CAICT 中国信通院　中国通信标准化协会 China Communications Standards Association

虚拟超级计算机探索与实践

唐宏伟
中国科学院计算技术研究所

2021 可信云大会
2021 TRUSTED CLOUD SUMMIT
数字裂变　可信发展

研究背景与动机

传统超级计算机的使用方式

- 运行时环境准备、程序编译
- 小规模试算
- 作业提交、排队、运行
- 运行中的问题排查：登录计算节点调试

研究背景与动机

存在的问题

- **应用部署效率低**：操作系统、编译器、通信库等可能都需要根据用户要求部署配置
- **用户使用不方便**：只能小规模试算、共享作业队列/计算节点对操作权限有限制
- **安全隔离性差**：不同用户的进程混跑在同一个节点、同一套网络，用户一定程度上可以访问其它用户的数据和作业
- **资源灵活性差**：物理分区、资源动态扩展性差

研究背景与动机

主要思路：把超级计算机虚拟化，给每个HPC用户一台虚拟独占的VSC

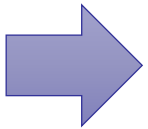
Virtual SuperComputer　Virtual SuperComputer　Virtual SuperComputer

研究背景与动机

主要思路：把超级计算机虚拟化，给每个HPC用户一台虚拟独占的VSC

传统方式

- 给用户分配登录账号
- 帮助用户部署和配置所需环境
- 用户上传数据
- 用户编译、试运行程序
- 提交作业、排队、运行
- 用户下载数据



VSC方式

- 用户选择资源规模、操作系统、运行时环境、作业系统等
- 通过软件定义方式根据用户的选择自动生成VSC
- 用户上传数据、运行作业
- 用户下载数据
- VSC自动销毁

主要技术挑战

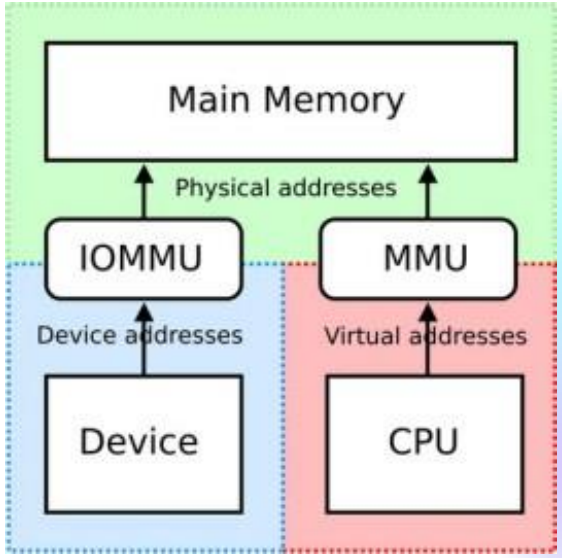
资源虚拟化引入了以下两个问题：

- **性能开销：** CPU根模式/非根模式上下文切换、I/O指令陷入模拟、内存虚拟化…
- **性能稳定性：**资源在VM之间共享和复用，VM相互之间会产生不确定性影响，比如cache争用、NUMA远程访存

