一、分离式内存概论

讲者：

上海交通大学 计算机系 SAIL实验室

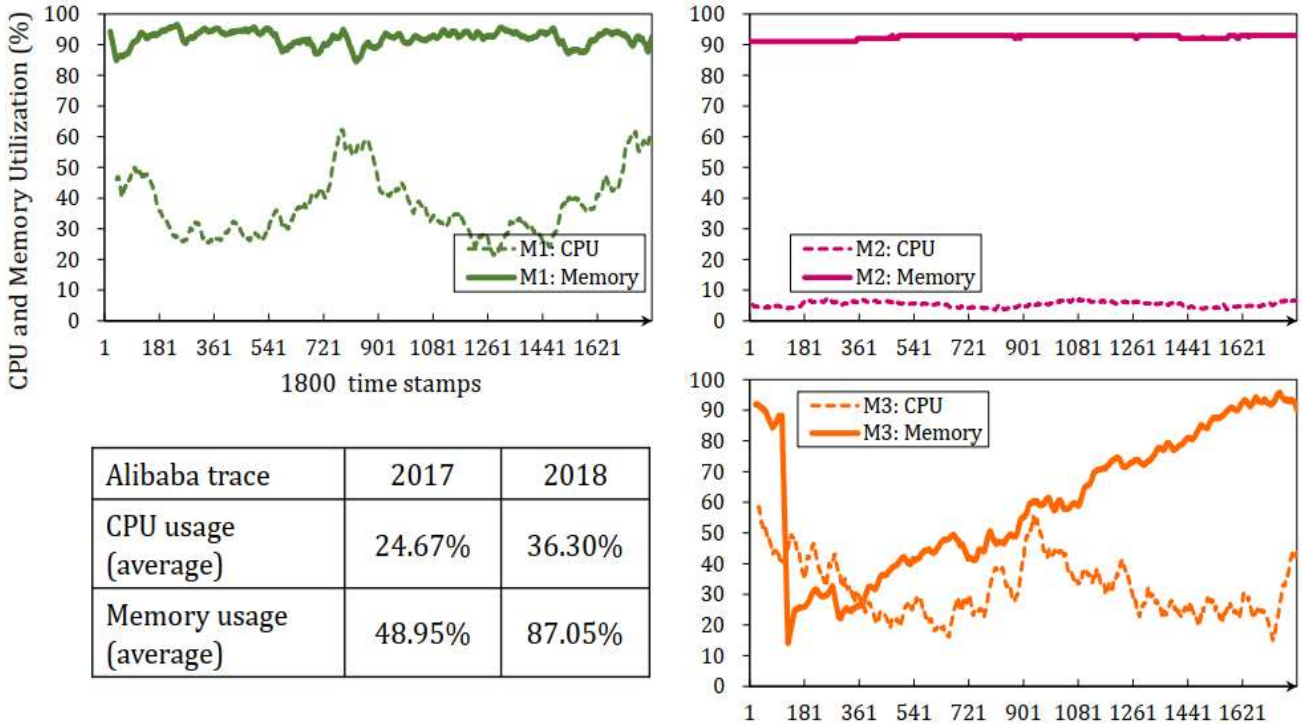
2023年2月

上海交通大学
SHANGHAI JIAO TONG UNIVERSITY

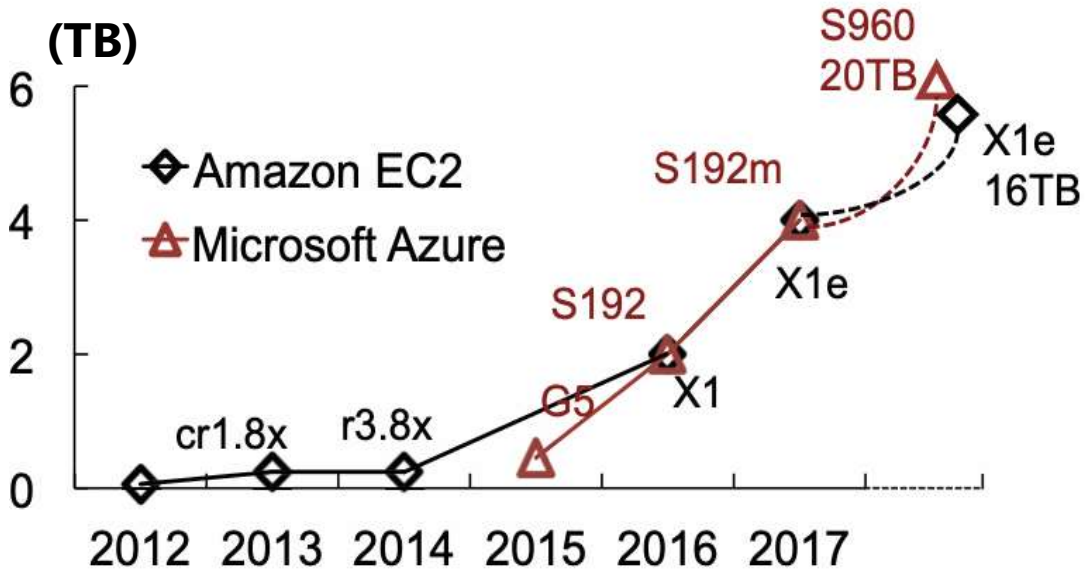
目录

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%



服务器内存扩展需求

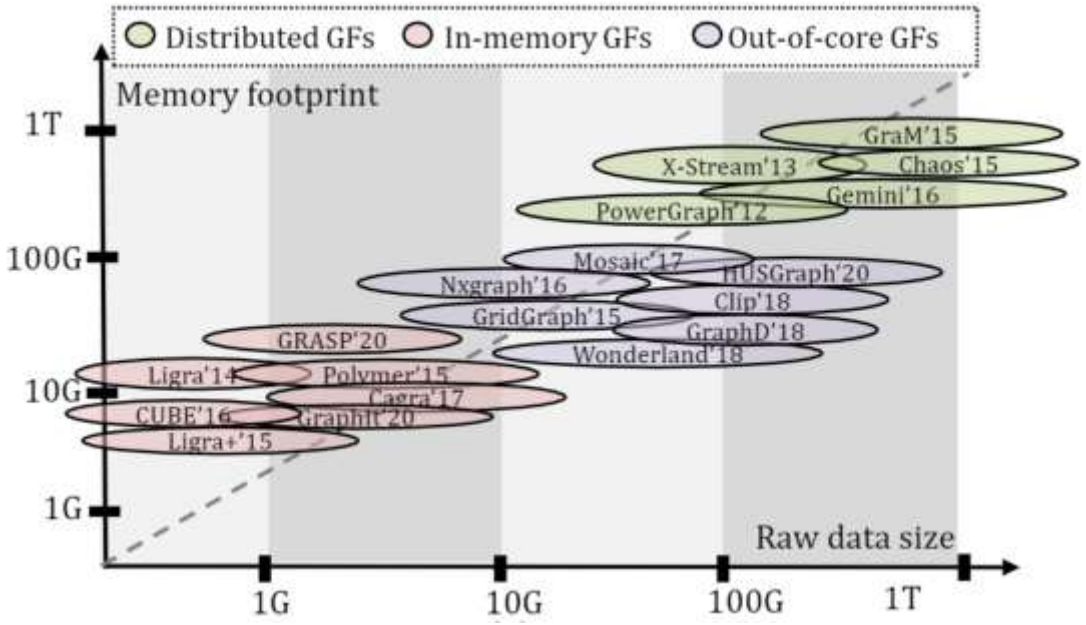
内存资源逐渐消耗殆尽，亟需进行内存资源内部和外部的拓展

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

背景2：应用的大内存需求

上海交通大學 SHANGHAI JIAO TONG UNIVERSITY

图计算：



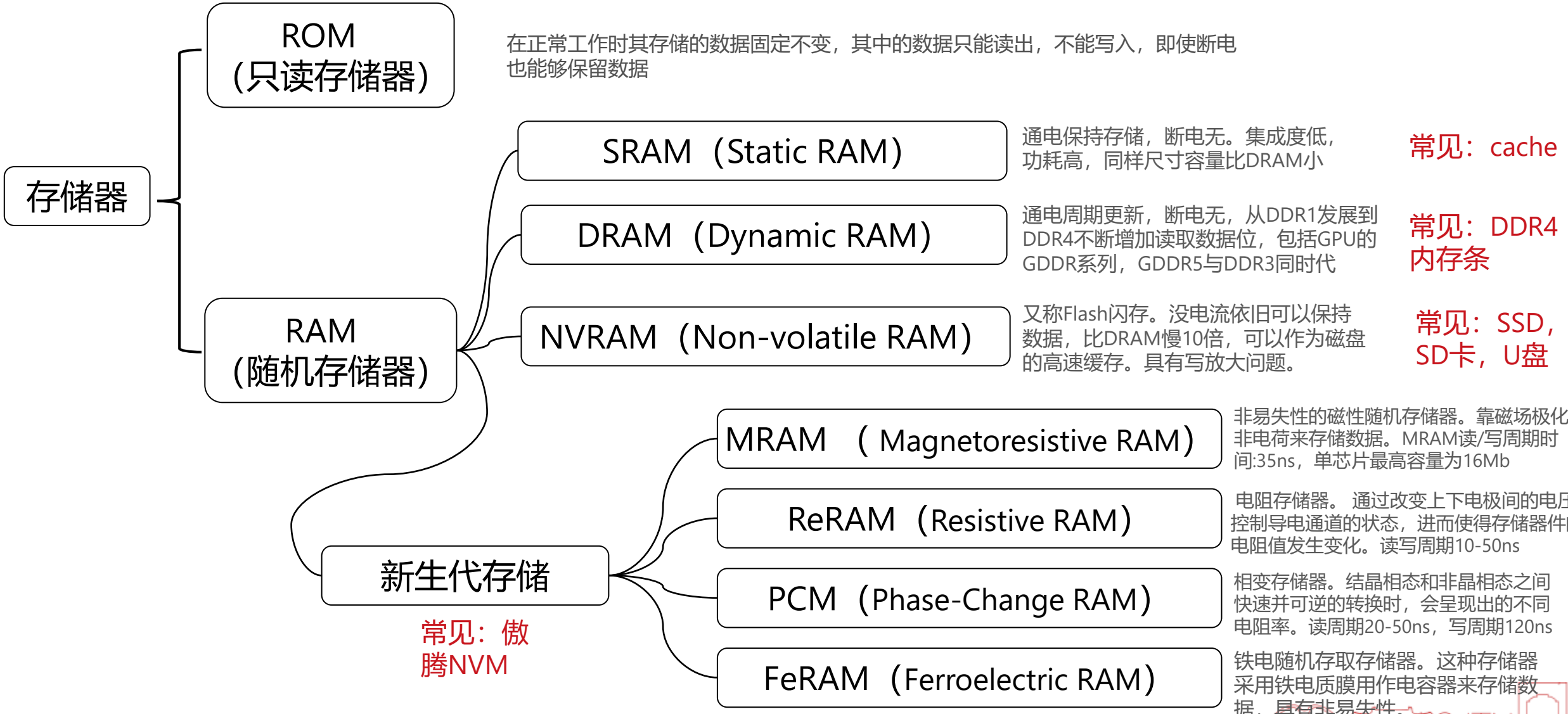
AI计算：

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

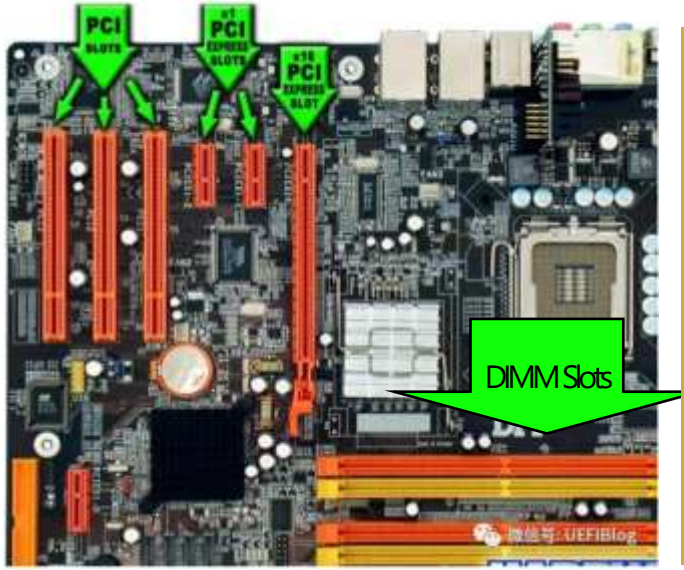
1.内存简介：硬件设备篇

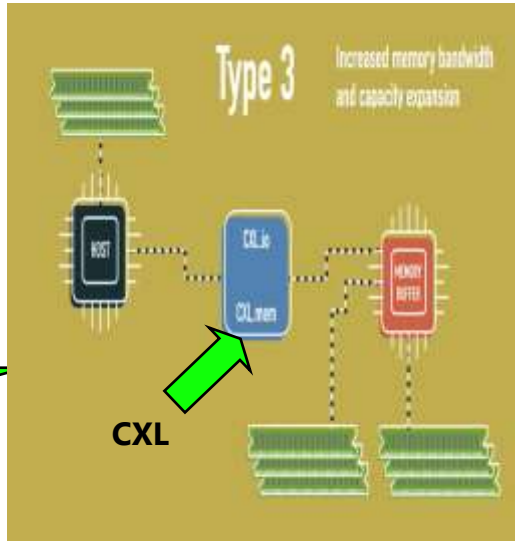

存储器
ROM
（只读存储器）
在正常工作时其存储的数据固定不变，其中的数据只能读出，不能写入，即使断电也能够保留数据
RAM
（随机存储器）
SRAM（Static RAM）
通电保持存储，断电无。集成度低，功耗高，同样尺寸容量比DRAM小
常见：cache
DRAM（Dynamic RAM）
通电周期更新，断电无，从DDR1发展到DDR4不断增加读取数据位，包括GPU的GDDR系列，GDDR5与DDR3同时代
常见：DDR4内存条
NVRAM（Non-volatile RAM）
又称Flash闪存。没电流依旧可以保持数据，比DRAM慢10倍，可以作为磁盘的高速缓存。具有写放大问题。
常见：SSD，SD卡，U盘
新生代存储
常见：傲腾NVM
MRAM （Magnetoresistive RAM）
非易失性的磁性随机存储器。靠磁场极化而非电荷来存储数据。MRAM读/写周期时间:35ns，单芯片最高容量为16Mb
