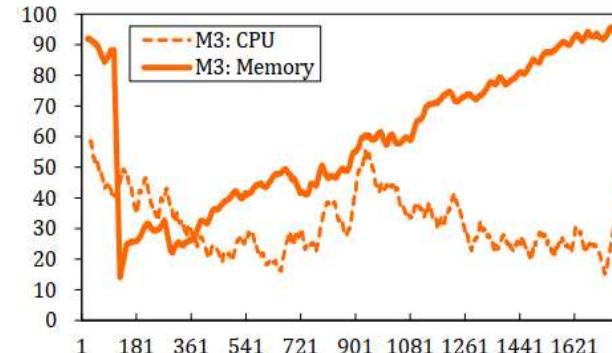
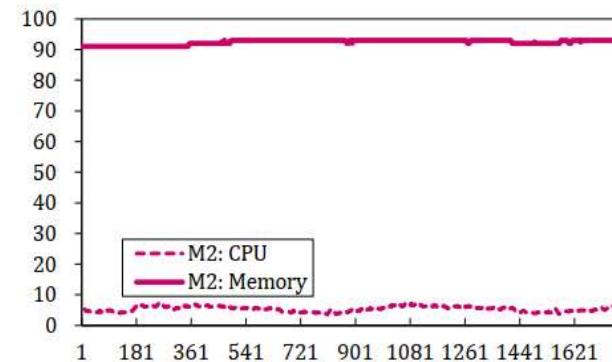


背景1：数据中心内存需求

根据对阿里云的数据中心trace分析，内存资源十分紧张



Alibaba trace	2017	2018
CPU usage (average)	24.67%	36.30%
Memory usage (average)	48.95%	87.05%



内存资源逐渐消耗殆尽，亟需进行内存资源内部和外部的拓展

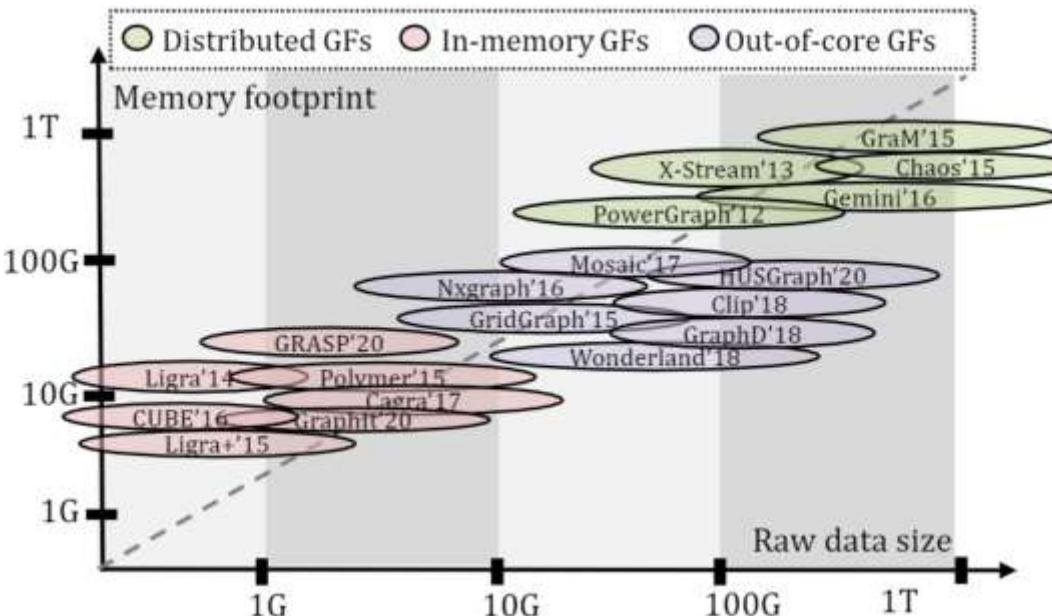
[1] Analyzing alibaba's co-located datacenter workloads, IEEE Big Data'18

星火计划



背景2：应用的大内存需求

图计算：



AI计算：

模型名称	模型内参数数量	数据规模
GPT	110M	4GB
BERT	340M	16GB
GPT-2	1.5B	40GB
RoBERTa	330M	160GB
T5	11B	800GB
GPT-3	175B	45TB

应用特点：

- 数据驱动型任务数量多，计算复杂，访存开销大
- 以机器学习、生物计算、图计算等的内存密集型应用体量庞大，可达TB级
- 实际系统中的内存占用远大于需要处理的应用数据

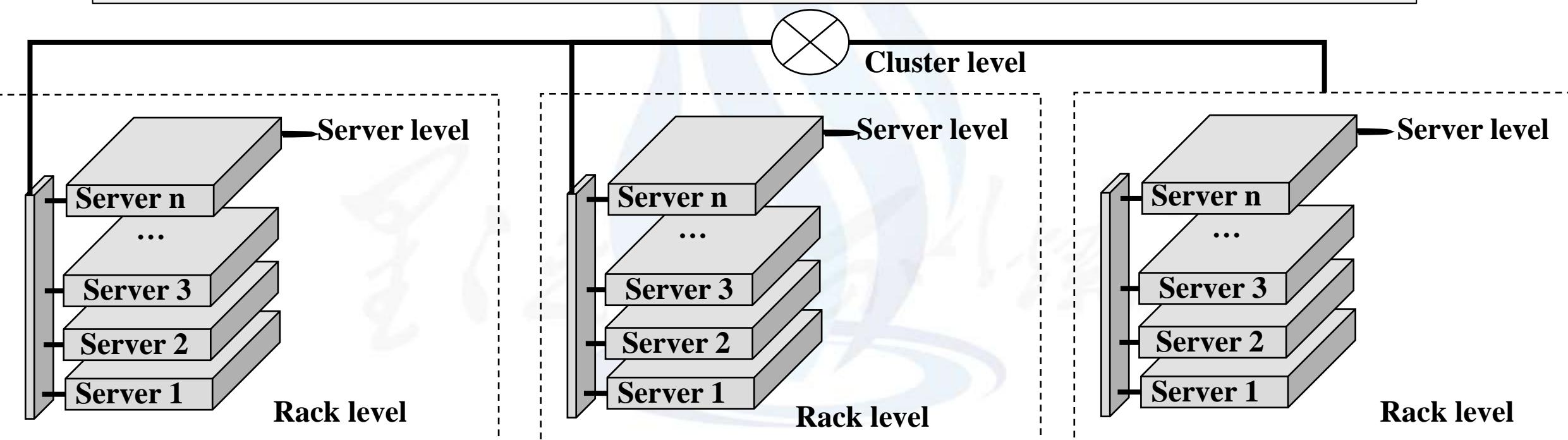
星火计划



分离式内存：助力数据中心协同资源管理

分离式内存存在数据中心中的应用：

- 协同内存资源管理：抽象并池化资源，统一高效管理
- 提升资源利用率：回收空闲内存资源并按需供给
- 提升任务运行性能：软硬件协同达成更高效的数据分布
- 降低能耗：按需关闭闲置资源，降低整体能耗



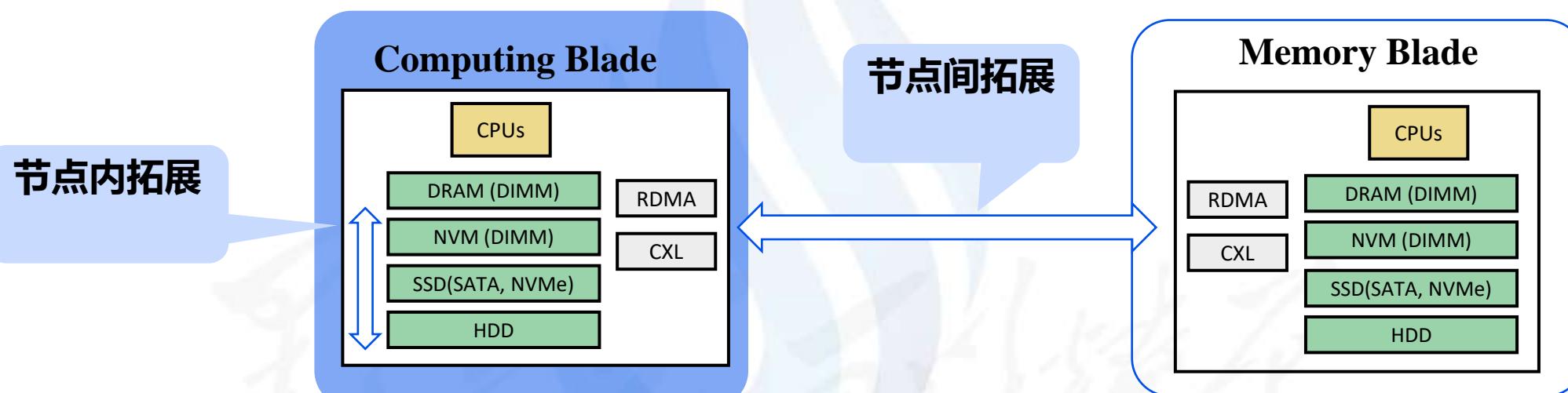
星火计划



分离式内存：助力数据中心服务器内存资源扩展

内存资源扩展方向：

- 节点内扩展：插入新型内存存储设备
- 节点间扩展：使用新型网络访问远端设备



星火计划

中国 电信 转型 专业 领军 人才 培养 项目

TSJU



分离式内存：主要目标

主要目标：

- 提高应用性能
- 提升资源使用效率
- 智能资源管理与调度
- 保证资源安全性

优化内容：



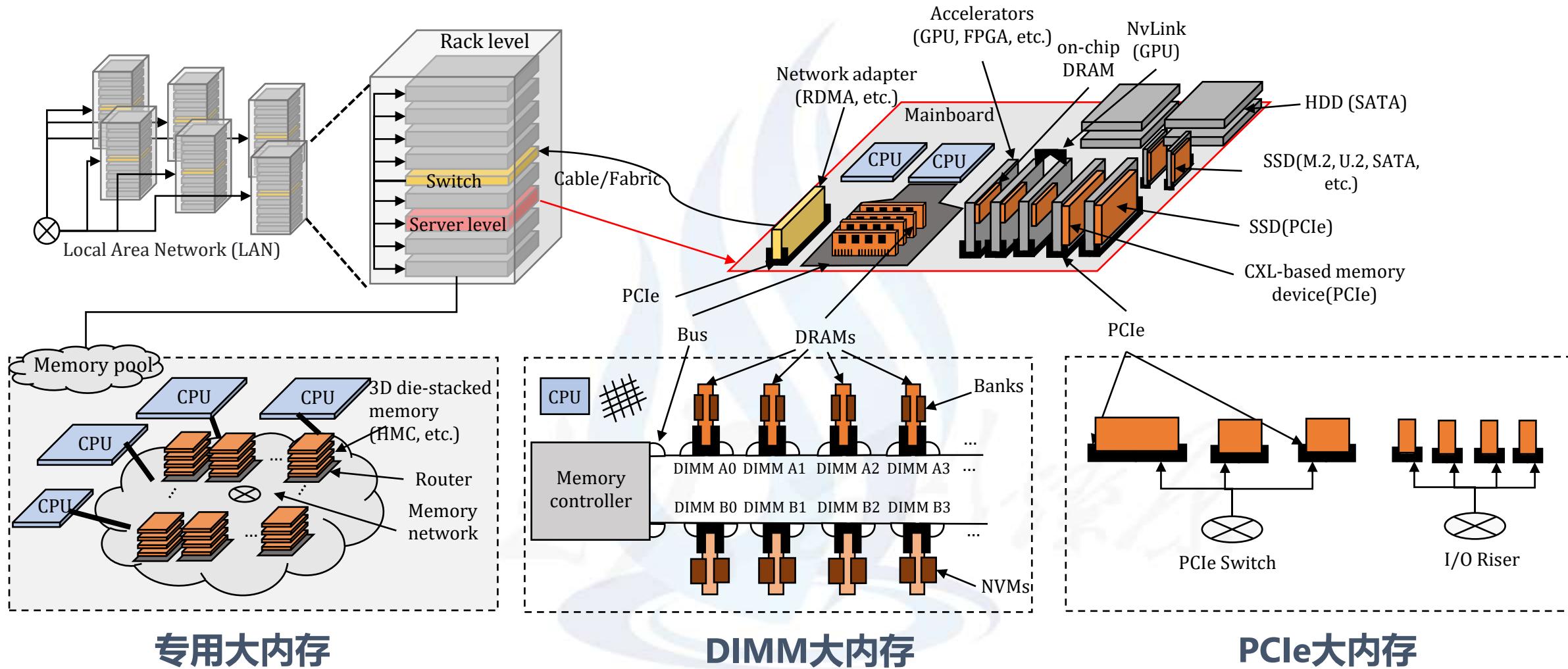
星火计划

中国 电信 转型 专业 领军 人才 培养 项目

电子科技大学



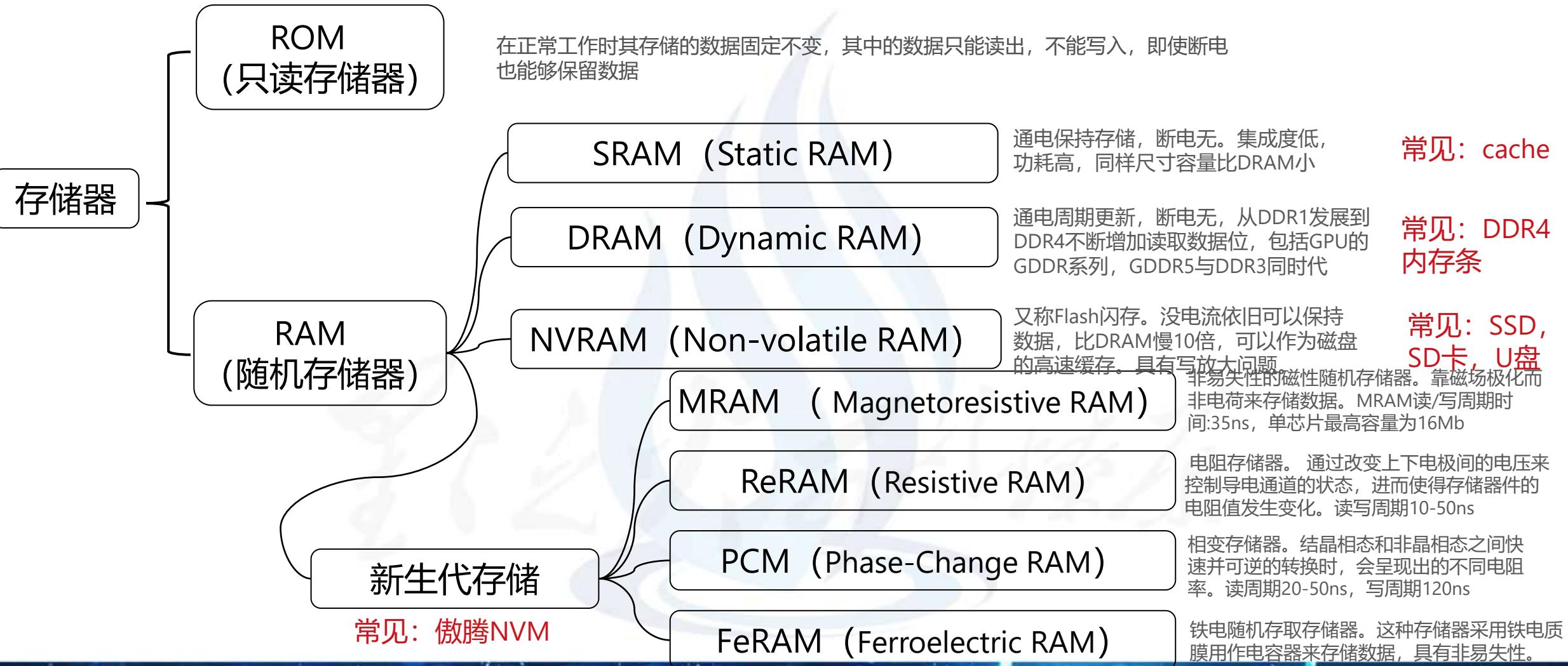
1. 内存简介：大容量高密度内存



星火计划



1. 内存简介：硬件设备篇



星火计划



1. 内存简介：互联篇

互联
接口

DIMM
(双列直插内存模块)

双列直插内存模块，这种接口模式的内存广泛应用于现在的计算机中，通常为84针，由于是双边的，所以一共有168针，体积较大，提供64位有效数据位。

PCIe
port

PCIe版本分为1.0/2.0/3.0/4.0/5.0，PCIe的连线是由不同的lane来连接的，这些lane可以合在一起提供更高的带宽。譬如两个1lane可以合成2lane的连接，写作x2。两个x2可以变成x4，最大直到x16

PCIe
(并行总线)

PCIe
switch

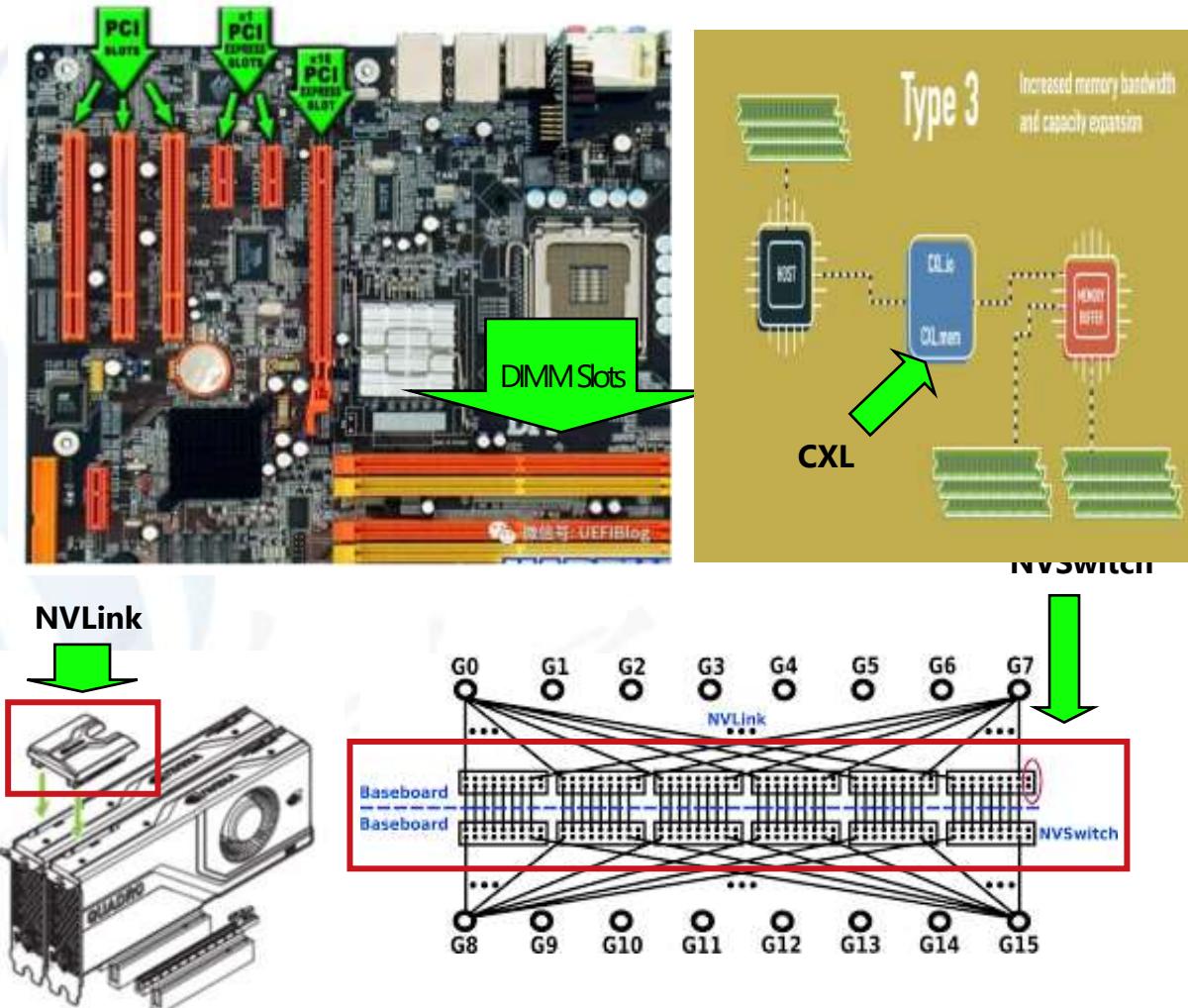
CXL是基于PCIe5.0的新互联协议

NVLink能在多GPU之间和GPU与CPU之间实现更高的连接带宽。如2016发布的P100是，单个GPU具有160GB/s的带宽，相当于PCIe Gen3 * 16带宽的5倍

NVLink

NVLink 模块

NVSwitch

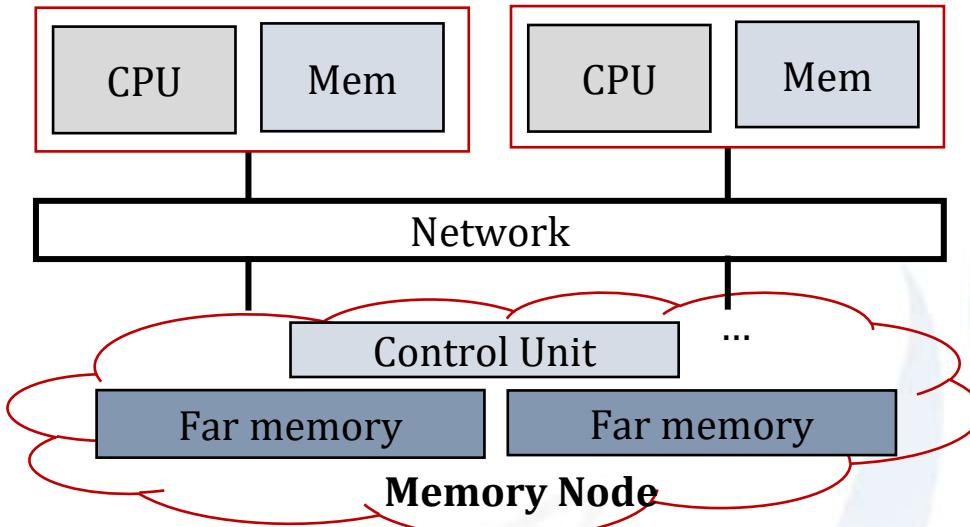


星火计划



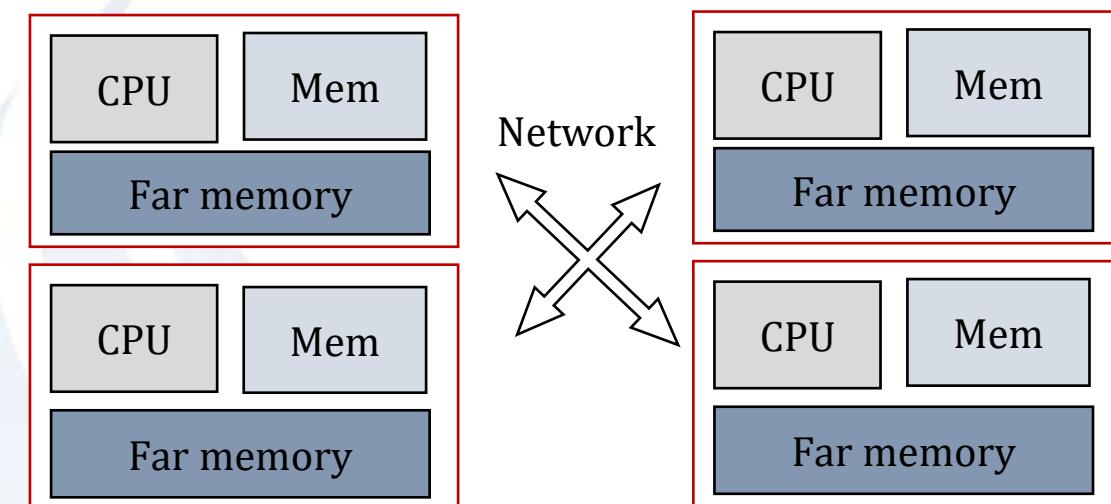
1. 内存简介：中心式与去中心式架构对比

Computing Node



(1). Centralized DM Architecture

Computing/Memory Node



(2). Decentralized DM Architecture

展望：

1. 中心内存池可以基于实体大节点，也可以虚拟化的方式构建
2. 中心化与去中心化架构结合发展
3. 资源管理方式随规模更新，类比于数据中心的集中式资源调度到共享视图的资源调度

星火计划



1. 内存简介：中心式内存池(Memory pool)

虚拟大节点的概念：

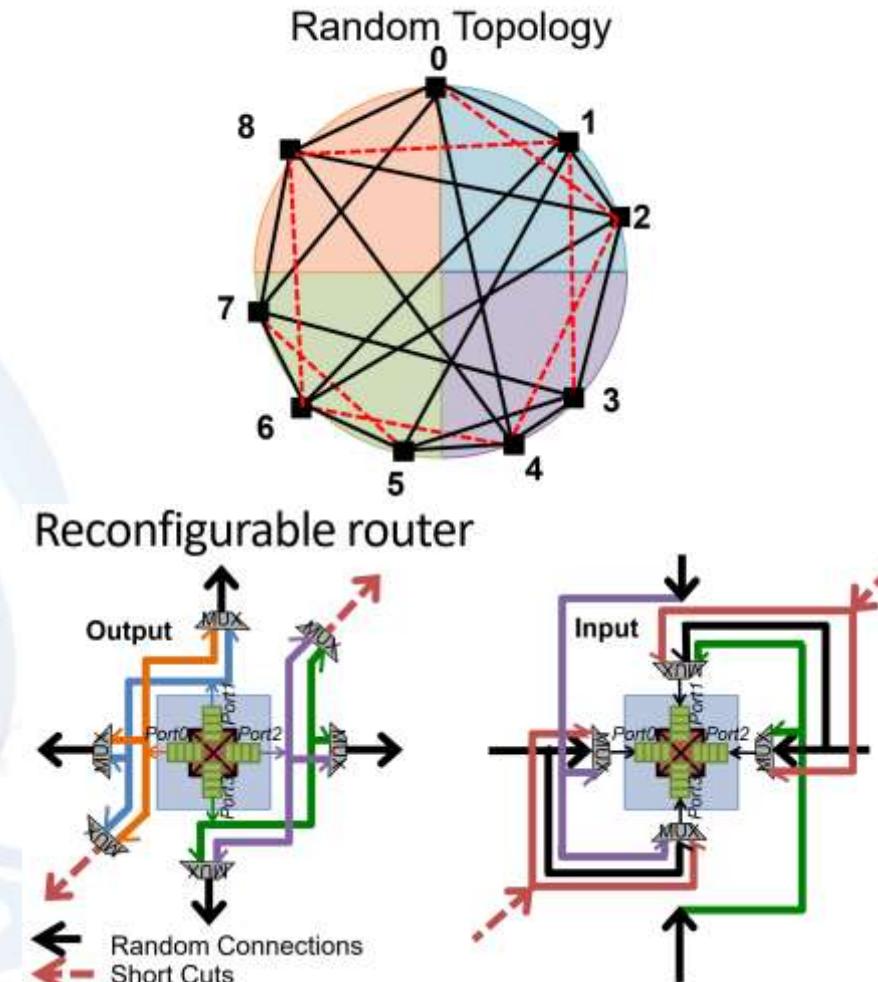
1989年，网络互联的共享虚拟内存概念便被提出
由中心资源监视器进行集中管理

设计目标：

- 良好的可扩展性
- 任意规模的高效互联

实体大节点的设计挑战：

- 复杂的芯片设计与节点拓扑分布
- 路由最短的数据访问路径算法
- 中心化方案中，监视器会成为性能瓶颈，难以建立超过1000个内存节点的内存池



星火计划



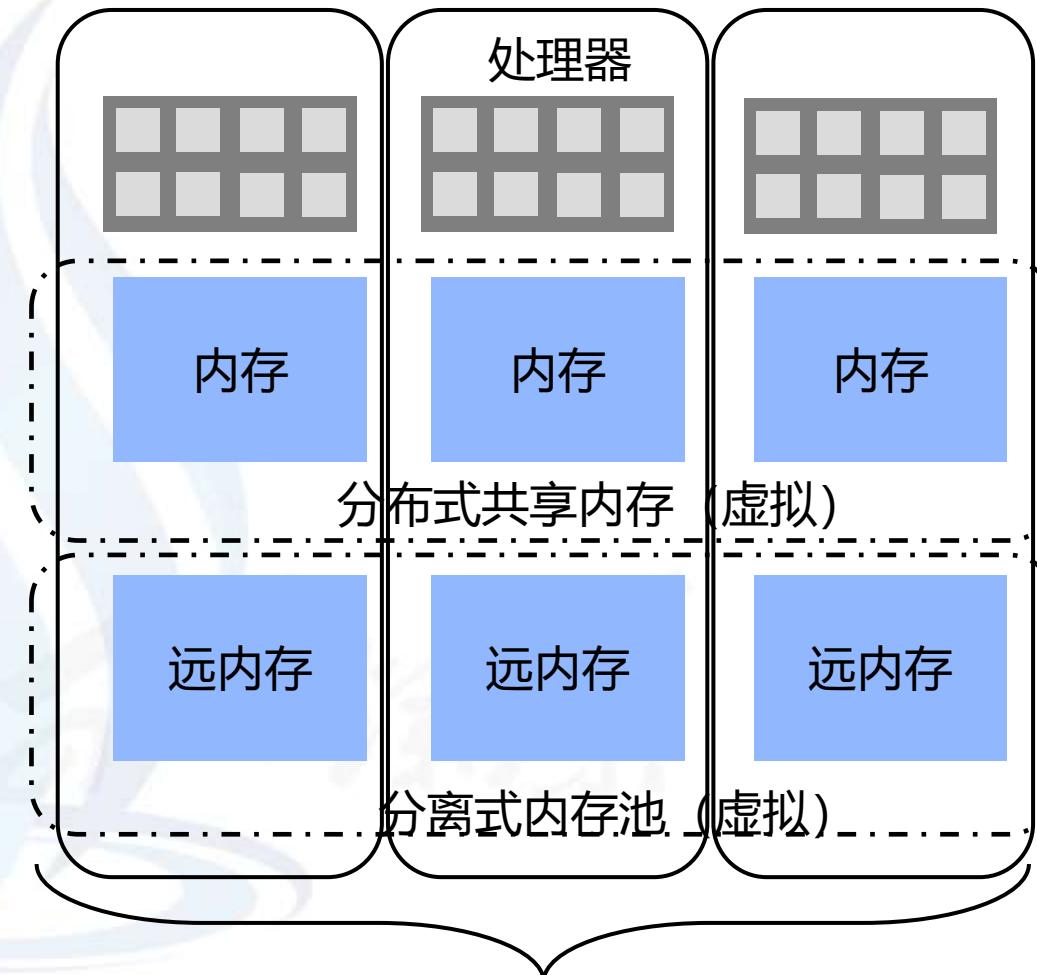
1. 内存简介：去中心式内存池(Memory pool)

基本概念：

- 节点内部构建远内存，可供自己本地内存使用，也可供给其他计算节点。处理器除本地内存外，可以访问其他处理器的空闲内存
- 采用一定的分布式管理策略（中心式管理，共享视图管理），构建虚拟的资源池，能够进行灵活的内存资源共享

设计挑战：

- 数据一致性处理
- 本地数据生命周期管理和安全保护
- 空闲内存的划分与冷热数据的迁移



星火计划



目 录

1

内存背景及需求：
大容量高密度内存简介

2

现代扩展后的内存层级

3

处理器和加速器间的
统一内存访问机制

4

现代服务器的存算分离架构

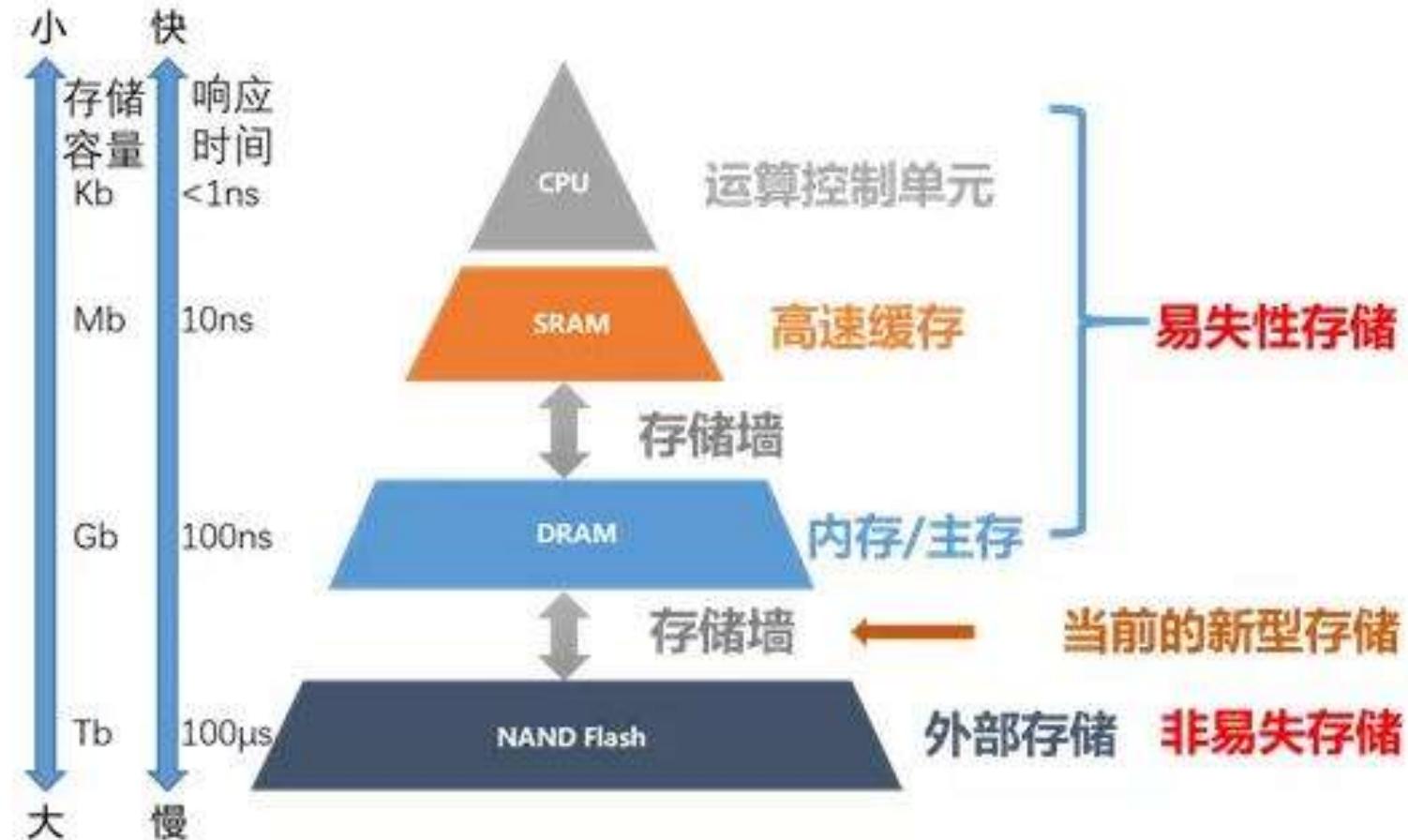
星火计划



2. 内存层级：传统层级

现代计算系统通常采取三级存储结构：

- 高速缓存(SRAM)，响应时间在纳秒级
- 主存(DRAM)，响应时间为100ns级，带宽在百GB级
- 外部存储(NAND Flash)，响应时间在100us级，带宽在MB~GB级



数据在这三级存储间传输时，后级的响应时间及传输带宽都将拖累整体的性能，形成“存储墙”。

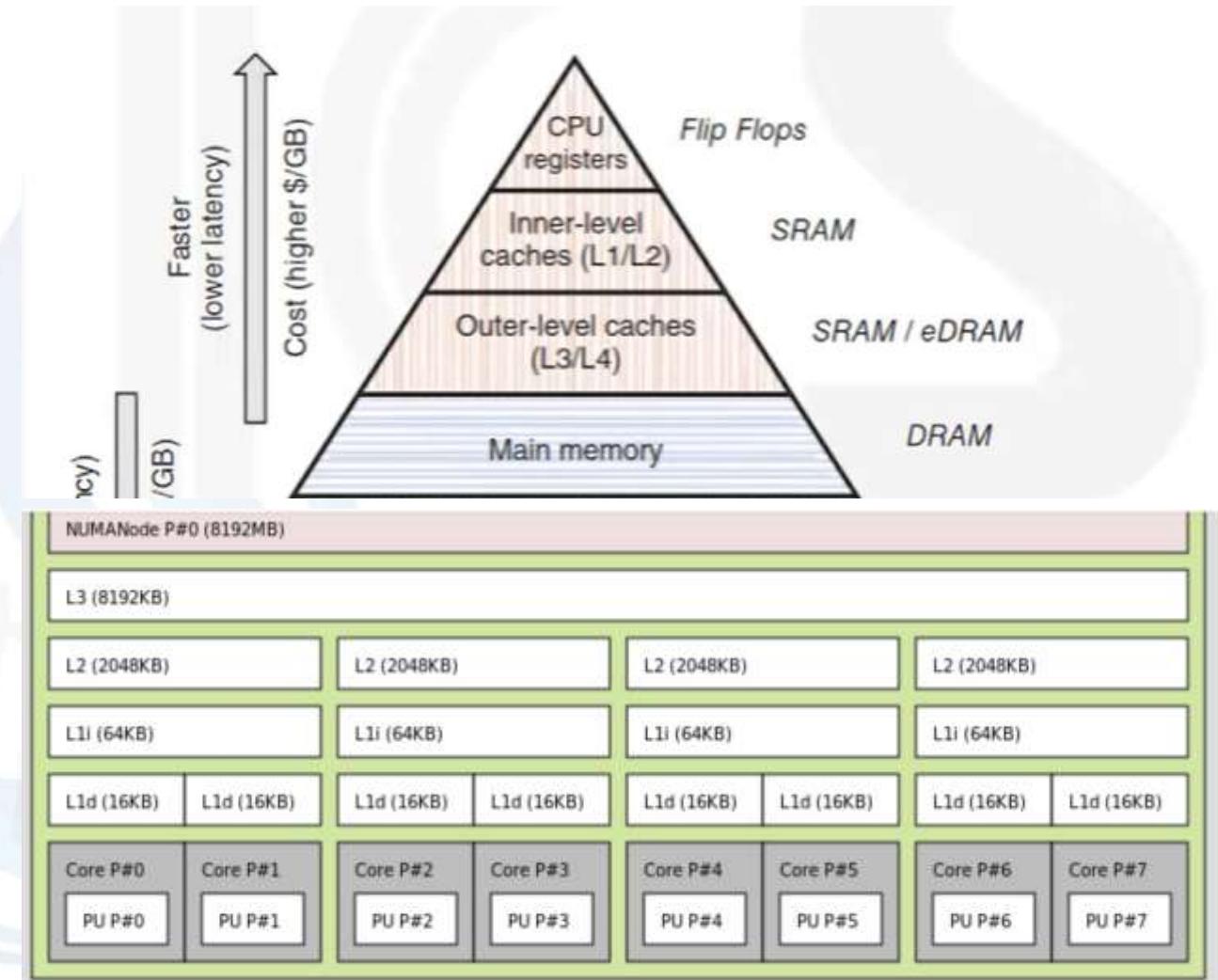
星火计划



2. 内存层级：拓展的Cache层级

缓存层级：

- L1缓存有分为L1i和L1d，分别用来存储指令和数据。每个CPU一个。
- L2缓存是不区分指令和数据的。每个CPU一个。
- L3缓存多个核心共用一个，通常也不区分指令和数据。
- TLB缓存，它主要用来缓存MMU使用的页表，以加快查询。
- Cache line大小与DDR3、4一次访存能得到的数据大小是一致的，即64Bytes