主要技术挑战（Cont...）

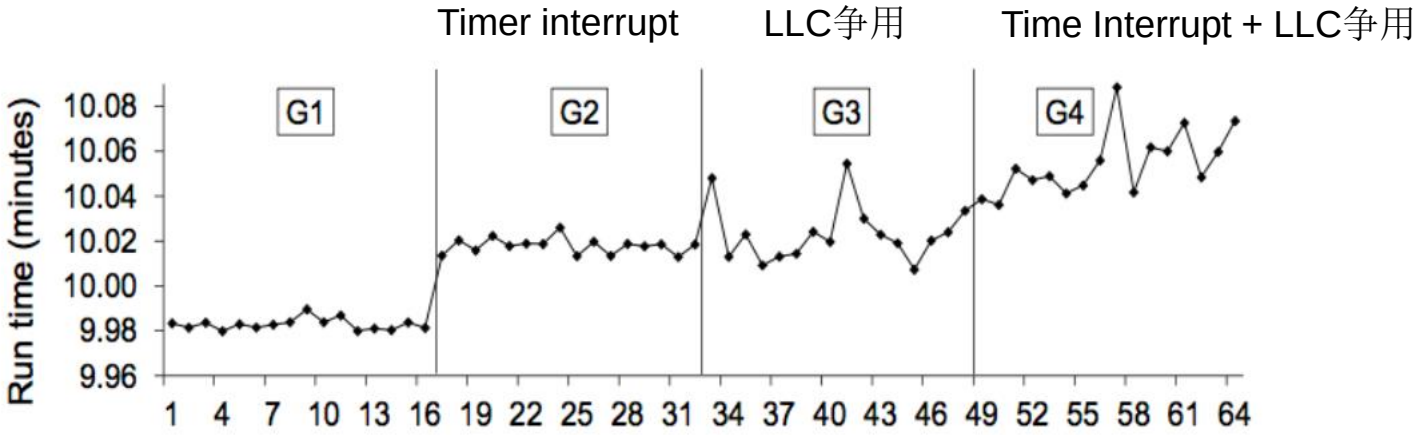
硬件直通是提高I/O虚拟化性能的有效手段，但是：

- **失去了资源弹性分配能力**
  - 直通硬件设备只能给特定虚拟机使用
  - 虚拟机内存需要在启动时全部分配到位：例如，128GB内存分配时间达42分钟



研究内容与关键技术

- KVM虚拟机性能优化
  - OS降噪、提高应用进程优先级

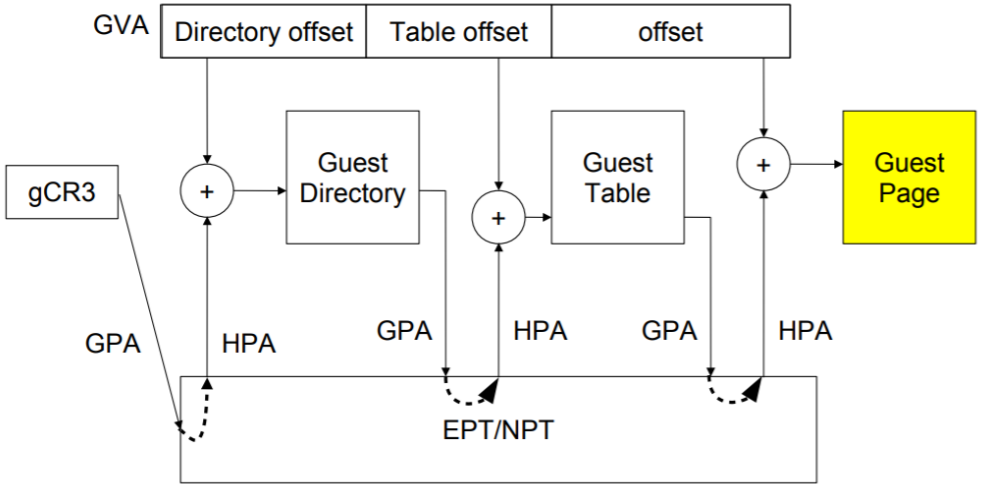


https://leogomes.github.io/engineering/2017/07/09/linux-os-jitter/

OS 噪声（OS Jitter）：操作系统因调度后台守护进程、处理异步事件（如中断）等对应用程序的运行造成的干扰，对大规模并 行程序的性能影响较为显著

研究内容与关键技术

- KVM虚拟机性能优化
  - 宿主机和客户机同时应用Huge Page (2MB)