ReRAM（Resistive RAM）
电阻存储器。 通过改变上下电极间的电压来控制导电通道的状态，进而使得存储器件的电阻值发生变化。读写周期10-50ns
PCM（Phase-Change RAM）
相变存储器。结晶相态和非晶相态之间快速并可逆的转换时，会呈现出的不同电阻率。读周期20-50ns，写周期120ns
FeRAM（Ferroelectric RAM）
铁电随机存取存储器。这种存储器采用铁电质膜用作电容器来存储数据，具有非易失性。

1.内存简介：互联篇

- 互联接口
  - DIMM（双列直插内存模块）
  - PCIe（并行总线）
    - PCIe port
    - PCIe switch
  - NVLink
    - NVLink 模块
    - NVSwitch

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。

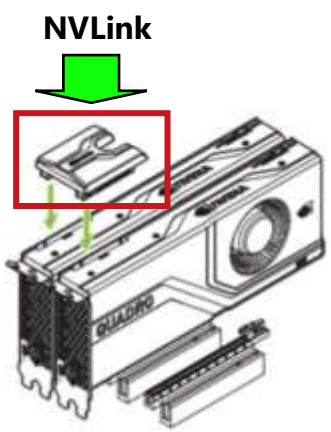


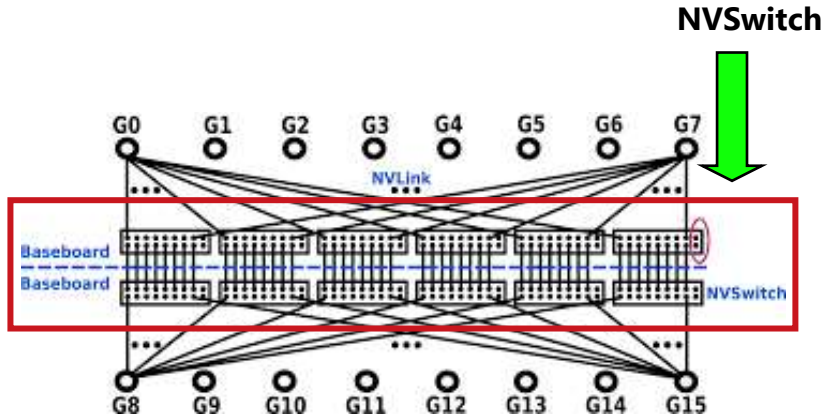


PCIe版本分为 1.0/2.0/3.0/4.0/5.0，PCIe的连线是由不同的lane来连接的，这些lane可以合在一起提供更高的带宽。譬如两个1lane可以合成2lane的连接，写作x2。两个x2可以变成x4，最大直到x16

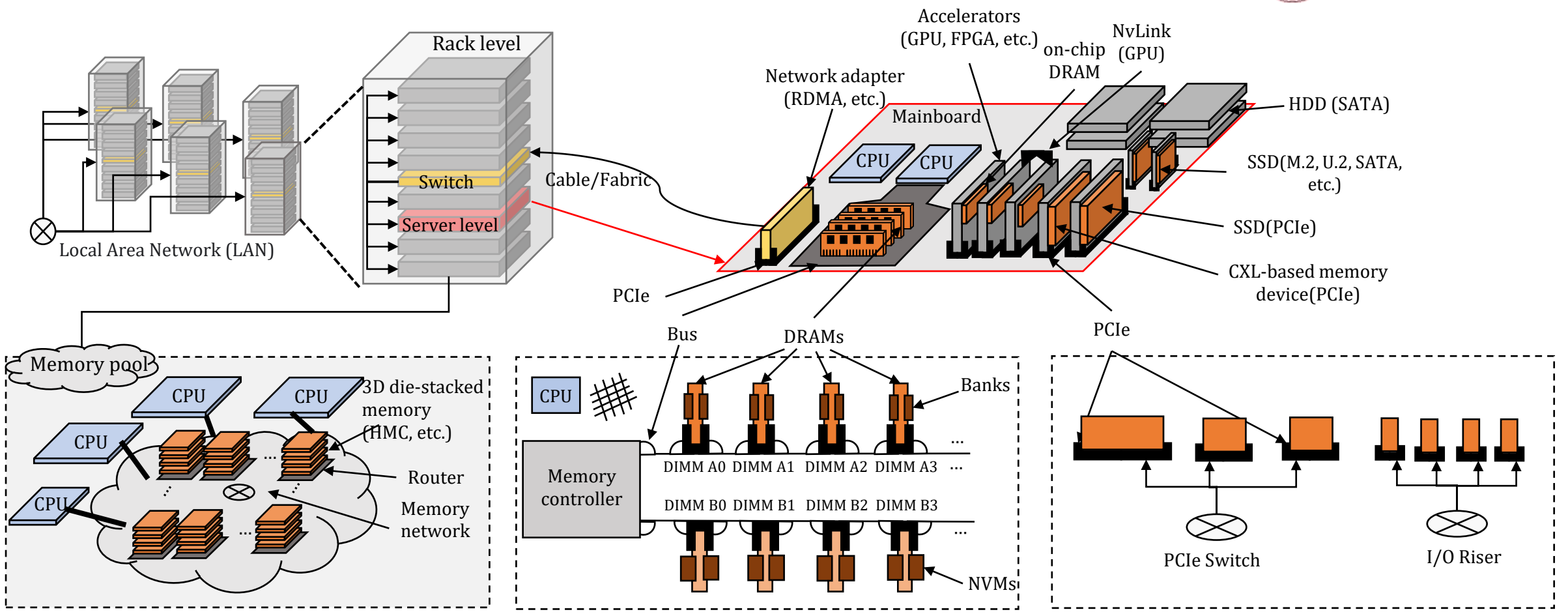
CXL是基于PCIe5.0的新互联协议

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



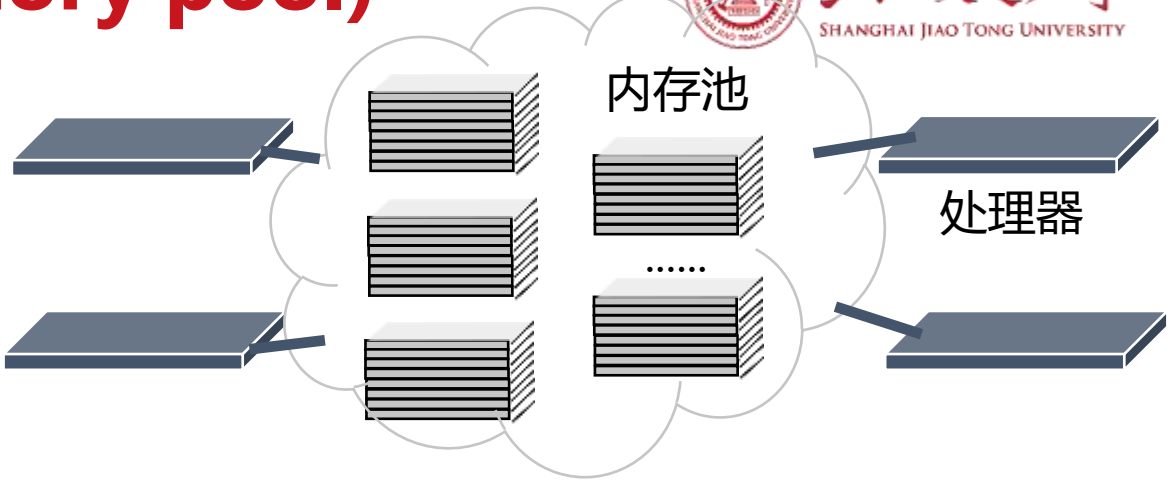


1.内存简介：大容量高密度内存



专用大内存　　DIMM大内存　　PCIe大内存

1.内存简介：中心式内存池(Memory pool)

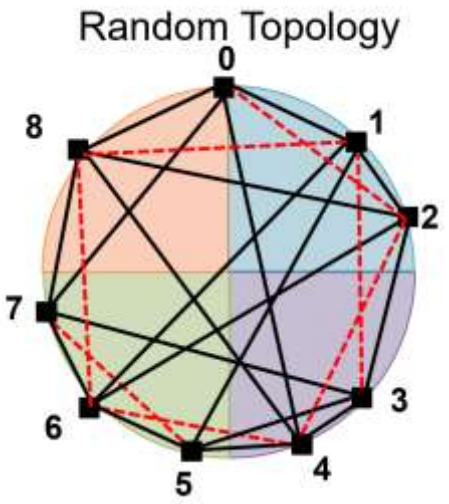


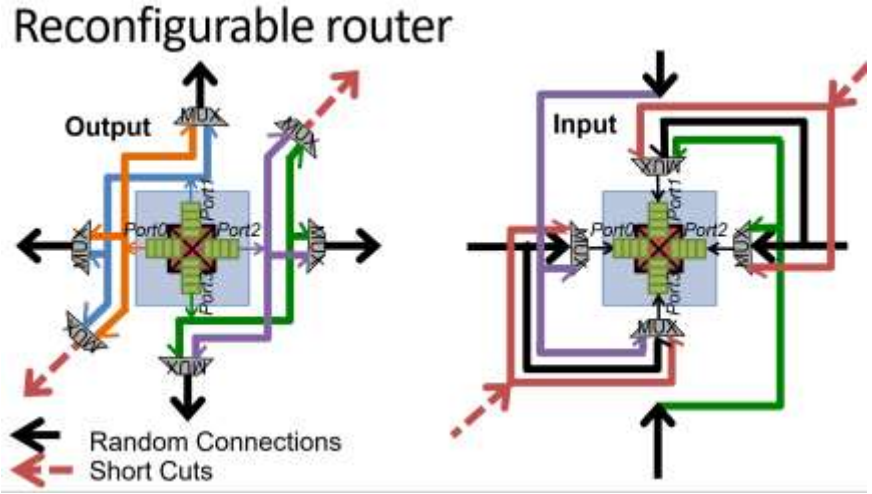
基本概念：

- 1989年，与之类似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或采用网络互联
- 设计目标：
  - 良好的可扩展性
  - 任意规模的高效互联