星火计划



2. 内存层级：拓展的Memory层级

内存层级：

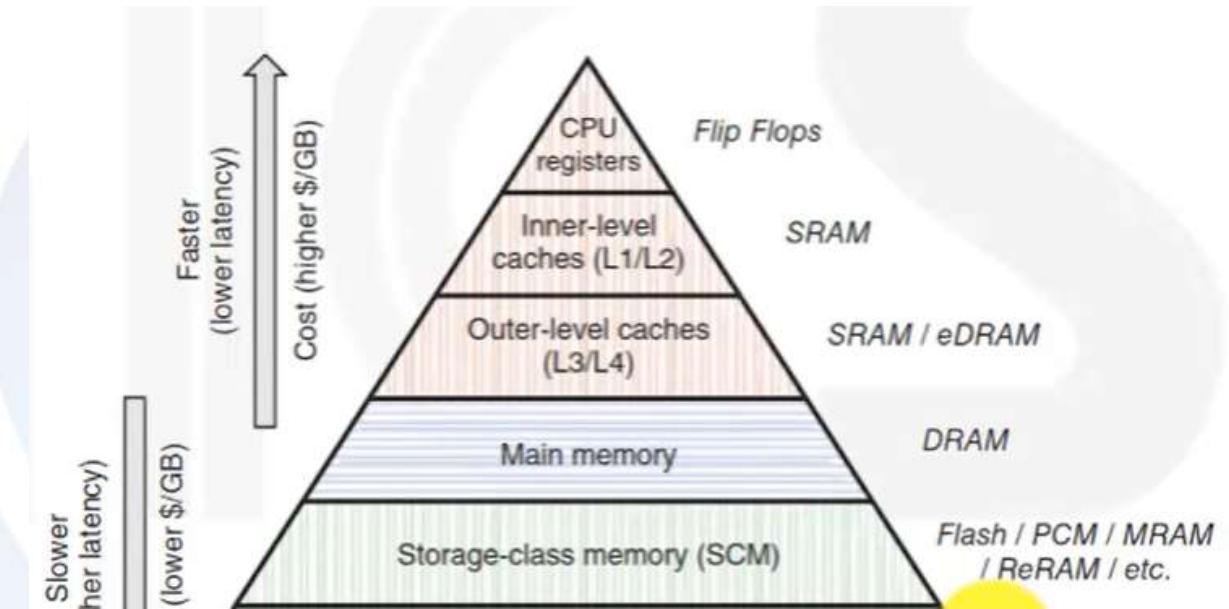
- 易失性存储层

- 以DRAM为例，DDR4单条容量为16GB/32GB/64GB, 单条带宽约为30GB/s, 响应时间在100ns级。

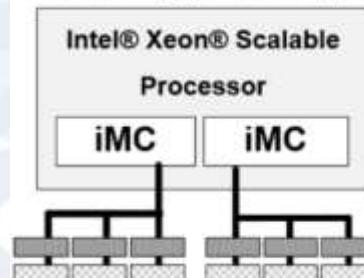
- 非易失性存储层

- 以NVM (PCM) 为例，需要附加额外电源模块用以刷新，搭配DRAM共同使用，内存容量大于DRAM，单条容量在128GB或256GB，单条带宽10GB/s, 读写时间比DRAM慢约2~8倍

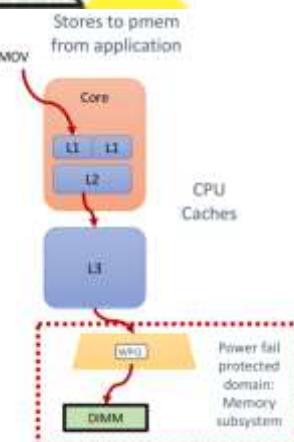
NVM采用专用编程模式，亦可被认为远内存。



Modes Supported: App Direct, Memory Mode



■ DRAM
■ Intel® Optane™ DC persistent memory



星火计划

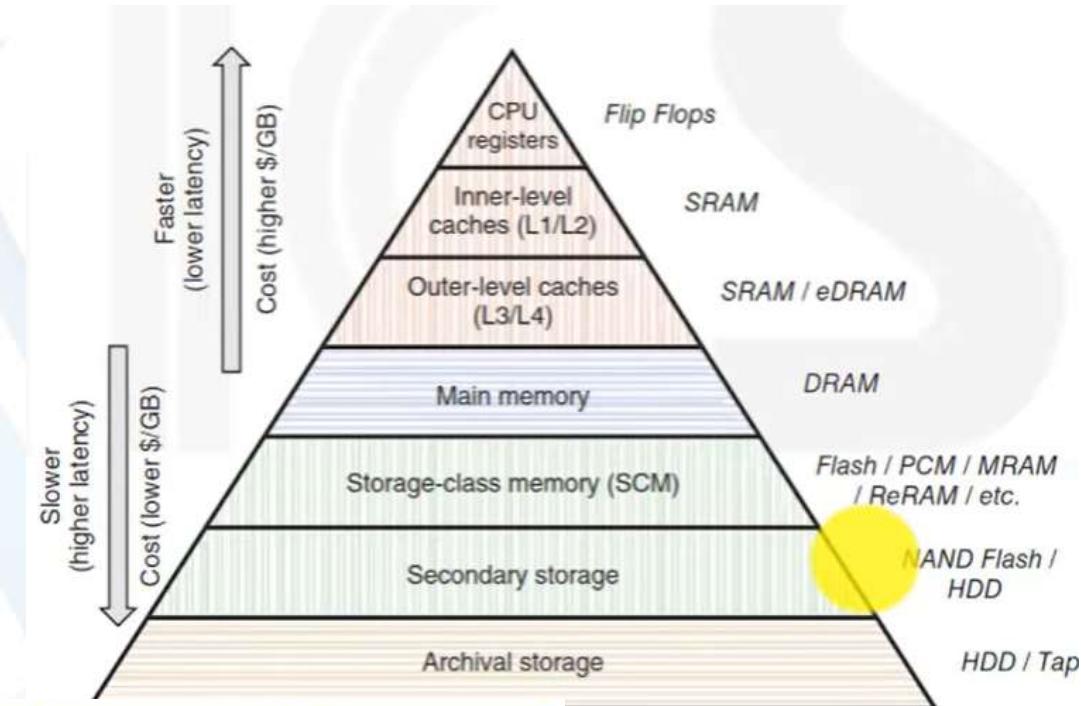


2. 内存层级：拓展的storage层级

存储层级：

- 固态硬盘存储层 (SDD)
 - 以SDD (=NAND Flash) 为例，单个容量为512GB/1TB/2TB, 接口版本包括SATA、M.2、U.2、PCIe x8等.带宽约为0.1~7GB/s, 响应时间在ms级。
- 机械硬盘存储层 (HDD)
 - 以HDD为例，单个存储容量高达18T，一般为SATA接口，带宽约为100~1000MB/s, 读写时间在ms级。

SDD具有高速存储功能，亦可被认为远内存。



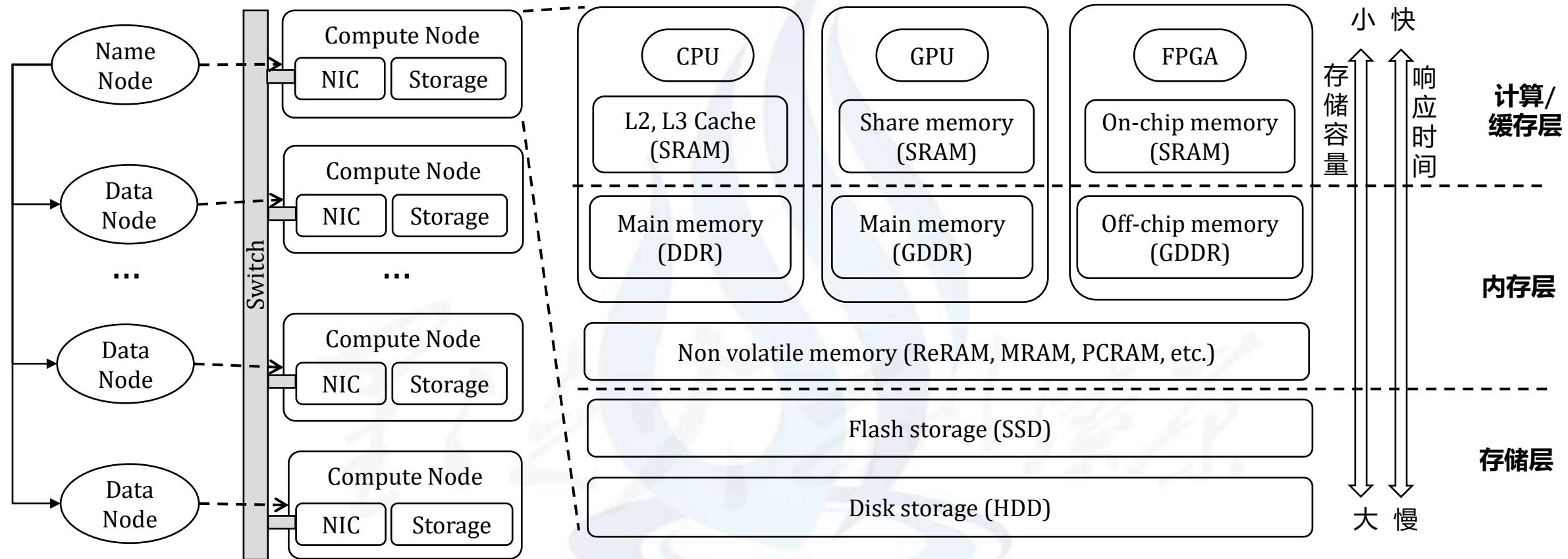
接口	协议	速度
SATA	AHCI	最高600MB/S
M.2	AHCI	最高600MB/S
	NVMe	最高3200MB/S
U.2	NVMe	最高3200MB/S
PCI-E	NVMe	最高3200MB/S



星火计划



2. 内存层级：拓展的存储层级总结



星火计划



目录

1

内存背景及需求：
大容量高密度内存简介

2

现代扩展后的内存层级

3

处理器和加速器间的
统一内存访问机制

4

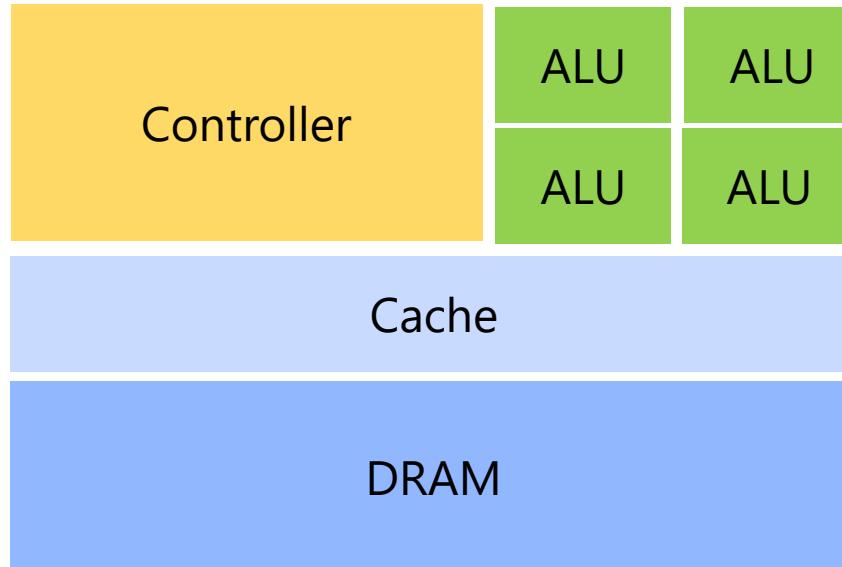
现代服务器的存算分离架构

星火计划

中国·电信·转型·专业·领军·人才·培养·项目



3.统一内存访问：处理器与加速器整体结构

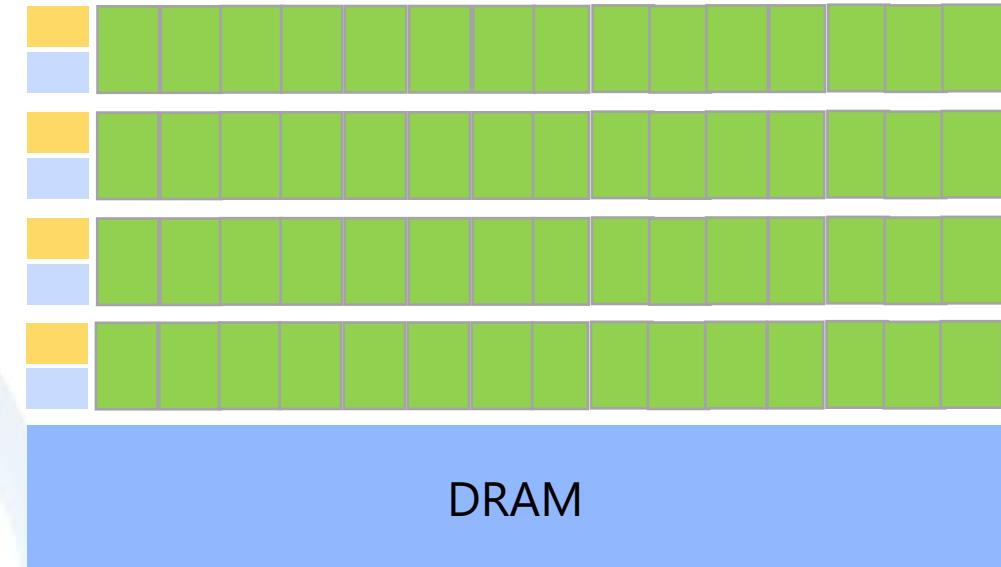


CPU

CPU有复杂控制逻辑，包含容量较大的Cache和Memory

GPU针对高吞吐设计，Cache容量较小，适合简单的并行任务

GPU需要访问CPU DRAM获取计算所需数据



GPU

星火计划



3.统一内存访问：显式访问-非固定内存

显式内存访问机制：

通过显式内存拷贝引擎(e.g. cudaMemcpy)

实现CPU和GPU之间的基本数据拷贝，以页为单位

可以手动优化数据传输，降低传输冗余

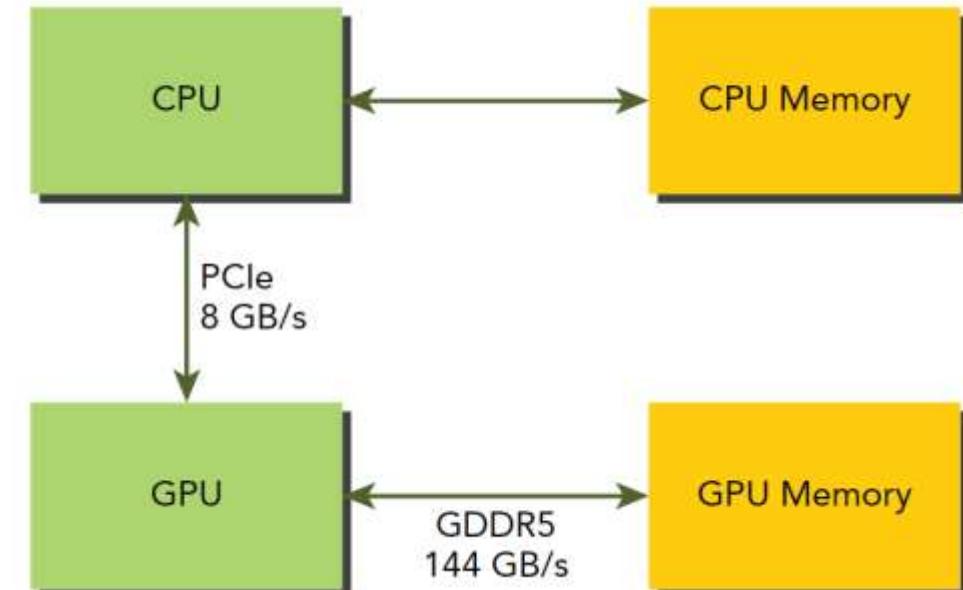
缺点：

使用不够方便，应用程序的修改更新难度大

应用只能看到虚拟的内存地址，操作系统可能随时更换物理地址的页，可能导致性能下降

应用（图计算）：

GraphReduce[SC 2015]、Graphie[ATC 2019]、Subway[EuroSys 2020]等工作都是针对显式访问进行优化的图处理框架



CUDA原语：cudaMemcpy(参数)

参数：

- cudaMemcpyHostToHost
- cudaMemcpyHostToDevice
- cudaMemcpyDeviceToHost
- cudaMemcpyDeviceToDevice

星火计划



3.统一内存访问：显式访问-固定内存

固定内存访问机制：

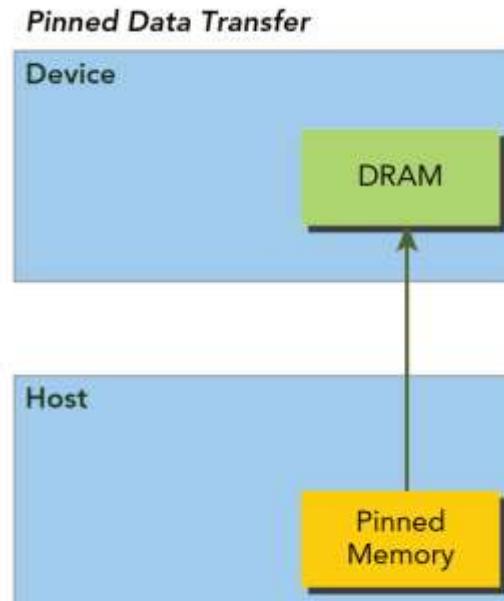
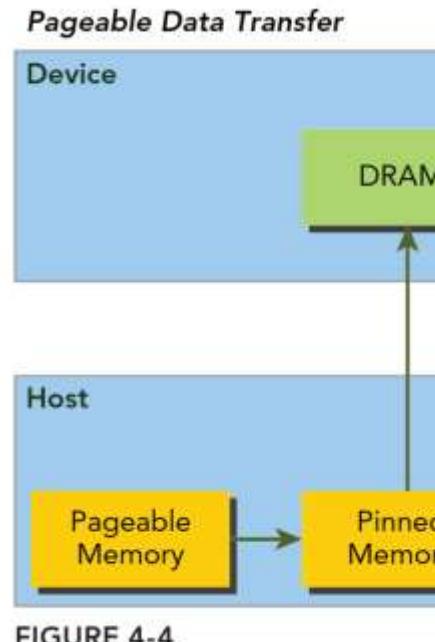
在数据传输之前，CUDA驱动会锁定页面，或者直接分配固定的主机内存，将主机源数据复制到固定内存上，然后从固定内存传输数据到设备上

传输速度更快，所以对于大规模数据，固定内存效率更高。

缺点：

固定内存的释放和分配成本比可分页内存要高很多，但是**应用（图计算）**：

GraphReduce[SC 2015]、Graphie[ATC 2019]、Subway[EuroSys 2020]等工作都是针对显式访问进行优化的图处理框架



CUDA原语：cudaHostAlloc(参数)

星火计划



3.统一内存访问：零拷贝

零拷贝访问机制：

当设备内存不足的时候，GPU设备直接访问主机内存的同一内存地址，使用固定内存，不可分页

允许线程以更小的粒度访问处理器内存

无需在加速器和处理器内存间进行页面迁移

优化：

内存**合并\对齐**访问

缺点：

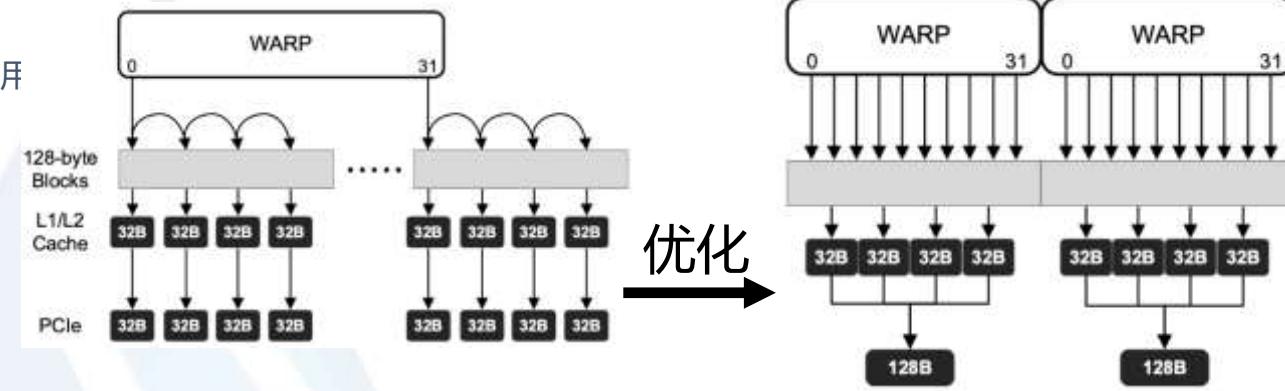
不提供数据重用功能，同一数据的多次访问会导致多次单独的数据传输

带宽利用率不稳定，请求不足会导致带宽浪费

比设备主存储器更慢，频繁读写时内存效率极低

应用（图计算）：

EMOGI[VLDB 2020]



CUDA原语：

cudaHostGetDevicePointer()
cudaHostAlloc(参数)

参数

- cudaHostAllocDefalt
- cudaHostAllocPortable
- cudaHostAllocWriteCombined

星火计划



3.统一内存访问：统一虚拟寻址UVA

UVA机制：

多台设备内存和主机内存被映射到同一虚拟内存地址中

分配的固定主机内存具有相同的主机和设备地址，可以直接将返回的地址传递给核函数

CUDA原语：

cudaMalloc()

cudaHostAlloc(参数)

参数

- cudaHostAllocMapped

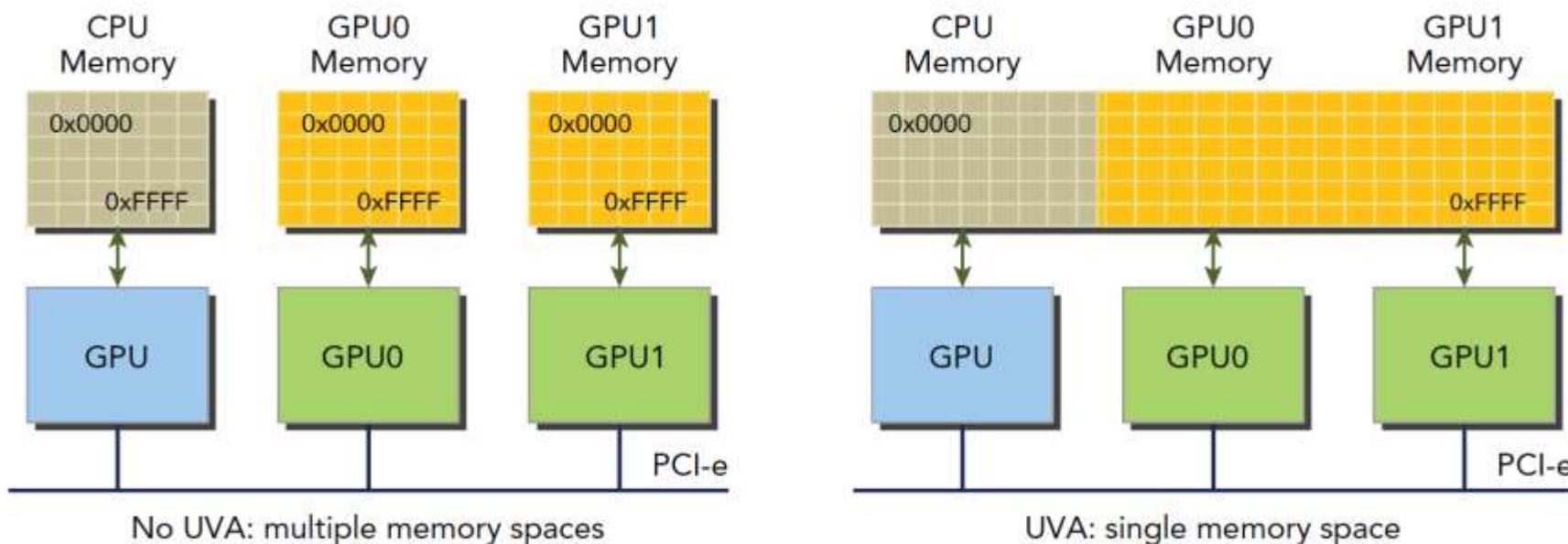


FIGURE 4-5

星火计划



3.统一内存访问：统一内存寻址UM

统一内存访问机制：

CPU和GPU可以共同访问具有连贯内存视图的单一地址空间

依靠GPU驱动和硬件来自动进行数据传输与迁移

应用开发友好

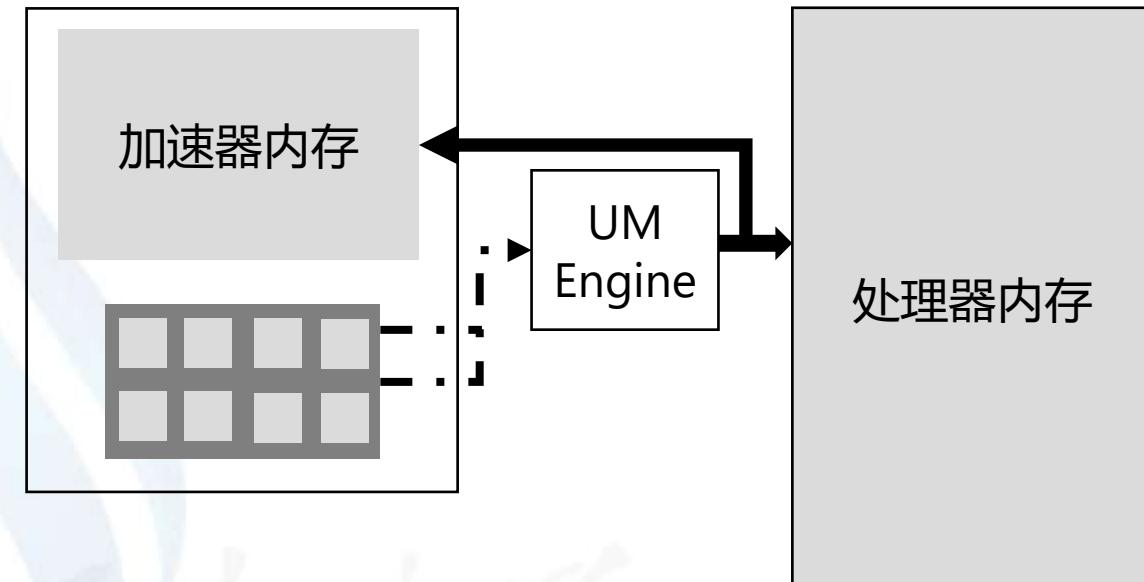
缺点：

自动迁移代价较大（传输开销以及发生页故障后页表更新开销等）

可能会引入冗余数据传输（传输粒度通常为页）

应用（图计算）：

HALO[VLDB 2020]、Grus[TACO 2021]等工作



CUDA原语：
`cudaMallocManaged()`

星火计划



目 录

1

**内存背景及需求：
大容量高密度内存简介**

2

现代扩展后的内存层级

3

**处理器和加速器间的
统一内存访问机制**

4

现代服务器的存算分离架构

星火计划

中国 电信 转型 专业 领军 人才 培养 项目



4. 存算分离架构：概述

分离式架构：

设计目标：资源灵活扩展，随用随取，按需分配

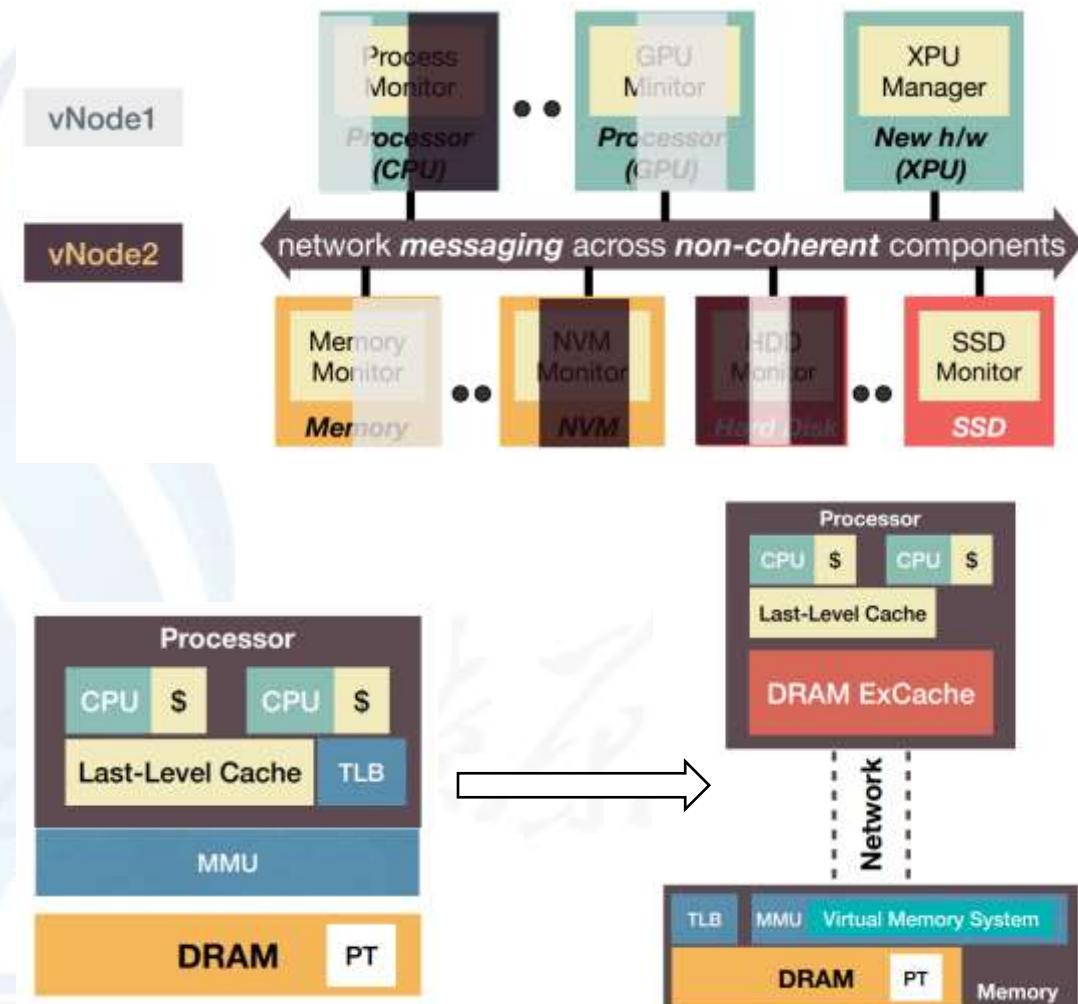
理想方式：将计算、内存、存储、加速器、网络等device完全分离，各自控制，形成统一网络互联的资源池

相似说法：分离可组合架构CDI (Composable Disaggregated Architecture)，超融合HCI (Hyper Converged Infrastructure)

分离式内存 Disaggregated memory (DM):

关注原因：应用面临内存大小和内存性能的瓶颈

理想方式：分离式大内存作为一个高效可用的部件，供传统计算单元使用，按需调取内存容量



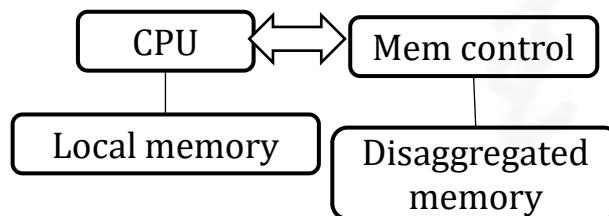
星火计划



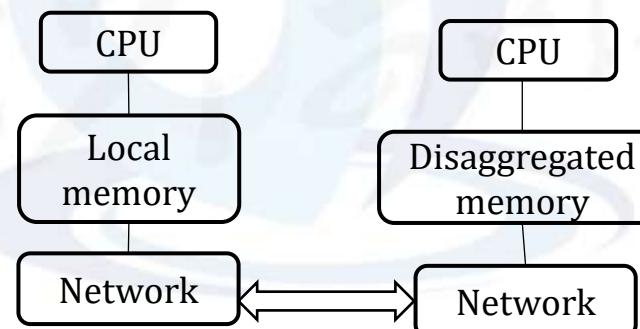
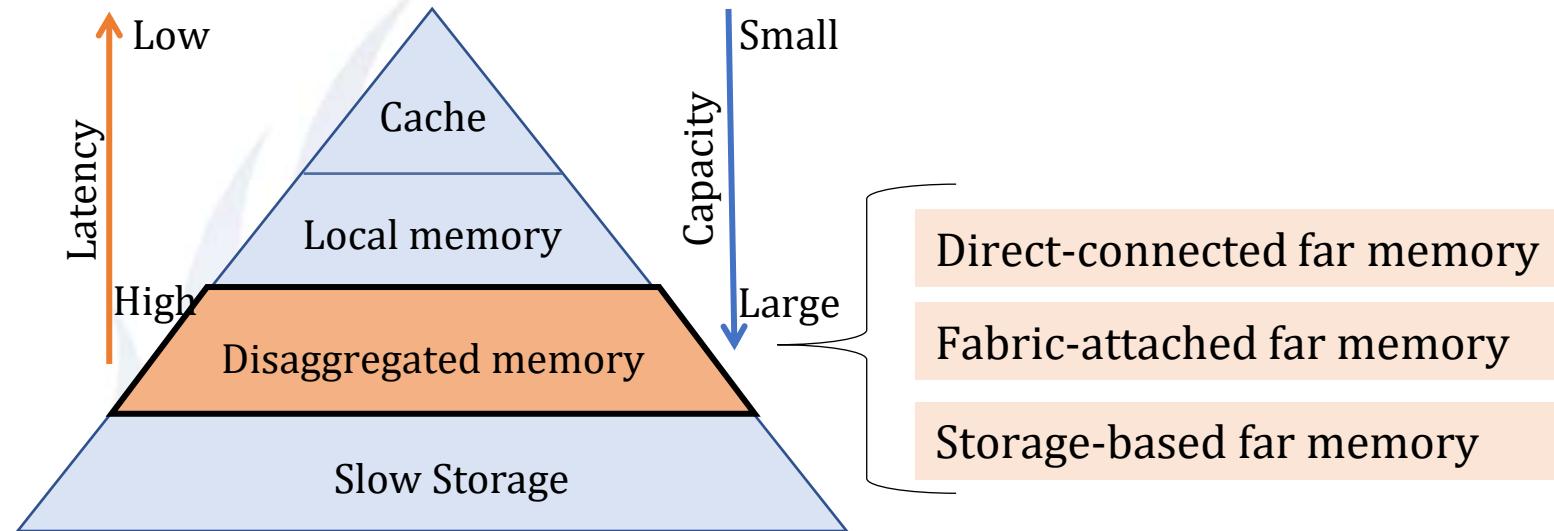
4. 存算分离架构：分离式内存和远内存

远内存 Far memory (FM):

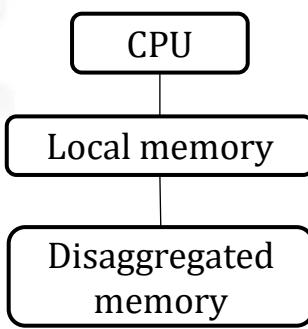
- 概念：更远的内存部件的访问方式
- 与DM关系： $DM \subseteq FM$
- 在一些论文中，DM代表着对专用DM部件的访问，而FM代表所有远于local memory 的内存系统的访问。
- Far memory 强调使用现有系统、硬件、架构实现远内存访问的方法与软件系统



Direct-connected far memory



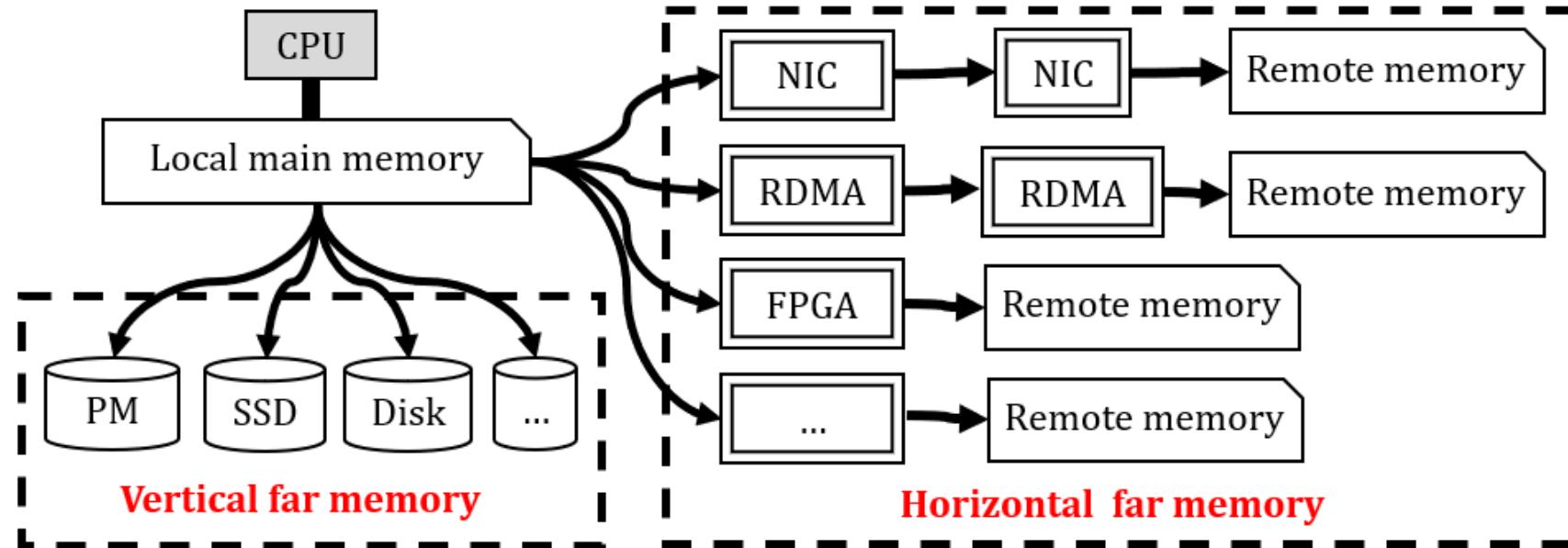
Fabric-attached far memory



Storage-based far memory



4. 存算分离架构：混合异构远内存



节点内（纵向）远内存

- 通过直连接口、链接服务器内部的内存或存储资源

节点间（横向）远内存

- 通过网络设备与网络协议连接服务器外部资源

星火计划



4. 存算分离架构：分离式内存系统设计

遇到的挑战：

RMA (remote memory access) 问题

- RMA latency is higher than local DRAM access latency
- RMA bandwidth is smaller than local memory bandwidth