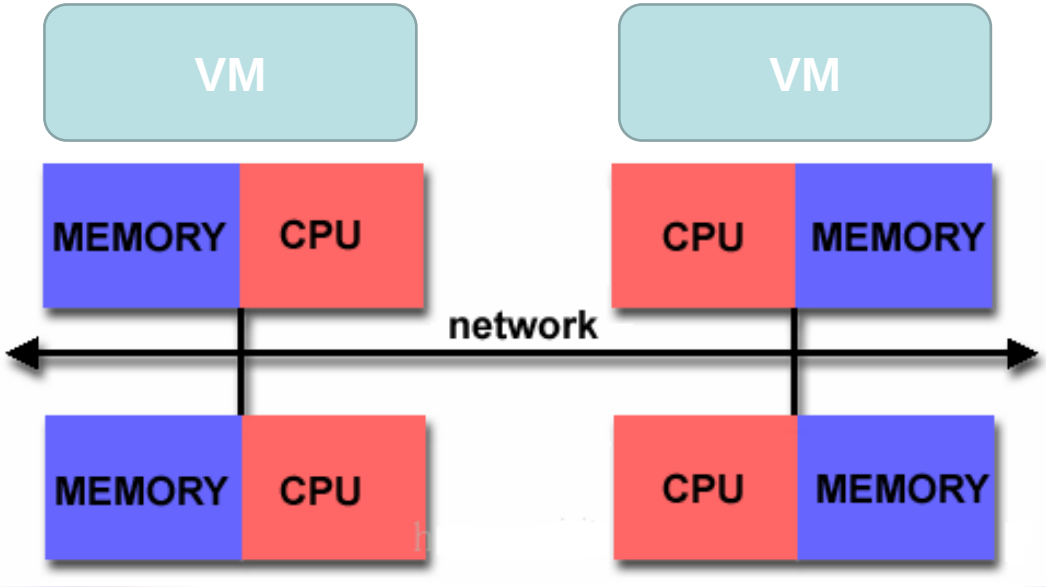


客户机虚拟地址翻译过程

PageSize Conf	G-nhp && H-nhp	G-nhp && H-hp	G-hp && H-hp
TLB Miss Cost	4 guest levels * 5 EPT access + 4 EPT access for final gpa->hpa translation = 24 memory access	4 guest levels * 4 EPT access + 3 access final gpa->hpa = 19 memory access	3 guest levels * 4 EPT access + 3 access final gpa->hpa = 15 memory access

研究内容与关键技术

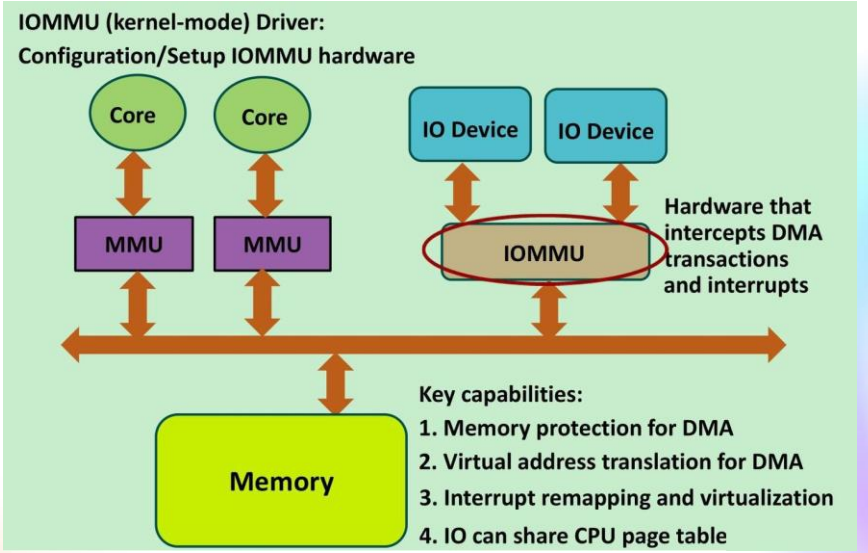
- KVM虚拟机性能优化
  - VCPU亲和性调度
  - NUMA资源调度优化：VCPU和内存协同分配