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





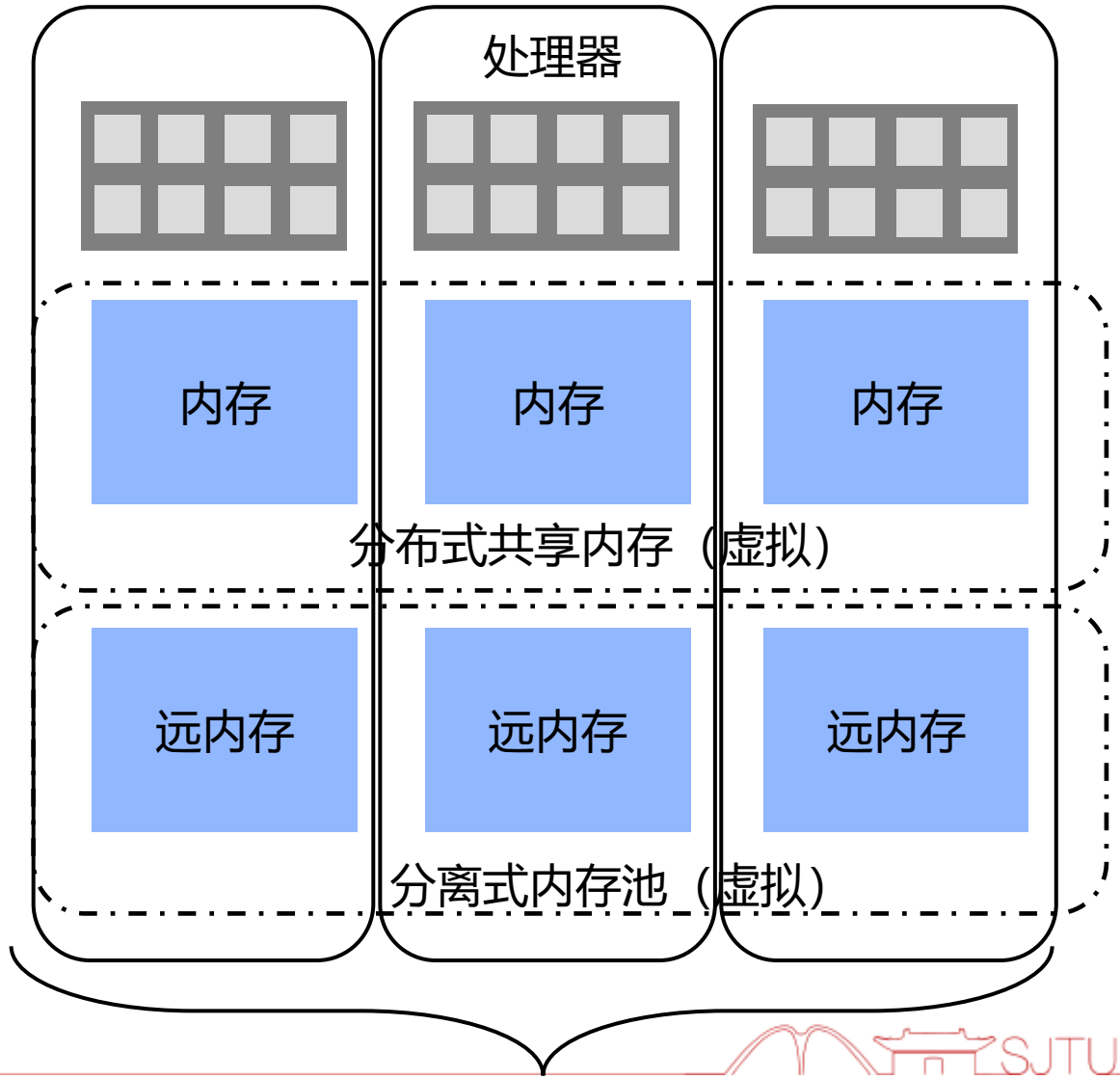
1.内存简介：去中心式内存池(Memory pool)

基本概念：

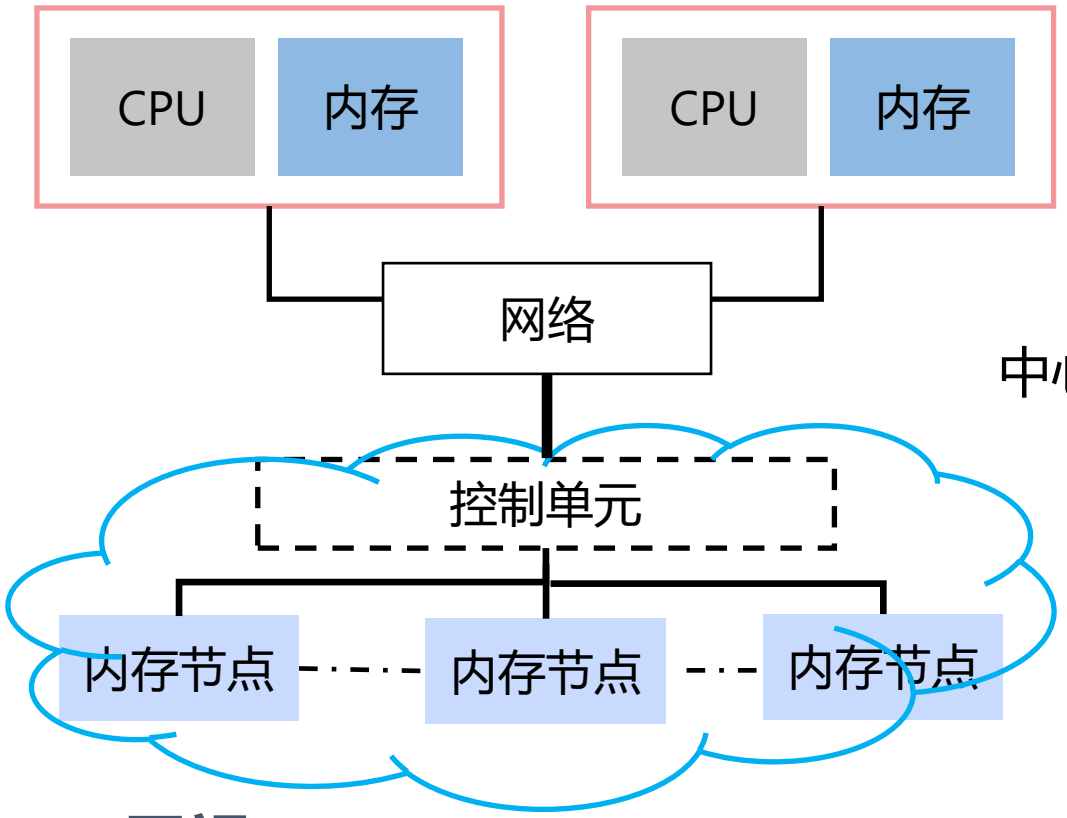
- 节点内部构建远内存，可供自己本地内存使用，也可供给其他计算节点。处理器除本地内存外，可以访问其他处理器的空闲内存
- 采用一定的分布式管理策略（中心式管理，共享视图管理），构建虚拟的资源池，能够进行灵活的内存资源共享

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

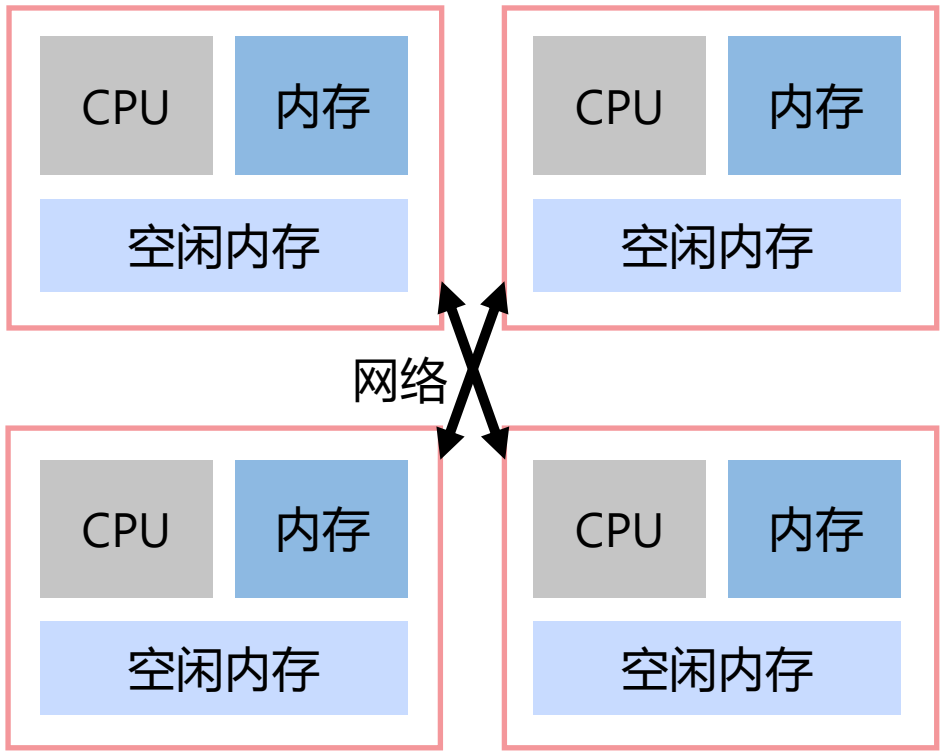


中心式与去中心式架构对比



中心化

去中心化



展望：

1. 内存池的发展趋势呈现混合化、异构化
2. 中心化与去中心化架构结合发展

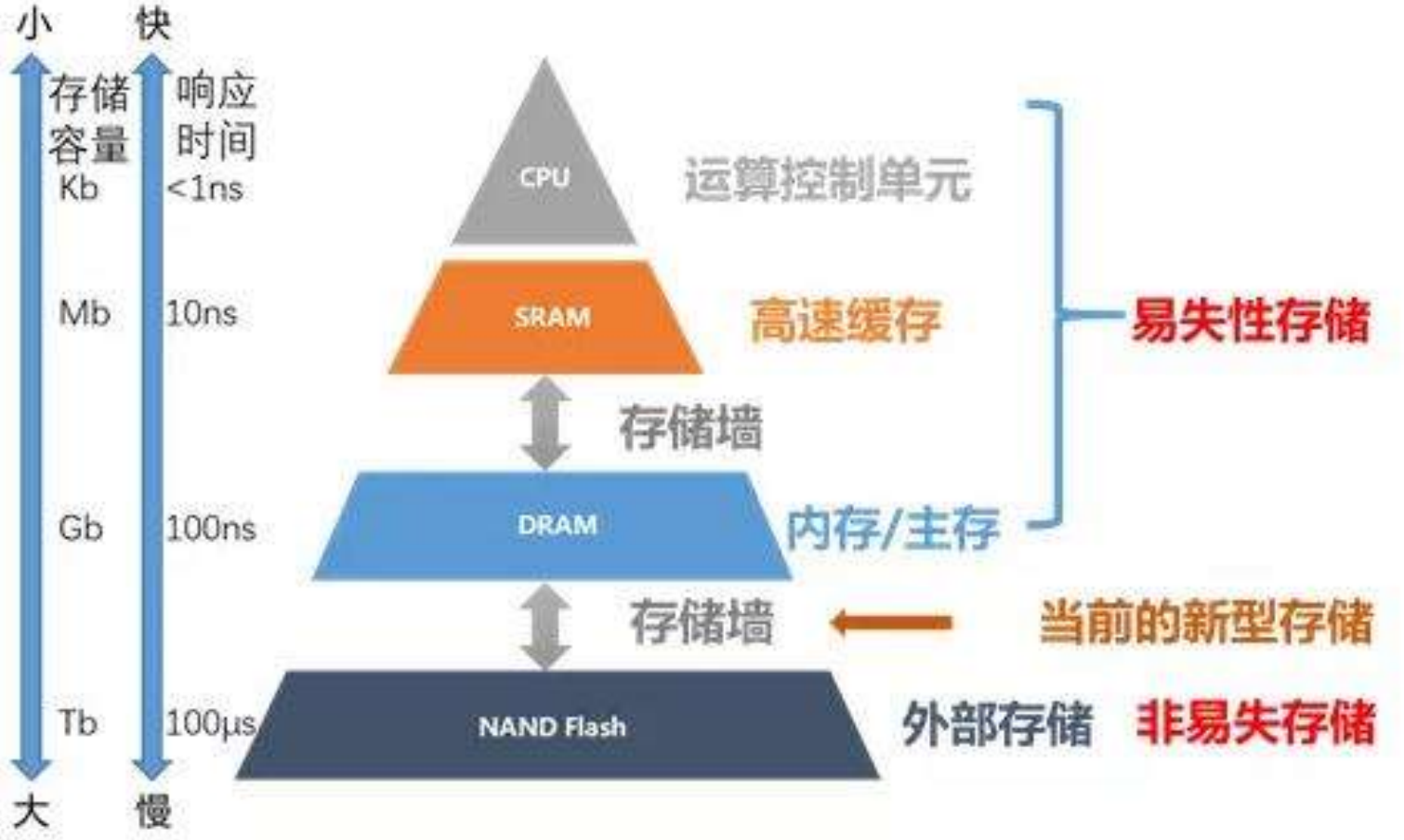
上海交通大学
SHANGHAI JIAO TONG UNIVERSITY

目录

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)， 响应时间位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

