第二十讲：分布式系统

第 4 节：LegoOS

向勇、陈渝、李国良

清华大学计算机系

xyong,yuchen,liguoliang@tsinghua.edu.cn

2021 年 5 月 10 日

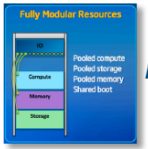
Ref:

- Paper: LegoOS
- Slides: LegoOS

分布式 I/O

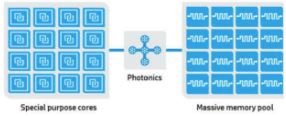
分布式 I/O 计算逐渐浮现在大型数据中心和移动终端领域。

- **Network is faster**
  - InfiniBand (*200Gbps, 600ns*)
  - Optical Fabric (*400Gbps, 100ns*)
- **More processing power at device**
  - SmartNIC, SmartSSD, PIM
- **Network interface closer to device**
  - Omni-Path, Innova-2

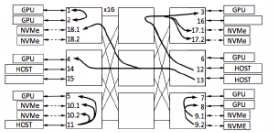


Intel Rack-Scale System

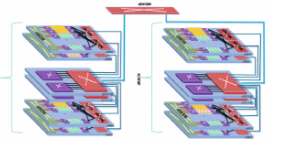
Berkeley Firebox



HP The Machine



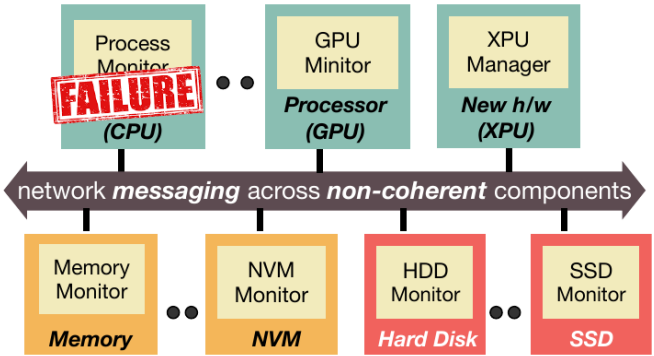
IBM Composable System



dReDBox

分布式 I/O

出现了新型的 splitkernel 架构



- Split OS functions into ***monitors***
- Run each monitor at h/w device
- Network messaging across non-coherent components
- Distributed resource mgmt and failure handling

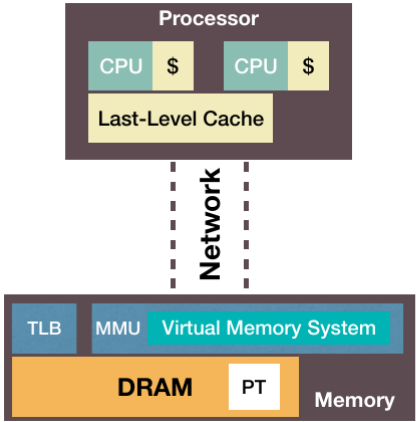
分布式 I/O

LegoOS：**把 CPU，Memory，Storage 分布在不同机器上，通过高速网络 RDMA 形成一个虚拟的大机器**
OS **抽象**：

- virtual Nodes (vNodes) –> hardware devices
- Similar semantics to virtual machines
- Unique vID, vIP, storage mount point
- Can run on multiple processor, memory, and storage components

分布式 I/O

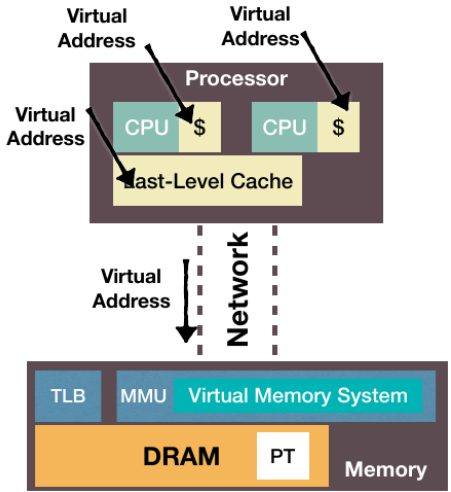
Separate Processor and Memory



Separate and move
virtual memory system
to memory component

分布式 I/O

Separate Processor and Memory



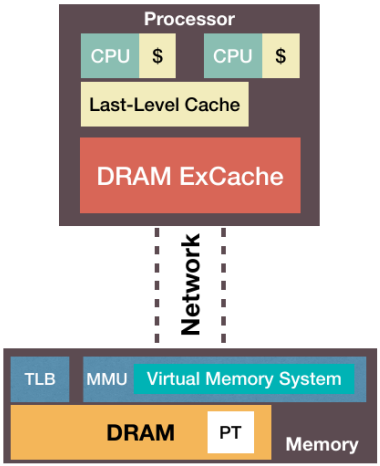
Processor components only see virtual memory addresses