系统支持问题

- 已有系统软件需要为分离式内存的实际访问做出更改
- 需要设计新的运行时接口等软件层面支持供上层应用使用

硬件设备问题

- 基于新型的分离式内存的研究其使用方式
- 基于现有服务器框架研究其系统软件适配

远内存应用层

- 应用访问接口
- 数据格式
- 内存管理的运行时

远内存系统层

- 内存管理的运行时
- 网络互联协议，CXL等
- 系统和驱动

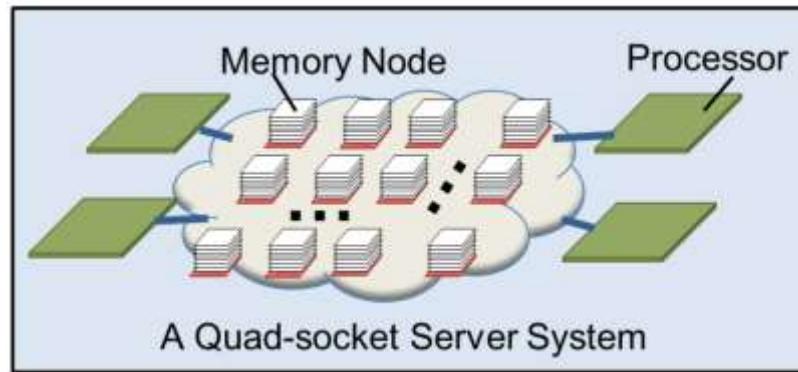
远内存硬件层

- 通用存储器件如SSD、PM、可编程存储器PIM等
- 网卡如RDMA、可编程网卡
- 高速通信电缆如铜缆

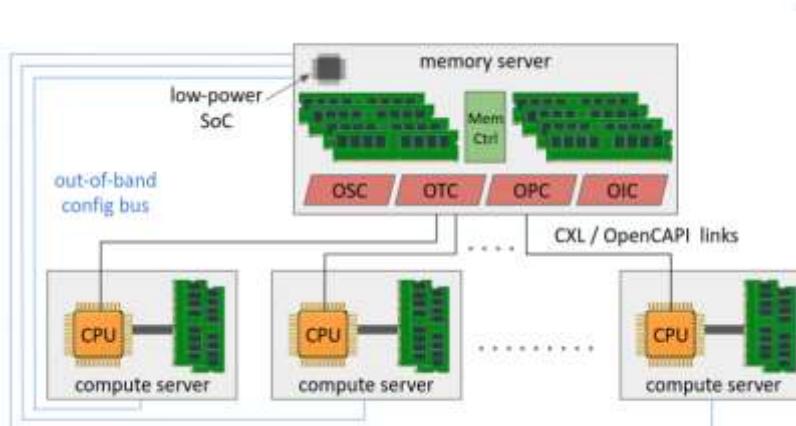
星火计划



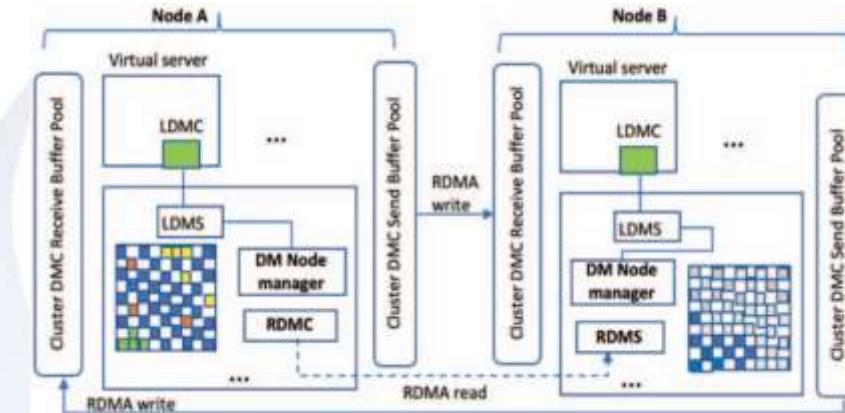
4. 存算分离架构：前沿架构



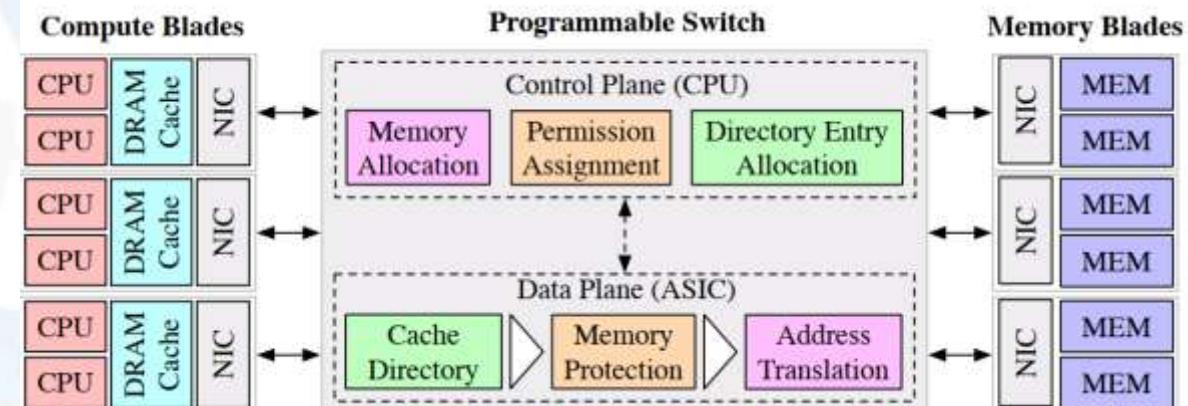
完全分离式架构



直连的分离式架构



基于网卡的分离式架构

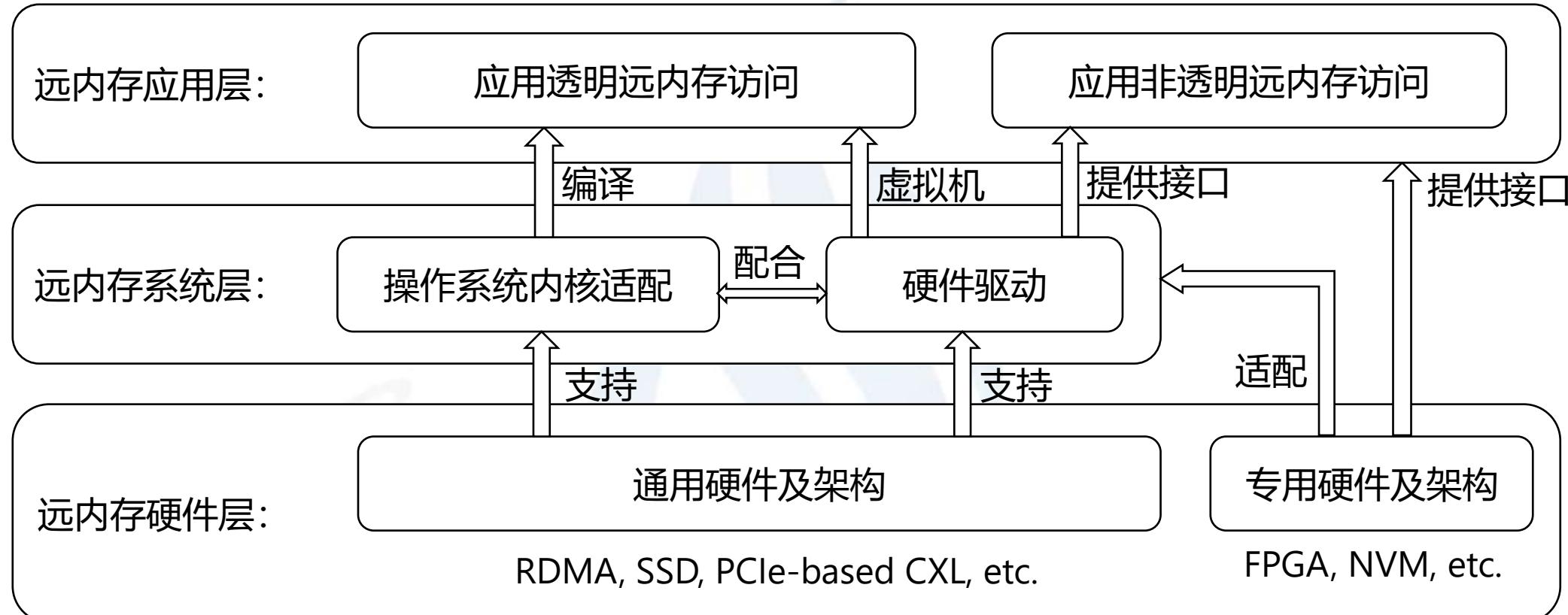


基于可编程switch的分离式Rack

星火计划



4. 存算分离架构：远内存系统层次



星火计划

中国电信转型专业领军人才培养项目

电子科技大学
SUSTech



4. 存算分离架构：远内存硬件层-通用型

通用远内存硬件

采用商业化的设备包括DRAM、SSD、RDMA等通用设备，在现有的服务器架构基础上组建远内存访问硬件层。

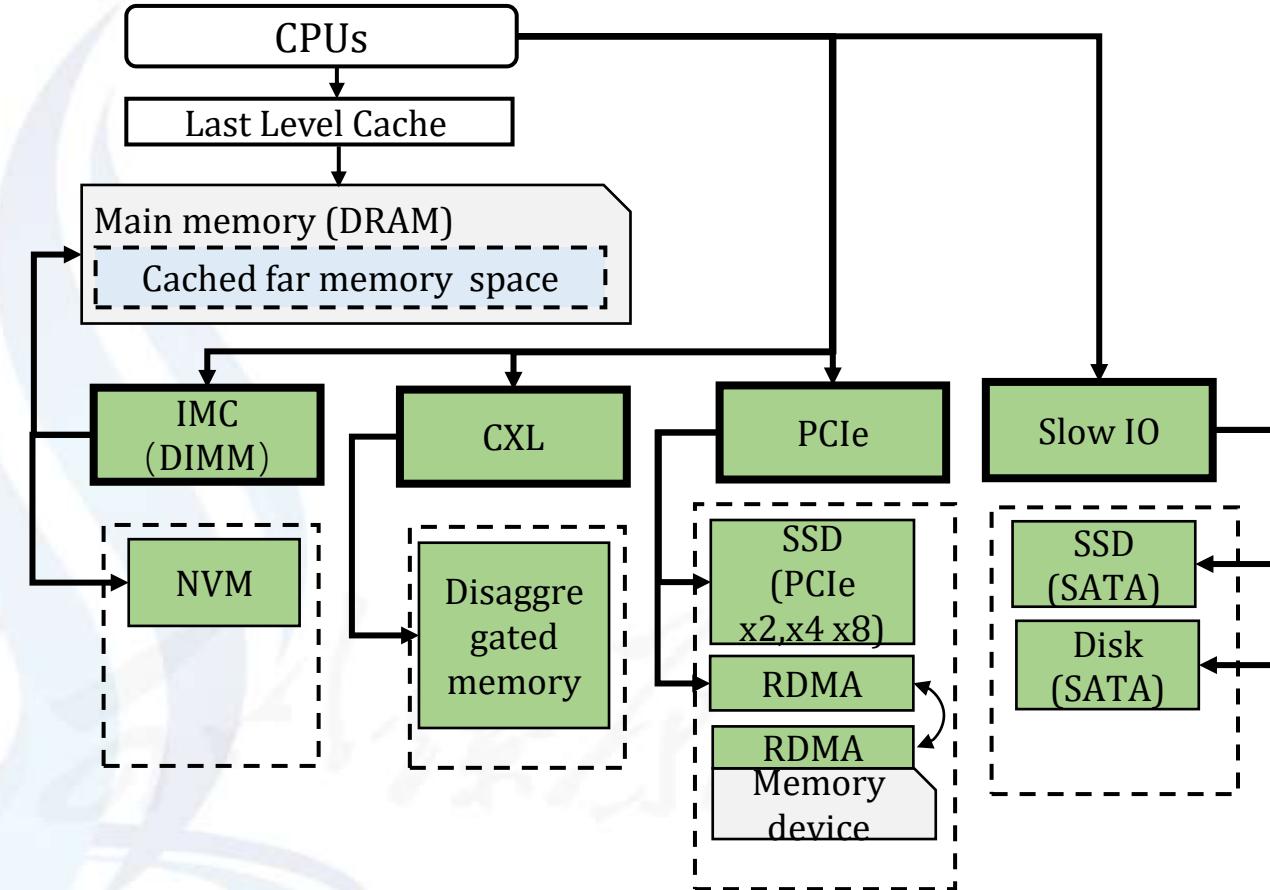
优势

- 硬件成本低
- 开发相对友好
- 快速适应现有架构

设计思想

- 系统层的配合设计
- 应用透明或应用感知

- 代表工作: Infiniswap, Fastswap, XMemPod, TMO, g-zswap, AIFM, pDPM, etc.



星火计划



4. 存算分离架构：远内存硬件层-专用型

专用远内存硬件

定制化的远内存硬件主要是采用基于新型内存器件设计、内存网络搭建、新型内存通信协议、新型网卡，组件专用远内存硬件层。

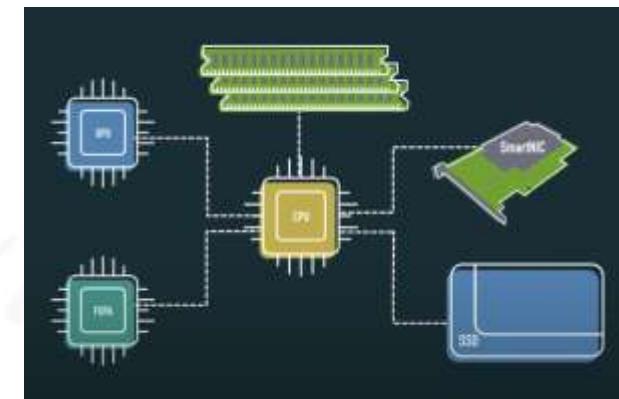
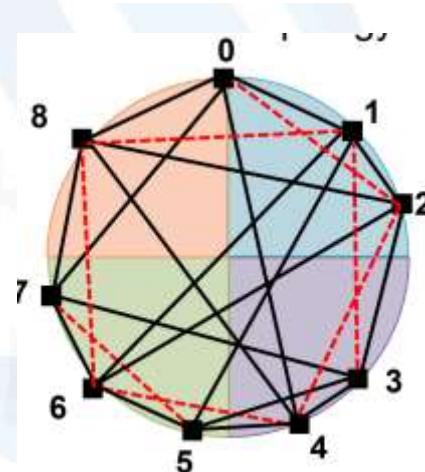
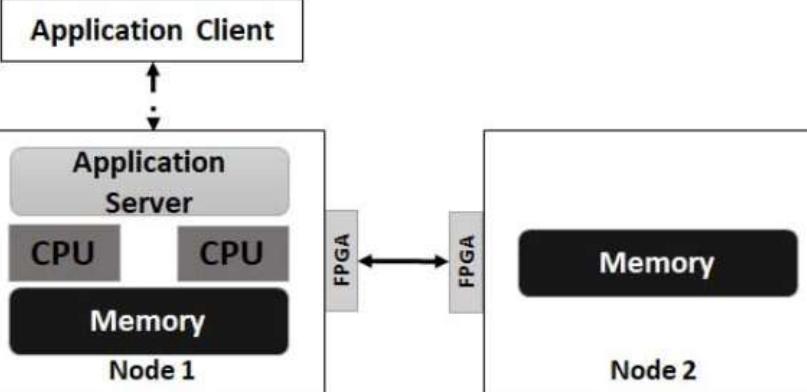
优势

- 性能好
- 创新性强
- 不断适应新的架构

设计思想

- 硬件逻辑设计
- 系统层的配合设计
- 应用透明或应用感知

- 代表工作: Cilo, StringFinger, CXL, MIND, ThymesisFlow, Zombieland, etc.



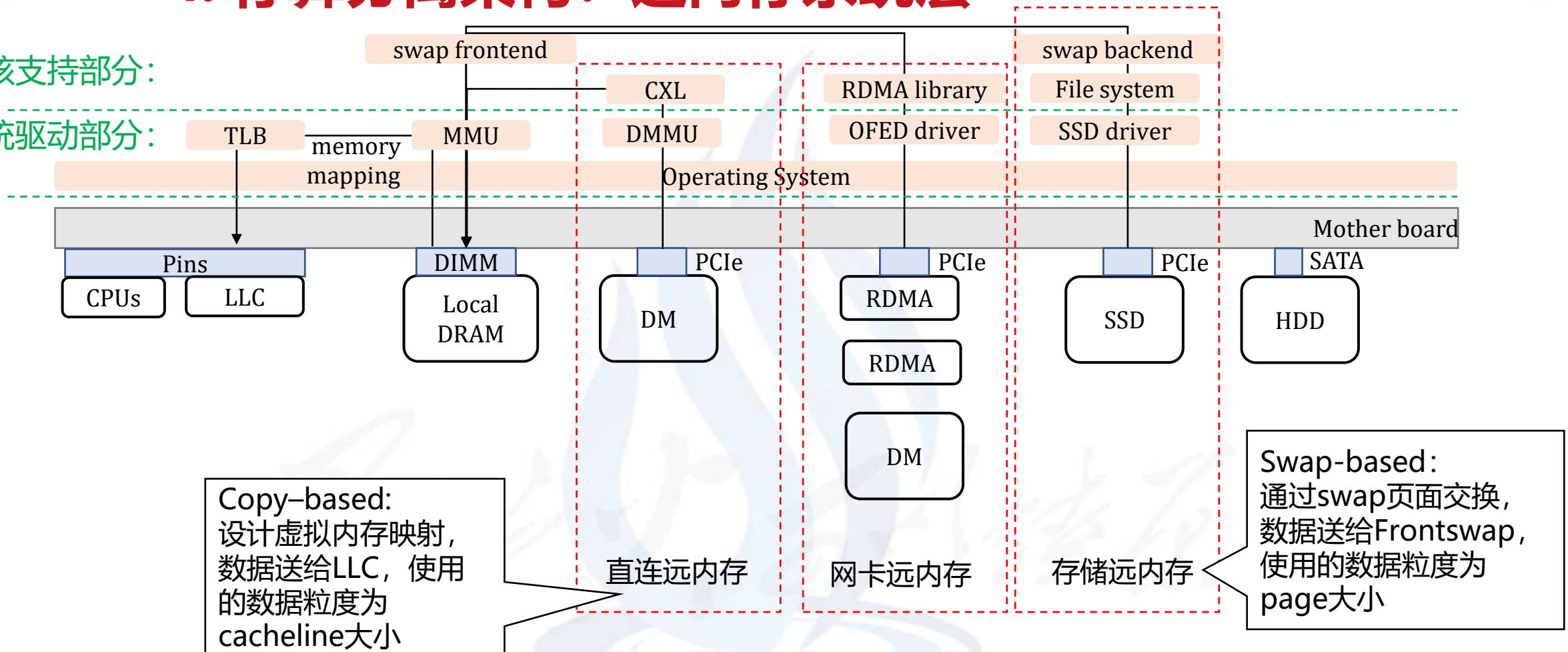
星火计划



4. 存算分离架构：远内存系统层

内核支持部分：

系统驱动部分：



- 代表工作: Infiniswap, Fastswap, XMeMPod, g-zswap, Kona, CXL-based disaggregation, etc.

星火计划



4. 存算分离架构：远内存应用层

应用透明的远内存

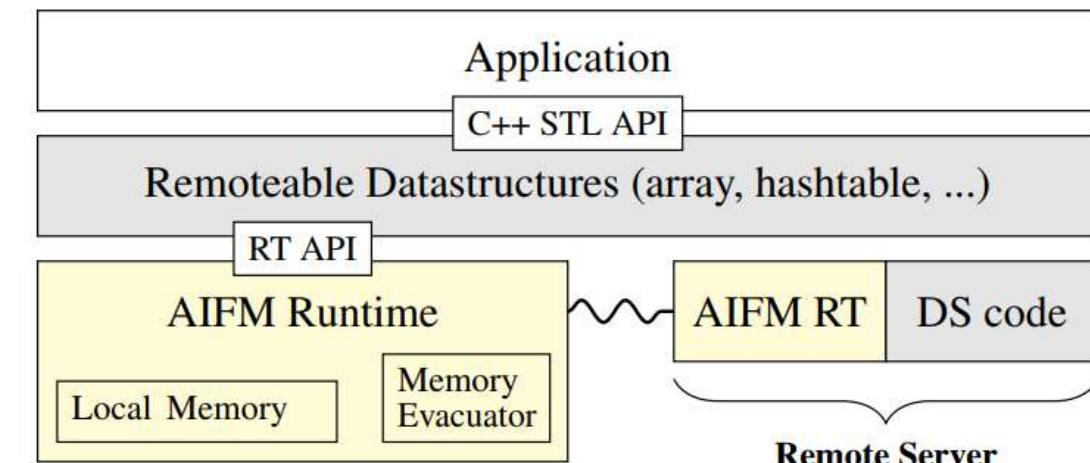
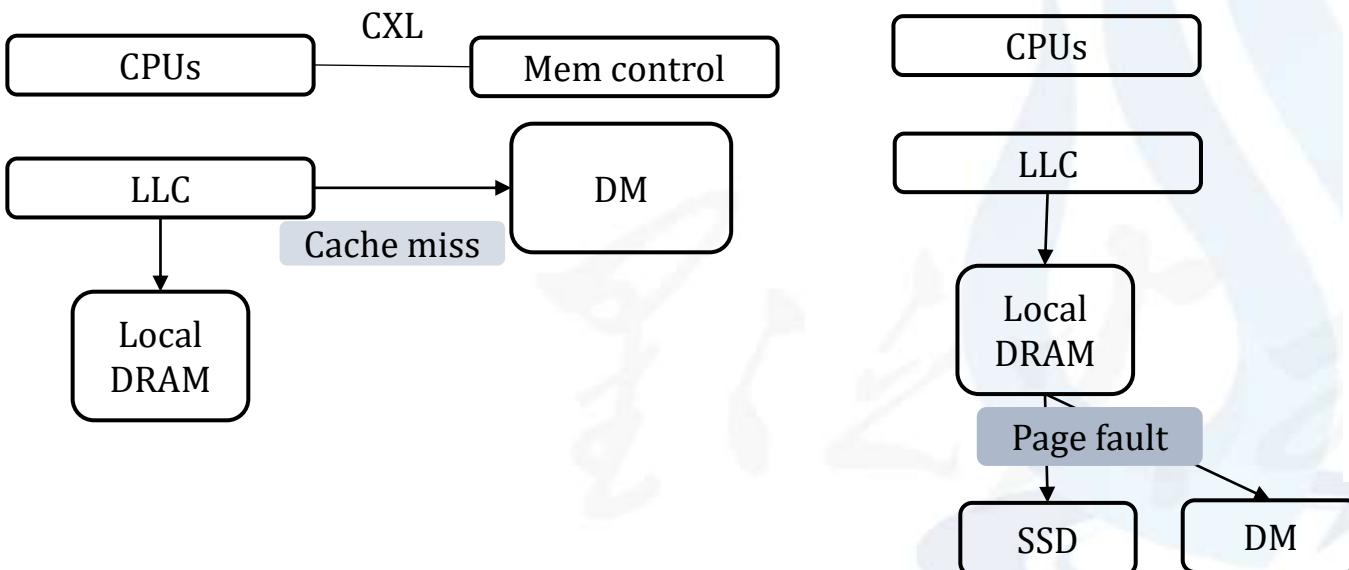
透明远内存主要是通过在系统层修改data mapping实现的：

- 修改LLC的cacheline 缓存
- 修改swap机制的后端

应用非透明的远内存

非透明远内存主要是通过在设计高效易用的编程接口：

- 设置不同的数据offloading大小格式
- 设置不同的冷热数据分布机制，数据压缩等



- 代表工作：FaRM, Remote regions, Lite, AIFM, FreeFlow, Fargraph, Kona, etc.

星火计划

4. 存算分离架构：分离式内存相关研究





二、分离式内存技术案例分析

讲者：李超

上海交通大学 计算机系 SAIL实验室

2023年2月

星火计划

中国电信转型专业领军人才培养项目



目录

1

国外企业技术案例介绍

2

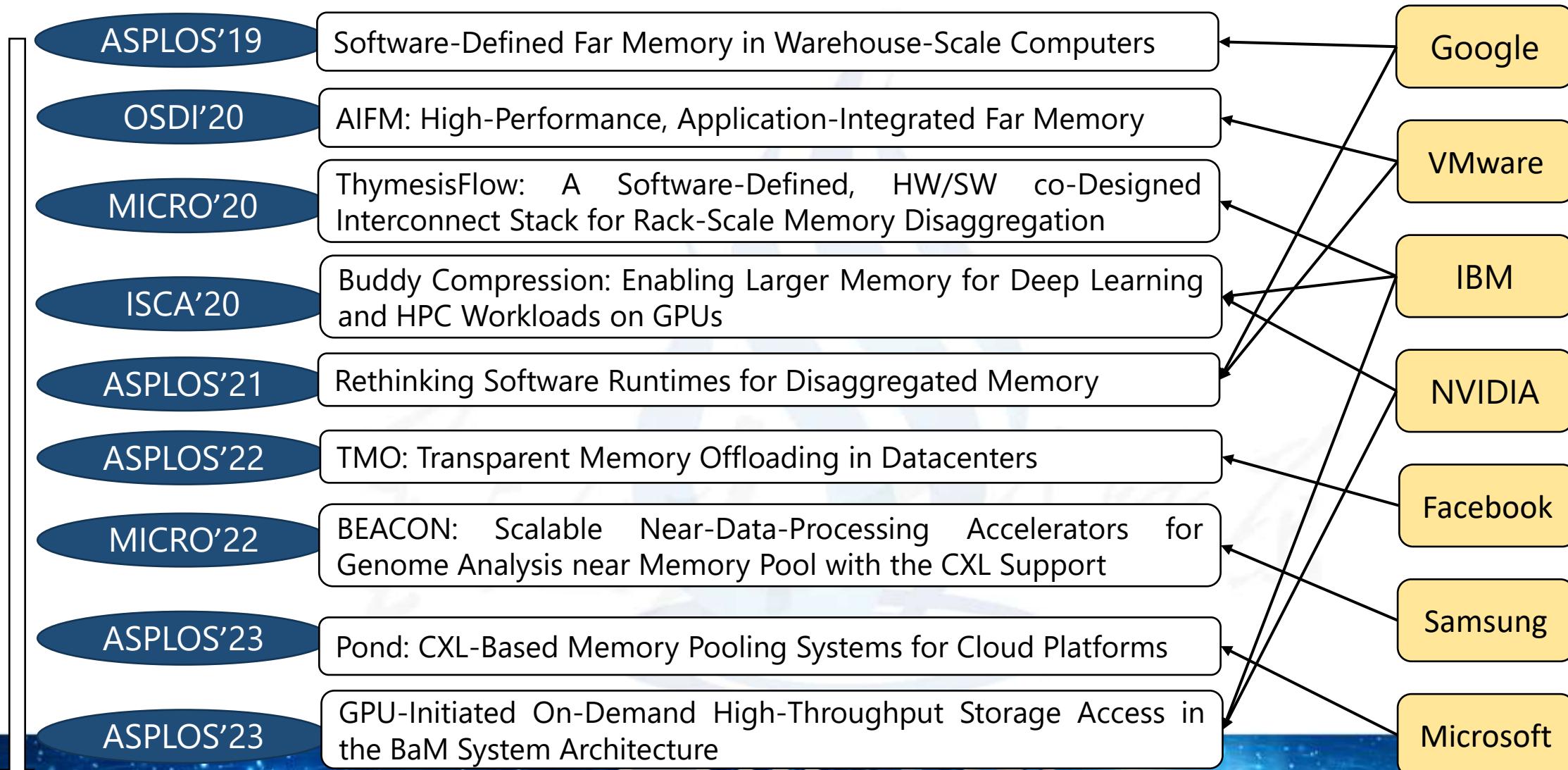
学术界研究分享

星火计划

中国电信转型专业领军人才培养项目



1.国外企业技术案例





1.国外企业技术案例：Google

ASPLOS'19

Software-Defined Far Memory in Warehouse-Scale Computers

主要发现：

存在大量的冷页碎片（32%）

冷页比例变化范围大（1~52%）

主要提出：

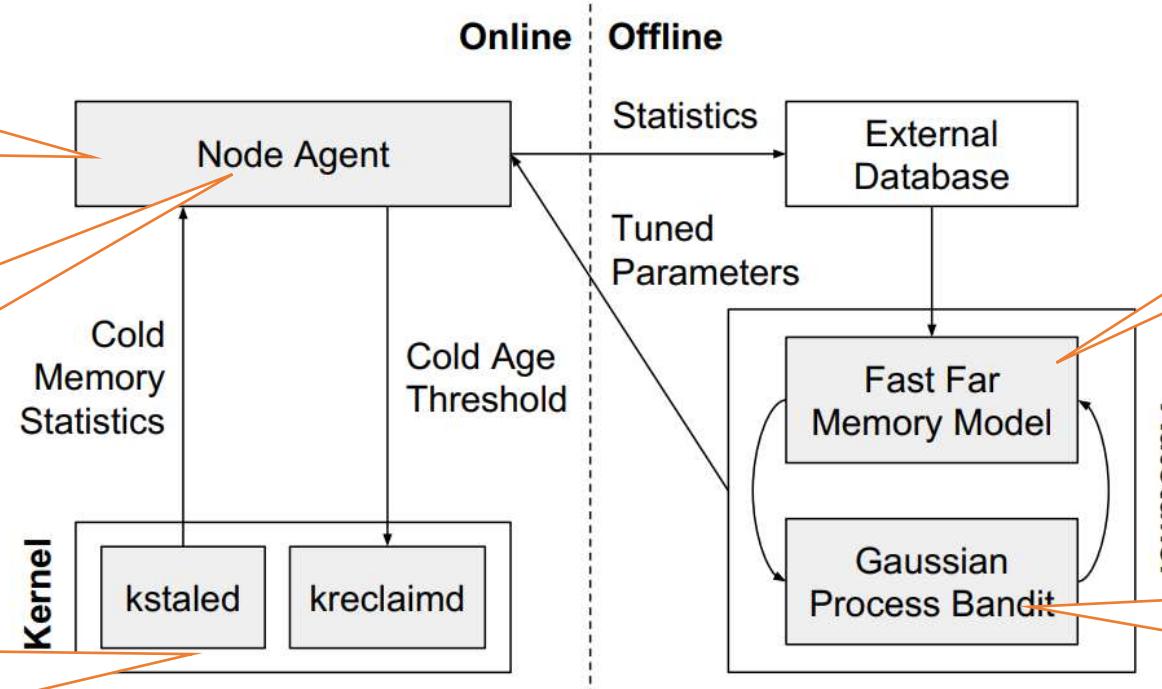
使用zswap将冷页压缩

按照任务SLO将冷页换出到远内存

周期性地更新冷页阈值，并收集数据供线下分析

Zswap在S秒内启动，设置下一秒的百分数为当前的第K百分位

周期性地扫描页表中的访问位，在DRAM和zswap中移动冷页面



并行化独立控制不同job，根据trace计算S和K

预测算法，寻找重要配置参数，尽量减少试验次数

Figure 4. Overall system design.



1.国外企业技术案例：Google

ASPLOS'19

Software-Defined Far Memory in Warehouse-Scale Computers

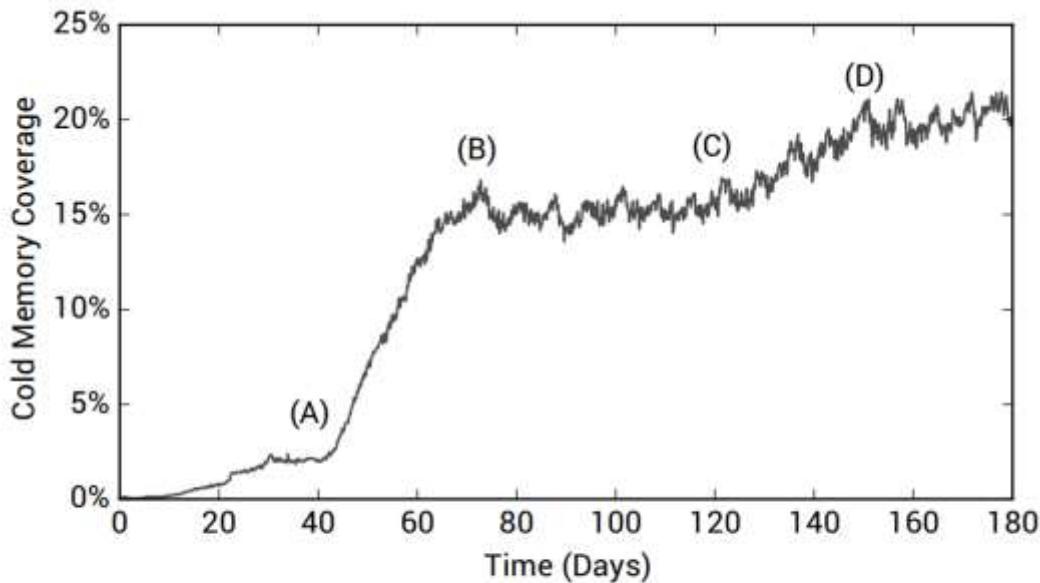


Figure 5. Cold memory coverage over time. zswap with hand-tuned parameters was rolled out during (A) to (B); the autotuner was rolled out during (C) to (D).

zswap : 15% of cold memory coverage.

ML-based autotuner: increased the cold memory coverage to 20%

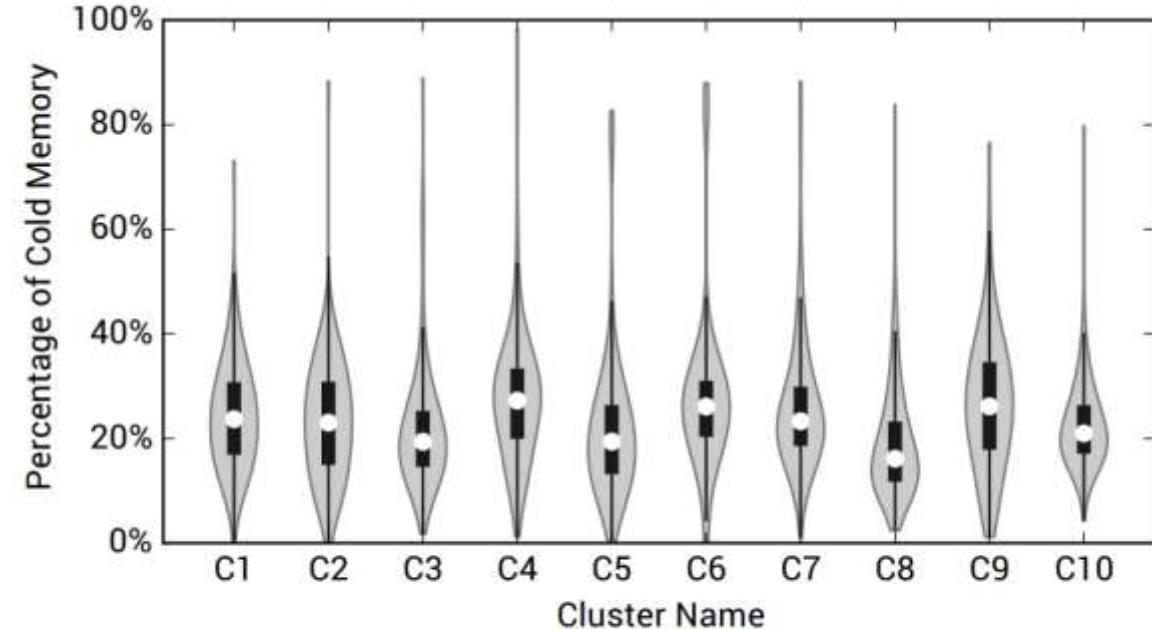


Figure 6. The distribution of the cold memory coverage across the machines in the top 10 largest clusters.