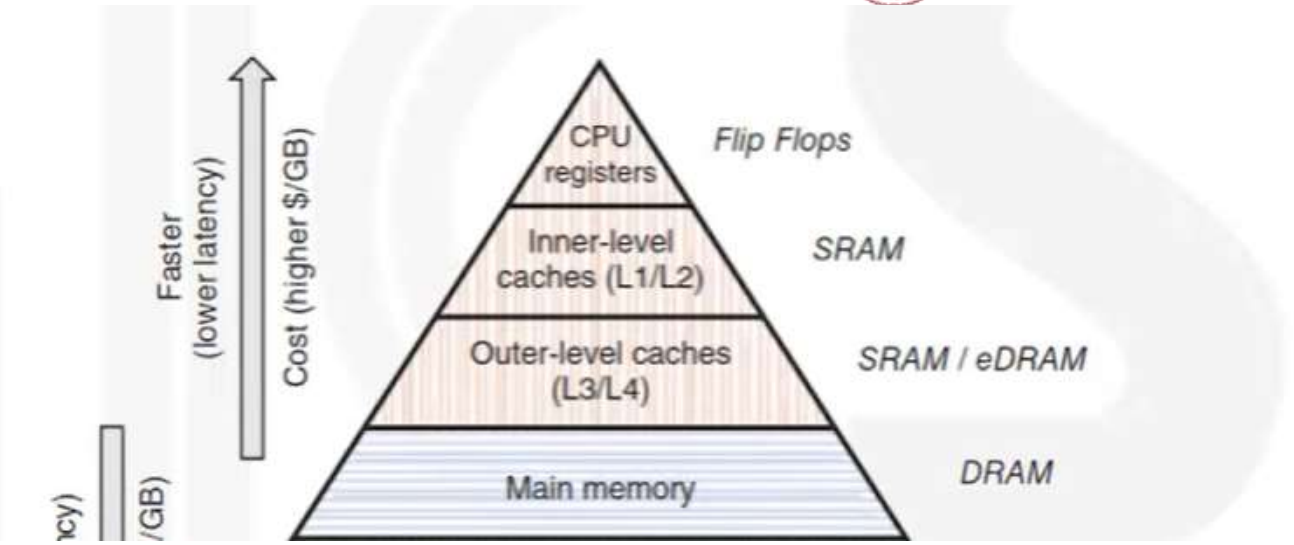


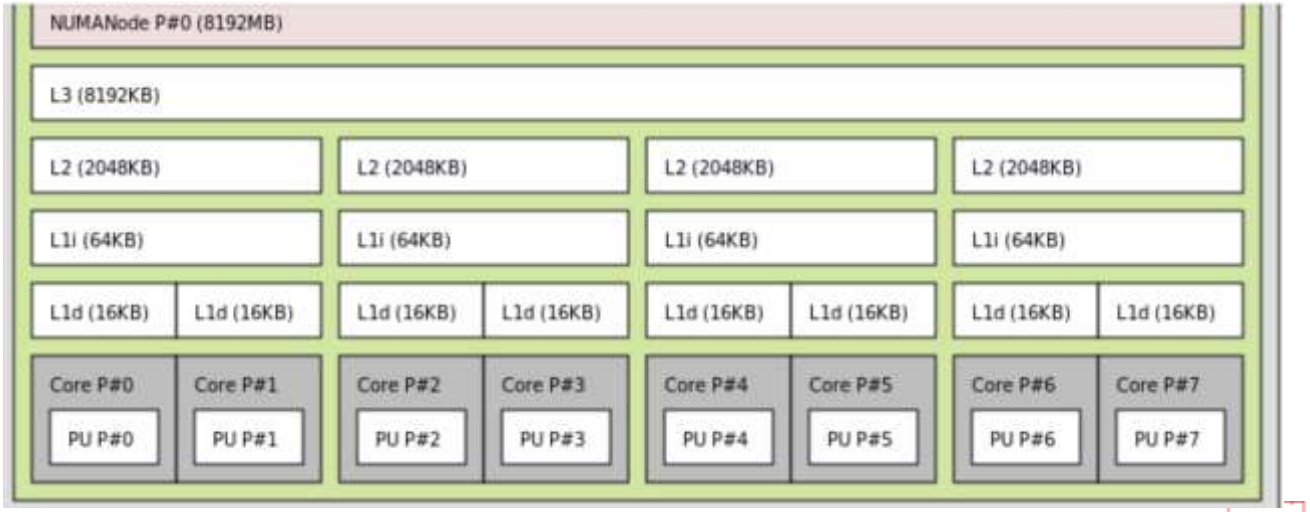
数据在这三级存储间传输时，后级的响应时间及传输带宽都将拖累整体的性能，形成“存储墙”。

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





2.内存层级：拓展的Memory层级

内存层级：

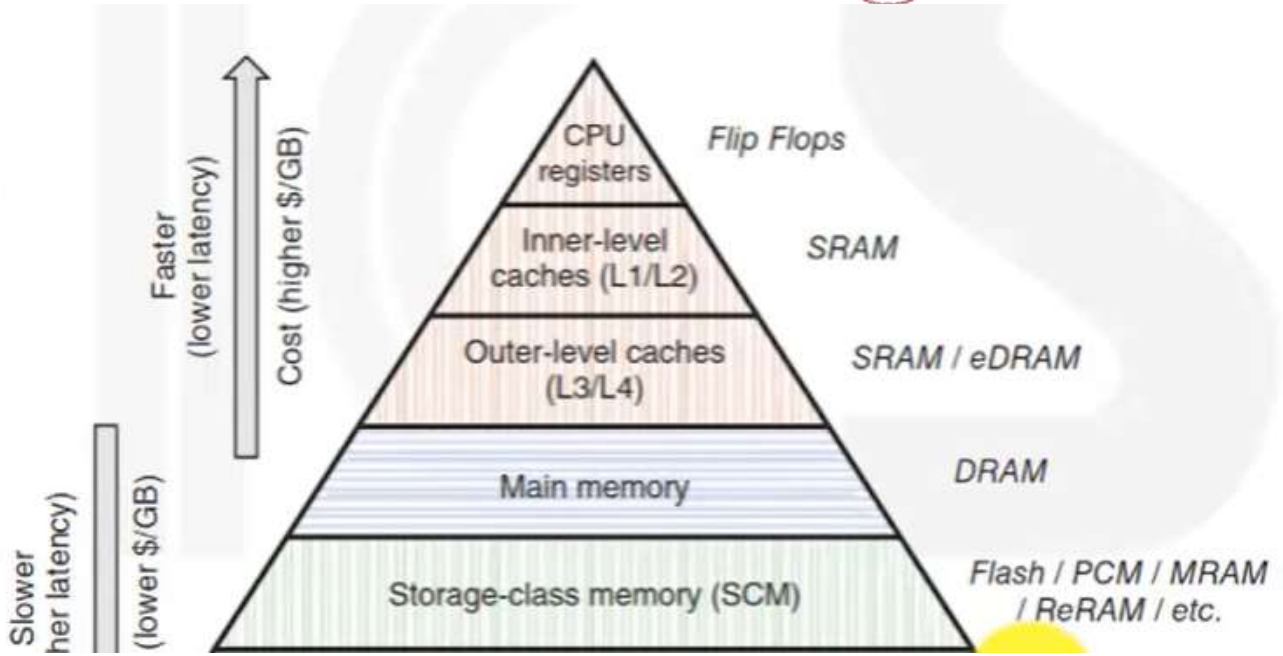
- 易失性存储层

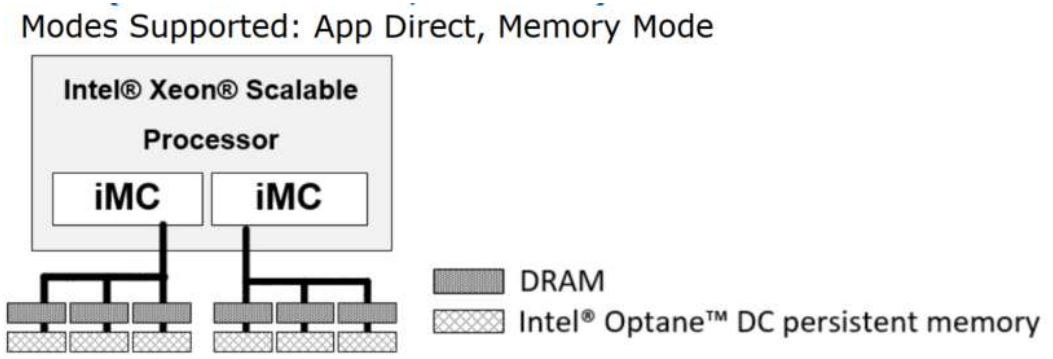
以DRAM为例，DDR4单条容量为16GB/32GB/64GB, 单条带宽约为30GB/s，响应时间在100ns级。

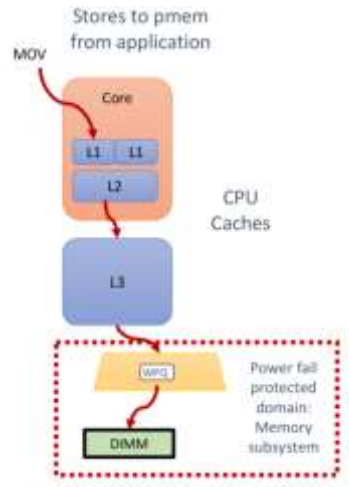
- 非易失存储层

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。







2.内存层级：拓展的storage层级

存储层级：

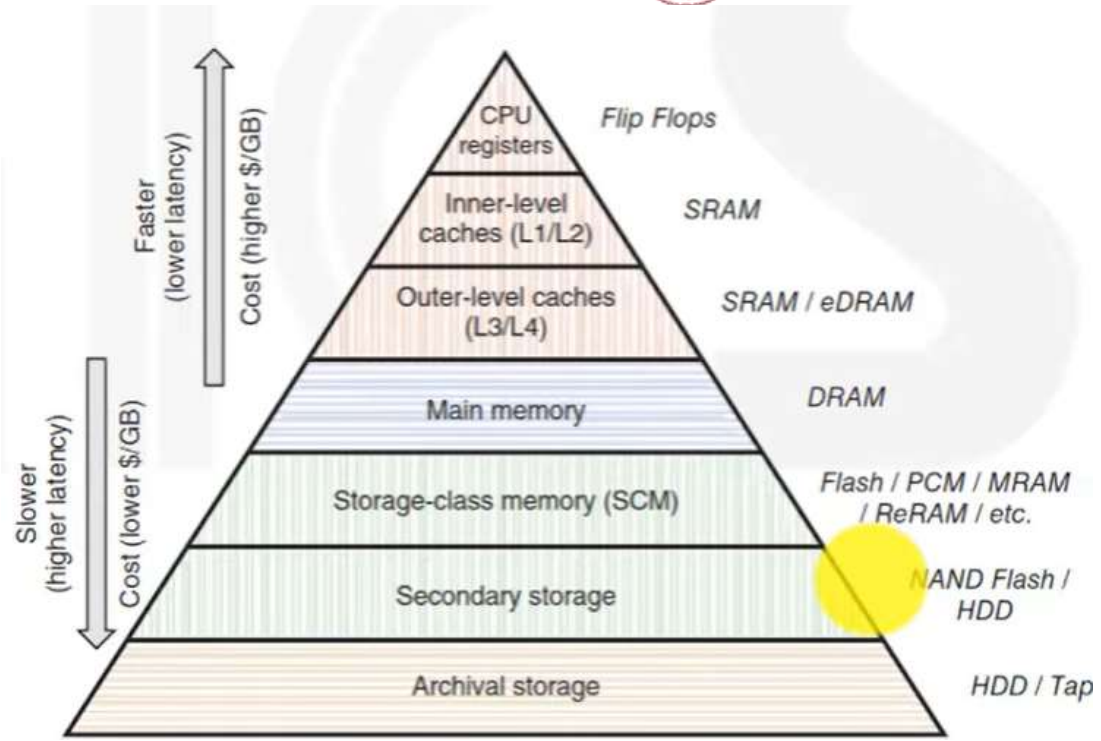
- 固态硬盘存储层（SDD）

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）

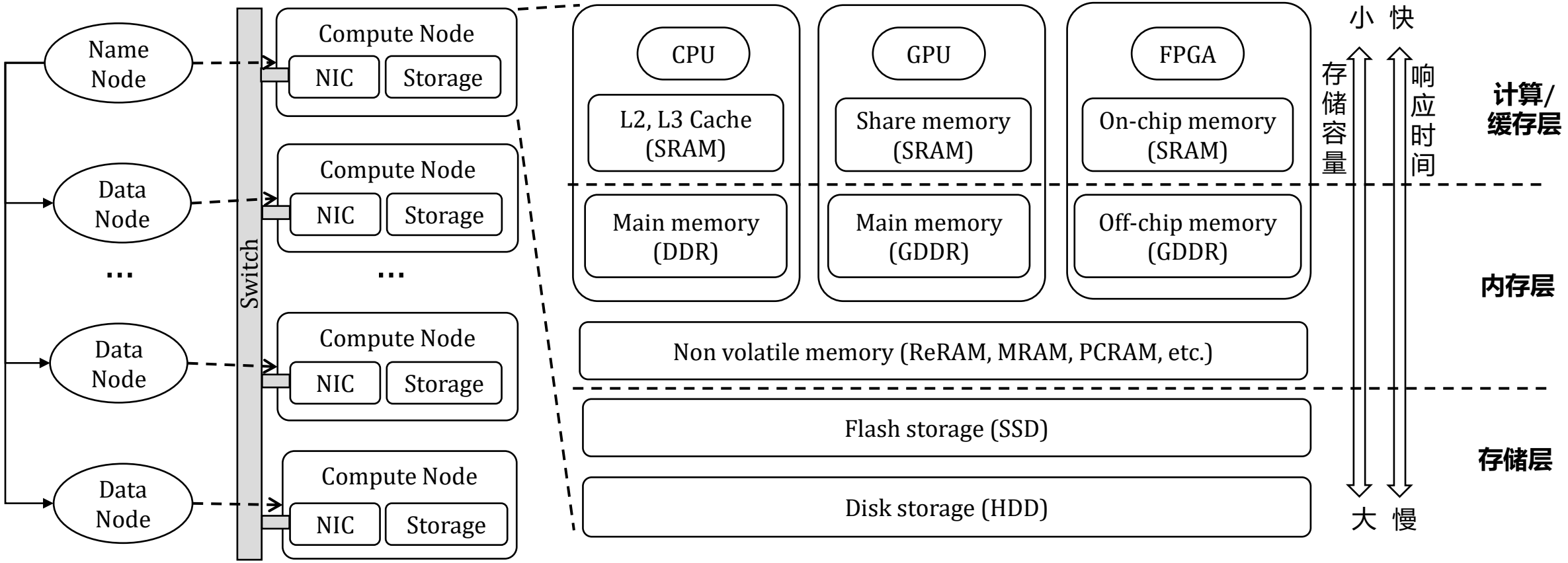
以HDD为例，单个存储容量高达18T，一般为SATA接口，带宽约为100~1000MB/s，读写时间在ms级。