跨NUMA节点的访存开销较节点内访存开销高1-2个数量级，KVM虚拟化的VCPU和内存分配策略并不考虑这一因素

研究内容与关键技术

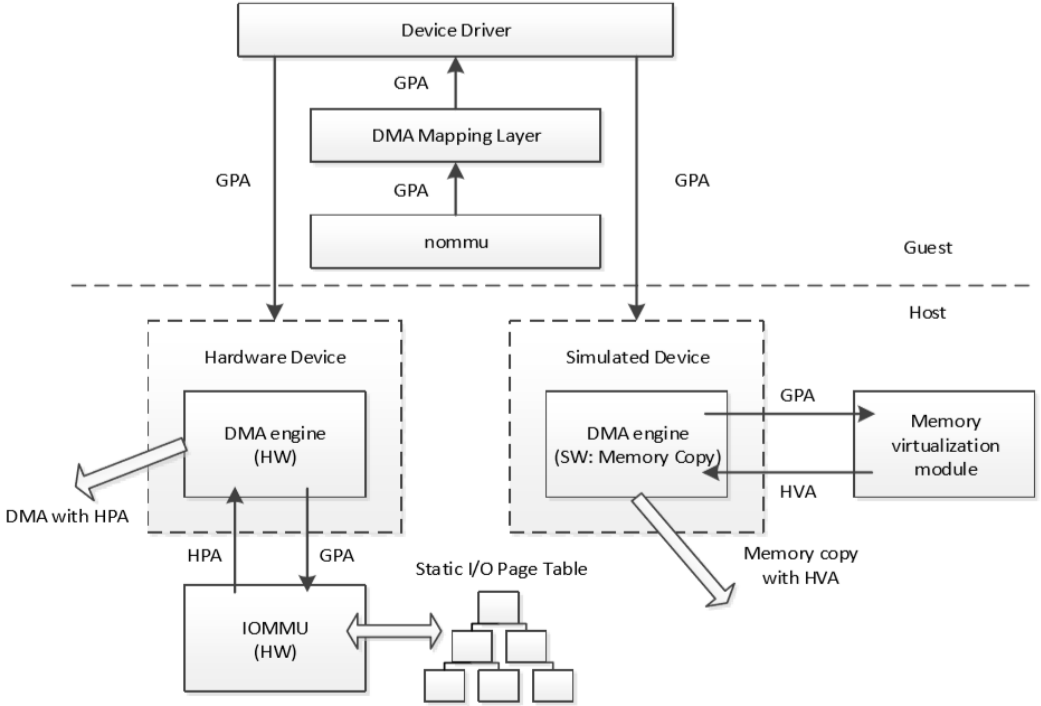
- **IOMMU半虚拟化——解决IB卡PassThrough内存弹性分配问题**
  - KVM虚拟机中没有IOMMU虚拟设备，设备I/O地址空间和系统内存地址空间相同
  - 利用EPT/NPT页表进行DMA内存地址翻译
  - DMA访存的前提是已经为将要访问的虚拟机内存已经分配了机器内存
  - 虚拟机启动过程中需要预先分配所有的机器内存，导致失去了弹性，启动缓慢
  - 缺少地址空间的隔离和访问控制，存在DMA安全风险