32% for the upper bound for cold memory ratio

67% cost reduction (3x compression ratio) for compressed pages



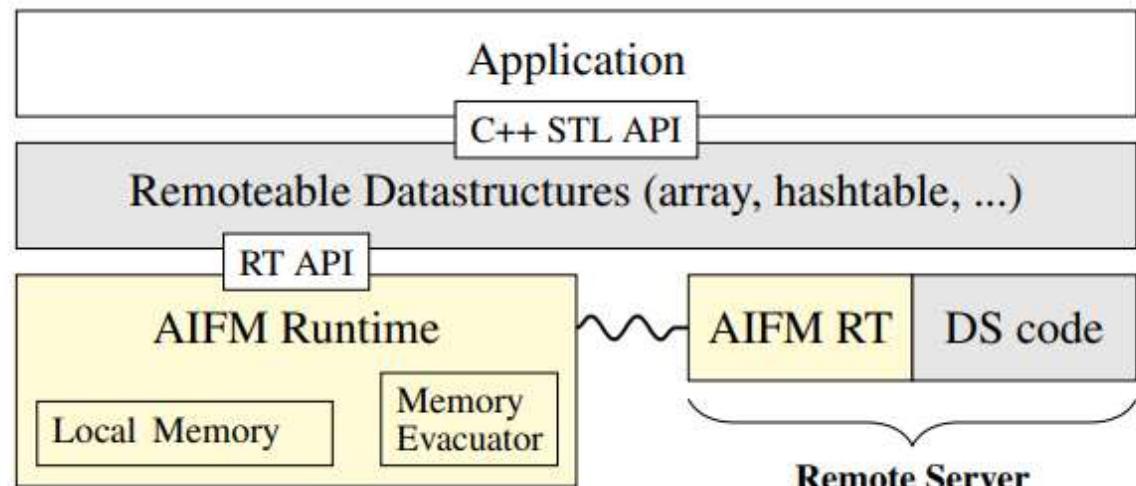
1.国外企业技术案例：VMware

OSDI'20

AIFM: High-Performance, Application-Integrated Far Memory

核心思想：

利用应用层的API实现本地内存和远内存的一致访问控制，需要用户根据AIFM提供的接口进行程序的适配



```
std::unordered_map<key_t, int> hashtable;
std::array<data_t> arr;

void print_data(std::vector<key_t>& request_keys) {
    int sum = 0;
    for (auto key : request_keys) {
        sum += hashtable.at(key);
    }
    std::cout << arr.at(sum) << std::endl;
}
```

The same code written using AIFM looks like this:

```
RemHashtable<key_t, int> hashtable;
RemArray<data_t> arr;

void print_data(std::vector<key_t>& request_keys) {
    int sum = 0;
    for (auto key : request_keys) {
        DerefScope s1; // Explained in Section 4.2.2.
        sum += hashtable.at(key, s1);
    }
    DerefScope s2;
    std::cout << arr.at(sum, s2) << std::endl;
}
```

星火计划



1.国外企业技术案例：IBM

MICRO'20

ThymesisFlow: A Software-Defined, HW/SW co-Designed Interconnect Stack for Rack-Scale Memory Disaggregation

主要贡献：

- 提出了带有软件控制平面的全硬件分离式内存硬件架构原型
- 可以动态创建NUMA节点，并和其Linux内核进行内存的实时卸载和读取
- 在现有商业硬件搭建原型，有着完整的软硬件系统

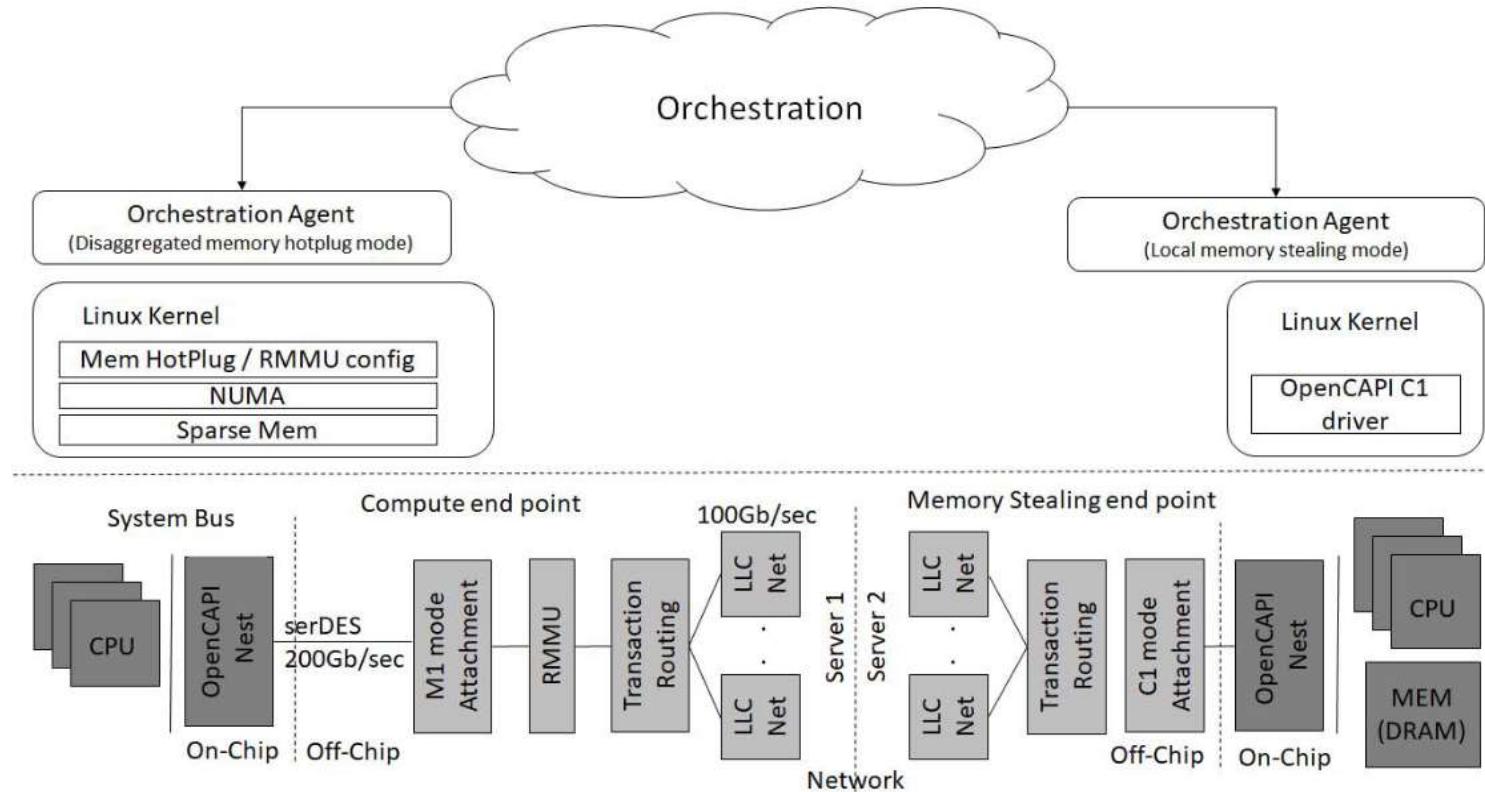


Fig. 2. ThymesisFlow overall architecture.

星火计划



1.国外企业技术案例：IBM

MICRO'20

ThymesisFlow: A Software-Defined, HW/SW co-Designed Interconnect Stack for Rack-Scale Memory Disaggregation

- 此外，ThymesisFlow 中为了实现动态加入内存节点，将地址空间映射到内存节点内核，也为内存池构建一致性的保存问题提供了解决方案

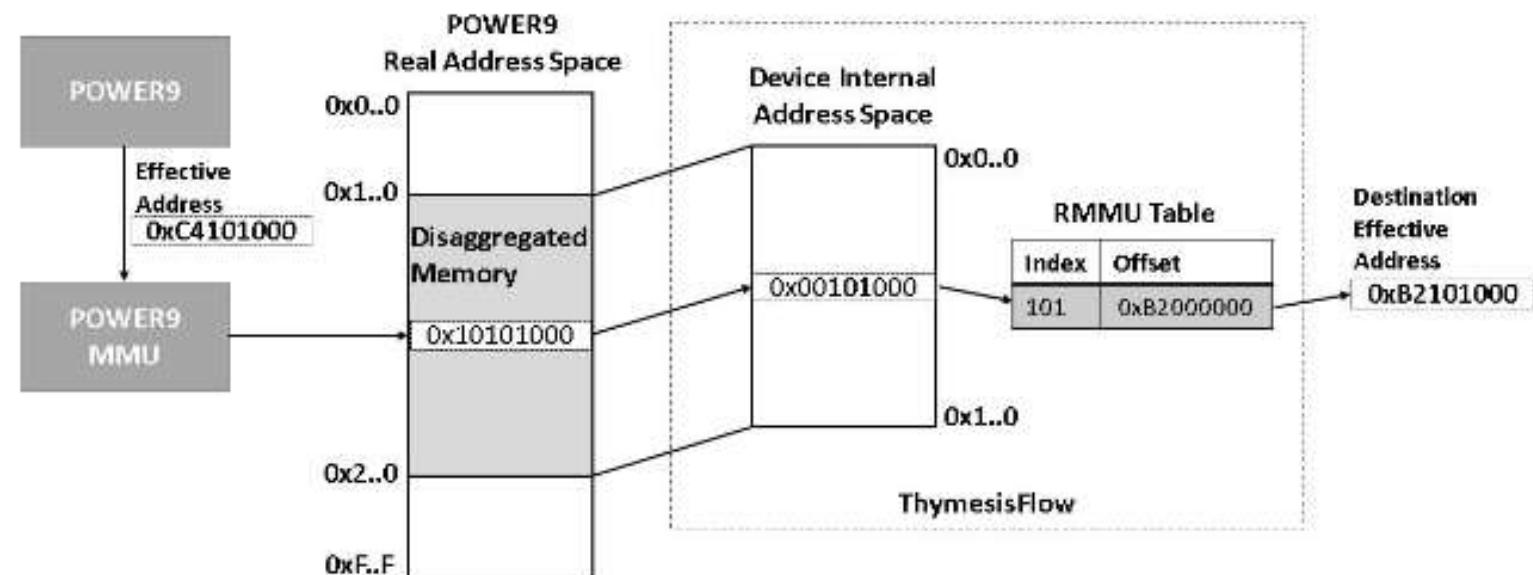


Fig. 3. ThymesisFlow address translation process.

星火计划

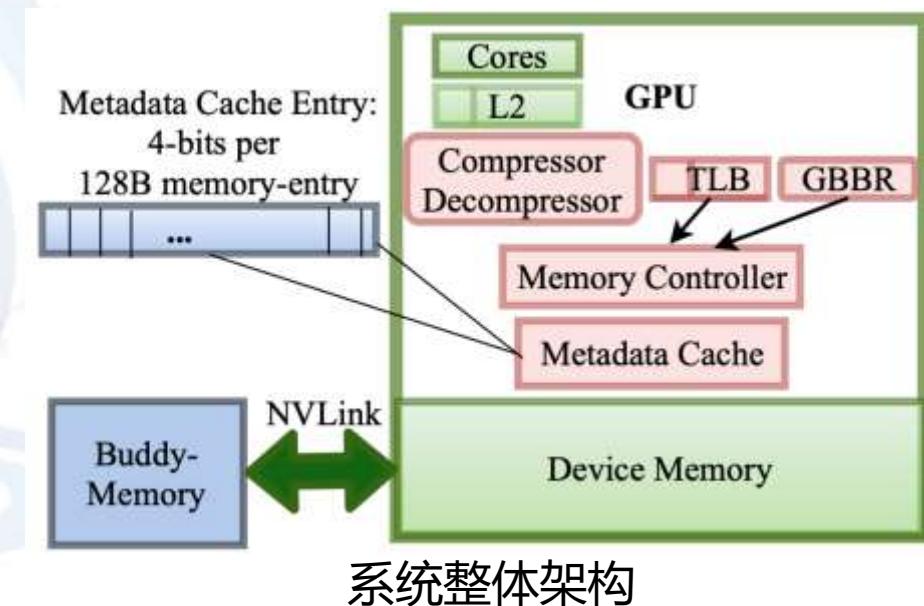
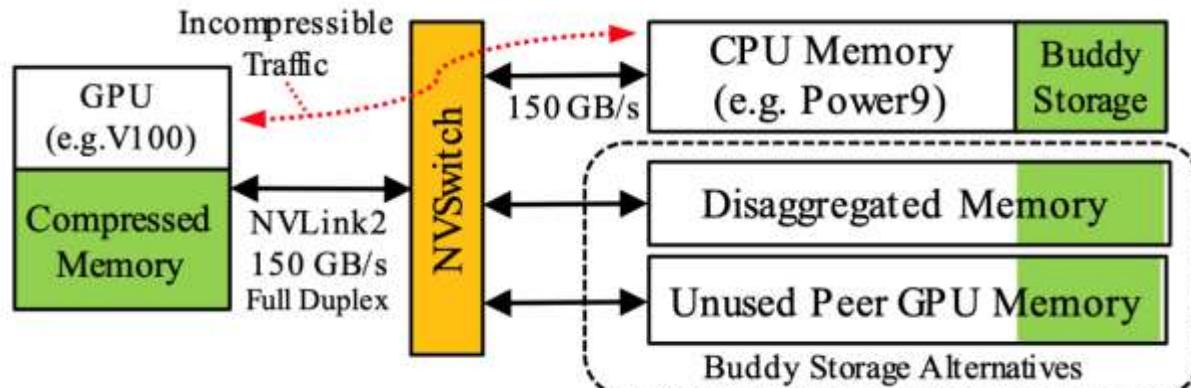


1.国外企业技术案例：NVIDIA

ISCA'20

Buddy Compression: Enabling Larger Memory for Deep Learning and HPC Workloads on GPUs

- 提出Buddy Compression方法，首次实现了使用压缩算法节省GPU内存使用
- 多数可压缩的内存条目完全由GPU内存访问，而少数的溢出则从GPU以外的内存中获取
- 理想情况下，远程存储可以使用系统内存，也可以使用分离式内存（如左图）
- 需对硬件进行相应更改



星火计划



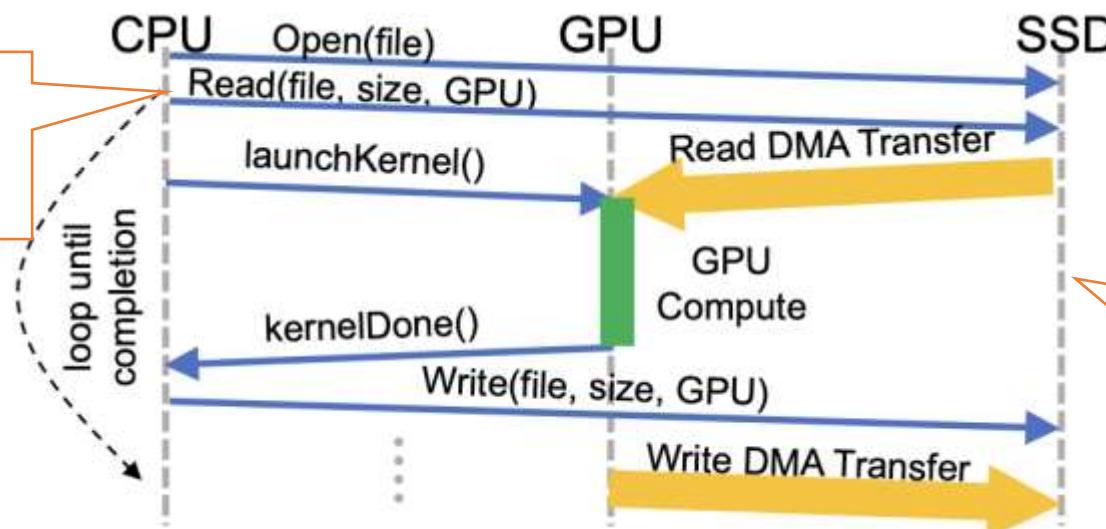
1.国外企业技术案例：NVIDIA

ASPLOS'23

GPU-Initiated On-Demand High-Throughput Storage Access in the BaM System Architecture

- 提出一种以加速器为中心的系统架构，GPU线程按需访问SSD中的数据（无需CPU参与）
- 实现高吞吐量的细粒度存储访问
- 提供软件定义缓存，用户利用局部性可以更好提升性能

缺点：传统处理器为中心的架构下，需要CPU控制数据分布



缺点：CPU与GPU频繁切换，会引入同步开销

缺点：某些计算任务中，CPU难以预知所需数据，可能预取大量无效数据

星火计划

中国 电信 转型 专业 领军 人才 培养 项目

TSJU



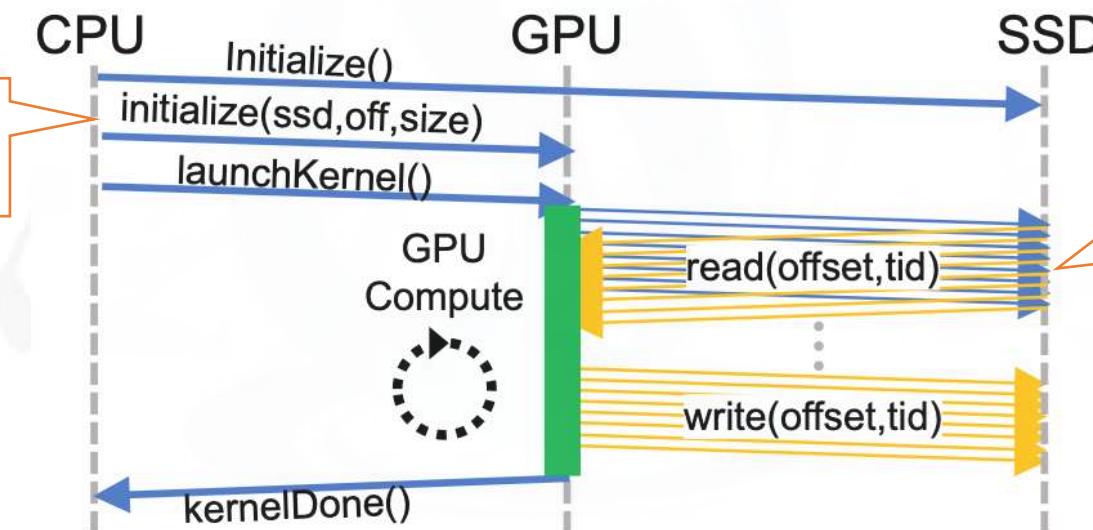
1.国外企业技术案例：NVIDIA

ASPLOS'23

GPU-Initiated On-Demand High-Throughput Storage Access in the BaM System Architecture

- 提出一种**以加速器为中心**的系统架构，GPU线程按需访问SSD中的数据（无需CPU参与）
- 实现高吞吐量的细粒度存储访问
- 提供软件定义缓存，用户利用局部性可以更好提升性能

优势：CPU只需进行相关初始化工作



优势：仅在计算需要某部分数据时读入，**减少无效数据传输**

优势：利用GPU的高并发性，**自动实现计算与数据传输阶段的重叠**



1.国外企业技术案例：Microsoft

ASPLOS'23

Pond: CXL-Based Memory Pooling Systems for Cloud Platforms

问题背景：

- 公有云的提供商需要平衡**性能要求**和**硬件成本**
- 现有的解决方案下，内存成本最多可占据总硬件成本的**50%**
- 对不同trace进行分析，**内存搁浅**（即某些资源全部租借但仍存在可用内存）是内存浪费和内存成本升高的重要来源
- 现有的解决方案（如压缩技术）会带来较大延时，无法满足用户要求

解决思路：

- CXL互连协议可以实现**ns**级别的内存访问
- 聚合过多socket的内存会增大时延，合理控制聚合socket数量可以达到时延和内存利用率的平衡
- 不同工作负载对于访问池化内存的**敏感性不同**，通过机器学习模型预测工作负载特性并对非敏感负载分配池化内存
- 引入**监控机制**，更新预测错误的负载的内存分配

星火计划



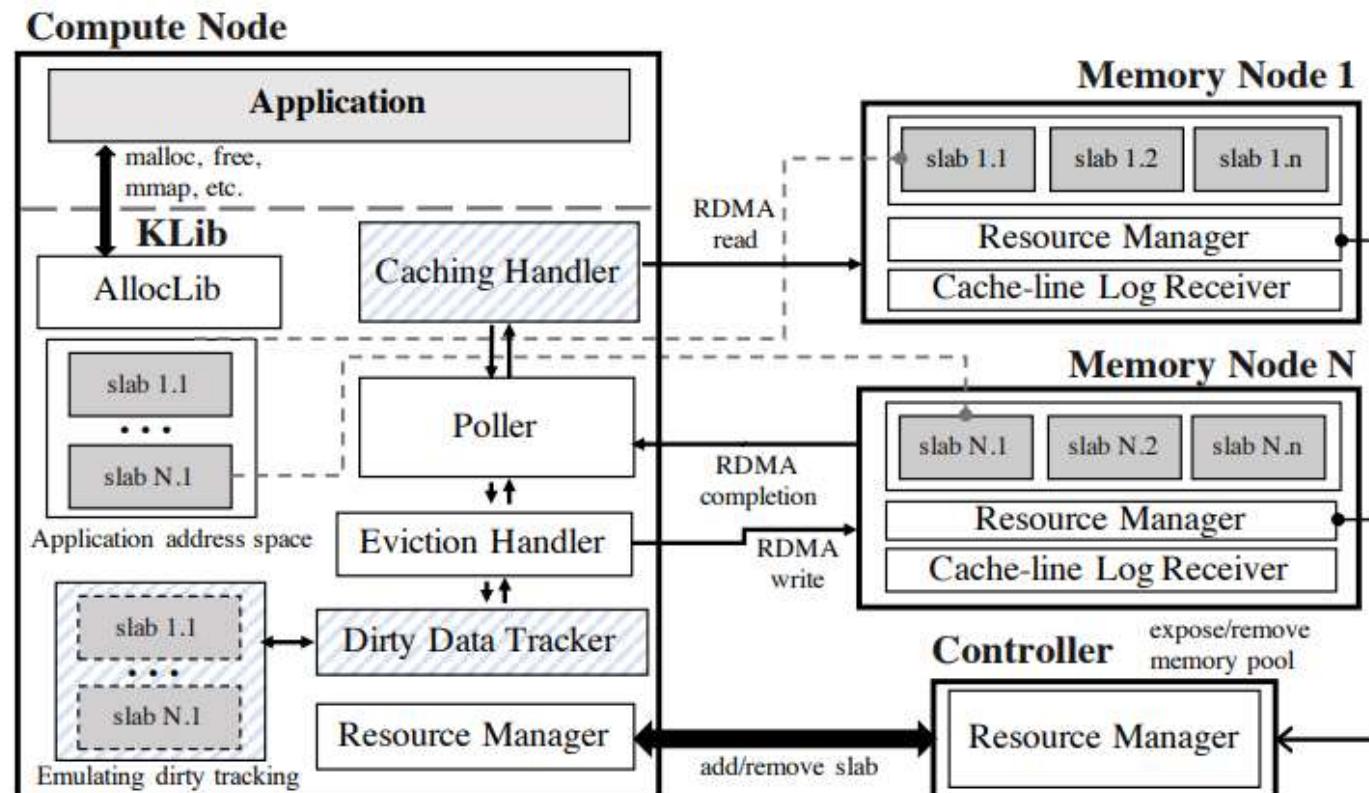
1.国外企业技术案例：Google

ASPLOS'21

Rethinking Software Runtimes for Disaggregated Memory

主要贡献：

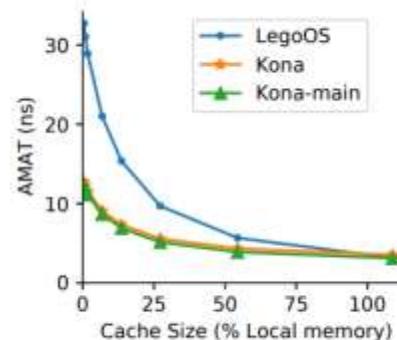
- 分析了现有分离式内存系统存在较大overhead和脏数据放大的缺点
- 提出了一种新的基于软件的分离式内存方案，利用硬件原语进行远程内存缓存和基于缓存一致性的缓存线脏数据跟踪
- 设计并实现了kona：一个使用新的硬件原语高效运行的软件框架

**星火计划**

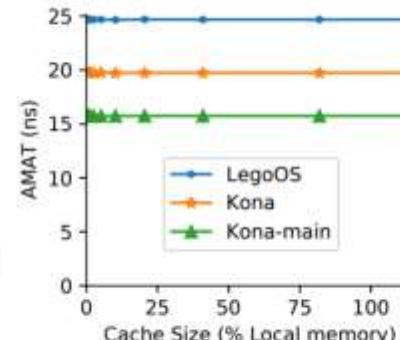
1.国外企业技术案例：Google

ASPLOS'21

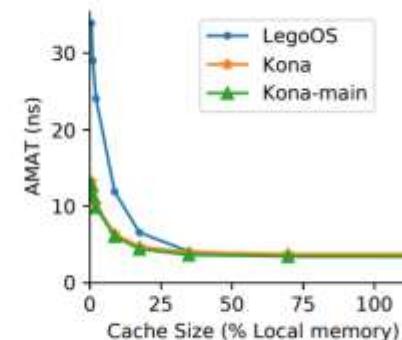
Rethinking Software Runtimes for Disaggregated Memory



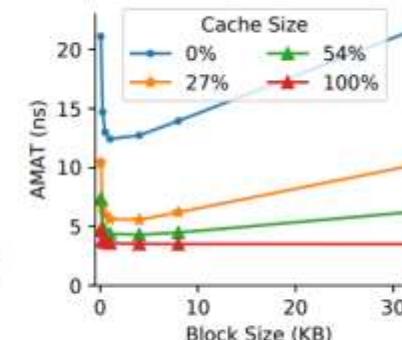
(a) Redis Rand



(b) Linear Regression



(c) Graph Coloring



(d) Redis Rand - Data fetch size

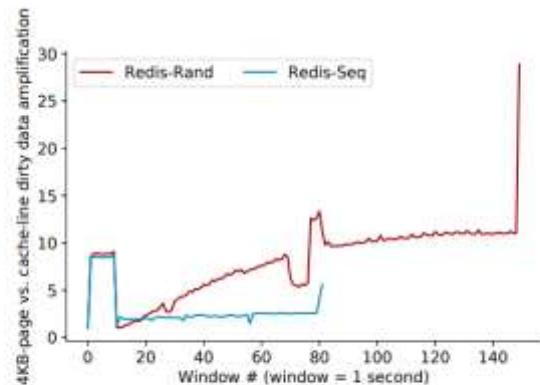


Figure 9: Dirty data amplification reduction.

✿ Kona在远端访问、脏数据卸载上有着更高的效率

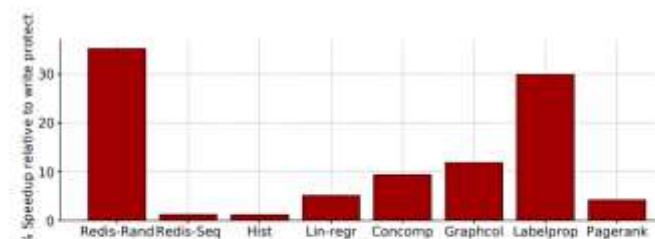


Figure 10: Speedup relative to write-protection.

星火计划



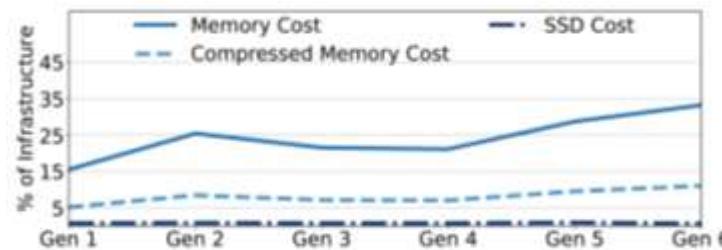
1.国外企业技术案例：Facebook

ASPLOS'22

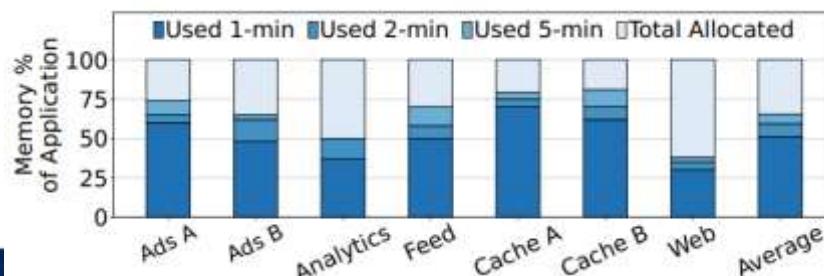
TMO: Transparent Memory Offloading in Datacenters

主要发现：

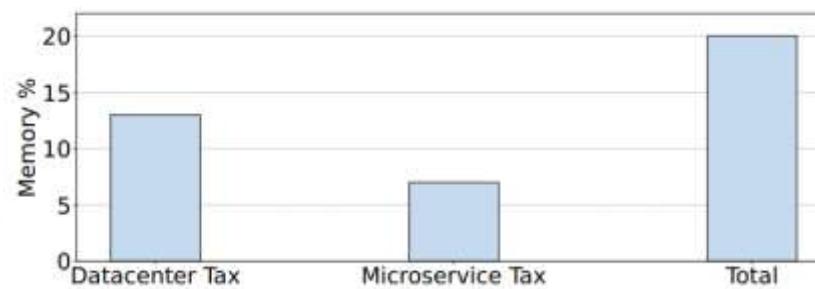
1. 内存成本逐渐升高；



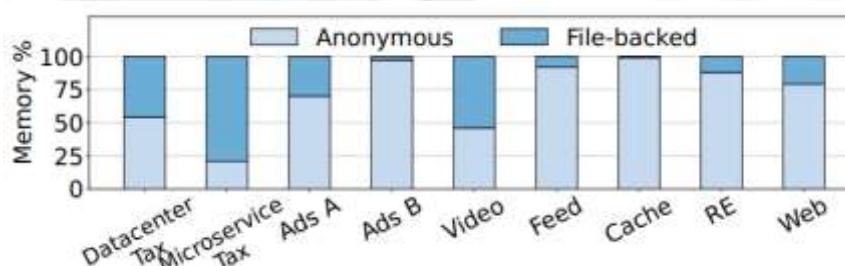
2. 不同应用冷热内存比例差异明显；



3. 内存税占总内存较大比例；



4. 不同应用页面类型差异明显；



主要贡献：

1. 提出PSI指标用以表征应用内存压力；
2. 提出用户态内核态结合的TMO架构；
3. 设计了基于压力的内存卸载算法并上传到Linux kernel

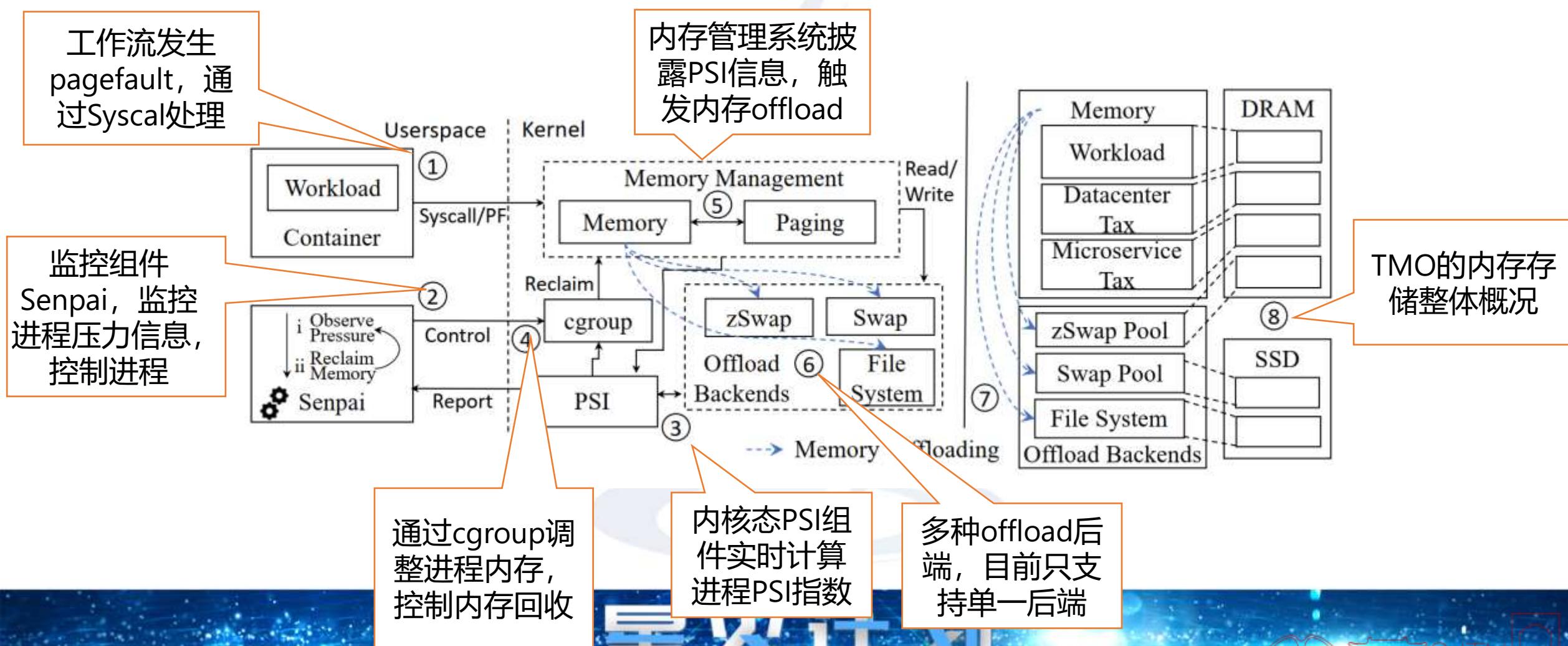




1.国外企业技术案例：Facebook

ASPLOS'22

TMO: Transparent Memory Offloading in Datacenters





1.国外企业技术案例：Samsung

MICRO'22

BEACON: Scalable Near-Data-Processing Accelerators for Genome Analysis near Memory Pool with the CXL Support

主要贡献：

- ④ 基于CXL设计了分离式内存池架构，用于基因分析运算加速
- ⑤ BEACON的架构使得未经修改的 CXL-DIMM 能够有效地按需扩展内存，并消除了通信的性能瓶颈

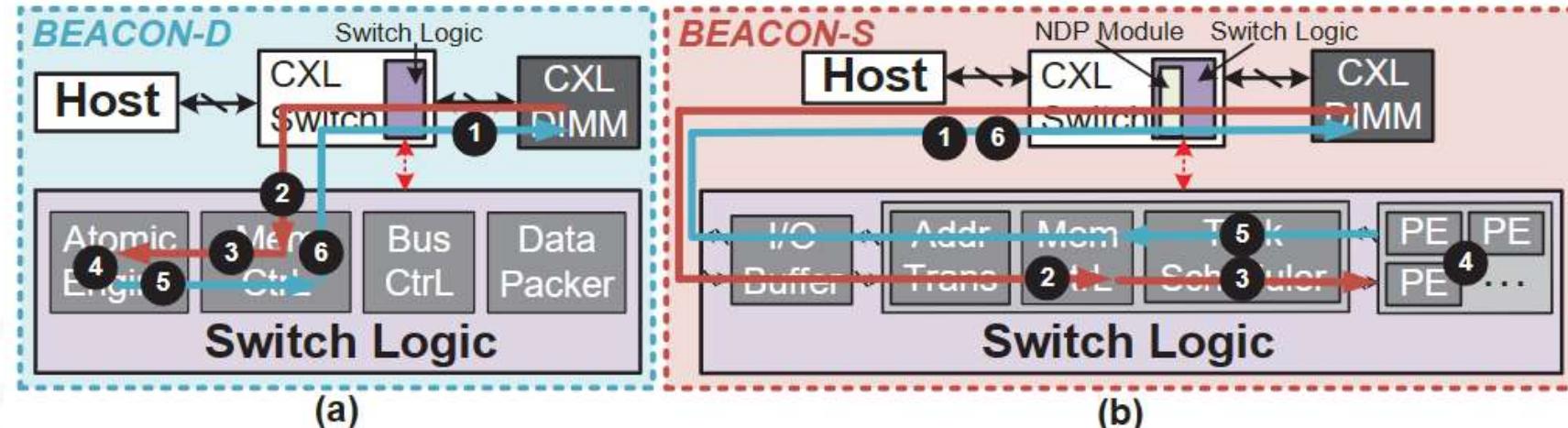


Figure 7. Workflow of performing atomic memory operations. (a) BEACON-D. (b) BEACON-S.