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

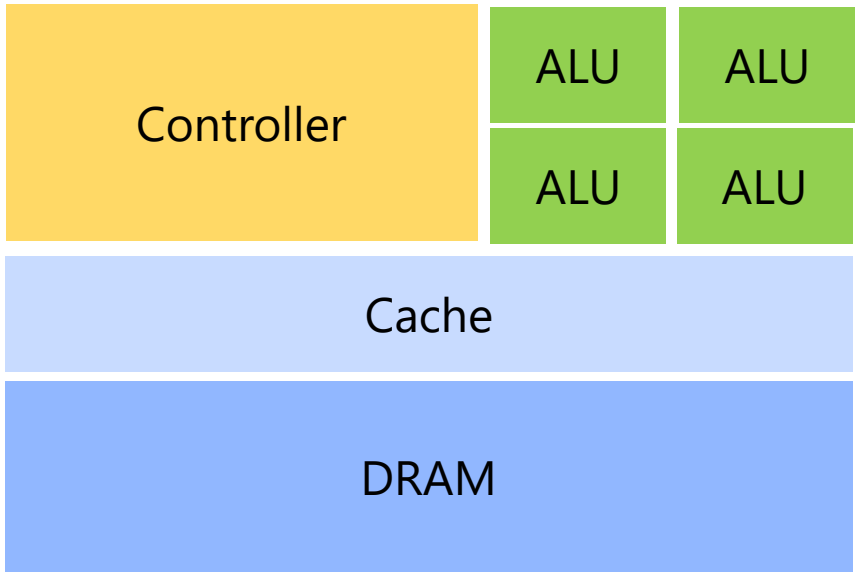
2.内存层级：拓展的存储层级总结



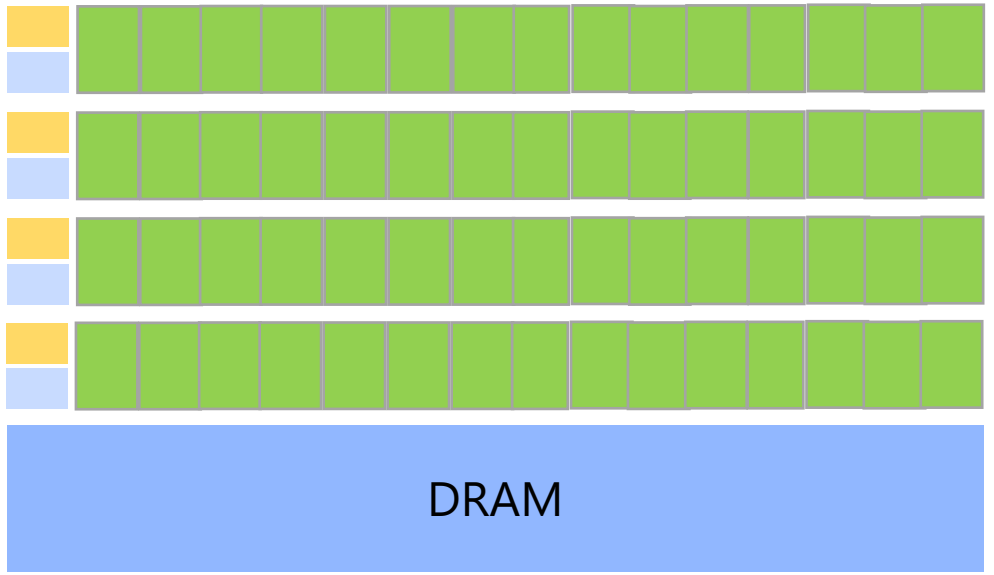
上海交通大学
SHANGHAI JIAO TONG UNIVERSITY

目录

3.统一内存访问：处理器与加速器整体结构



CPU



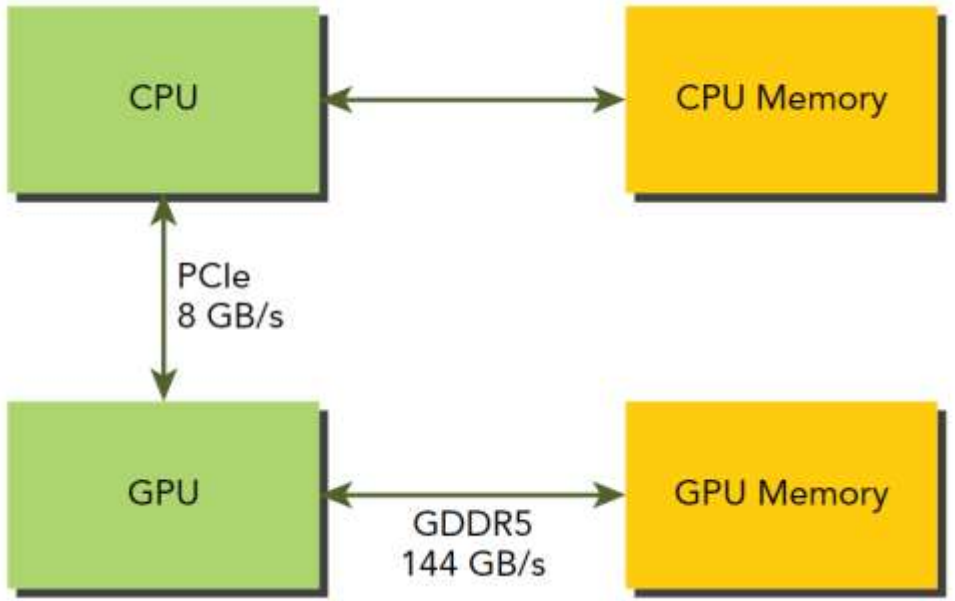
GPU

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

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

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

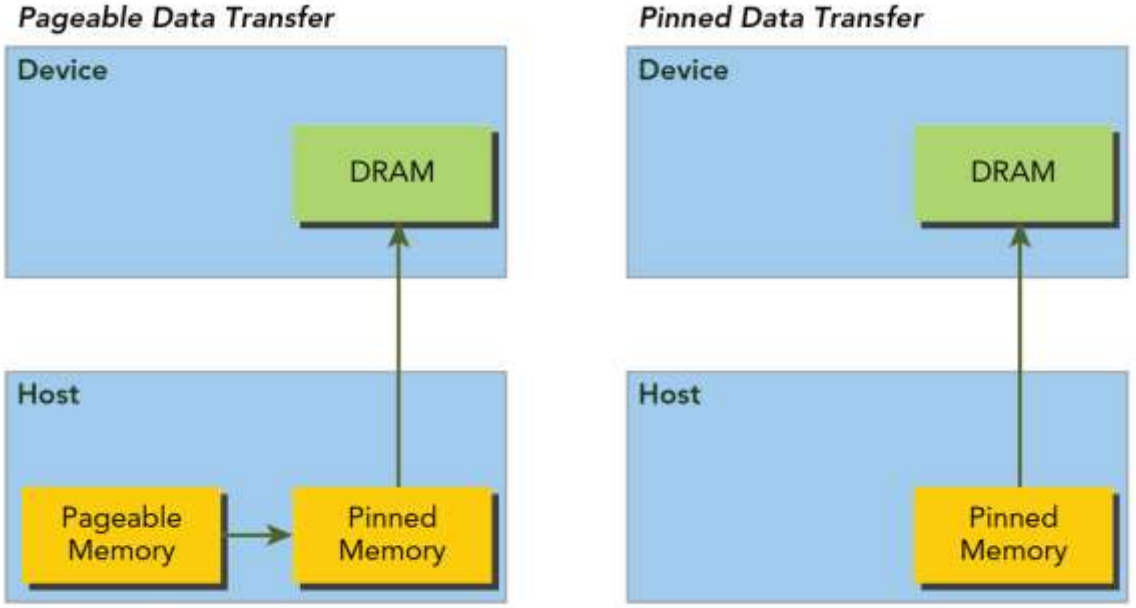


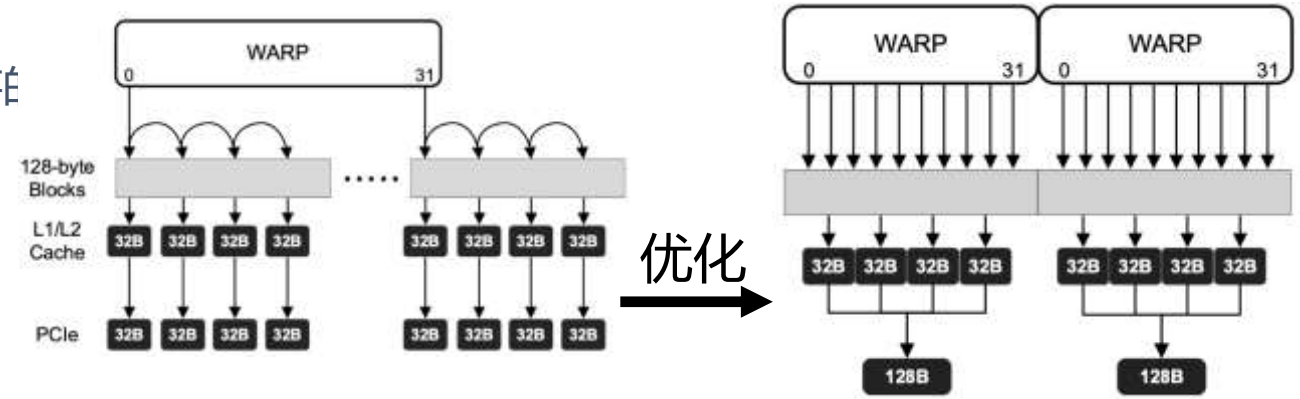
FIGURE 4-4

CUDA原语：cudaHostAlloc(参数)

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

3.统一内存访问：统一虚拟寻址UVA

上海交通大学 SHANGHAI JIAO TONG UNIVERSITY

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：
cudaMelloc()
cudaHostAlloc(参数)