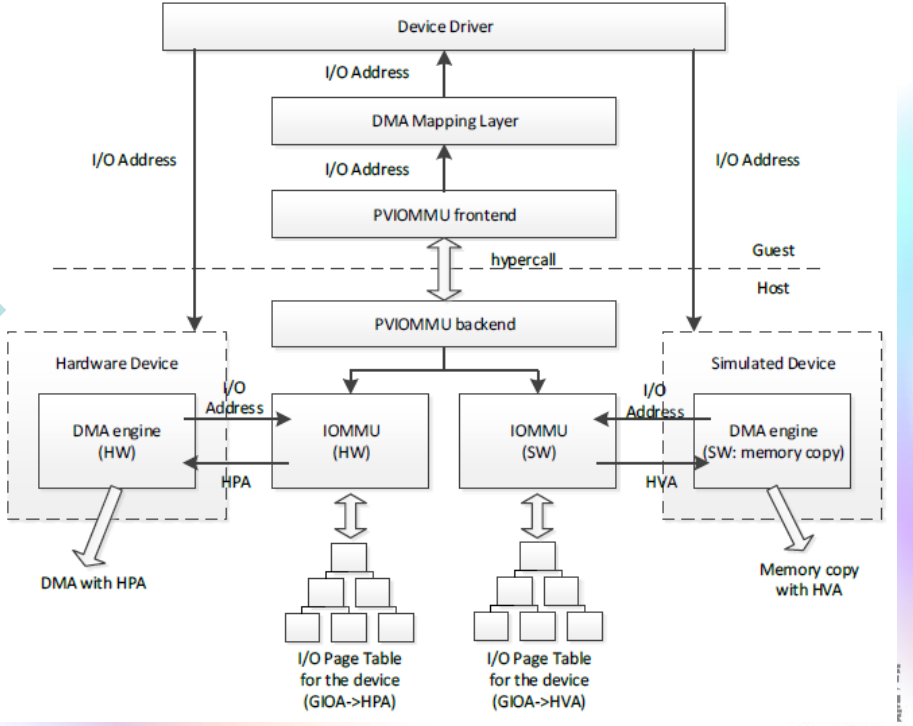
研究内容与关键技术

由Guest设备驱动触发动态建立DMA映射

- IOMMU半虚拟化——解决IB卡PassThrough内存弹性分配问题



没有虚拟IOMMU



IOMMU半虚拟化

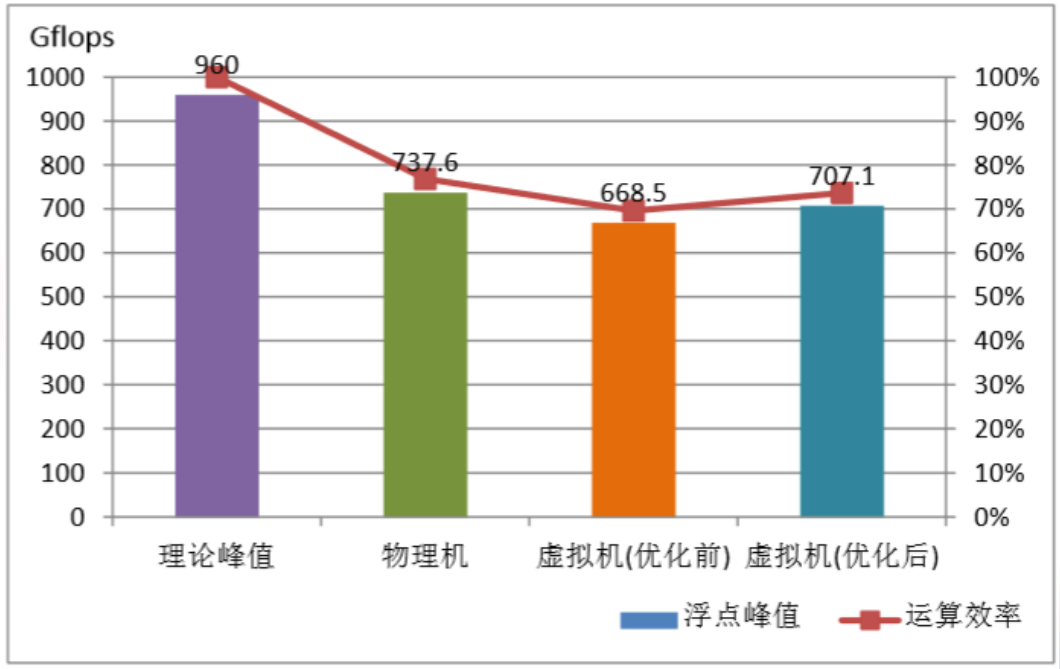
性能评测

- 虚拟超算性能测试（24VCPU/120GB）
  - 测试地：上海超算中心机房
  - 测试平台：曙光TC4600刀片服务器

软件环境	说明
Linpack版本	hpl-2.1
编译器	Intel composer_xe_2013_sp1.0.080
操作系统	CentOS 6.6
内核	Linux 2.6.32

硬件环境	说明
CPU	Intel E5-2680 V3 (12核) * 2
内存	8×16GB
硬盘	10Krpm SAS
网络	InfiniBand/Mellanox FDR 56Gbps