- ① MC向CXL-DIMM发送内存请求
- ② 数据带回到MC
- ③ 数据流向计算单元
- ④ 数据在单元内完成计算
- ⑤ 数据流回MC
- ⑥ MC写回内存池

星火计划

中国 电信 转型 专业 领军 人才 培养 项目

TSJU



目录

1

国外企业技术案例介绍

2

学术界研究分享

星火计划

中国电信转型专业领军人才培养项目



2. 学术界研究分享

硬件架构

大内存节点

- String finger, HPCA'19
- pDPM, ATC'20

可编程网卡

- Cilo, ASPLOS'22

智能交换机

- MIND, SOSP'21

系统设计

非透明远内存

- LITE, SOSP'17
- MemLiner, OSDI'22

透明远内存

- Infiniswap, NSDI'17

混合内存

- Hybrid^{^2}, HPCA'20

应用加速

图计算加速

- Fargraph, IPDPS'22

DL应用加速

- COARSE, HPCA'22

Severless管理

- Jiffy, EuroSys'22

资源管理

资源压力感知

- Canvas, NSDI'23

敏感性分析

- HyFarM, ICCD'22

功耗管理

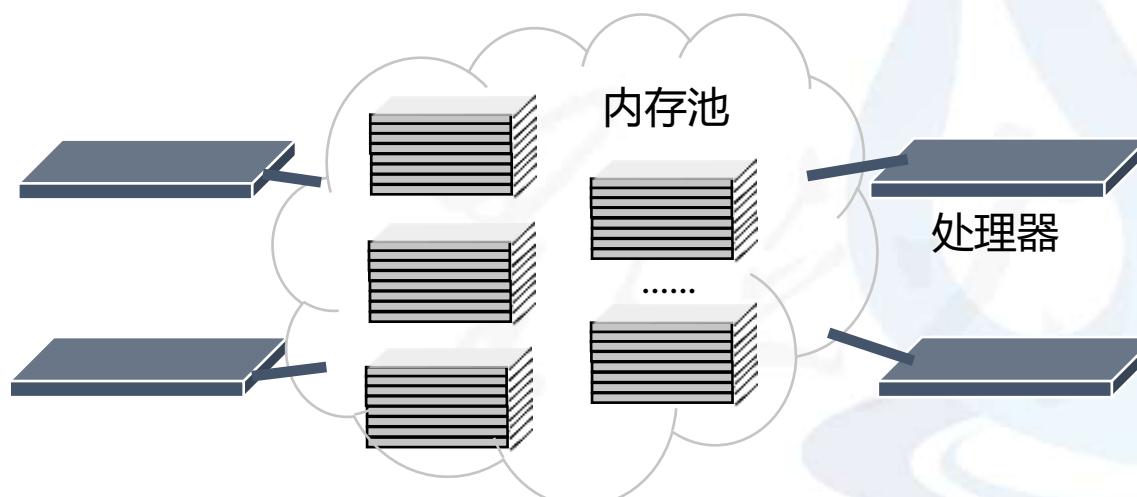
- Zombieland, EuroSys'18

全网计算

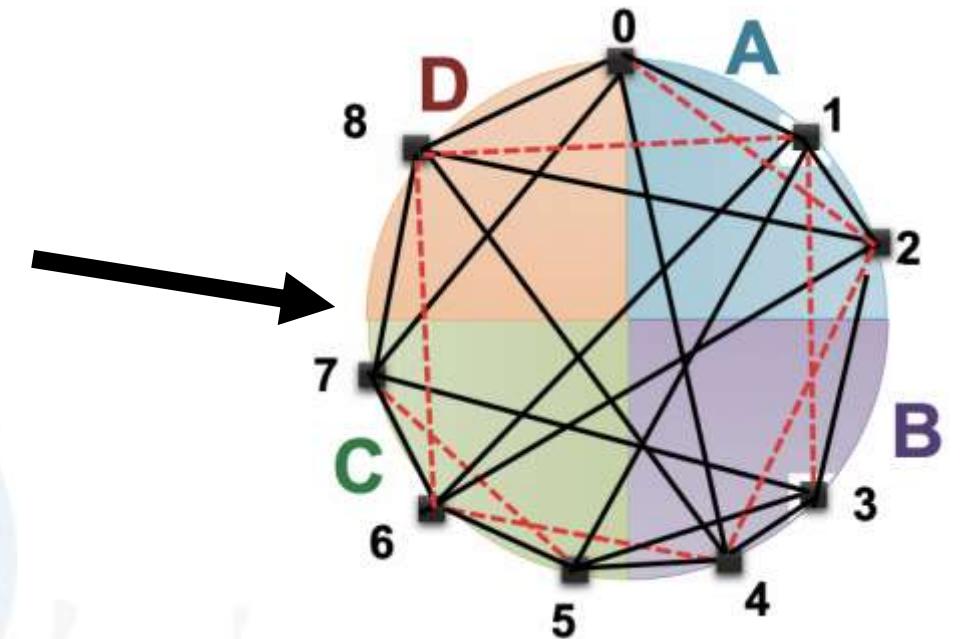
2. 学术界研究分享之硬件架构：大内存节点

String Figure:

- 实现了弹性可扩展的内存网络（中心化内存池）
- 提出了随机拓扑算法，以维持任意数量的存储节点的高吞吐量互连
- 提出混合路由协议，保证了路由开销的亚线性增加
- 实验表明，String Figure可以支持至多1296个内存节点的高吞吐互连



[1] String Figure: A Scalable and Elastic Memory Network Architecture, HPCA'19



9个四端口路由的内存节点互连示意图

星火计划



2. 学术界研究分享之硬件架构：大内存节点

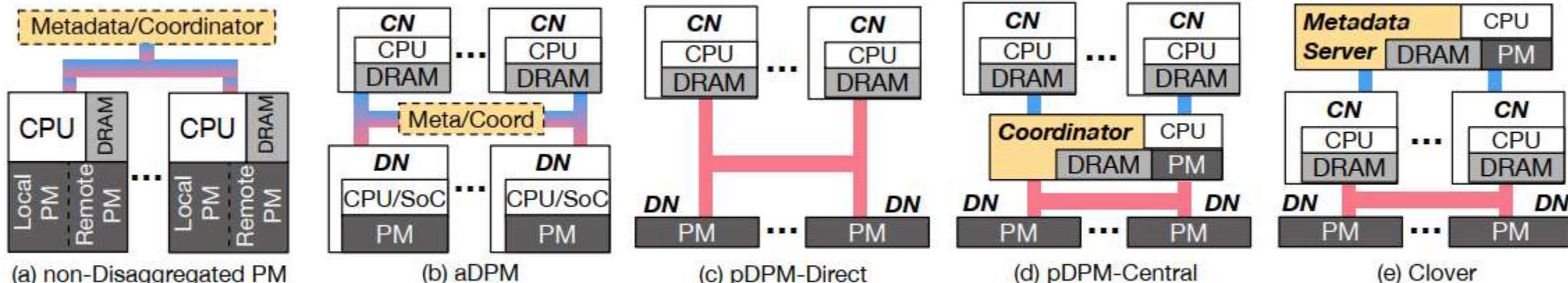


Figure 1: **PM Organization Comparison.** Blue bars indicate two-way communication and pink ones indicate one-way communication. Bars with both blue and pink mean support for both. Dashed boxes mean some but not all existing solutions adopt centralized metadata server (or a coordinator).

- 从计算服务器远程管理持久式内存构成的分离式内存节点
- 允许所有计算节点直接访问和管理存储节点，并使用中央协调器来协调计算节点和存储节点之间的通信。同时分离了数据平面和元数据/控制平面的位置、通信机制和管理策略。
- pDPM显著降低了货币和能源成本，并避免了存储服务器上的可伸缩性瓶颈

[2] Disaggregating Persistent Memory and Controlling Them Remotely: An Exploration of Passive Disaggregated Key-Value Stores, ATC'2020

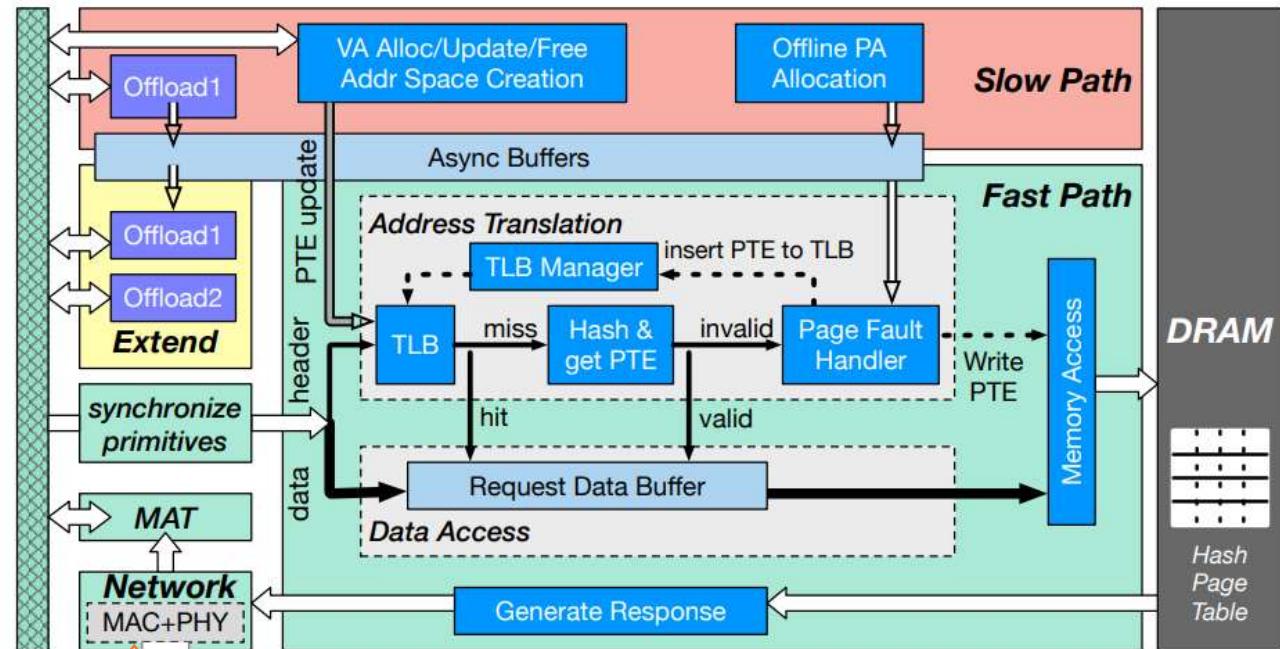
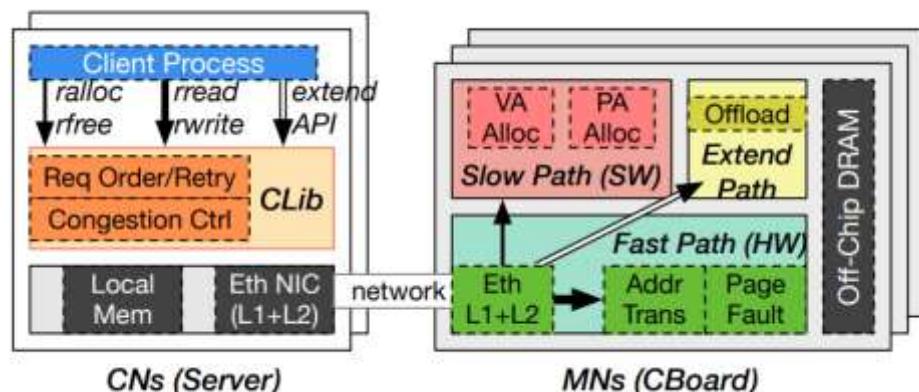
星火计划

2. 学术界研究分享之硬件架构：可编程网卡

Clio:

核心思想：提出一种软硬件协同的分离式内存方案，在内存节点和单纯的内存节点之间取得均衡表现

实现方式：构建了操作系统功能结构，硬件架构，网络系统，计算节点和内存节点



其中NM单元由Cboard实现其功能，Cboard由FPGA实现原型的快速路径，ARM处理元数据和控制的慢路径，FPGA承载内存卸载的计算扩展路径组成

[3] Clio: a hardware-software co-designed disaggregated memory system, ASPLOS'2022

星火计划



2. 学术界研究分享之硬件架构：智能交换机

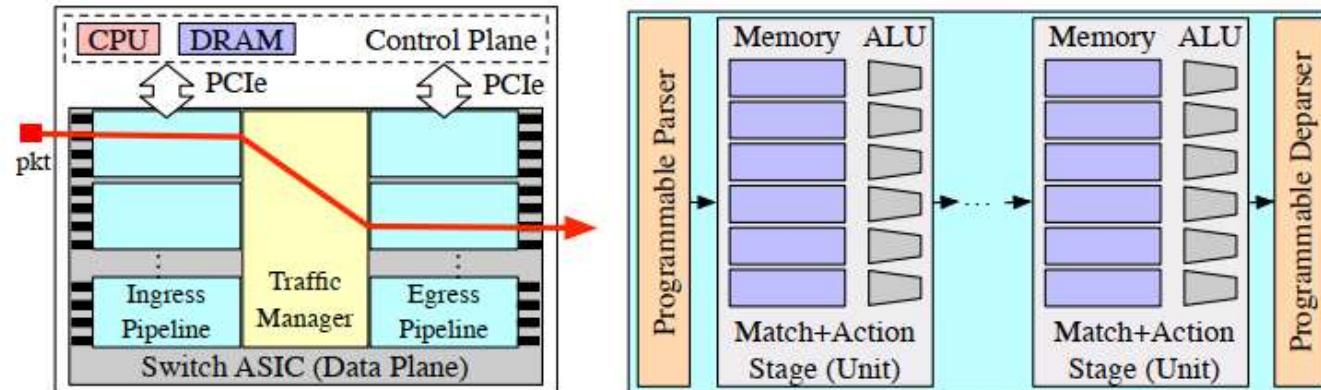
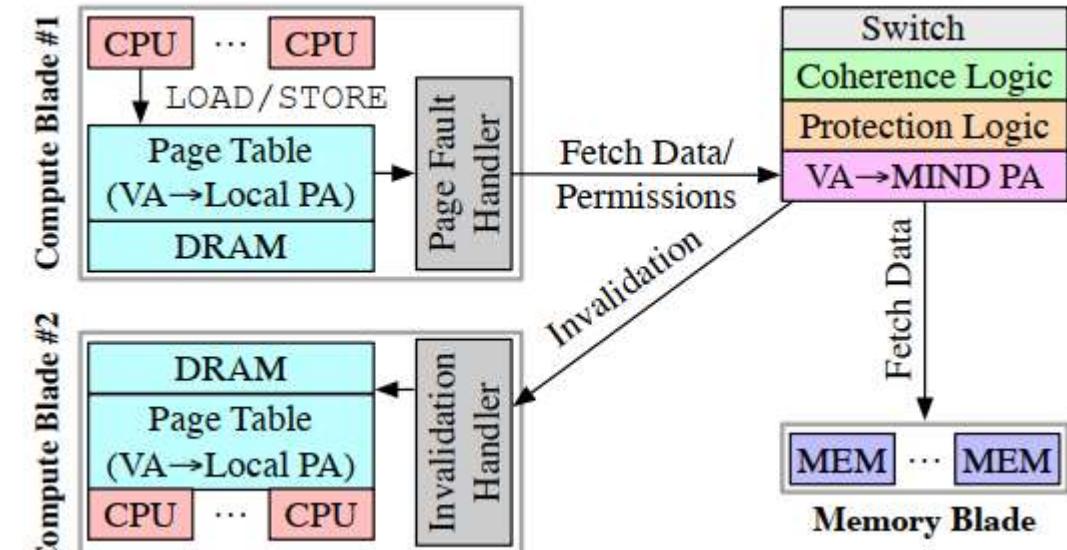


Fig. 1. Enabling technologies for MIND. (left) Programmable switch architecture and (right) Switch ingress/egress pipeline.



- MIND采用所有进程共享的全局虚拟地址空间，使用物理内存分配机制，可以跨内存刀片负载均衡
- 将内存权限的存储与地址转换项分离，实现了细粒度和灵活的保护，同时减少了ASIC切换的开销。
- 利用交换机ASIC中的多播等网络中心硬件原语来有效地实现其一致性协议

[4] MIND: In-Network Memory Management for Disaggregated Data Centers, SOSP'2021

星火计划



2. 学术界研究分享

硬件架构

大内存节点

- String finger, HPCA'19
- pDPM, ATC'20

可编程网卡

- Cilo, ASPLOS'22

智能交换机

- MIND, SOSP'21

系统设计

非透明远内存

- LITE, SOSP'17
- MemLiner, OSDI'22

透明远内存

- Infiniswap, NSDI'17

混合内存

- Hybrid^{^2}, HPCA'20

应用加速

图计算加速

- Fargraph, IPDPS'22

DL应用加速

- COARSE, HPCA'22

Severless管理

- Jiffy, EuroSys'22

资源管理

资源压力感知

- Canvas, NSDI'23

敏感性分析

- HyFarM, ICCD'22

功耗管理

- Zombieland, EuroSys'18

全网计算



2. 学术界研究分享之系统设计：非透明远内存

- 指出在数据中心环境中使用本机RDMA的三个主要问题：
 - RDMA不易使用，它的抽象和数据中心应用程序的需求之间的不匹配。
 - SRAM有限的情况下，RDMA性能基本上很难在MRs, MRs的总大小, qp的总数进行扩展
 - RDMA不提供任何机制来安全地共享资源，如qp、cq、内存缓冲区等
- LITE提出了自己的一套API并解决了，在提升易用性的同时也提升了性能。

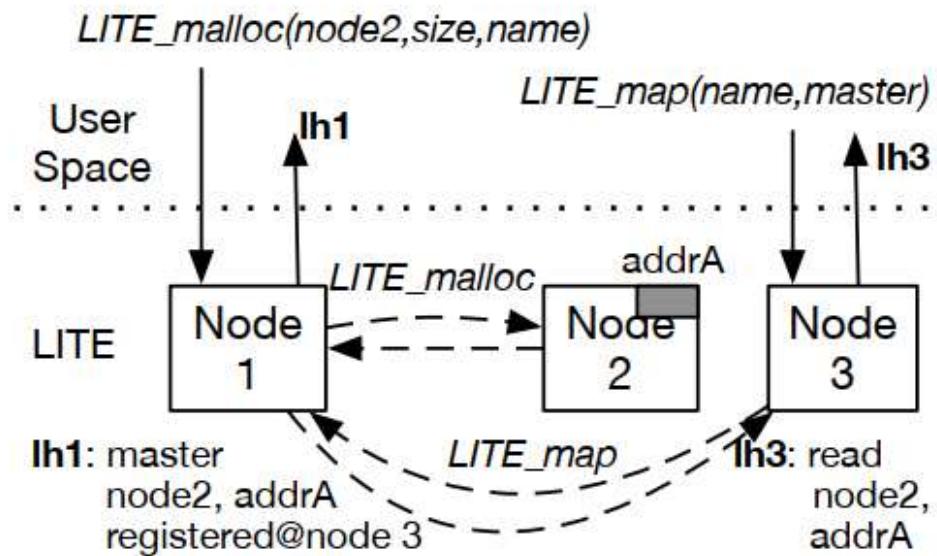


Figure 3: LITE *lh* Example. *Node1 allocates an LMR from node2 and is the master of it. Node3 maps the LMR with read permission by contacting Node1.*

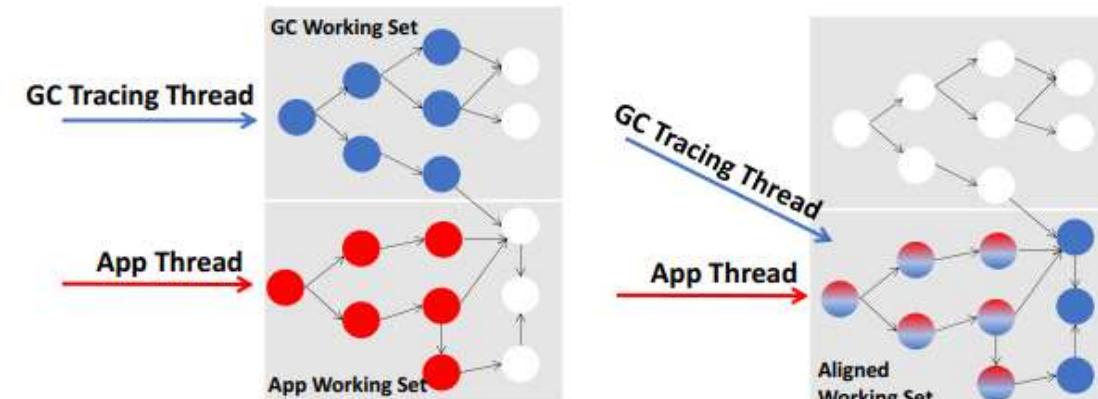
[5] LITE Kernel RDMA Support for Datacenter Applications, SOSP'2017

星火计划

2. 学术界研究分享之系统设计：非透明远内存

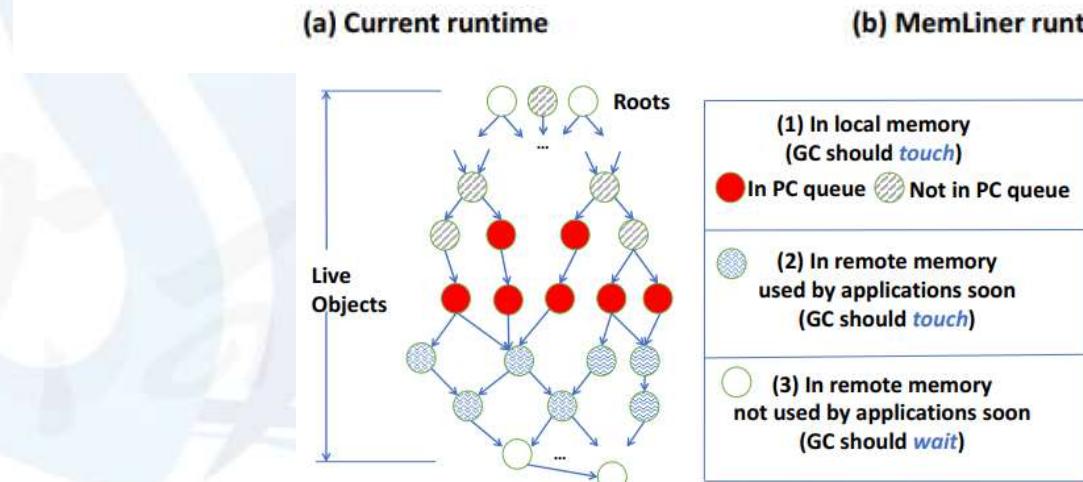
核心思想：

- 设计了基于Java虚拟机的远内存访问运行时优化
- 在Java中垃圾回收GC机制中实现，保留了一定的透明性



实现方式：

- 将来自应用程序和GC的内存访问进行“排列”和分类，使它们遵循相似的内存访问路径，从而减少本地工作及并提升预取性能



[6] MemLiner: Lining up Tracing and Application for a Far-Memory-Friendly Runtime, OSDI'2022

星火计划

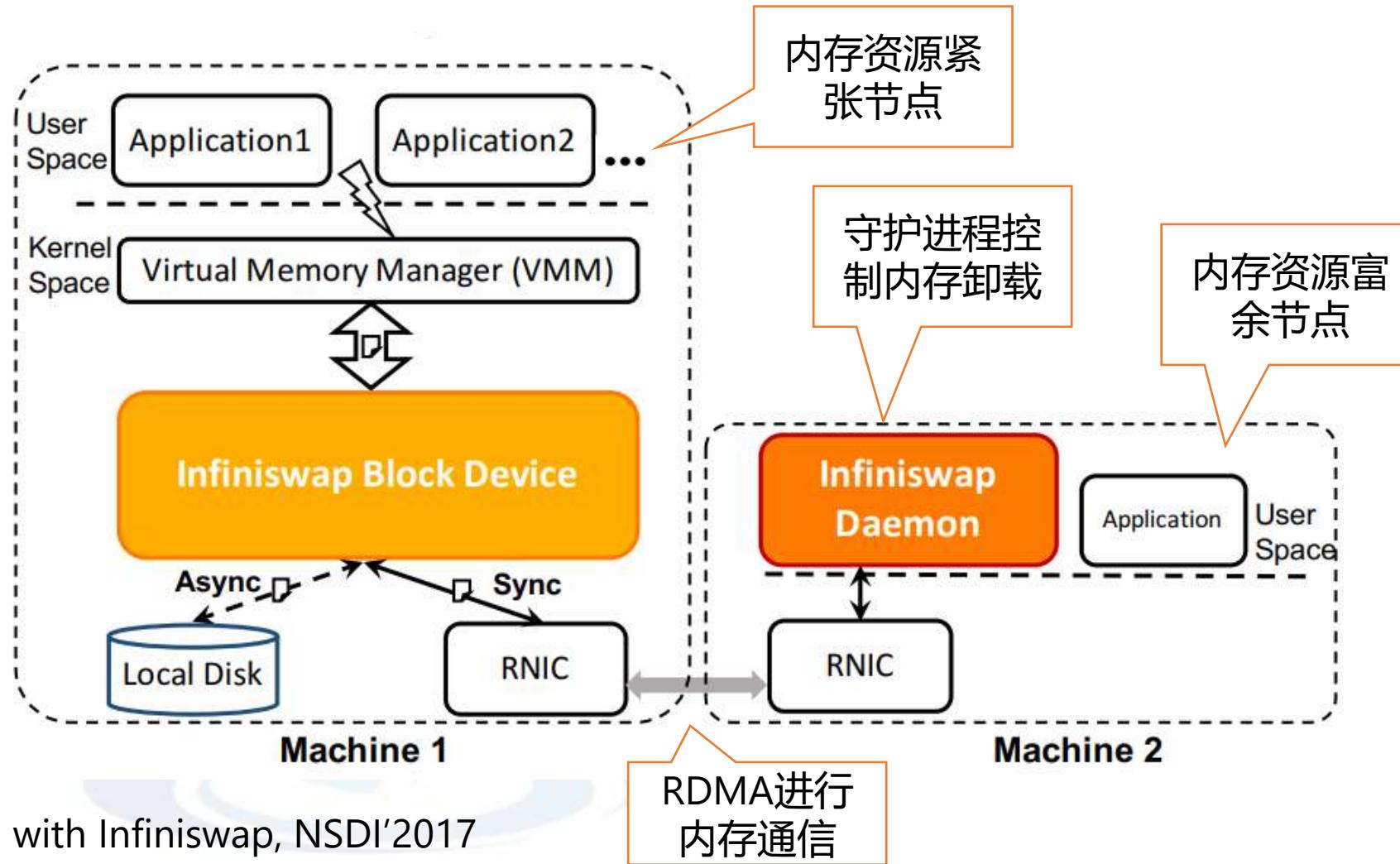


2. 学术界研究分享之系统设计：透明远内存

Infiniswap:

核心思想：将服务器中的空闲内存资源提供给集群中的其它服务器

实现方式：block device实现地址空间的管理、去中心化的slab分布管理、处理I/O请求、处理内存回收、处理远程中断，INFINISWAP daemon进程，管理内存资源。



[7] Efficient Memory Disaggregation with Infiniswap, NSDI'2017

星火计划

中国 电信 转型 专业 领军 人才 培养 项目

交通大学



2. 学术界研究分享之系统设计：混合内存

Hybrid2:

核心思想：组织不同具有不同特性的内存，在性能和容量上取得平衡性能

实现方式：使用速度快，容量小，成本高的NM和速度较慢，容量大，成本低的FM组成内存结构，NM的一部分用于缓存，剩余的和FM构成同一地址空间，不损失性能的情况下拥有较大容量

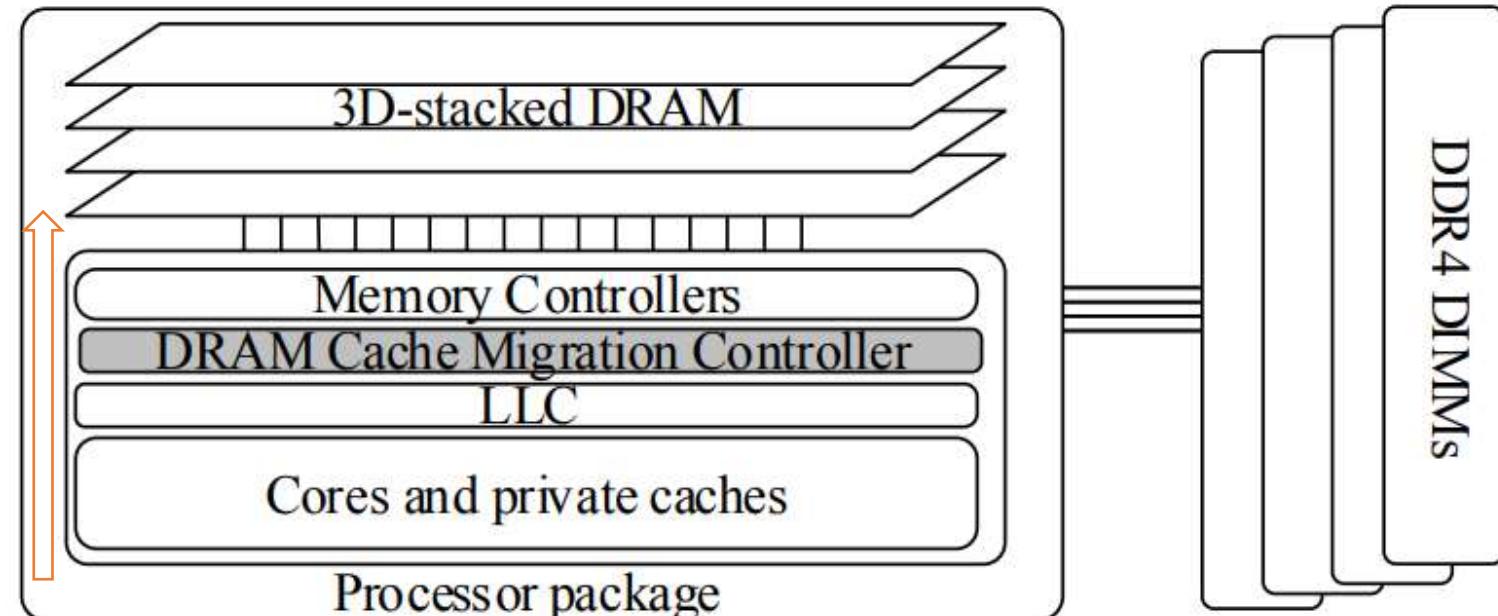


Figure 3: System Overview.

[8] Hybrid2: Combining Caching and Migration in Hybrid Memory Systems, HPCA'2020

星火计划

中国 电信 转型 专业 领军 人才 培养 项目

TSJU



2. 学术界研究分享

硬件架构

大内存节点

- String finger, HPCA'19
- pDPM, ATC'20

可编程网卡

- Cilo, ASPLOS'22

智能交换机

- MIND, SOSP'21

系统设计

非透明远内存

- LITE, SOSP'17
- MemLiner, OSDI'22

透明远内存

- Infiniswap, NSDI'17

混合内存

- Hybrid^{^2}, HPCA'20

应用加速

图计算加速

- Fargraph, IPDPS'22

DL应用加速

- COARSE, HPCA'22

Severless管理

- Jiffy, EuroSys'22

资源管理

资源压力感知

- Canvas, NSDI'23

敏感性分析

- HyFarM, ICCD'22

功耗管理

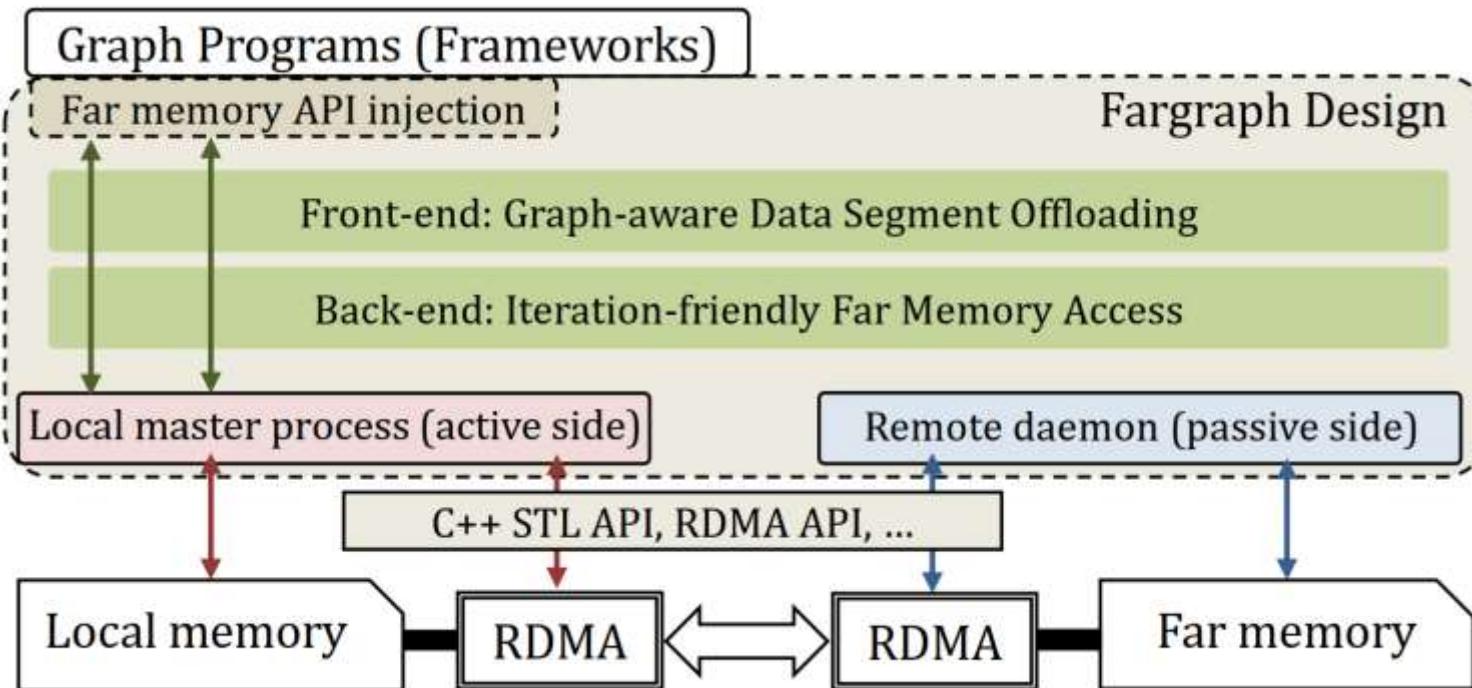
- Zombieland, EuroSys'18



2. 学术界研究分享之应用加速：图计算加速

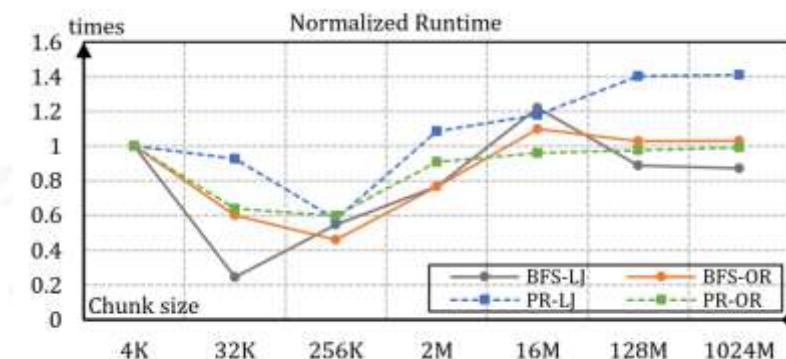
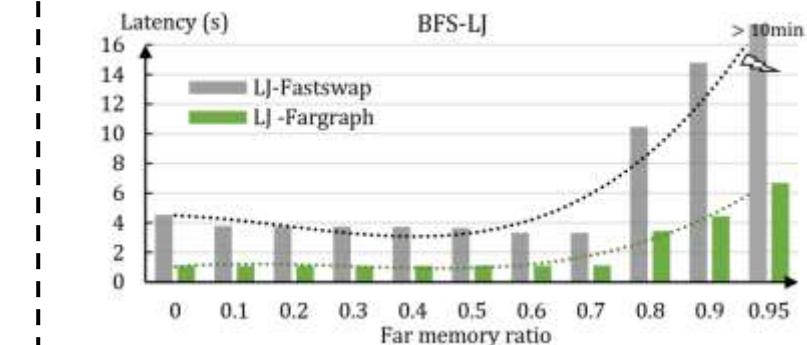
主要贡献：

- 第一个针对图计算任务的远内存系统
- 提出图感知的memory offloading策略
- 利用计算与传输的overlap提高远内存访问性能



[9] Excavating the Potential of Graph Workload on RDMA-based Far Memory Architecture, IPDPS'2022

取得的性能提升



星火计划



2. 学术界研究分享之应用加速：DL应用加速

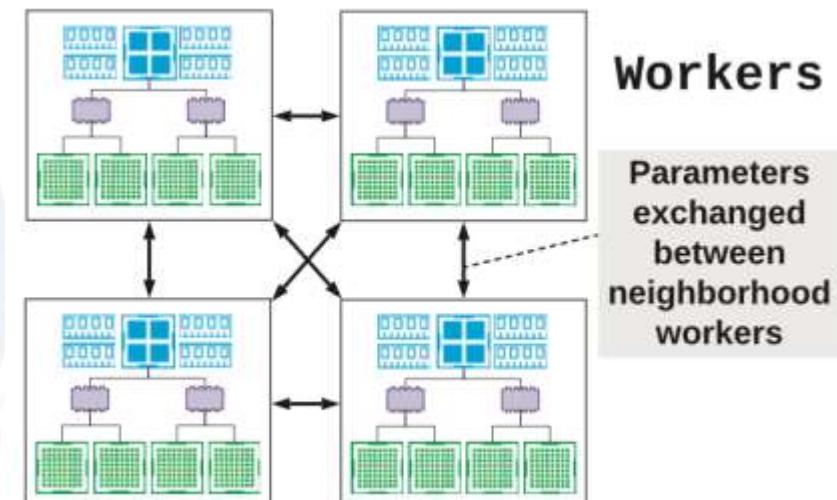
问题背景：

现有分布式DL训练受到 (1)单机并行度和片上内存容量 (2)跨节点通信开销 的限制

跨设备通信协议（如PCIe协议）带宽比设备内存带宽低一个量级，过度增加并行度会增加通信瓶颈

本文贡献：

- 使用缓存一致互连CCl协议，通过分离式内存系统加速DL训练
- 提出分散式参数通信方案，将参数同步和本地化参数存储分离
- 提出张量路由和分区方案，以充分利用串行总线带宽



分布式DL训练通信示意图

[10] Enabling Efficient Large-Scale Deep Learning Training with Cache Coherent Disaggregated Memory Systems, HPCA'2022

星火计划

2. 学术界研究分享之应用加速：Serverless管理

创新点：

小内存块的粒度上执行资源分配

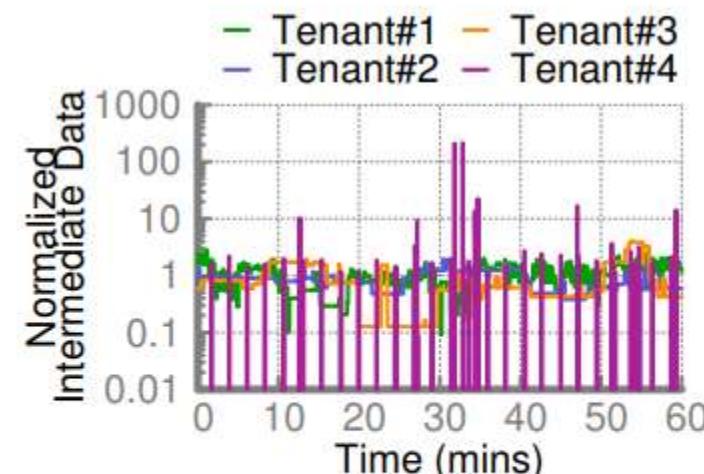
弹性资源分配

对中间数据进行显式生命周期管理

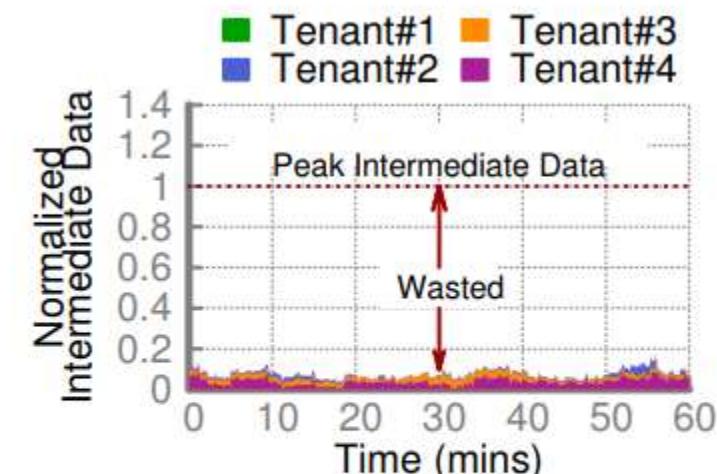
效果：

在并发运行的作业之间复用内存容量，减少了读取和写入较慢的持久存储的开销，任务执行时间提高了1.6–2.5倍。

无服务计算任务有大量数据卸载的潜力



(a) Intermediate data (normalized by mean usage)



(b) Cumulative intermediate data (normalized by peak usage)