参数
•cudaHostAllocMapped

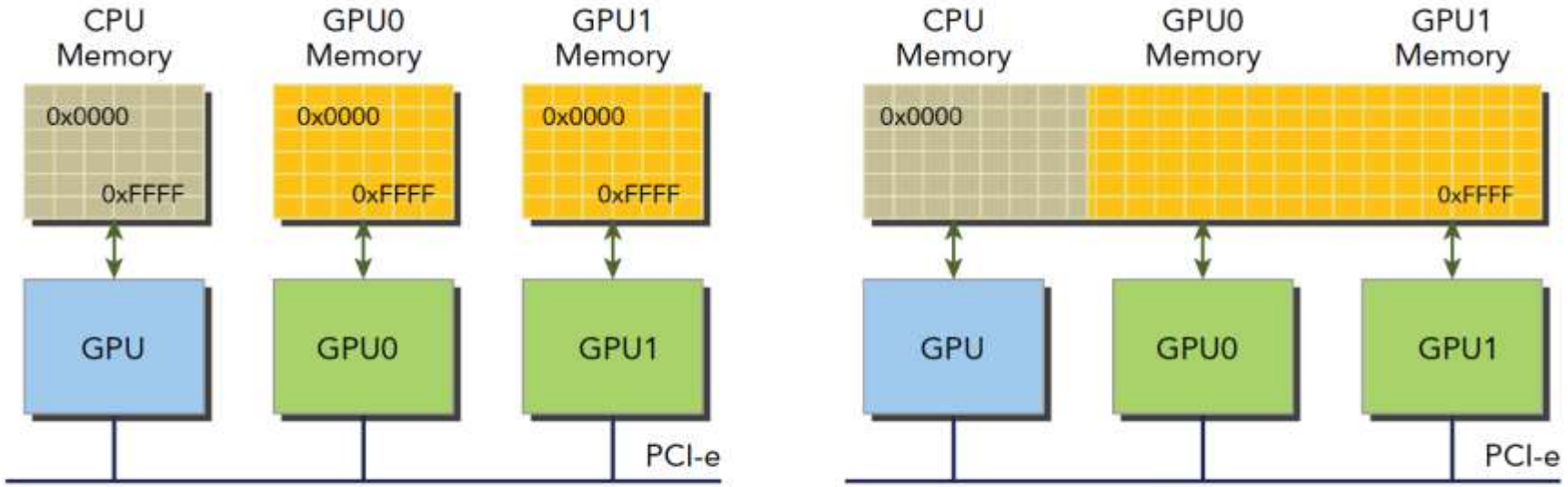


FIGURE 4-5

3.统一内存访问：统一内存寻址UM

统一内存访问机制：

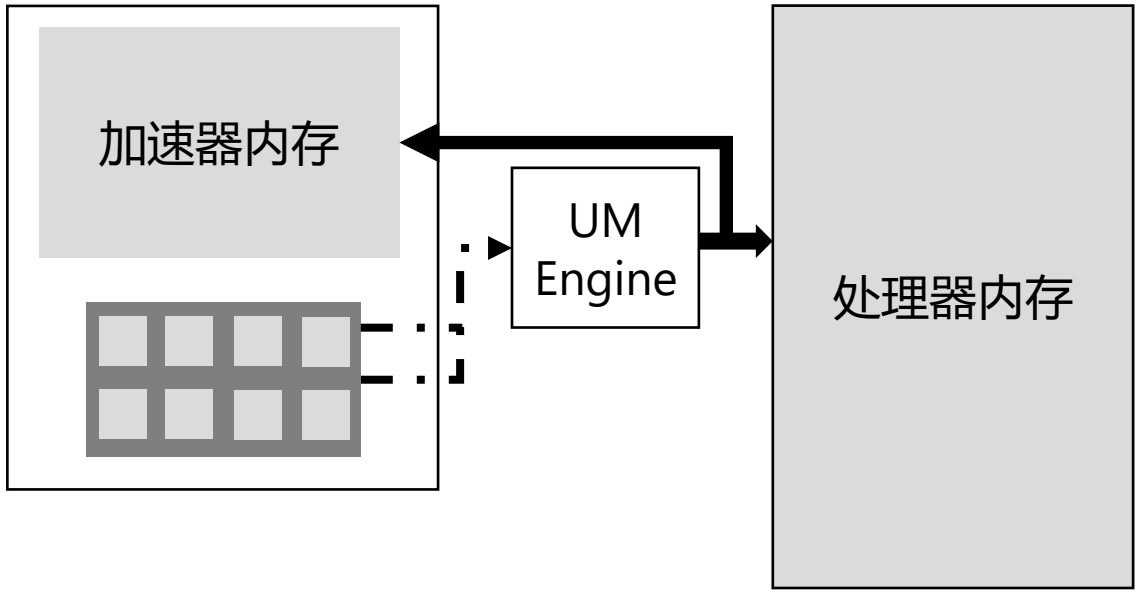
- CPU和GPU可以共同访问具有连贯内存视图的单一地址空间
- 依靠GPU驱动和硬件来自动进行数据传输与迁移
- 应用开发友好

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：
cudaMellocManaged()

上海交通大学
SHANGHAI JIAO TONG UNIVERSITY

目录

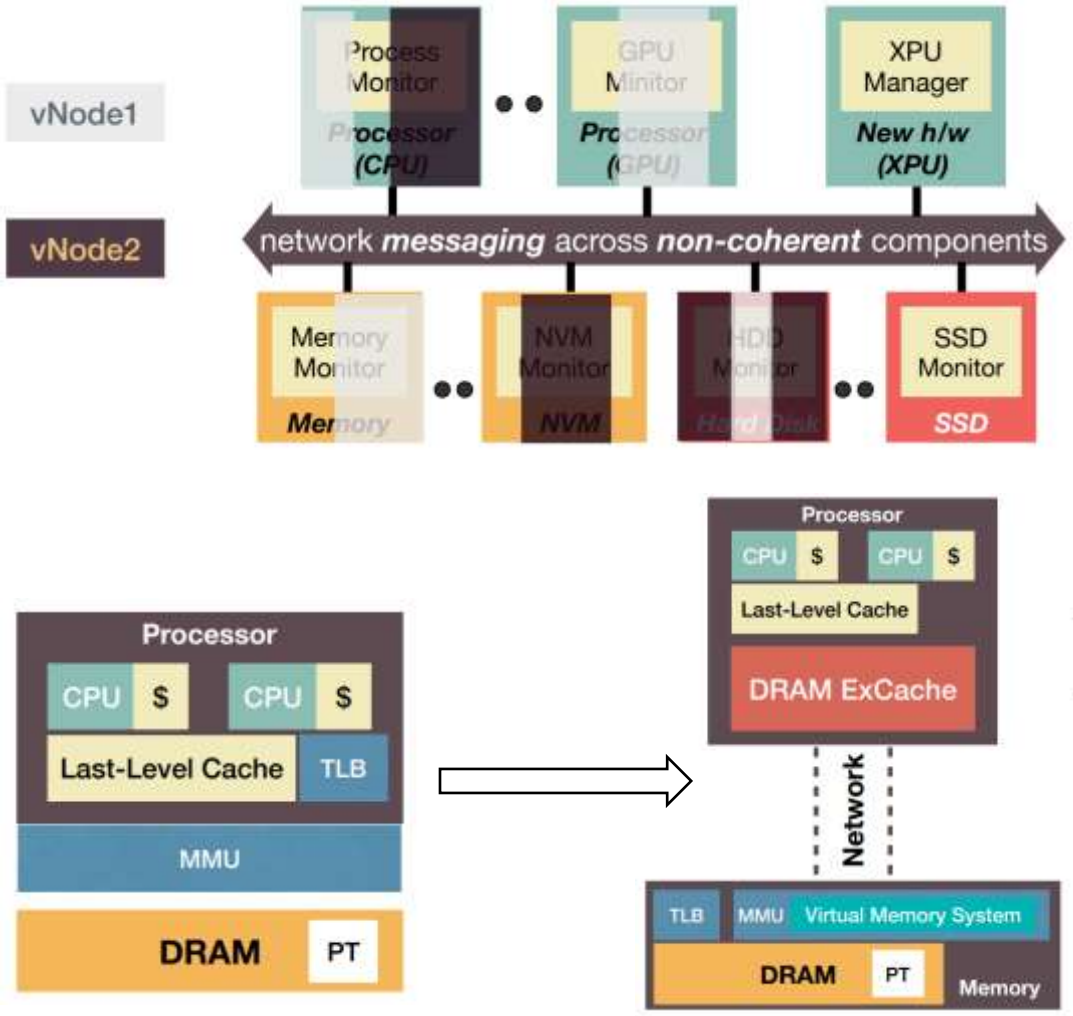
4. 存算分离架构：概述

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):

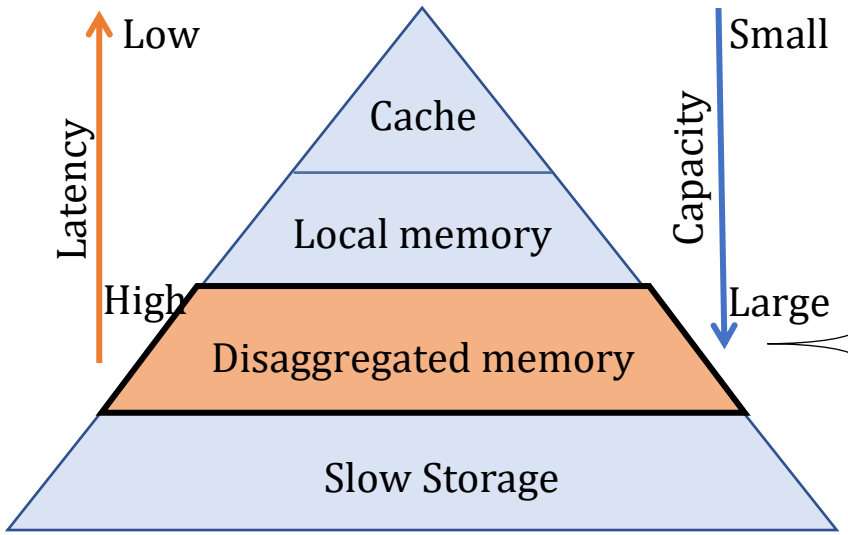
- 关注原因：应用面临内存大小和内存性能的瓶颈
- 理想方式：分离式大内存作为一个高效可用的部件，供传统计算单元使用，按需调取内存容量



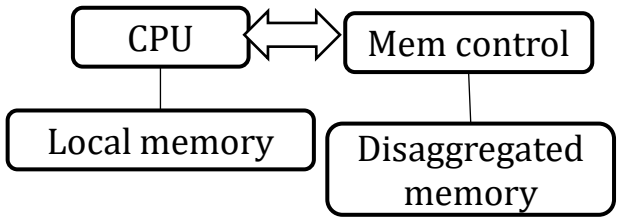
4. 存算分离架构：分离式内存和远内存

远内存 Far memory (FM):

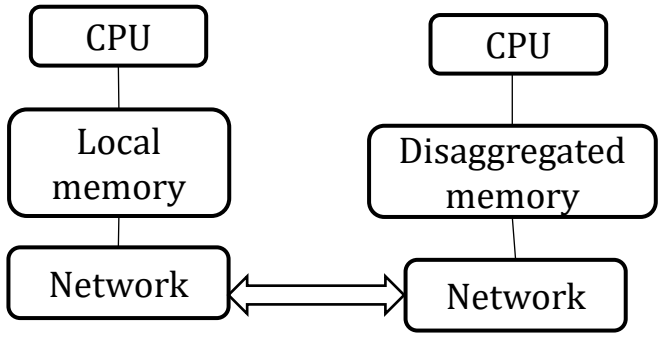
- 概念：更远的内存部件的访问方式
- 与DM关系：DM⊆FM
- 在一些论文中， DM代表着对专用DM部件的访问，而FM代表所有远于local memory 的内存系统的访问。
- Far memory 强调使用现有系统、硬件、架构实现远内存访问



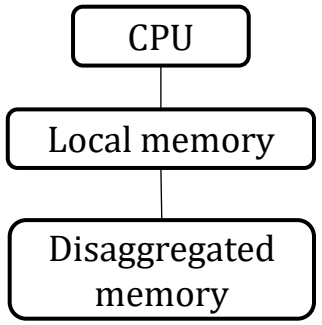
- Direct-connected far memory
- Fabric-attached far memory
- Storage-based far memory



Direct-connected far memory

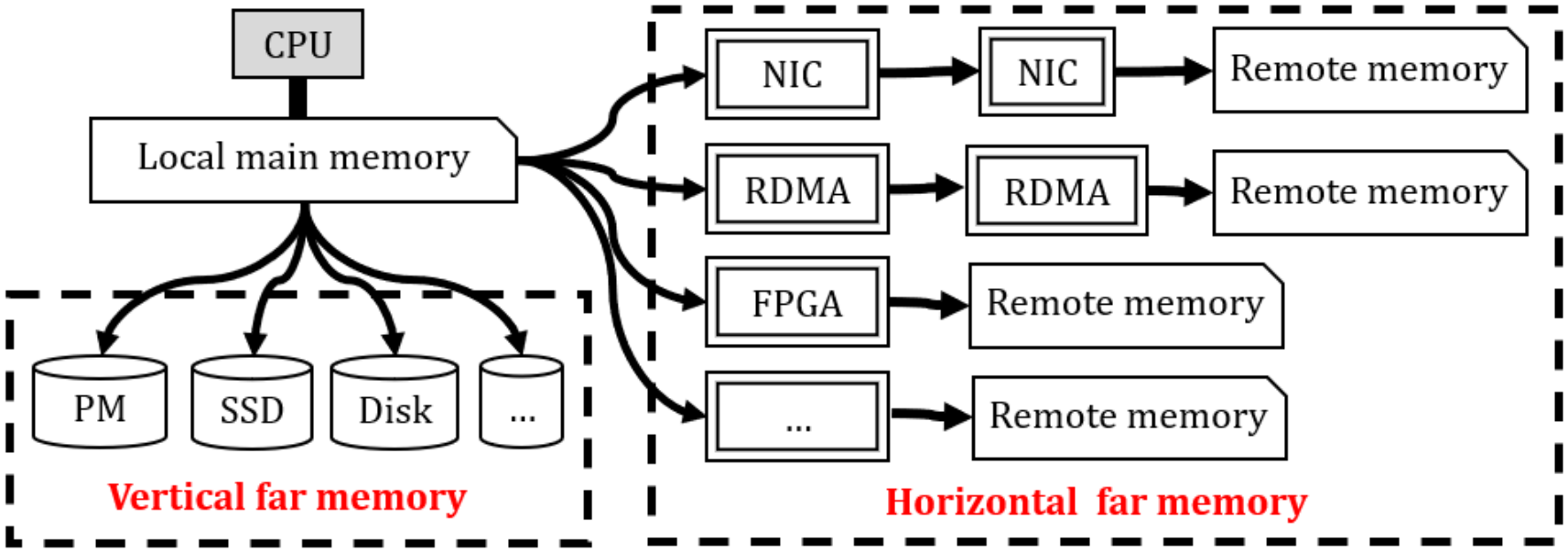


Fabric-attached far memory



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

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
  - 分离式内存的单独器件还在研发中
  - 分离式内存的架构目前还未有统一定论