All levels of cache are ***virtual cache***

Memory components manage virtual and physical memory

分布式 I/O

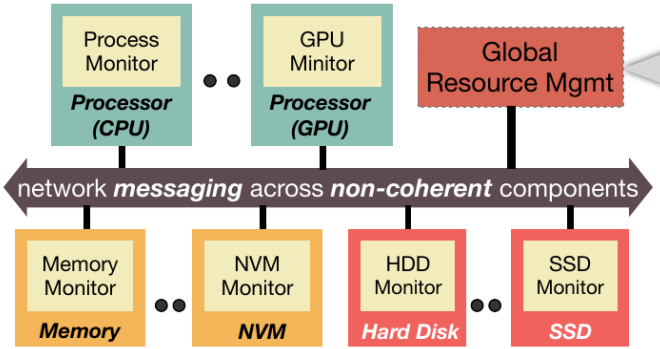
Add Extended Cache at Processor



- Add small DRAM/HBM at processor
- Use it as Extended Cache, or ***ExCache***
  - Software and hardware co-managed
  - Inclusive
  - Virtual cache

分布式 I/O

Distributed Resource Management



Global Process Manager (*GPM*)

Global Memory Manager (*GMM*)

Global Storage Manager (*GSM*)

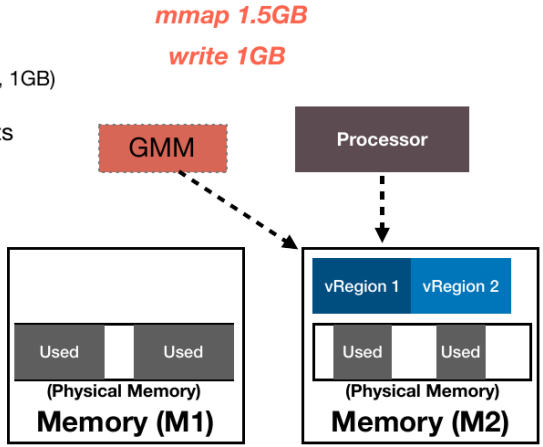
1. Coarse-grain allocation
2. Load-balancing
3. Failure handling

分布式 I/O

Distributed Memory Management

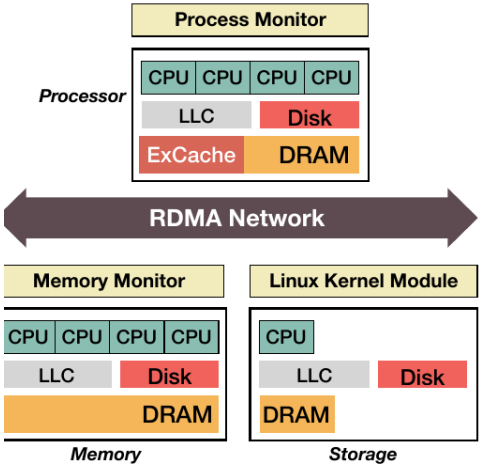
fix-sized, **coarse-grain** virtual region (***vRegion***) (e.g., 1GB)

- GMM assigns vRegions to mem components
  - On virtual mem alloc syscalls (e.g., `mmap`)
  - Make decisions based on global loads
- Owner of a vRegion
  - Fine-grained virtual memory allocation
  - **On-demand** physical memory allocation
  - Handle memory accesses



分布式 I/O

Implementation



- **Status**
  - 206K SLOC, runs on x86-64, **113** common Linux syscalls
- **Processor**
  - Reserve DRAM as ExCache (4KB page as cache line)
  - h/w only on hit path, s/w managed miss path
- **Memory**
  - Limit number of cores, kernel-space only
- **Storage/Global Resource Monitors**
  - Implemented as kernel modules on Linux
- **Network**
  - RDMA RPC stack based on LITE [*SOSP'17*]