[11] Jiffy: Elastic Far-Memory for Stateful Serverless Analytics, EuroSys'2022

星火计划



2. 学术界研究分享

硬件架构

大内存节点

- String finger, HPCA'19
- pDPM, ATC'20

可编程网卡

- Cilo, ASPLOS'22

智能交换机

- MIND, SOSP'21

系统设计

非透明远内存

- LITE, SOSP'17
- MemLiner, OSDI'22

透明远内存

- Infiniswap, NSDI'17

混合内存

- Hybrid^{^2}, HPCA'20

应用加速

图计算加速

- Fargraph, IPDPS'22

DL应用加速

- COARSE, HPCA'22

Severless管理

- Jiffy, EuroSys'22

资源管理

资源压力感知

- Canvas, NSDI'23

敏感性分析

- HyFarM, ICCD'22

功耗管理

- Zombieland, EuroSys'18

全网计算

2. 学术界研究分享之资源管理：资源压力感知

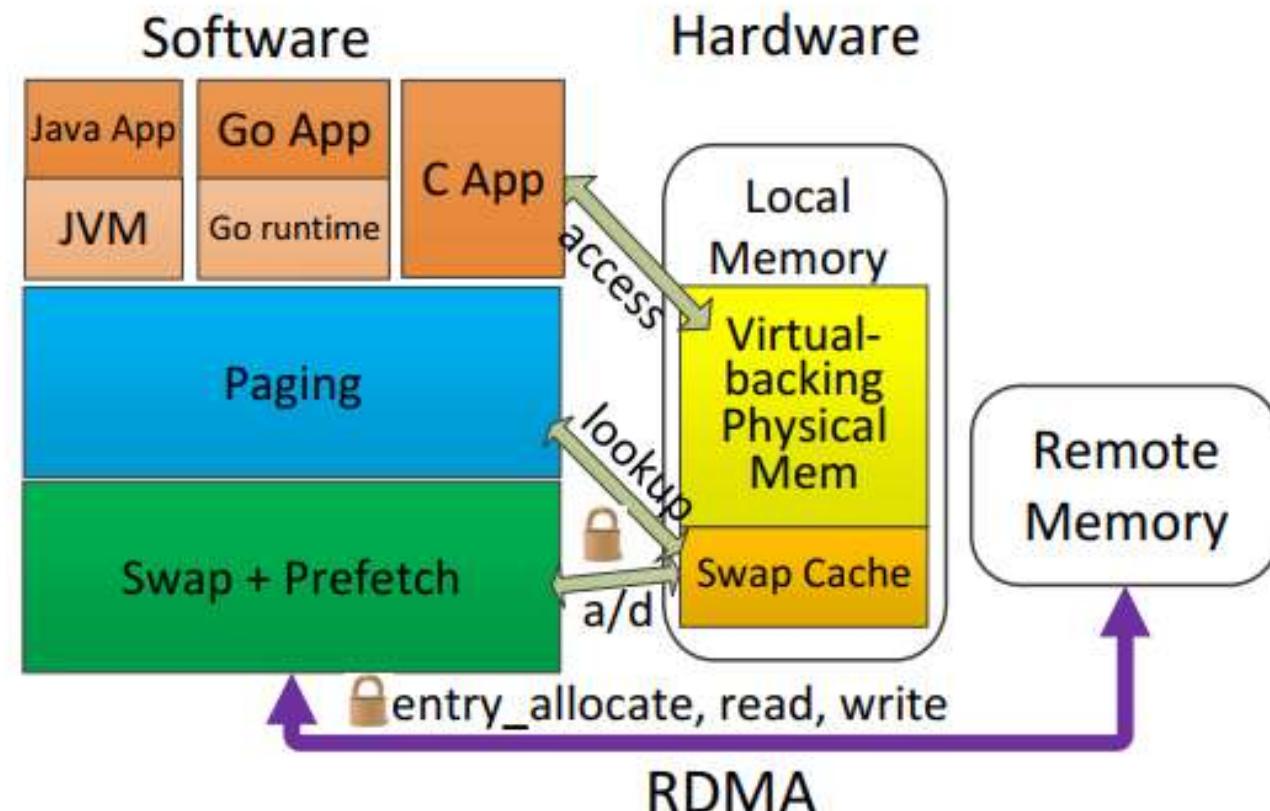
研究内容

创新点：

- 研究远内存swap区的争抢，隔离了远程内存应用程序的交换路径。
- 允许每个应用程序拥有其专用的交换分区、交换缓存，根据资源限制设计预取器和分配RDMA带宽

实验结果：

- 可以将最先进的固结技术的能源效率提高86%，而额外的复杂性最小。



[12] Canvas: Isolated and Adaptive Swapping for Multi-Applications on Remote Memory, NSDI'23

星火计划

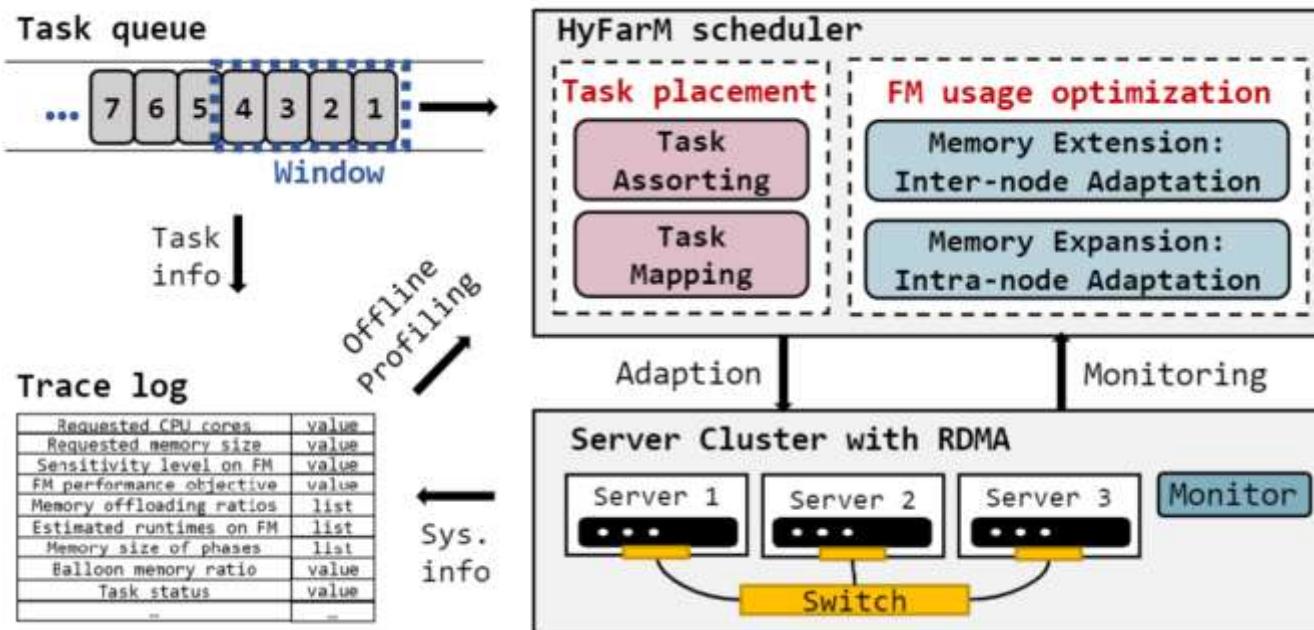
中国 电信 转型 专业 领军 人才 培养 项目

TSJU

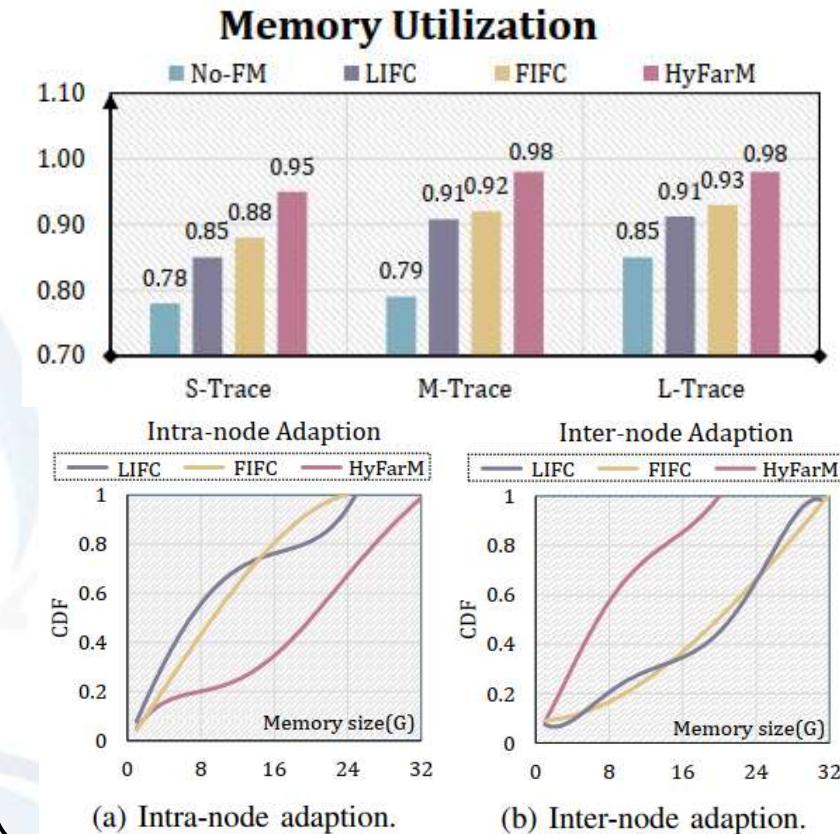
2. 学术界研究分享之资源管理：敏感型分析

主要贡献：

- 第一个基于混合远内存架构的任务调度框架
- 设计了不同远内存敏感性的应用混部部署策略
- 动态控制并调整横纵远内存的使用



取得的效率提升



[13] HyFarM: Task Orchestration on Hybrid Far Memory for High Performance Per Bit, ICCD'2022

星火计划



2. 学术界研究分享之资源管理：功耗管理

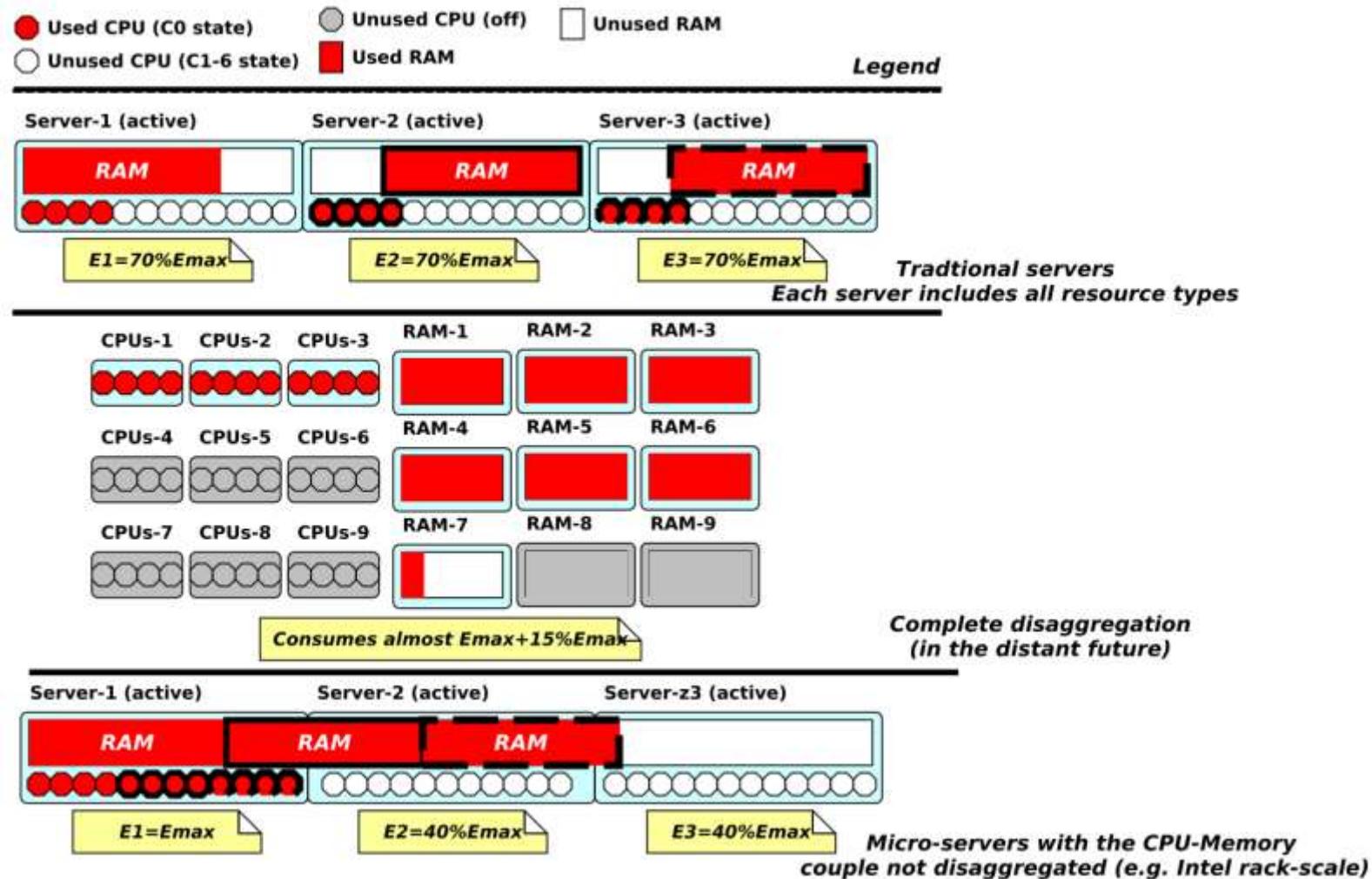
研究内容

创新点：

- 设计僵尸节点，允许暂停服务器从而节省能源)，同时使其内存远程访问
- 允许整个机架资源的透明利用(避免资源浪费)。

实验结果：

- 可以将最先进的固结技术的能源效率提高86%，而额外的复杂性最小。



[14] Welcome to zombieland: Practical and Energy-efficient memory disaggregation in a datacenter, Eurosyst'2018

星火计划



总结：分离式内存相关研究





三、分离式内存的发展趋势

讲者：李超

上海交通大学 计算机系 SAIL实验室

2023年2月

星火计划

中国电信转型专业领军人才培养项目



目录

1

**网络技术
对分离式内存的影响**

2

**软件技术
对分离式内存的影响**

3

**硬件技术
对分离式内存的影响**

4

**分离式内存
与超融合基础设施**

星火计划

中国 电信 转型 专业 领军 人才 培养 项目

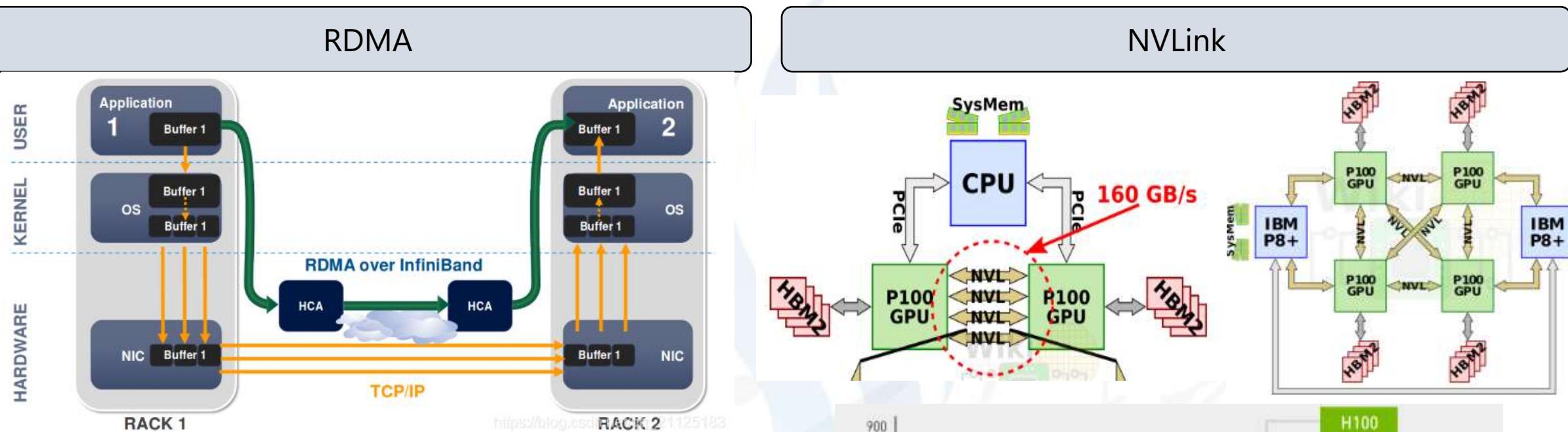


1. 网络技术：整体趋势

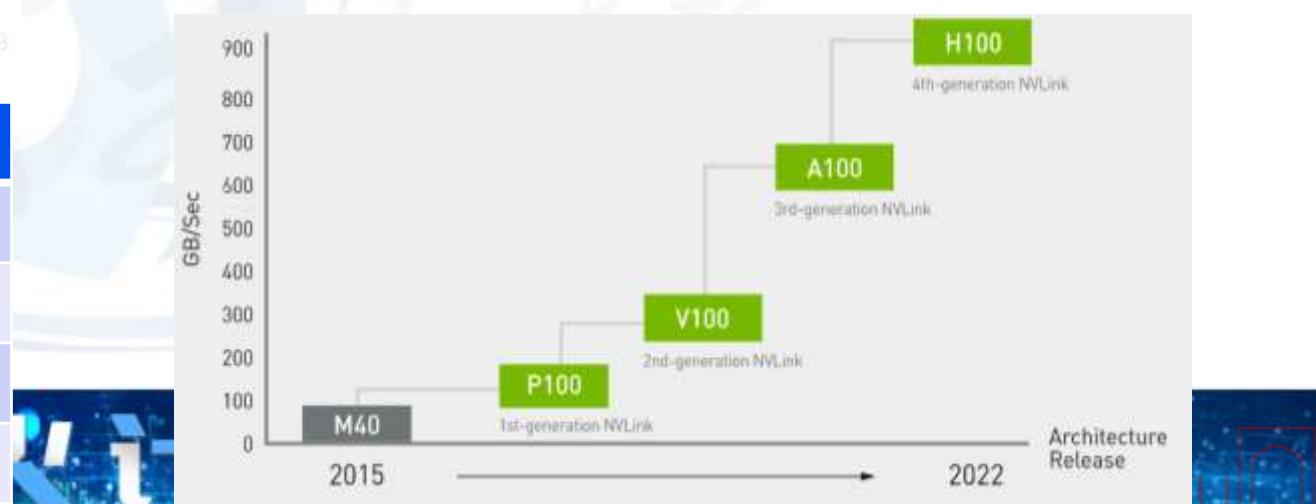


星火计划

1. 网络技术：先进网络技术及性能



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）

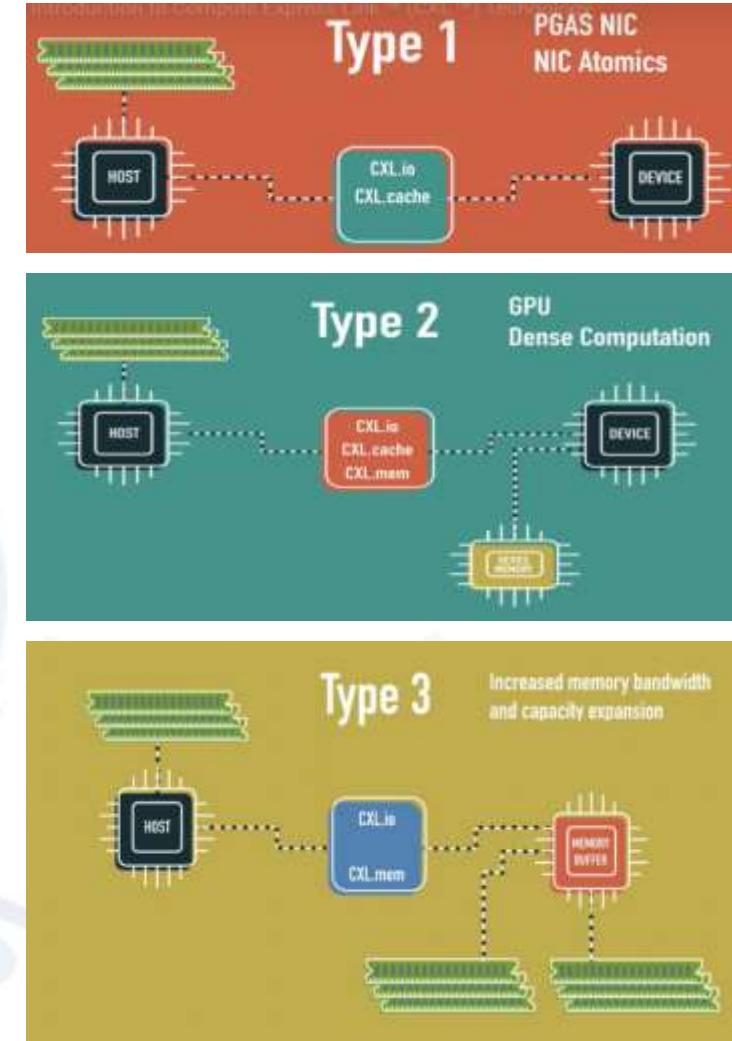
发展状况：CXL 3.0版本在 2022 年 8 月发布

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



星火计划

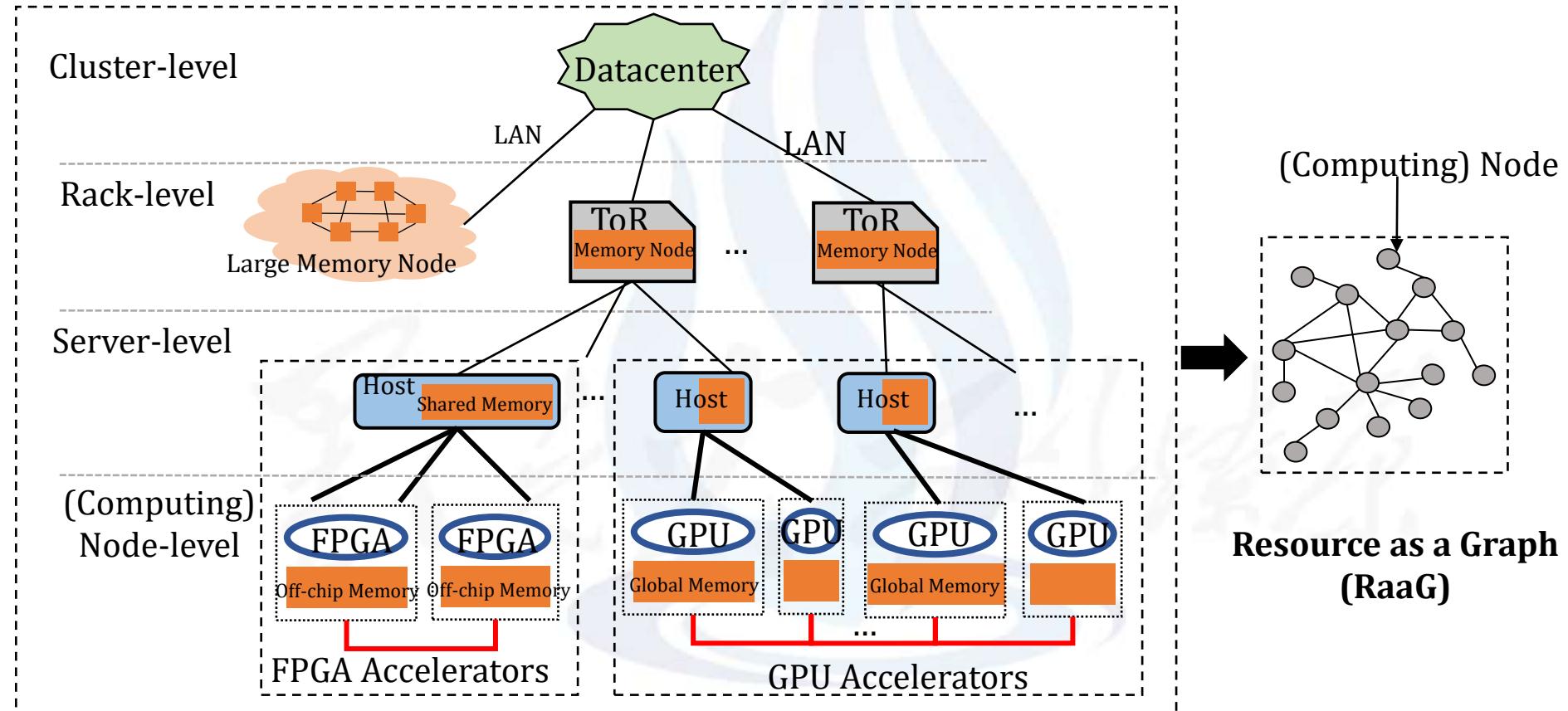


1. 网络技术：对内存互联结构的影响

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

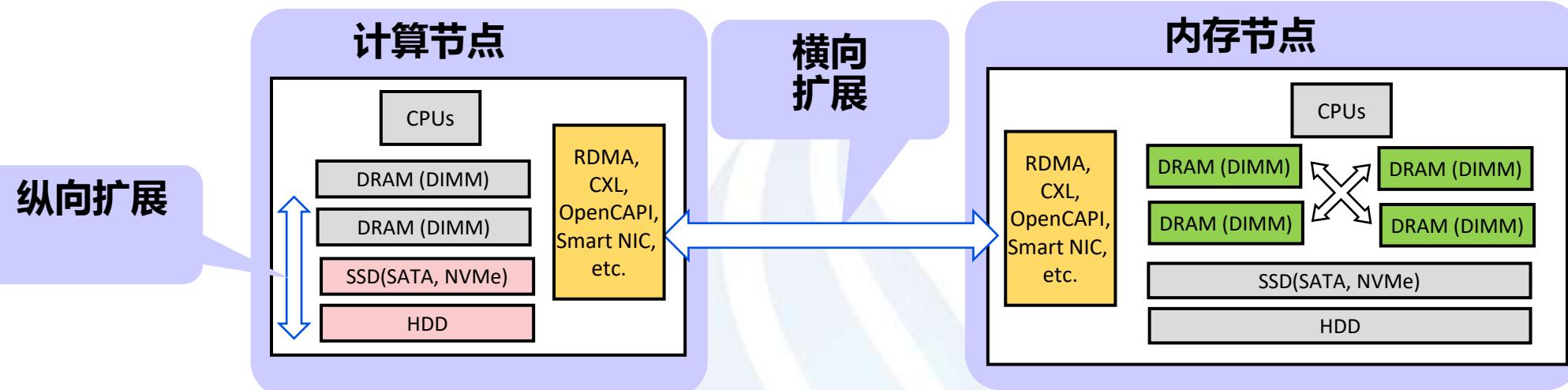
高速互联->多层级内存池构建



星火计划



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

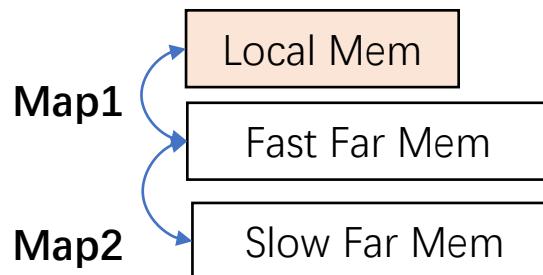
[2] Can far memory improve job throughput?, EuroSys'20

星火计划

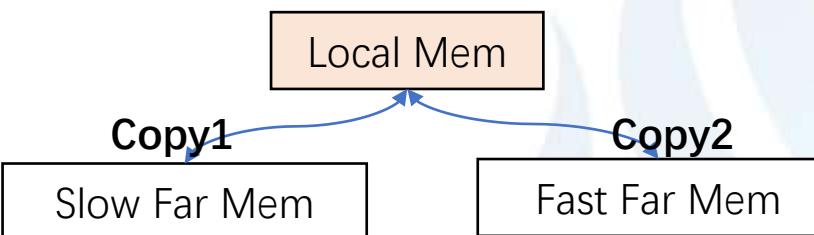


1. 网络技术：对内存层级的影响

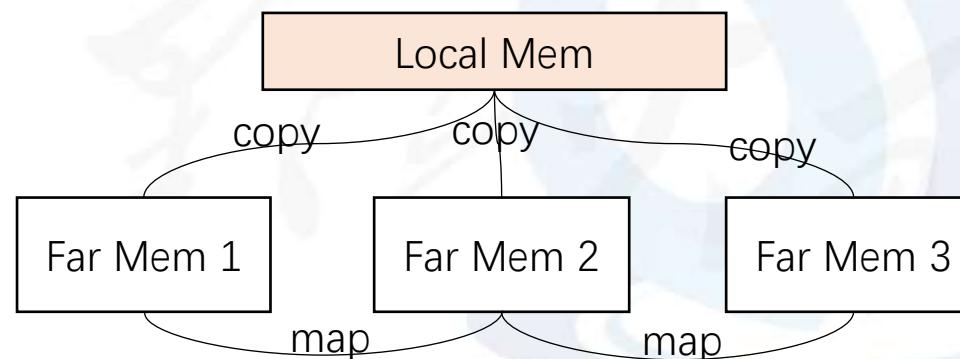
基于swap的传统层级映射式[1]



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:

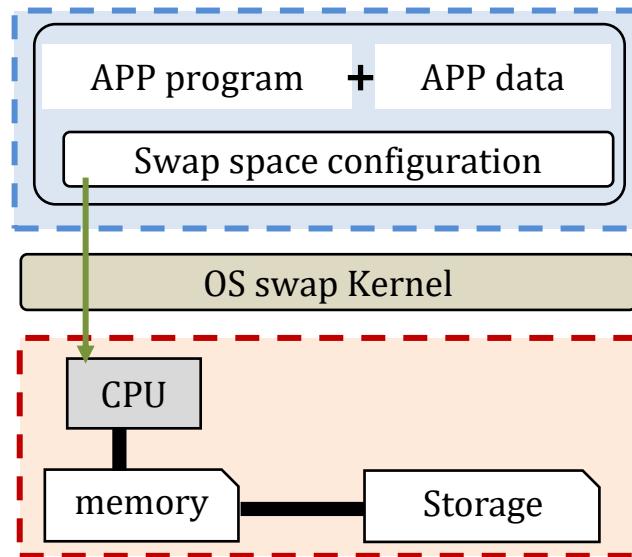
- 性能好
- 空间合理利用

[1] Transparent and Lightweight Object Placement for Managed Workloads atop Hybrid Memories, VEE 22

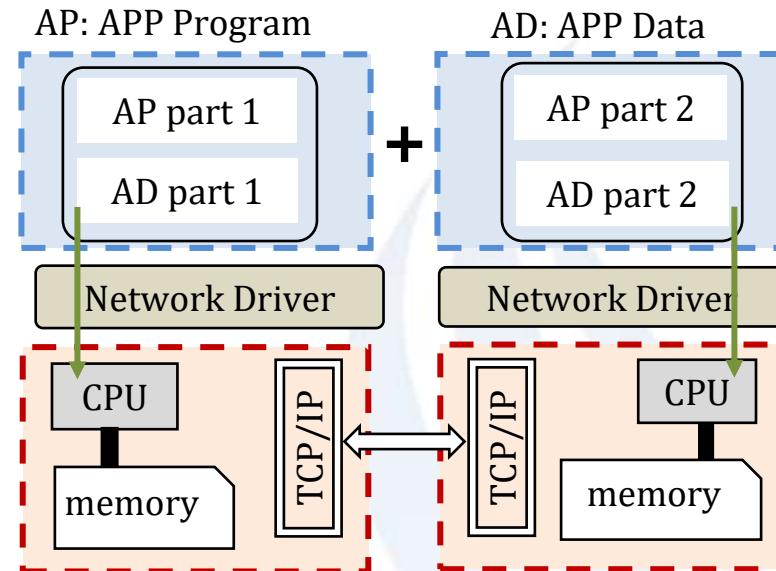
[2] Panthera: Holistic Memory Management for Big Data Processing over Hybrid Memories, PLDI 19



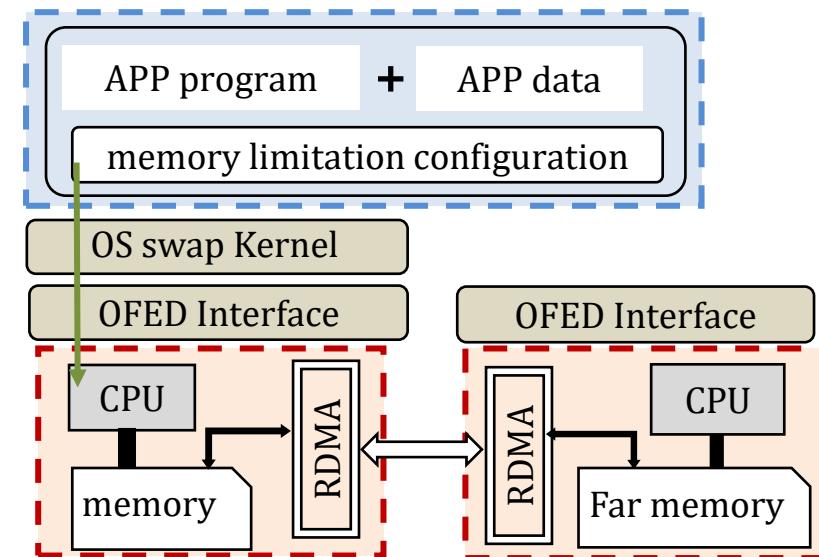
1. 网络技术：对应用执行模式的影响



单节点处理



分布式处理



远内存处理

- 使用本地的存储空间
- 额外数据访问依赖I/O
- 应用透明
- 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
- 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