远内存应用层

- 应用访问接口
- 数据格式
- 内存管理的运行时

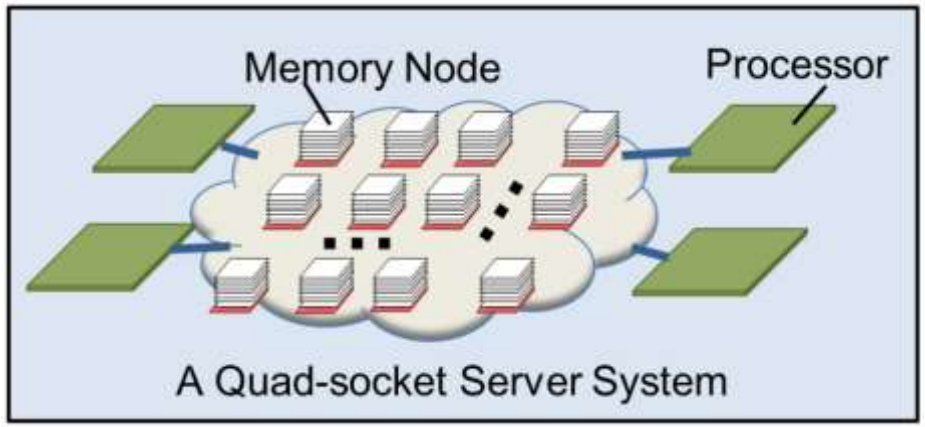
远内存系统层

- 内存管理的运行时
- 网络互联协议，CXL等
- 系统和驱动

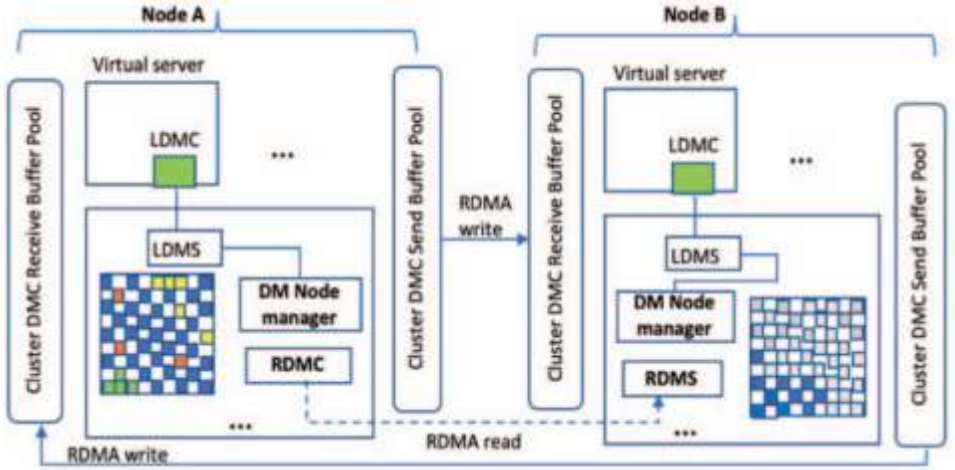
远内存硬件层

- 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

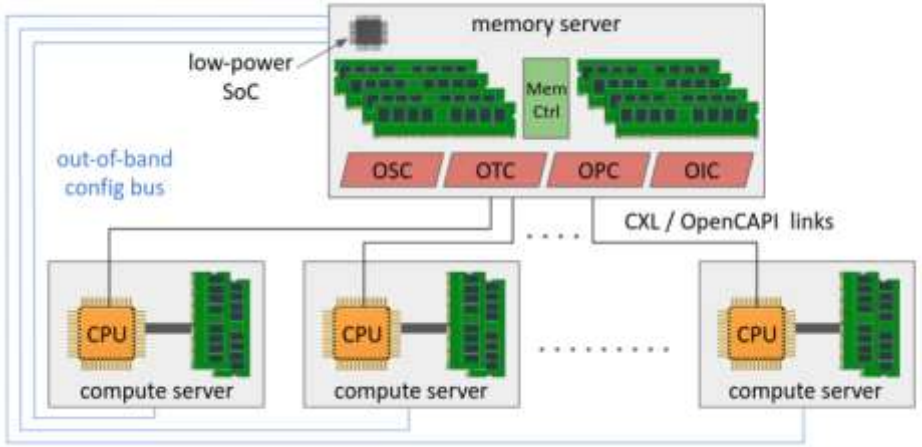
4. 存算分离架构：前沿架构



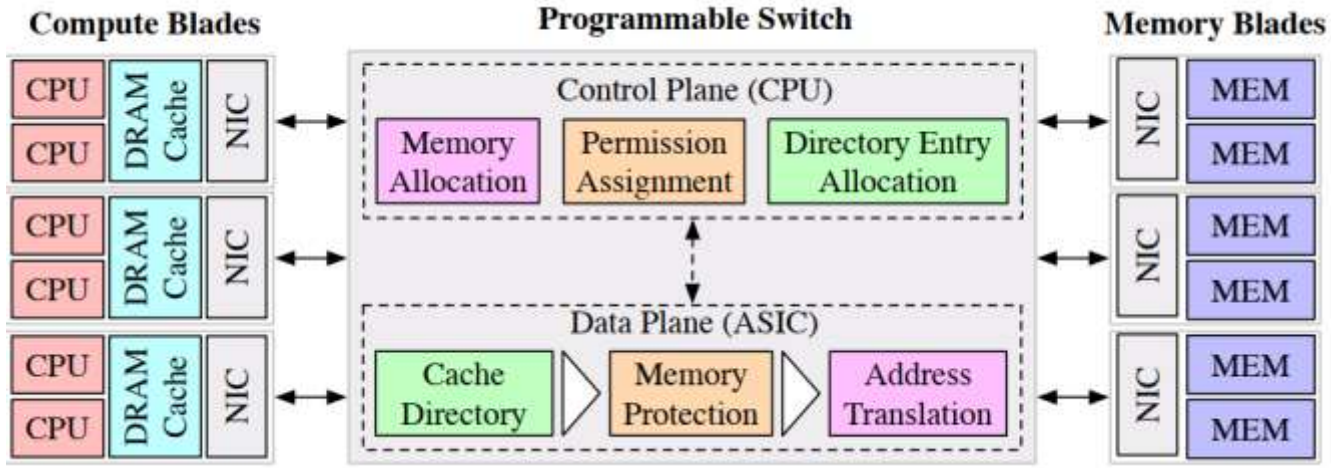
完全分离式架构



基于网卡的分离式架构



直连的分离式架构



基于可编程switch的分离式Rack

4. 存算分离架构：远内存硬件层-通用型

上海交通大学 SHANGHAI JIAO TONG UNIVERSITY

通用远内存硬件

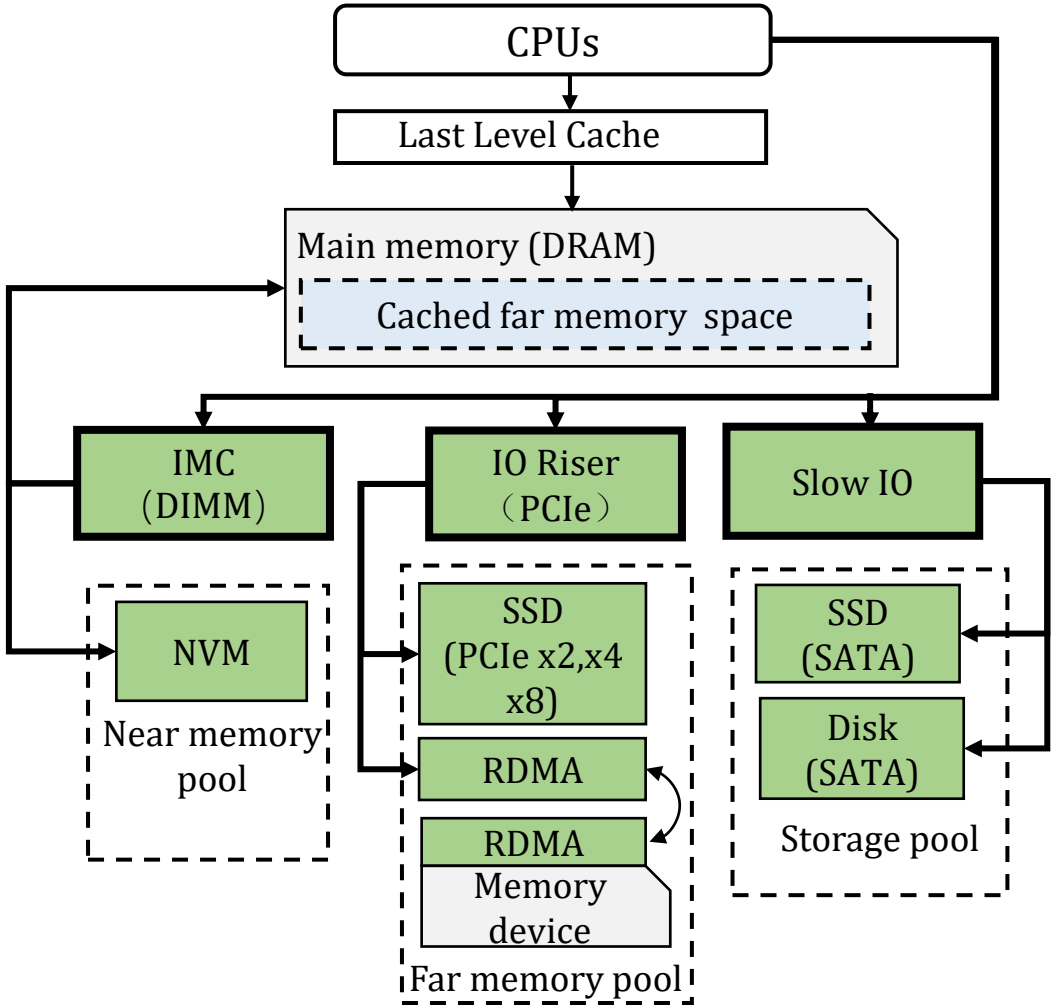
采用商业化的设备包括DRAM、SSD、RDMA等通用设备，在现有的服务器架构基础上组建远内存访问硬件层。

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



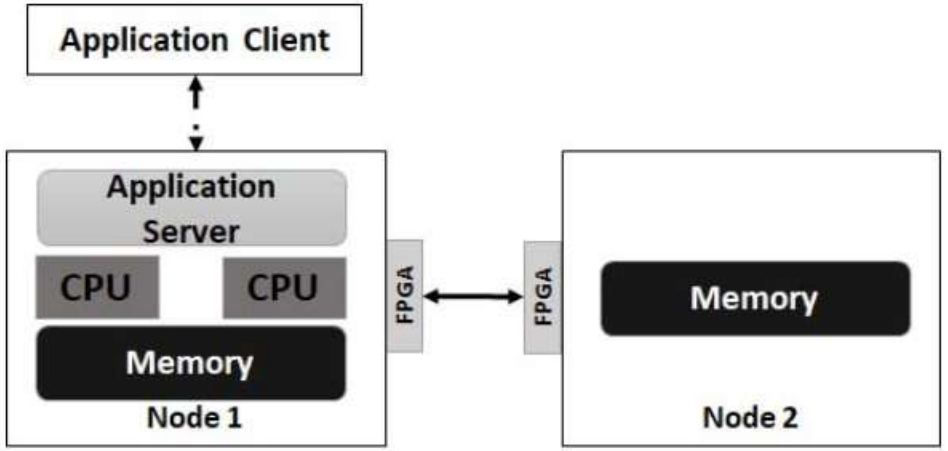
- 代表工作：Infiniswap， Fastswap，XMemPod，TMO，g-zswap，AIFM, pDPM，etc.