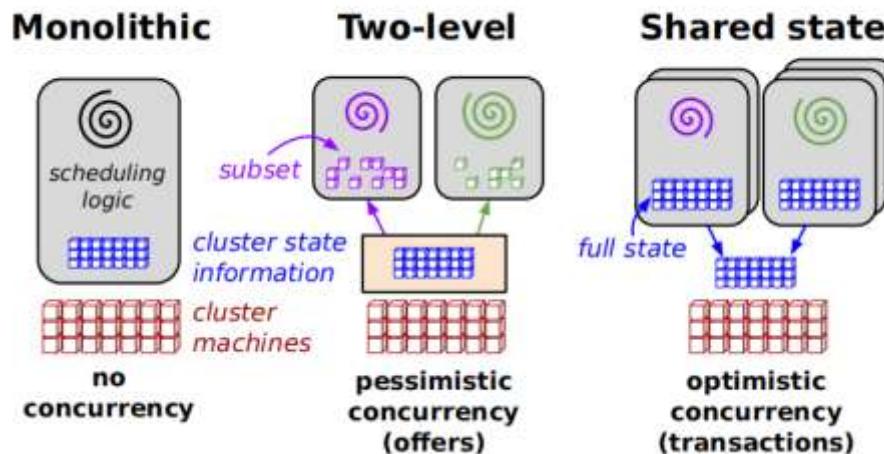
星火计划



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

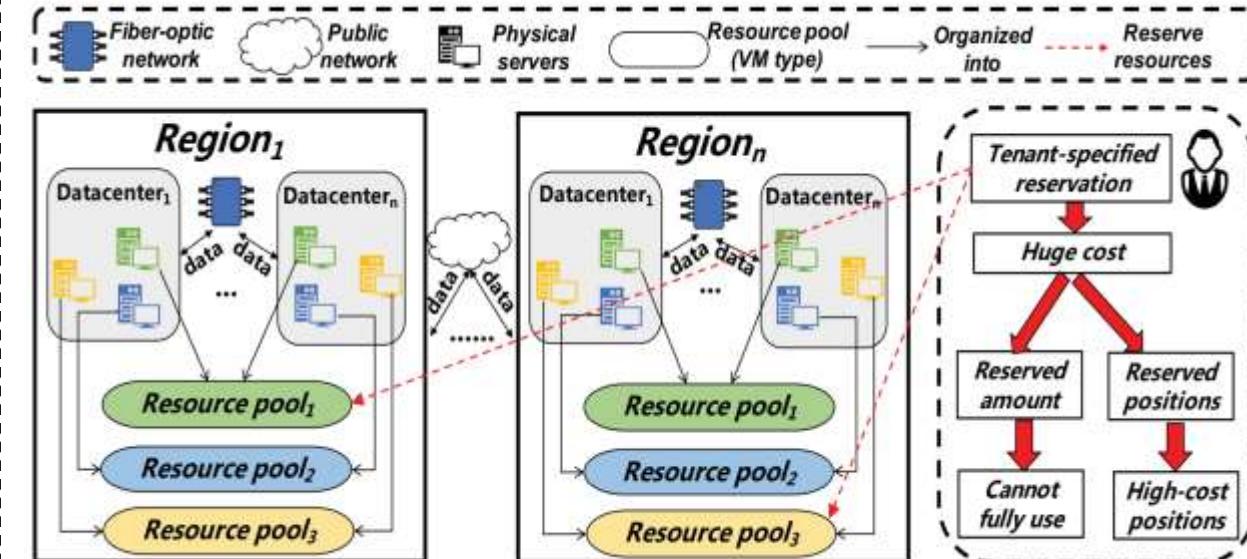
规模过大->分布式共享视图



- 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

星火计划

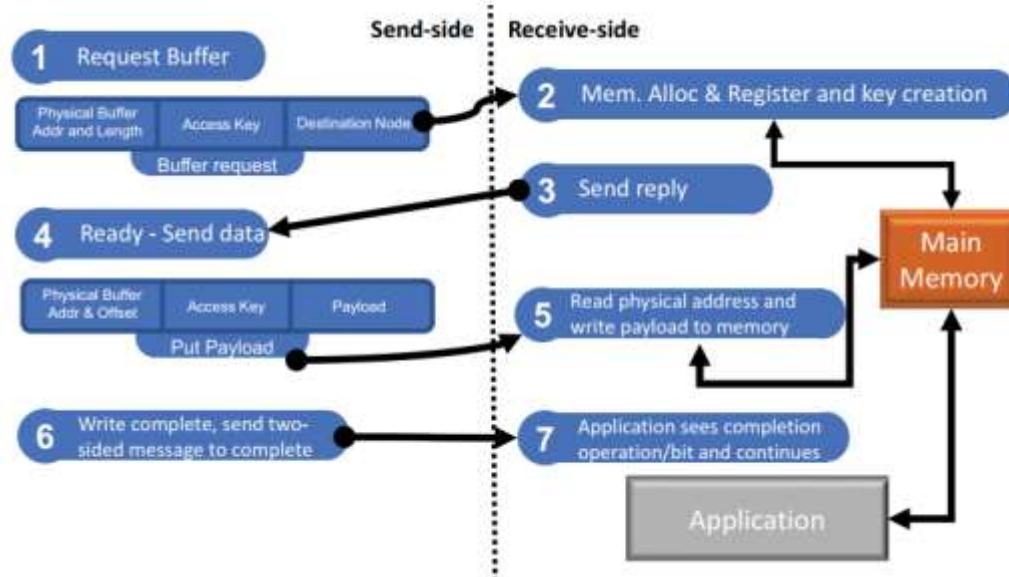


1. 网络技术：对内存访问框架设计的影响

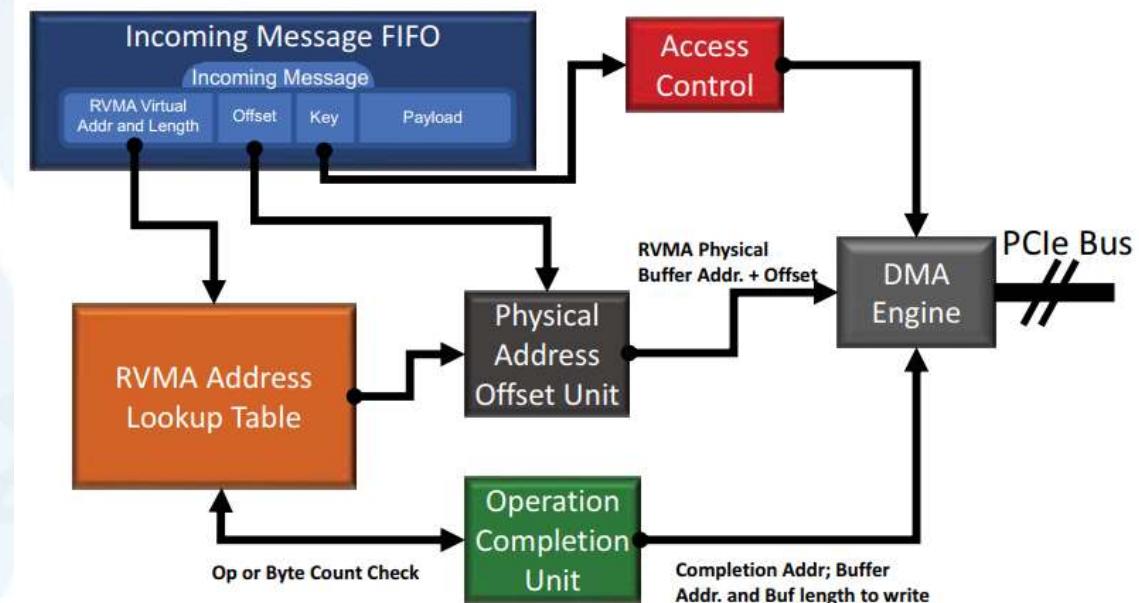
为了网络通道的安全性，人们设计了保证数据传输鲁棒性和安全性的远内存访问框架

- 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

星火计划



目录

1

**网络技术
对分离式内存的影响**

2

**软件技术
对分离式内存的影响**

3

**硬件技术
对分离式内存的影响**

4

**分离式内存
与超融合基础设施**

星火计划

中国 电信 转型 专业 领军 人才 培养 项目



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

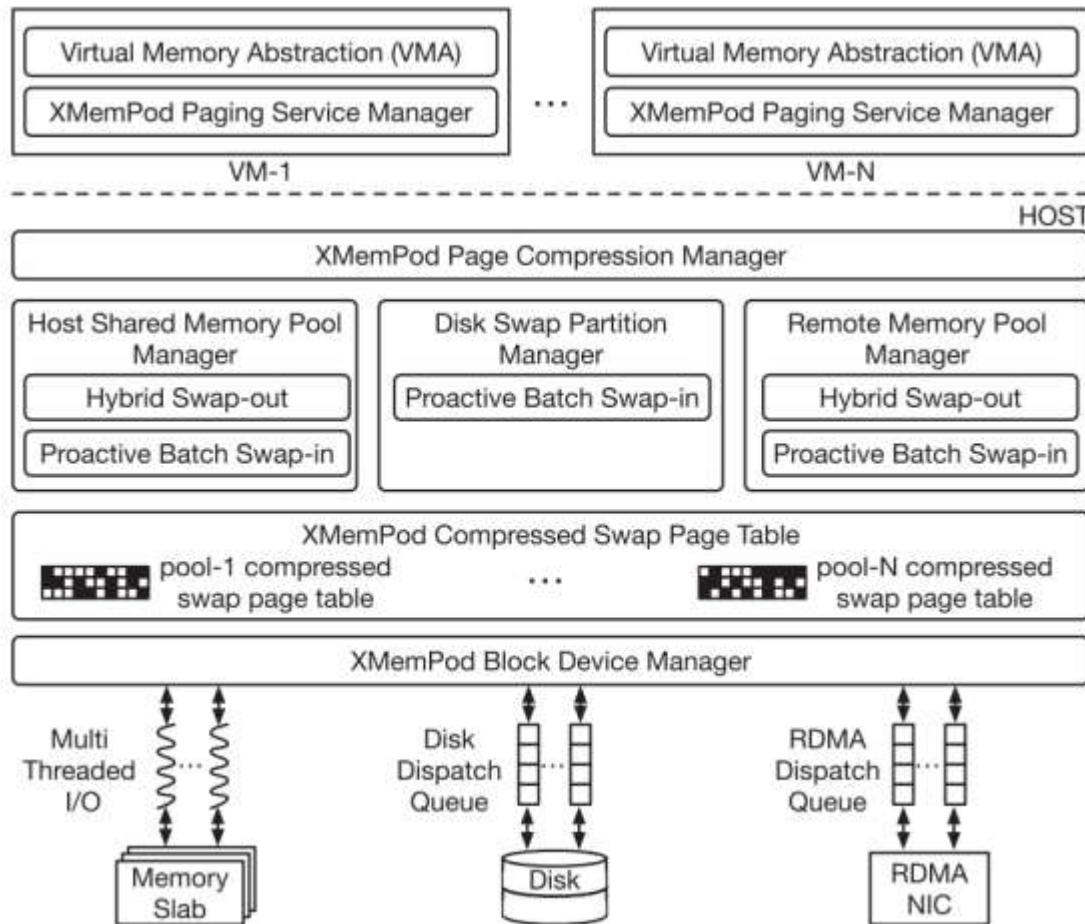
性能更高

资源分配合理

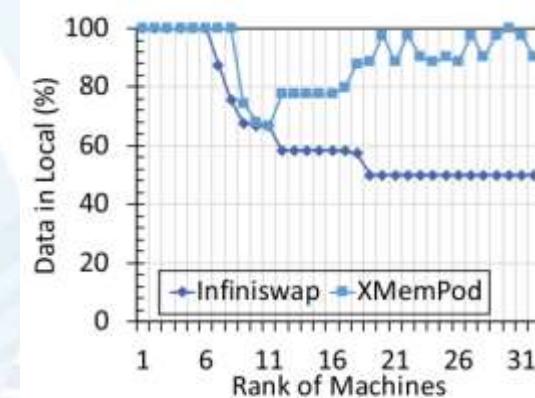
星火计划



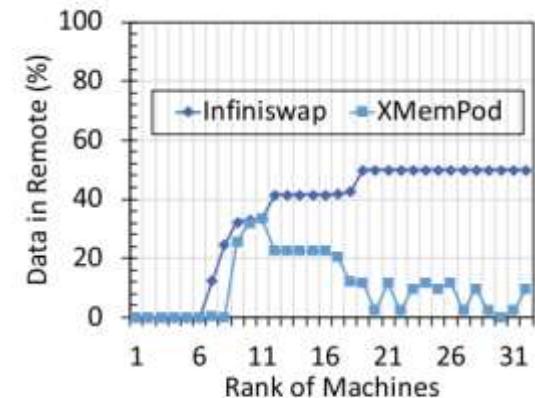
2. 软件技术：硬件虚拟化部署->高并行度



- 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 (%)



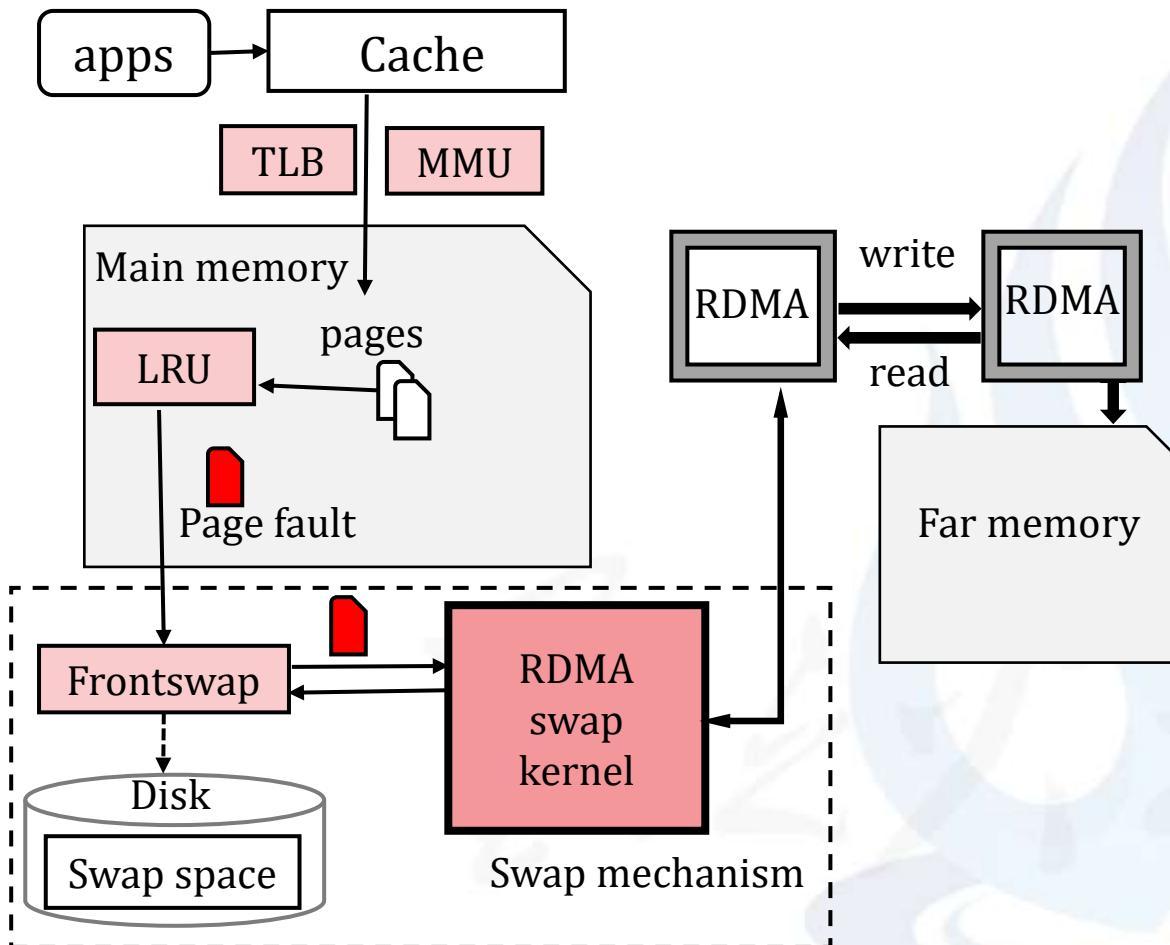
(b) Data Stored in Remote (%)

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

星火计划



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

星火计划



2. 软件技术：应用框架部署->编程简单

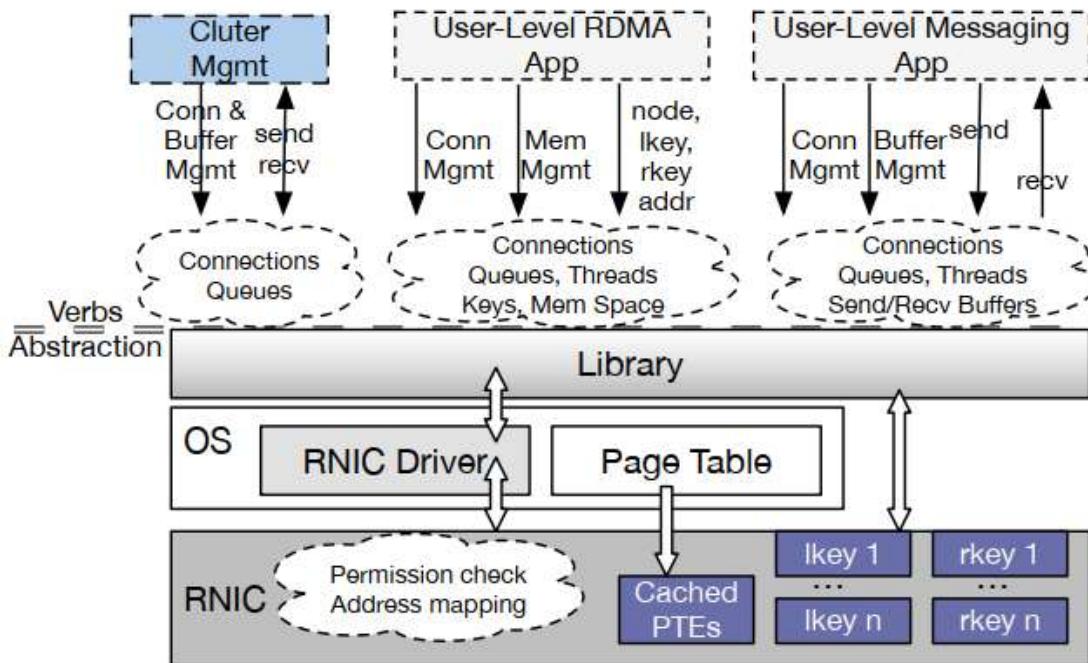


Figure 1: Traditional RDMA Stack.

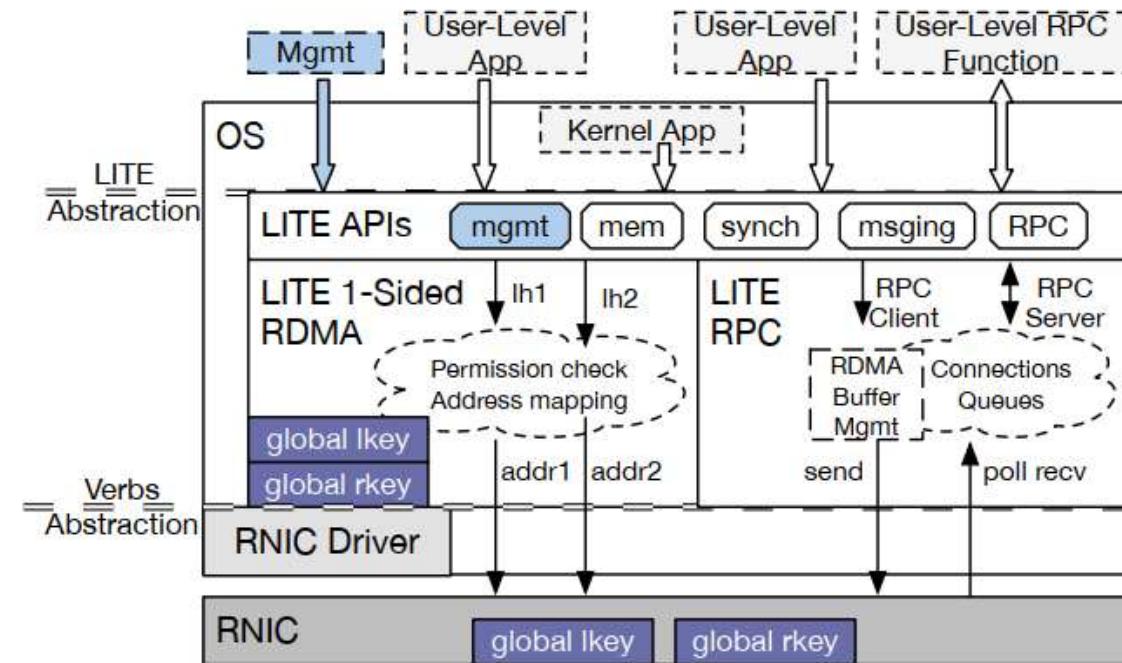
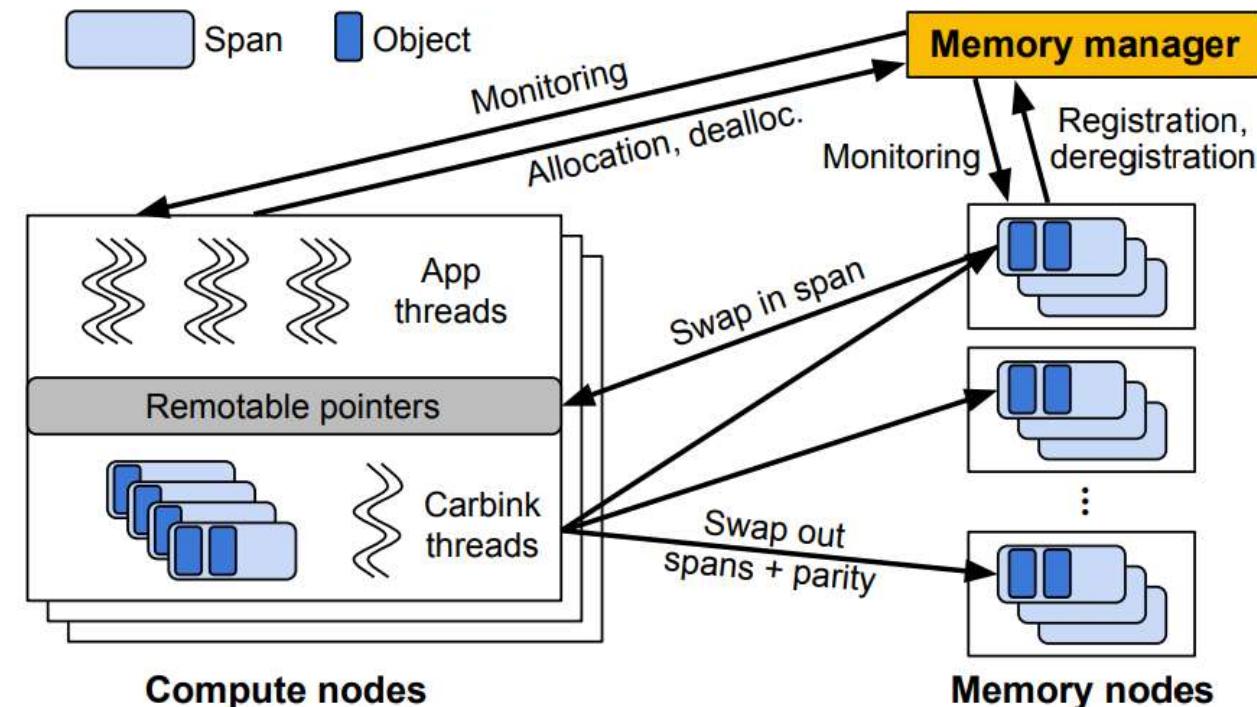
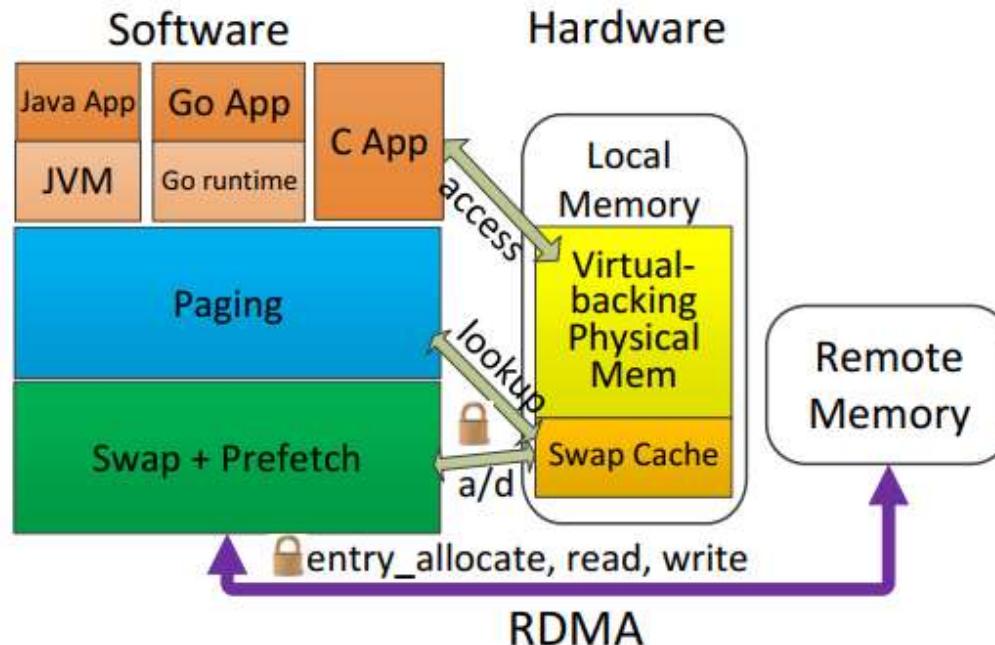


Figure 2: LITE Architecture.

- 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。

星火计划

2.软件技术：数据安全->隔离与容错



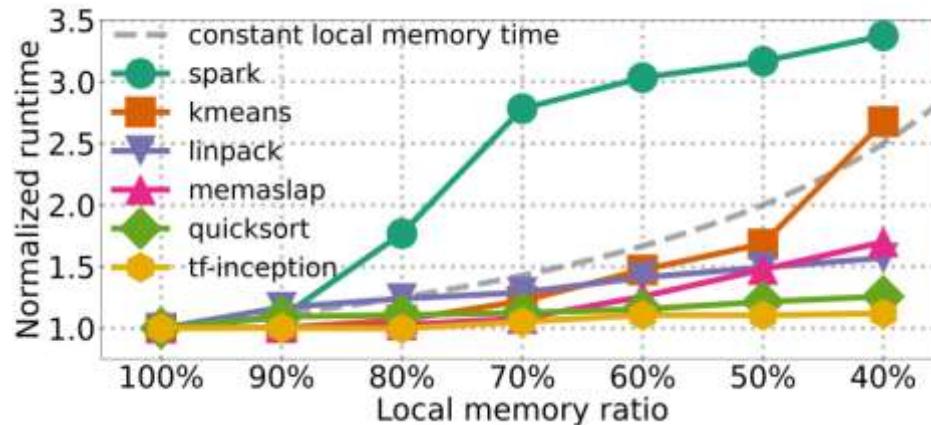
数据隔离

容错处理-提供冗余

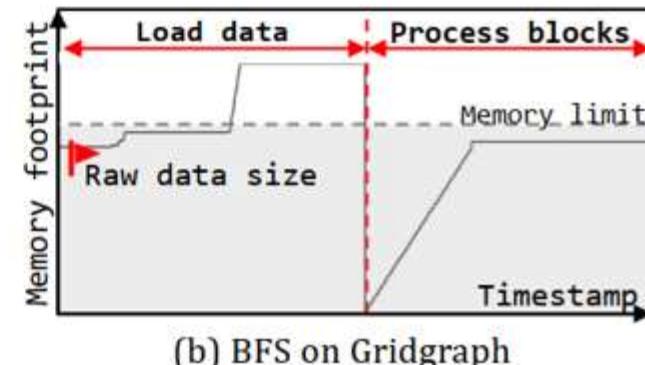
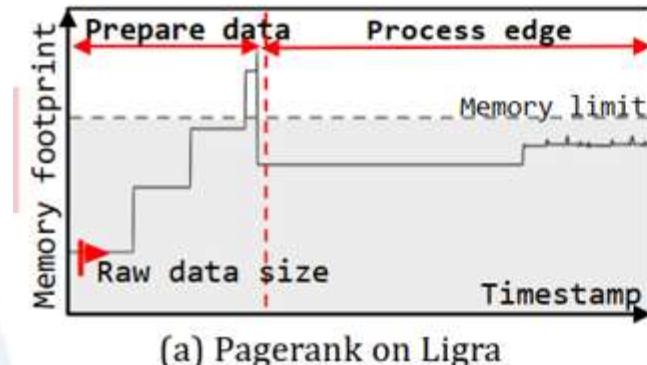
需要一种安全的方式来更新远程内存中保存的数据，为数据提供空间、时间和引用上的安全性

星火计划

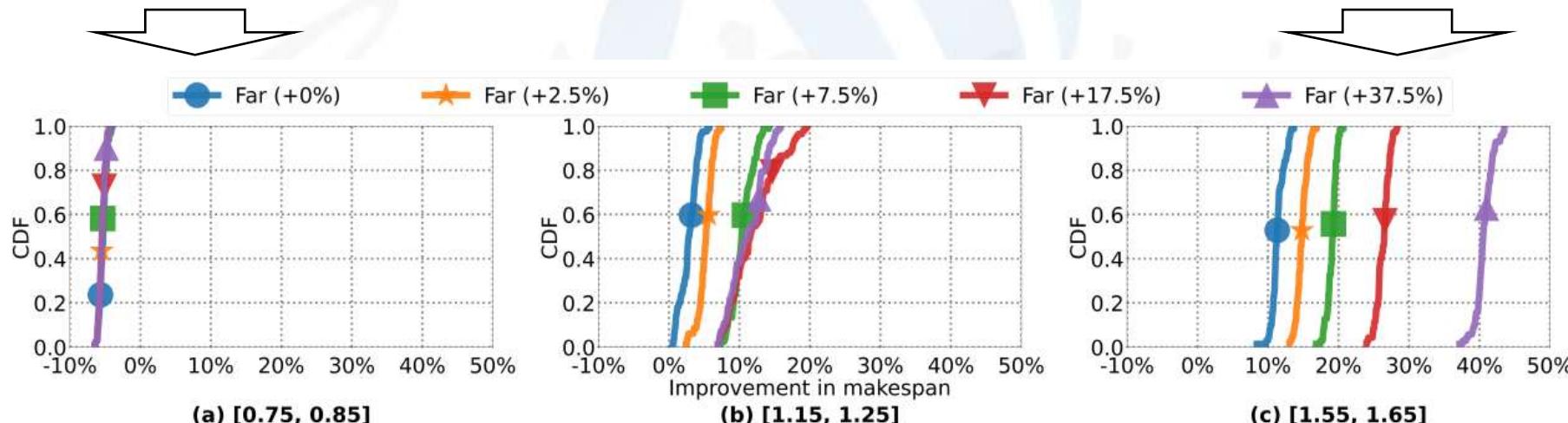
2. 软件技术：资源管理->高利用率



任务在远内存系统上表现有差异

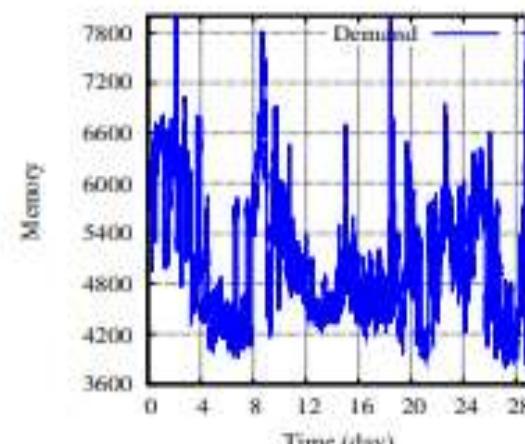
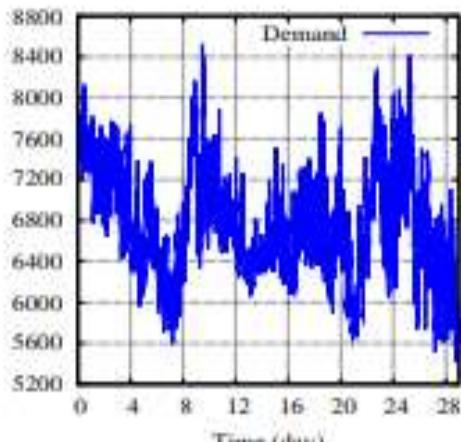


任务在远内存系统上的差异具有动态性

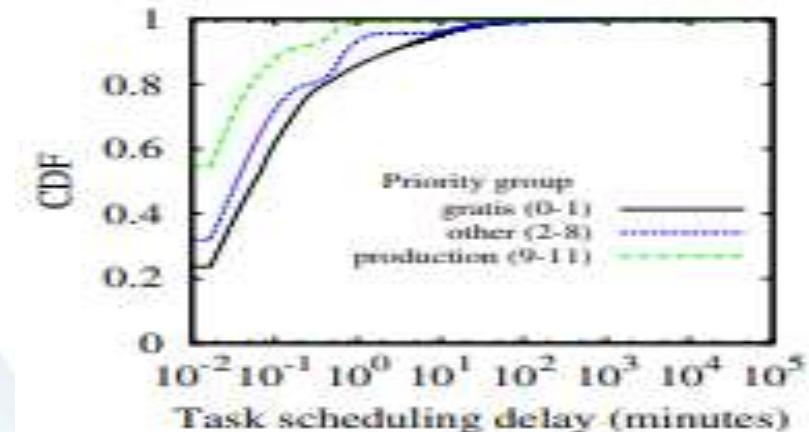


根据任务表现限制其内存并触发远内存访问，可以提升利用率和任务吞吐量

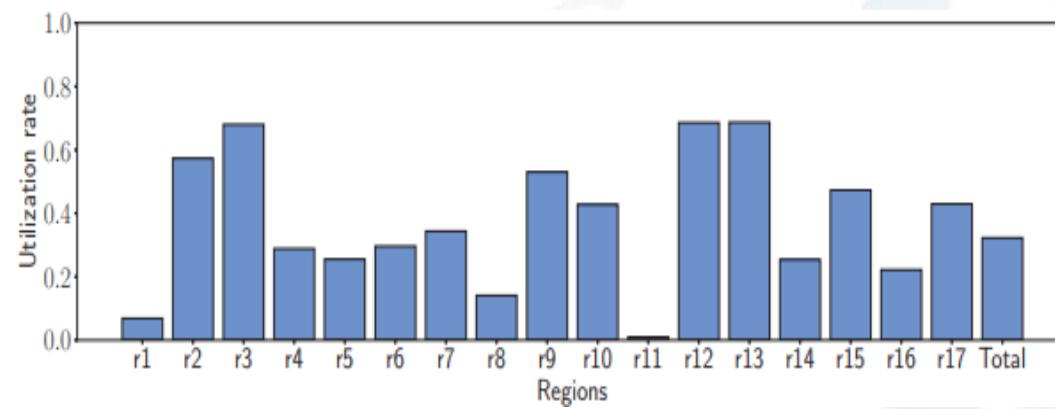
2.软件技术：异构资源调度



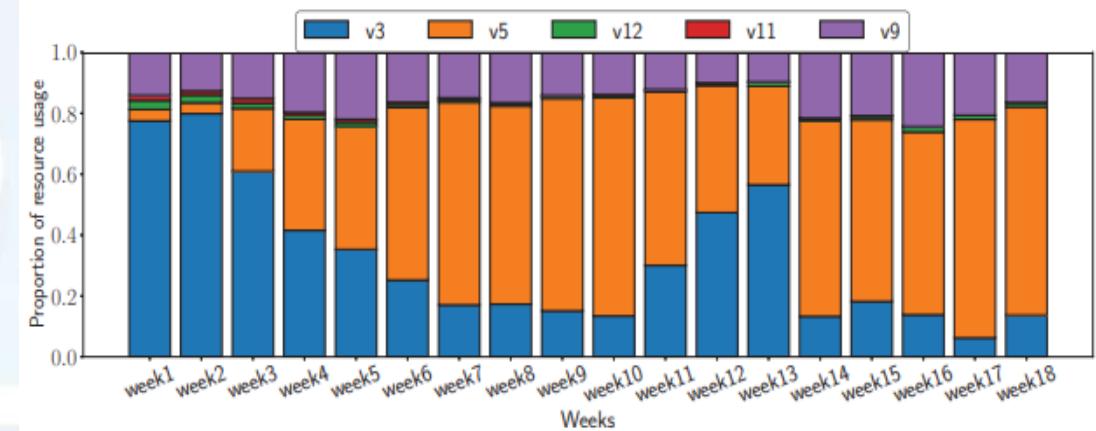
任务对于异构资源需求差异



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



目录

1

网络技术
对分离式内存的影响

2

软件技术
对分离式内存的影响

3

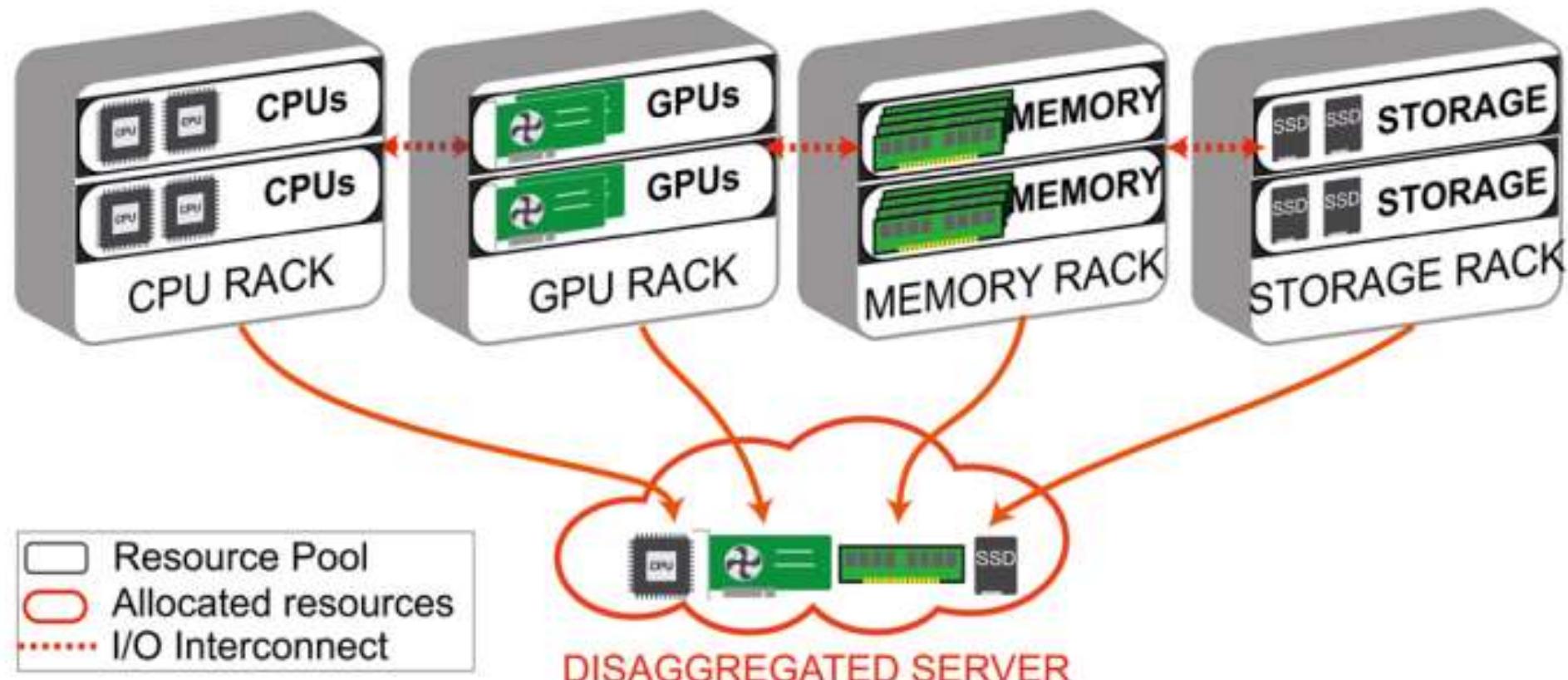
硬件技术
对分离式内存的影响

4

分离式内存
与超融合基础设施

星火计划

3.硬件技术：分离式架构主要部件



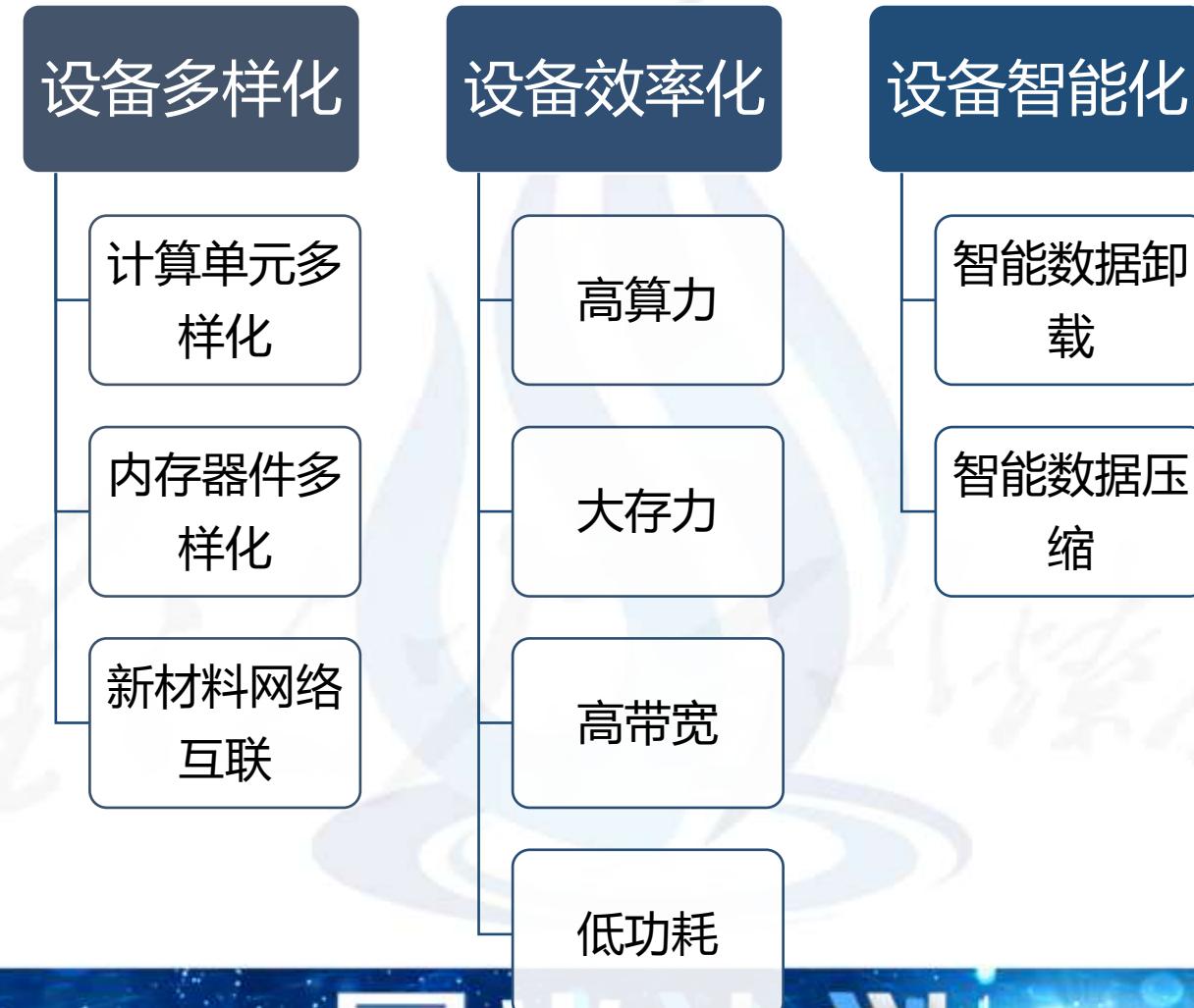
星火计划

中国 电信 转型 专业 领军 人才 培养 项目

交通大学



3.硬件技术：整体趋势

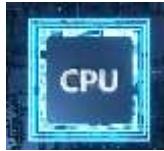


星火计划



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

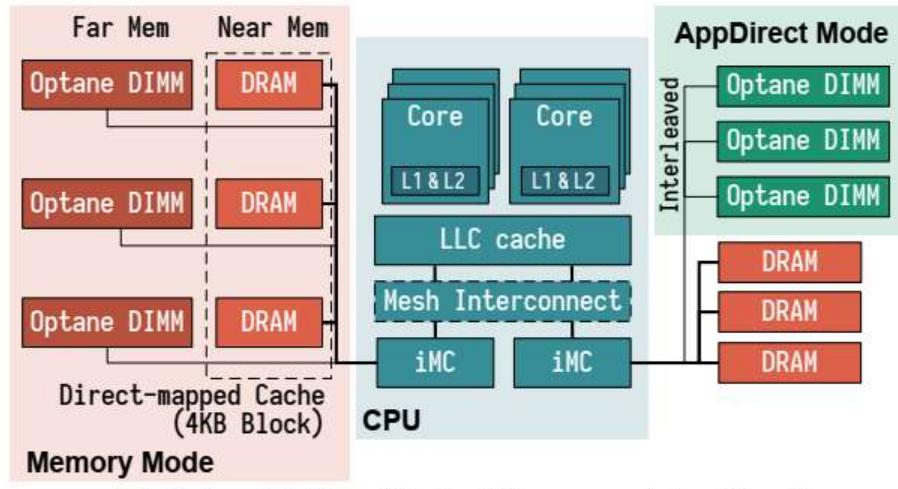
星火计划



3.硬件技术：设备多样化

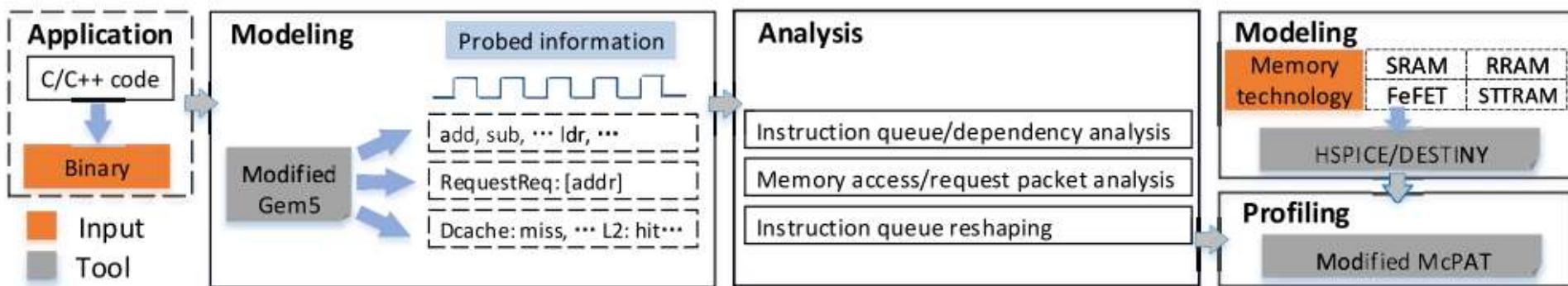
内存设备多样化

持久式内存



(a) Optane Platform Modes (Memory and AppDirect)

存内计算



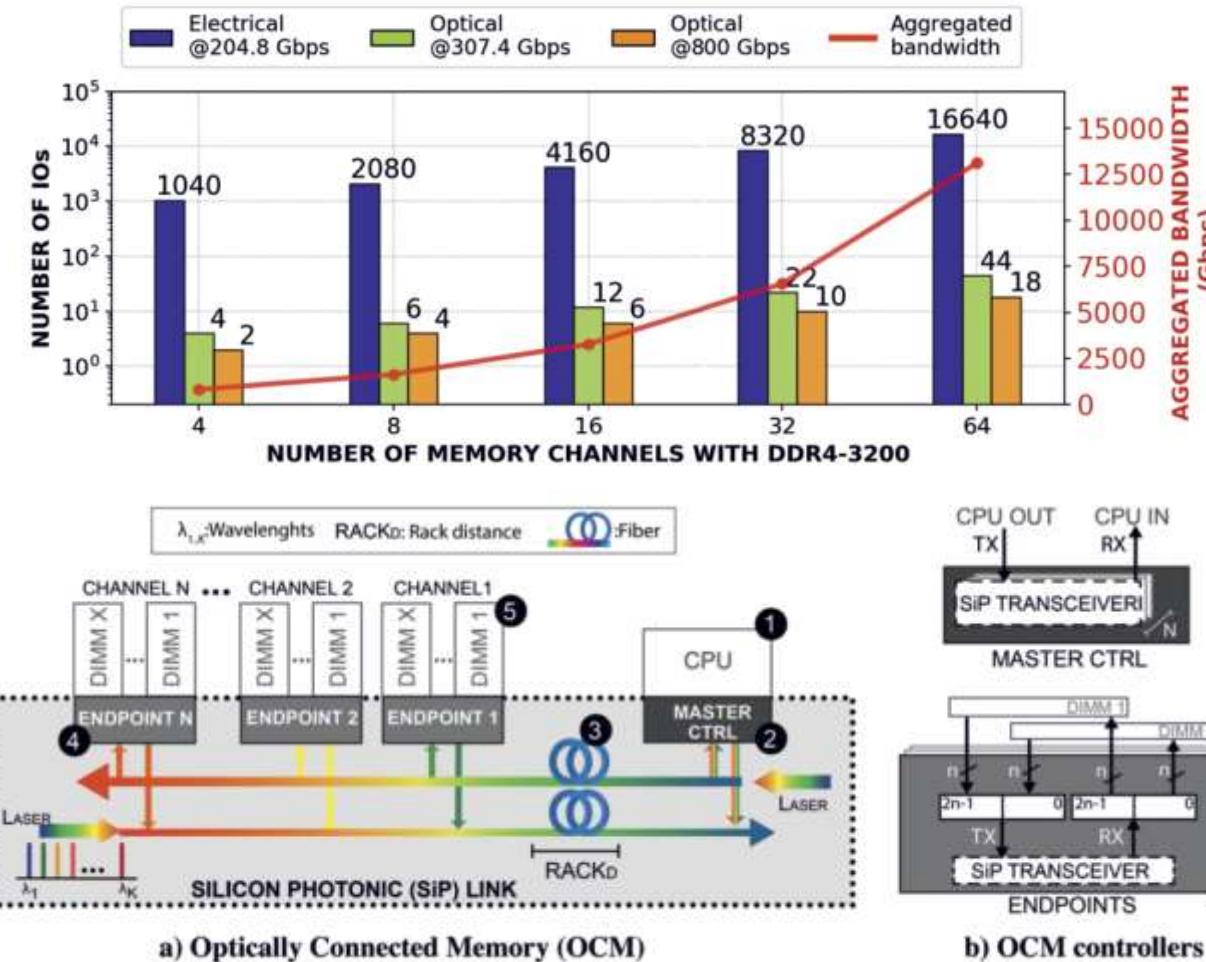
- 基于硬件语言编程
- 基于仿真设计

- 基于专用原语编程
- 基于仿真设计

星火计划



3.硬件技术：设备多样化



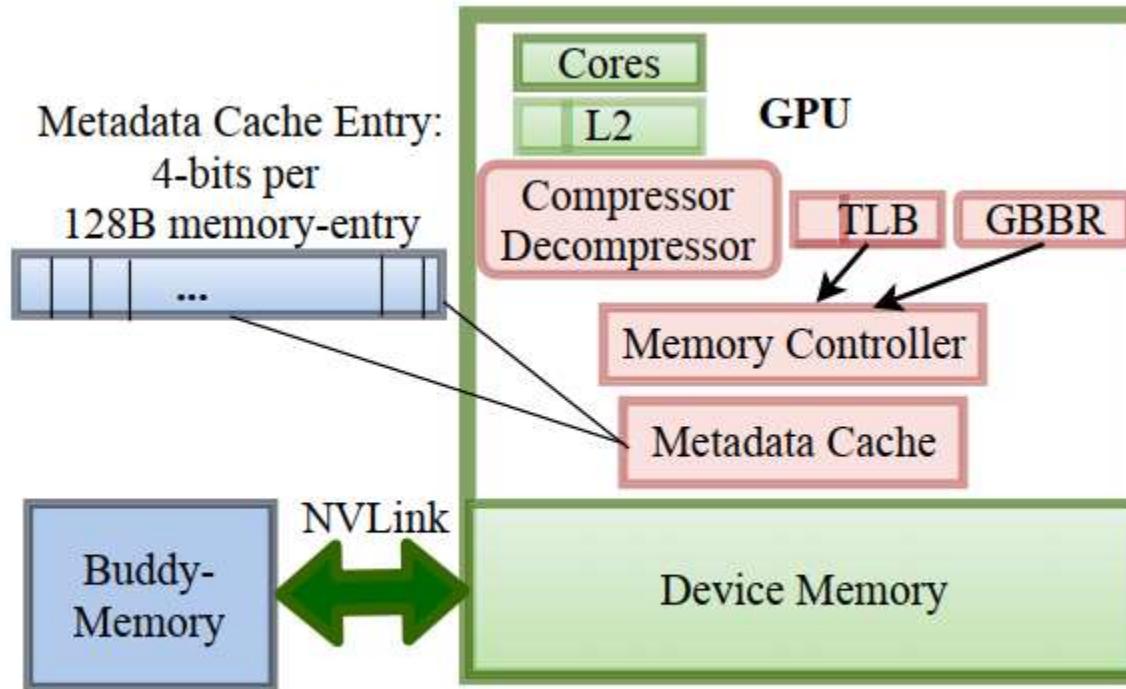
- Optically connected memory (OCM)
- 采用光学互联内存的新材料
- 超高带宽
- 超低功耗
- 执行速度比基于nic的分解内存快5倍

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

星火计划



3.硬件技术：设备智能化-智能压缩



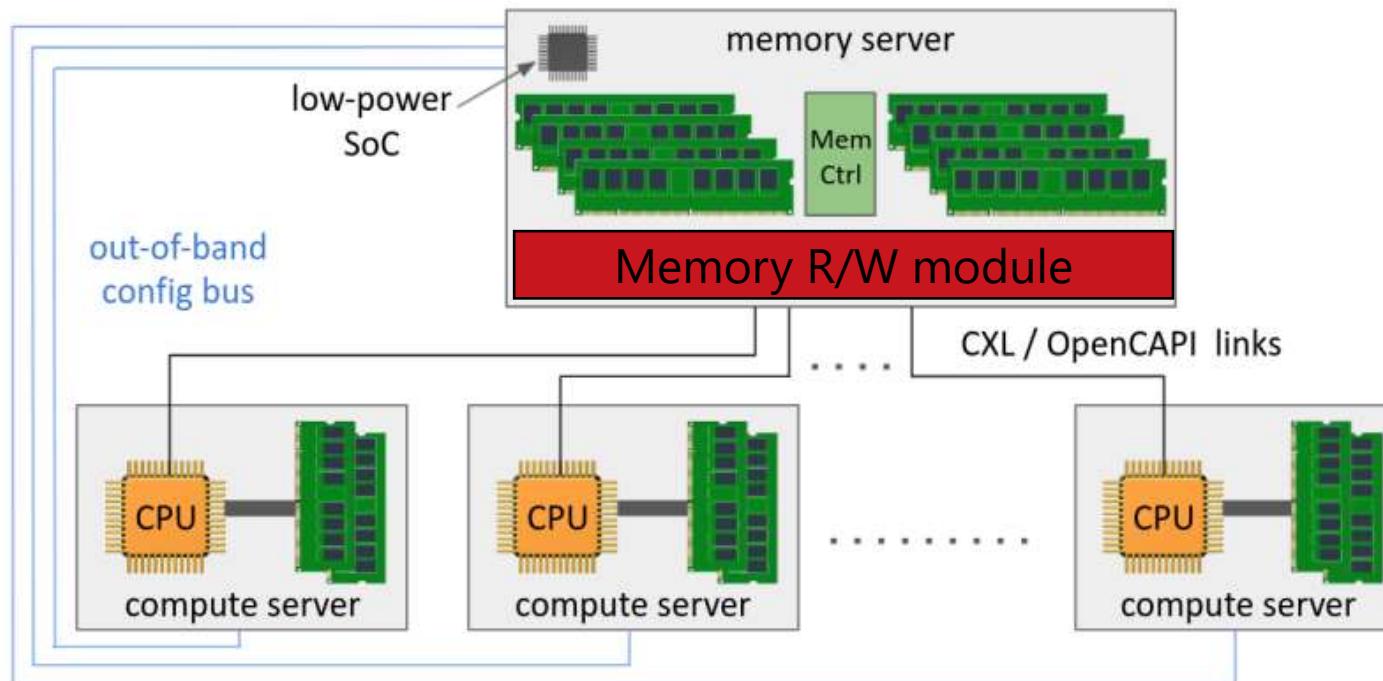
(a) Architectural overview

- 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

星火计划



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

星火计划



目 录

1

网络技术
对分离式内存的影响

2

软件技术
对分离式内存的影响

3

硬件技术
对分离式内存的影响

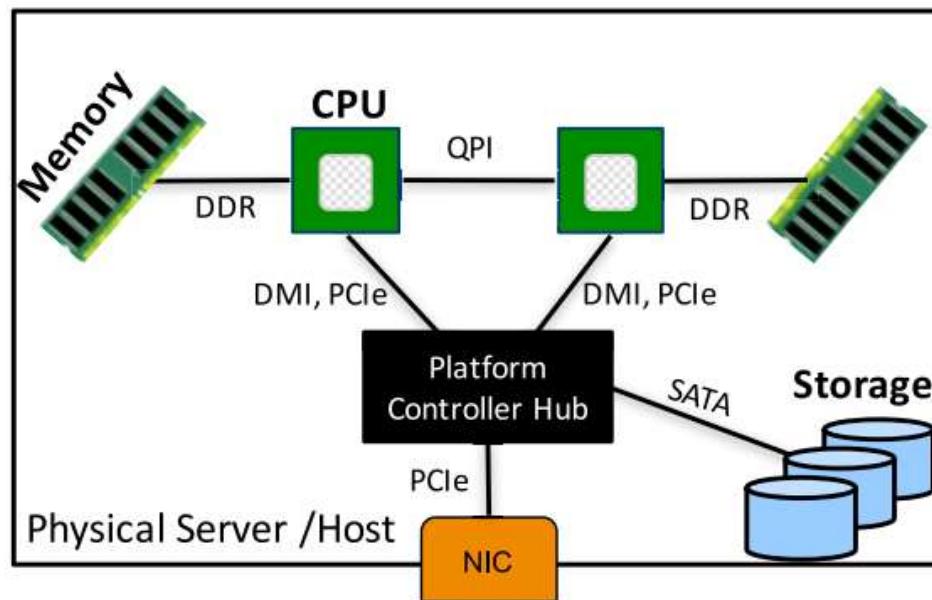
4

分离式内存
与超融合基础设施

星火计划

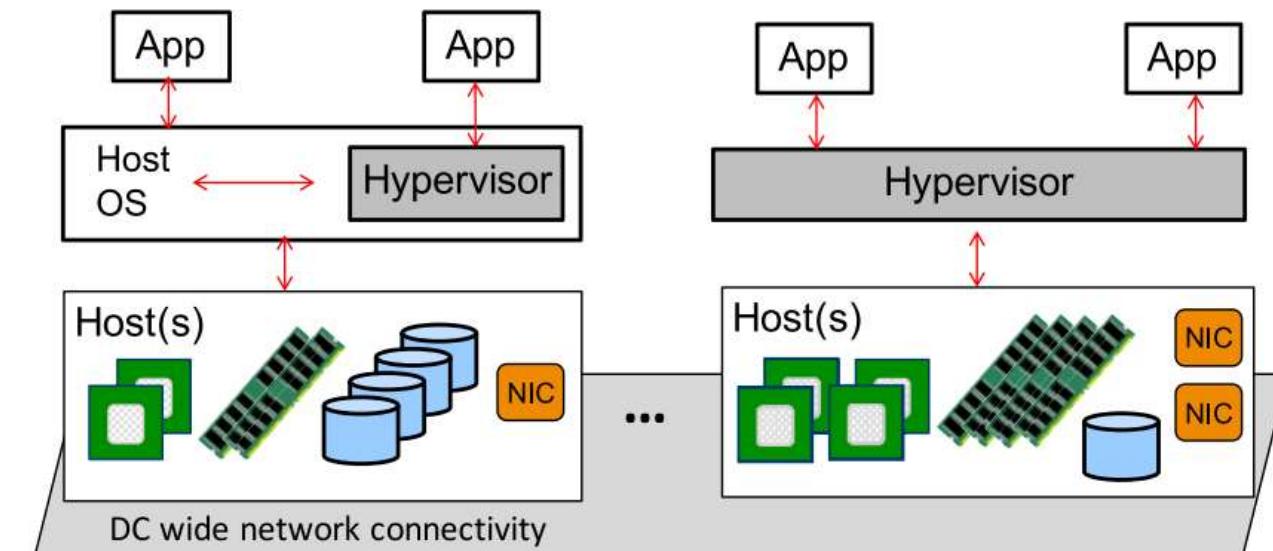
4. 超融合基础设施：软件定义硬件 (SDHI)

Hardware infrastructures



实际硬件结构

Software defined infrastructures (SDI)



Server oriented SDI

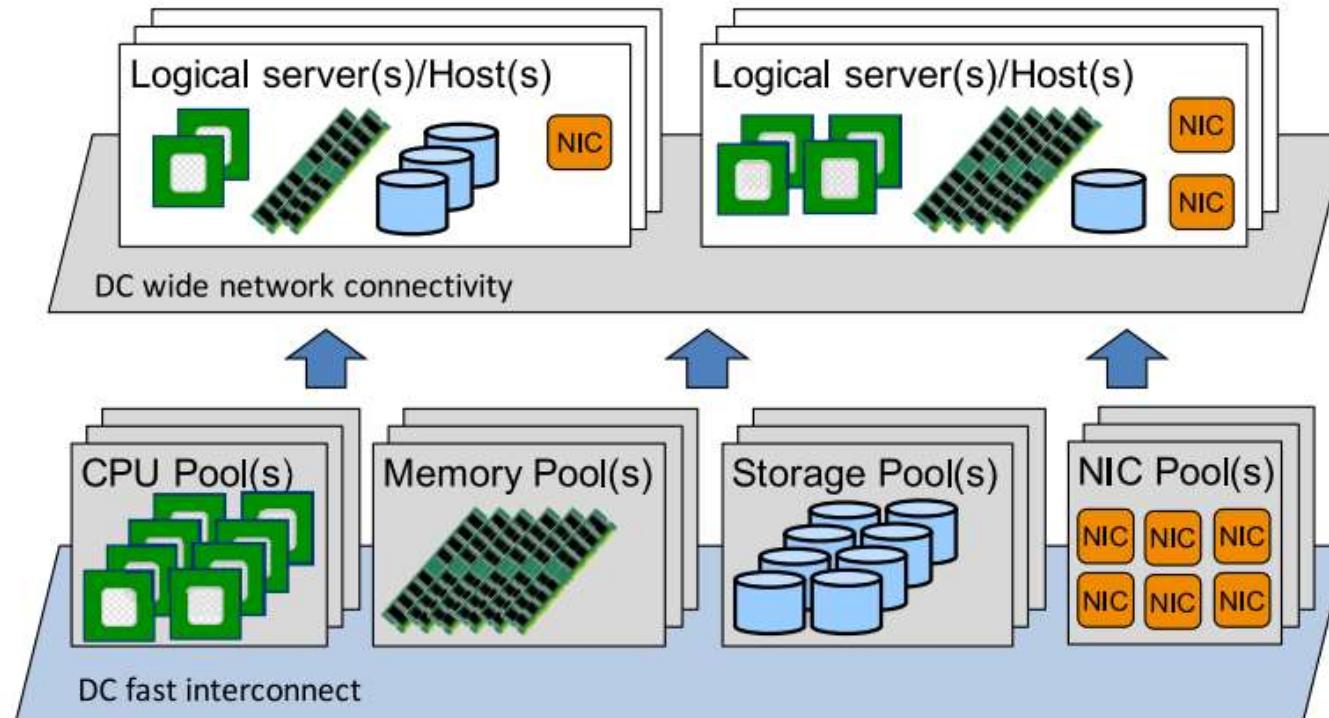
软件定义的架构

星火计划



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）

星火计划



4. 超融合基础设施：软件定义硬件 (SDI)

(SDI)

Software-Defined Networking	SDN [17]	Separates the network control planes from data planes and physical network entities to improve programmability, efficiency, and extensibility of network.
Software-Defined Storage	SDS [21]	Separates the control planes from the data plane of a storage system enabling heterogeneous storage to respond dynamically to changing workload demands.
Software-Defined Computing	SDC [11]	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

星火计划

中国电信转型专业领军人才培养项目

TSJU



4. 超融合基础设施： HCI简介

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

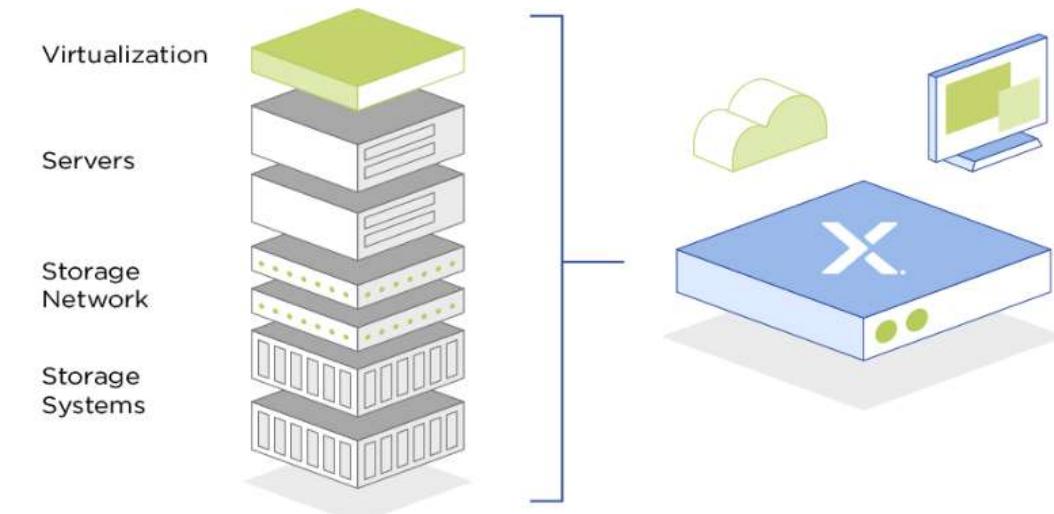
超融合基础架构是一个软件定义的 IT 基础架构，它可虚拟化常见“硬件定义”系统的所有元素。

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。

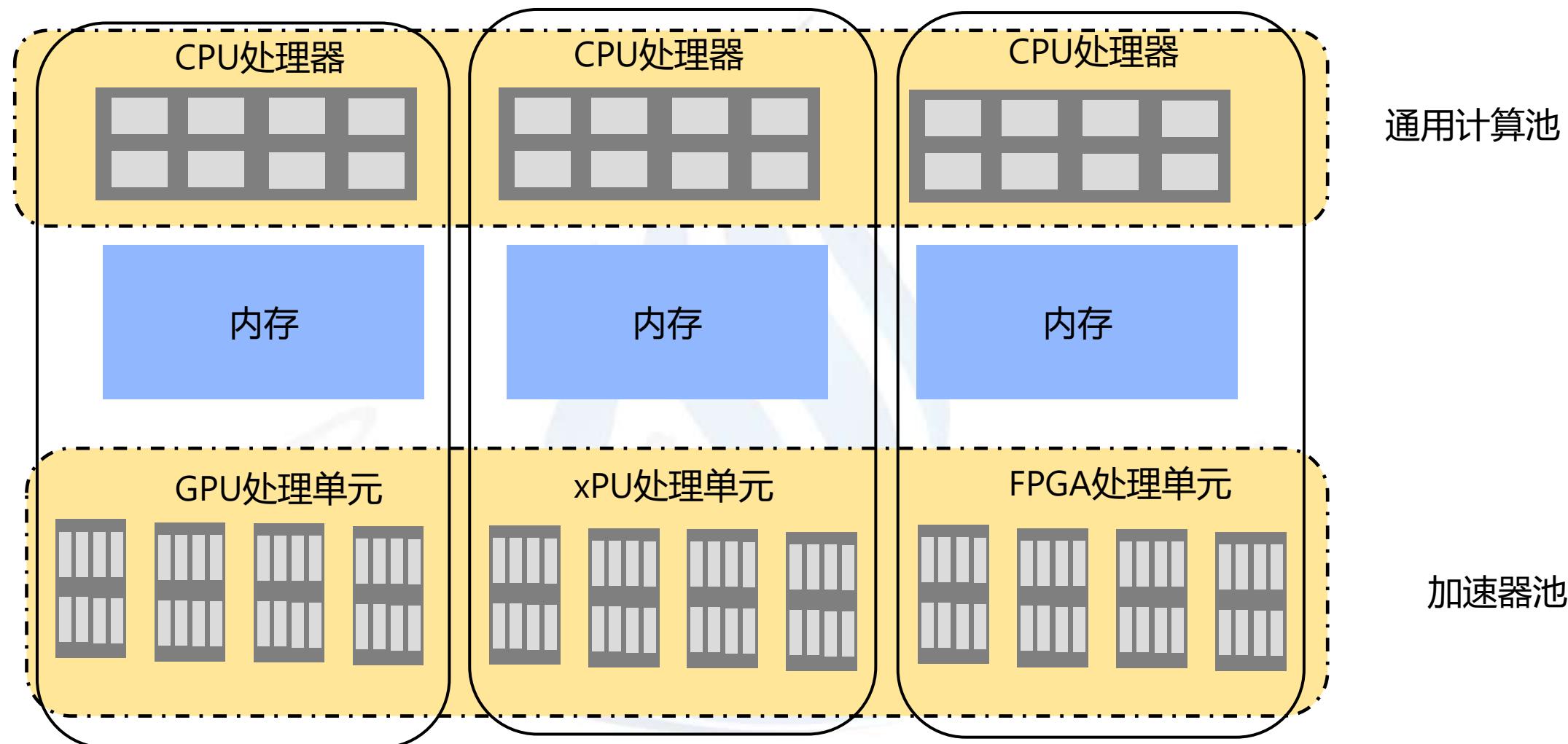
HCI 由两个主要组件组成：**分布式平面和管理平面**。

- **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。



星火计划

4. 超融合基础设施：异构计算架构



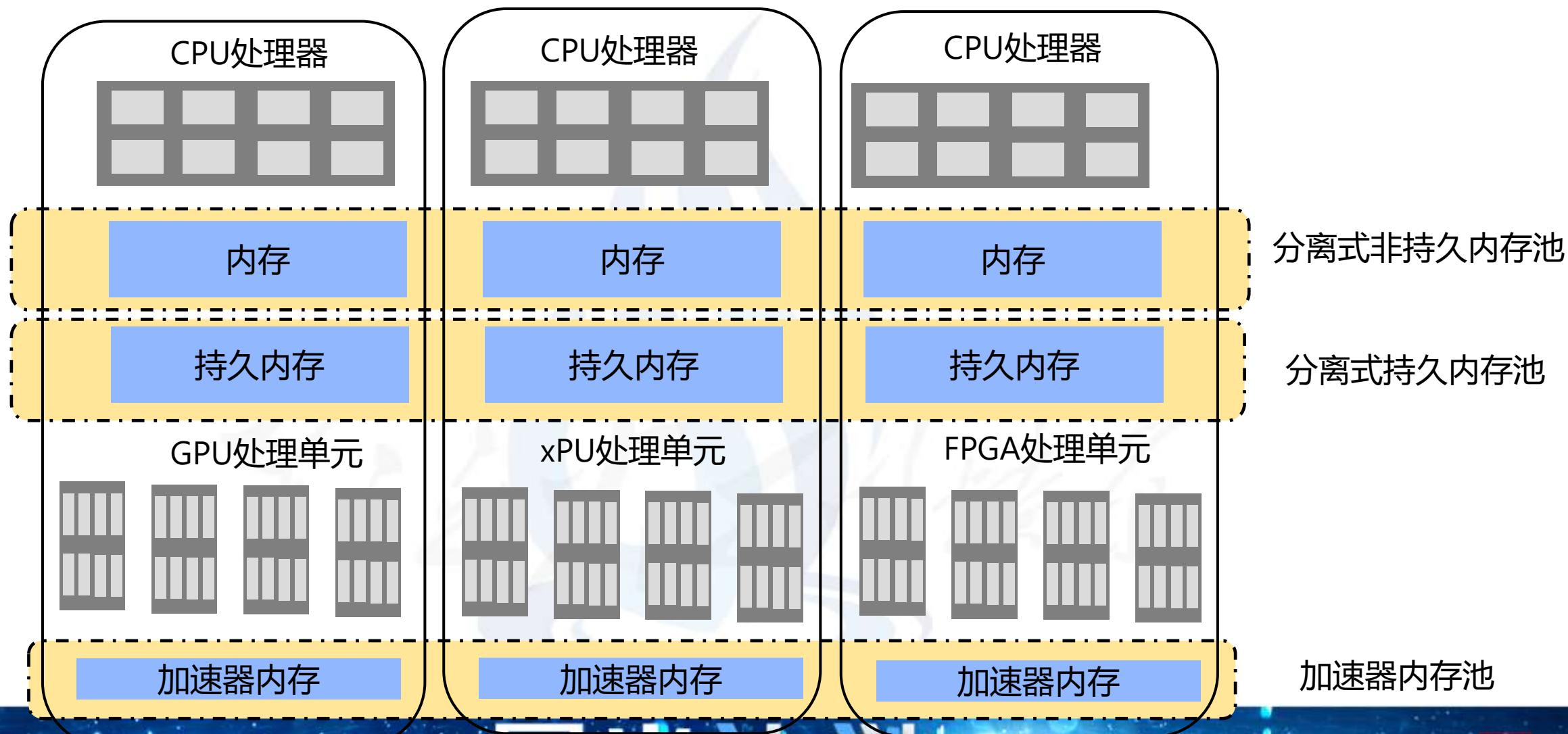
星火计划

中国 电信 转型 专业 领军 人才 培养 项目

交通大学

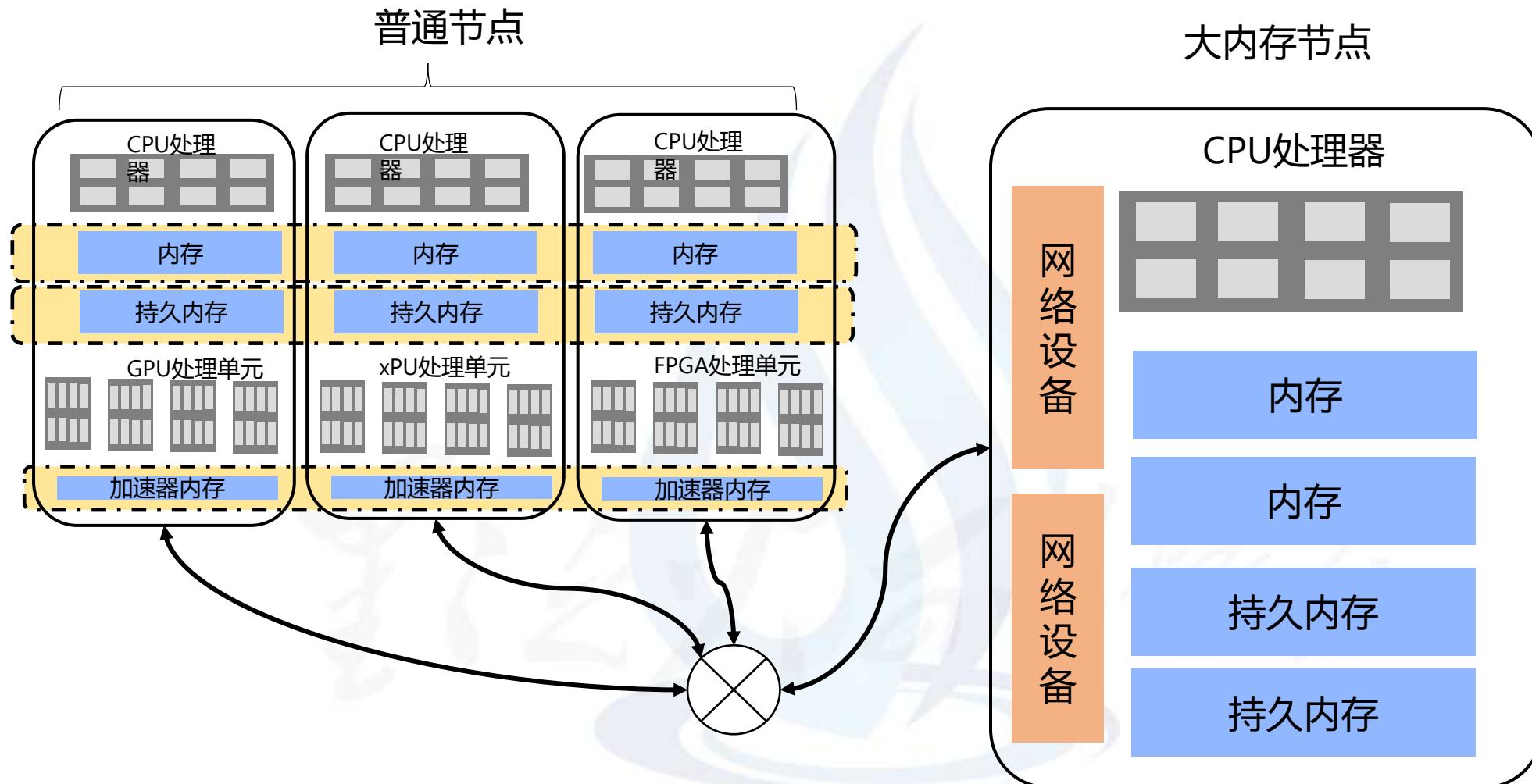


4. 超融合基础设施：内存层次



星火计划

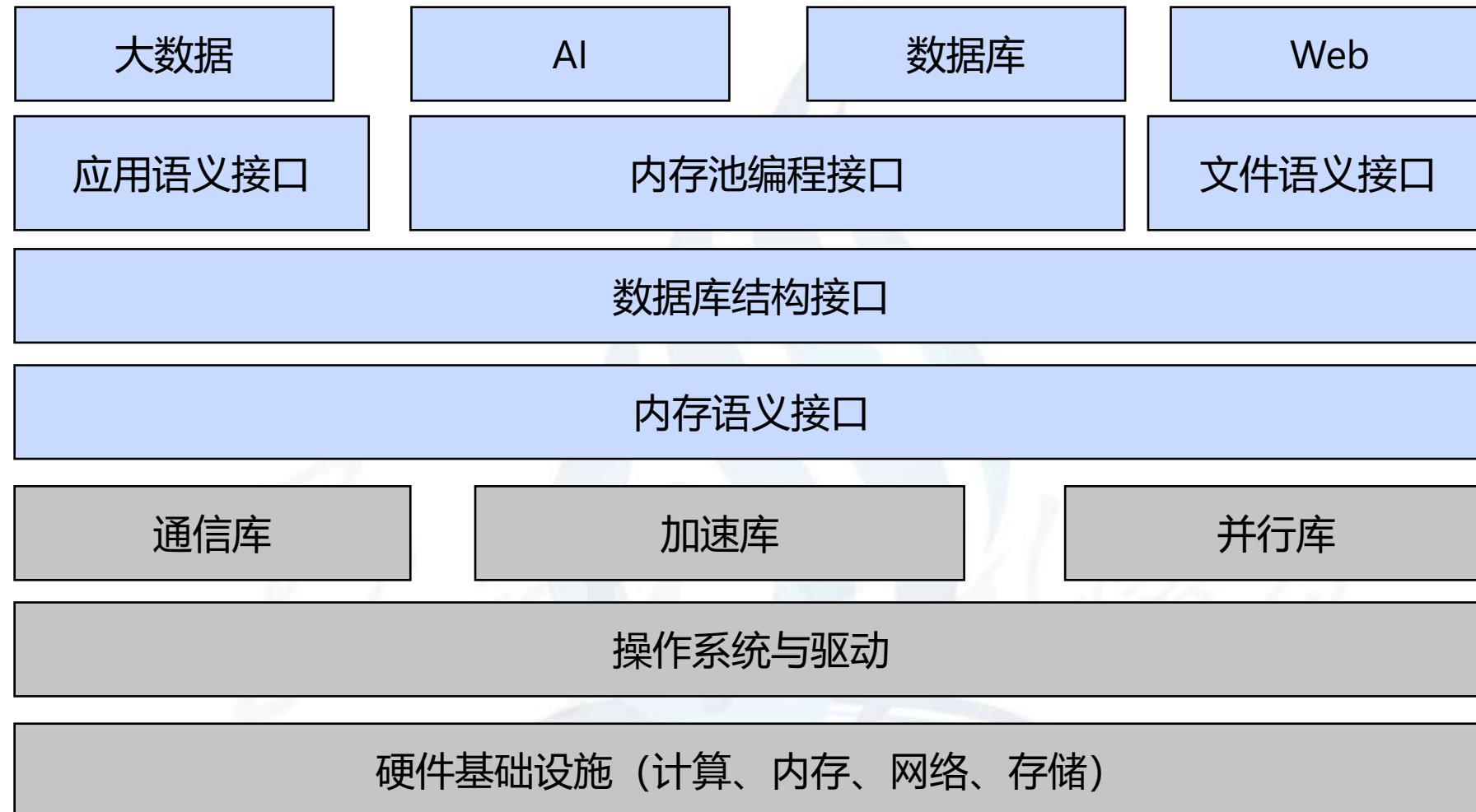
4. 超融合基础设施：网络互联资源池



星火计划



4. 超融合基础设施：应用使用接口



星火计划



谢谢大家！

星火计划

中国电信转型专业领军人才培养项目