4. 存算分离架构：远内存硬件层-专用型

上海交通大学 SHANGHAI JIAO TONG UNIVERSITY

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。

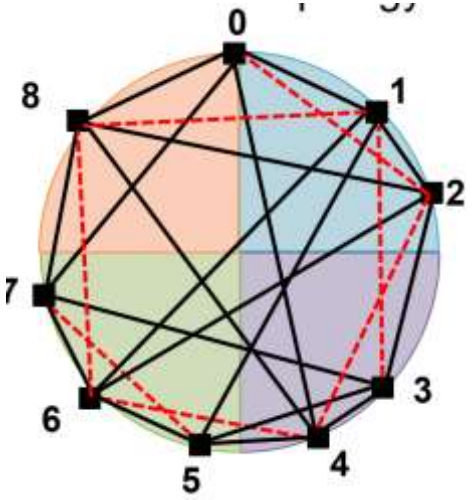


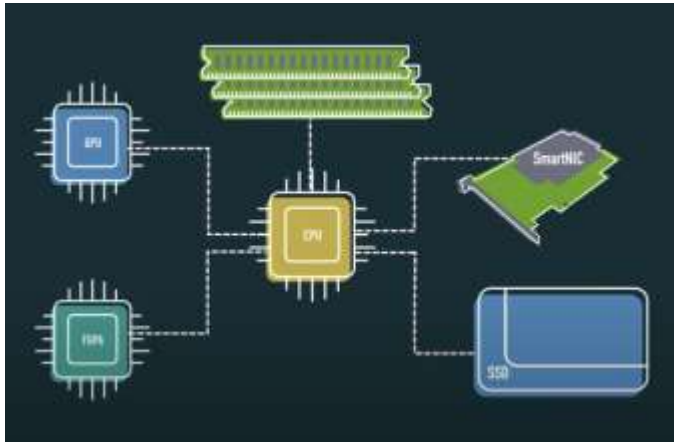
优势

- 性能好
- 创新性强
- 不断适应新的架构

设计思想

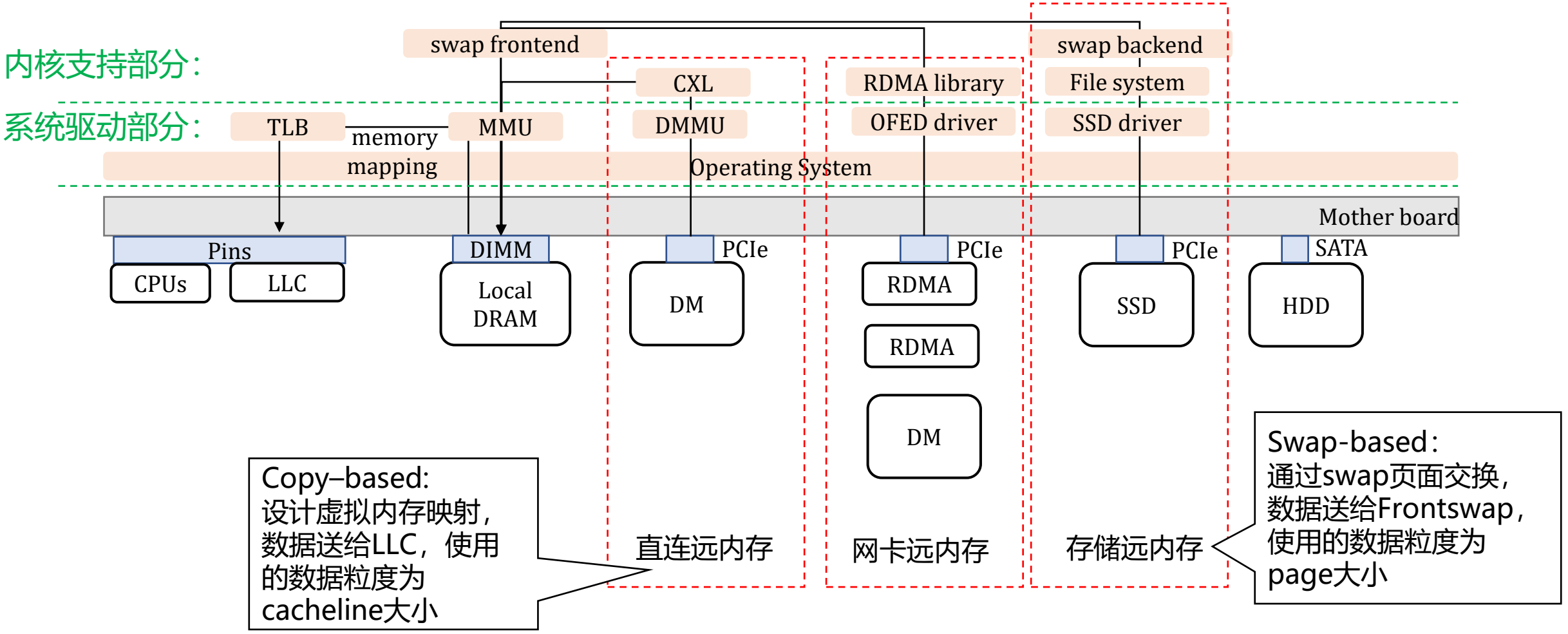
- 硬件逻辑设计
- 系统层的配合设计
- 应用透明或应用感知





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

4. 存算分离架构：远内存系统层



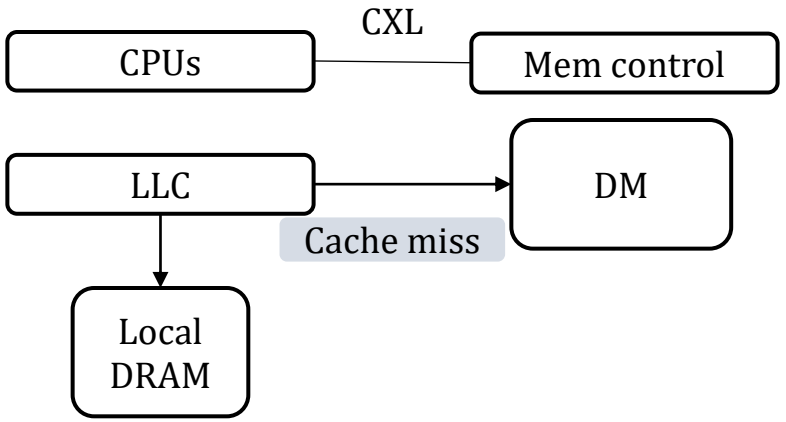
- 代表工作：Infiniswap, Fastswap，XMeMPod, g-zswap, Kona, CXL-based disaggregation, etc.

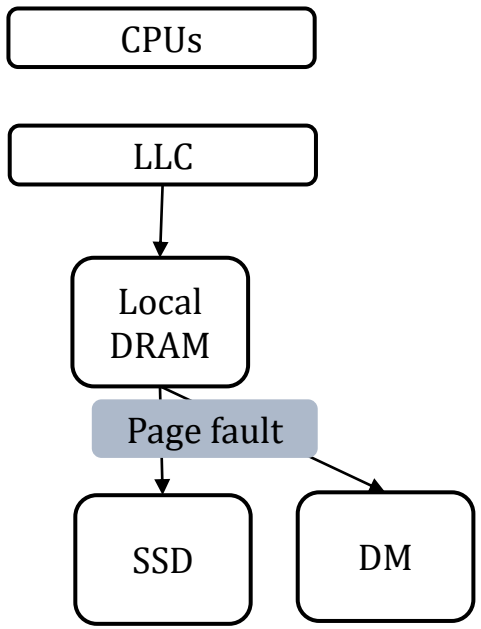
4. 存算分离架构：远内存应用层

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

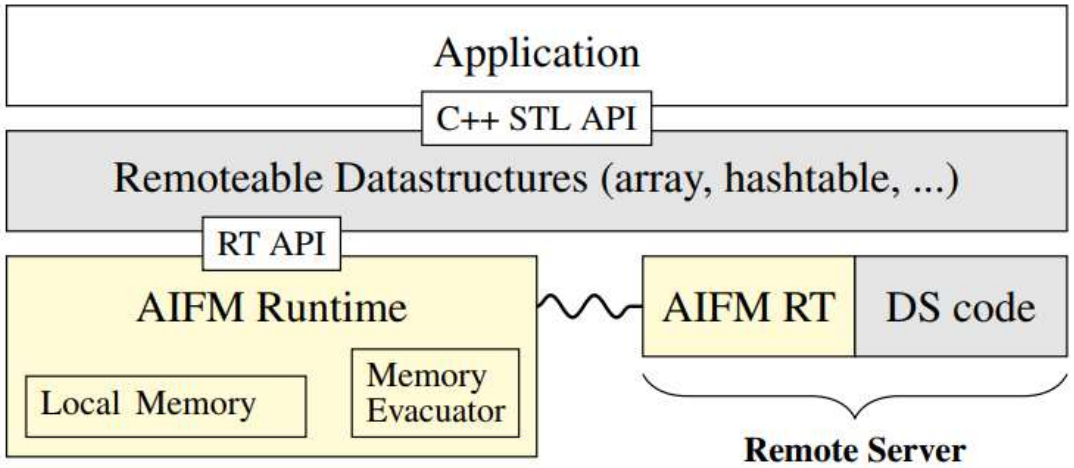




应用非透明的远内存

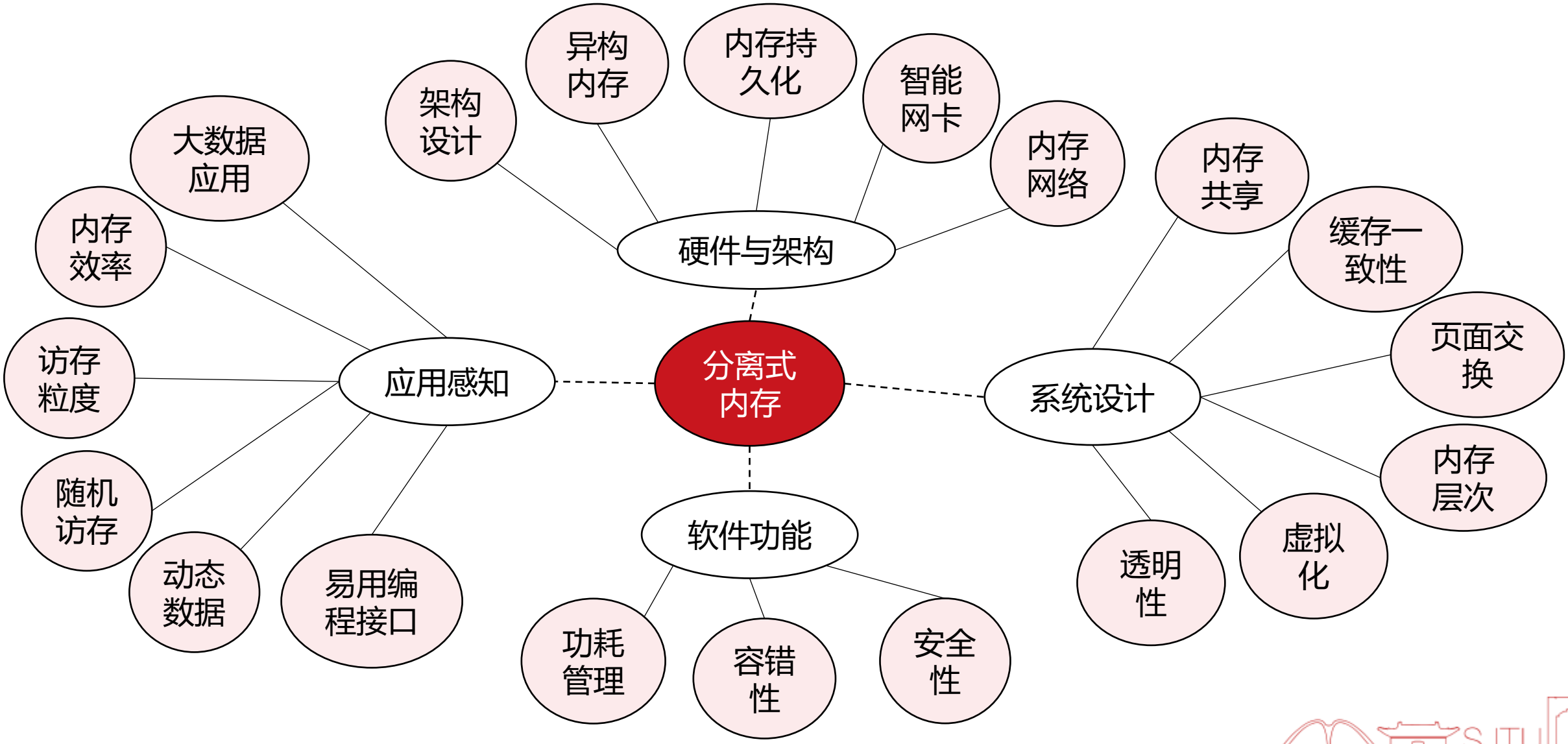
非透明远内存主要是通过在设计高效易用的编程接口：

- 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.

4. 存算分离架构：分离式内存相关研究



上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Thank You