性能评测



Linpack效率

24VCPU/120GB单节点Linpack效率

宿主机： Linpack效率76.83%

虚拟机优化前： Linpack效率69.64%

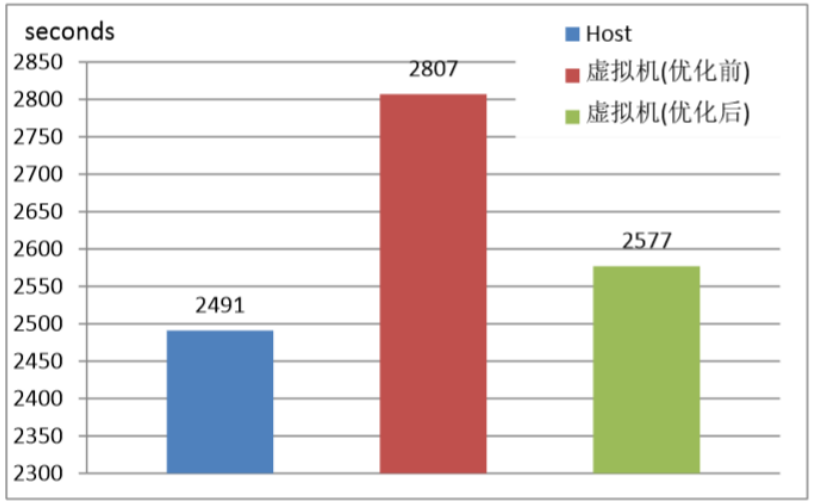
虚拟机优化后： Linpack效率73.66%

性能评测

24VCPU/120GB单节点、两节点应用性能

ABAQUS

(算例内存规模为100GB)

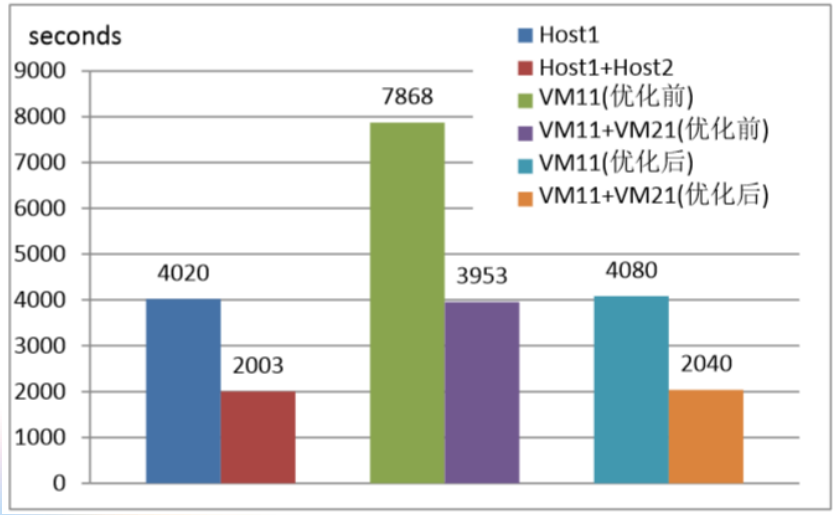


优化前： VM性能相比Host下降13%

优化后： VM性能相比Host下降3%

FLUENT

(算例模型为飞机外流场规模为1400万网格)

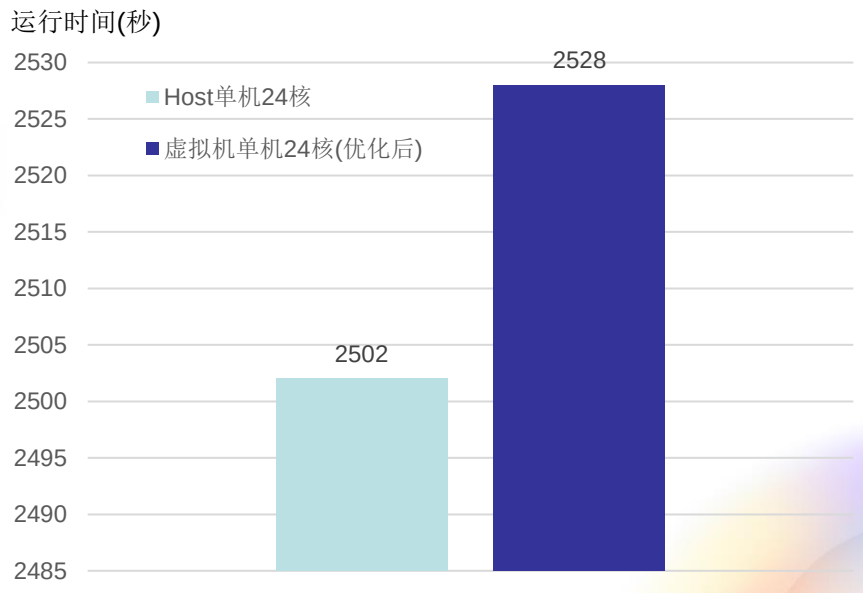


相对于物理机而言，优化前虚拟机下FLUENT运行时间增加近一倍， 优化后运行时间与宿主机相当

性能评测

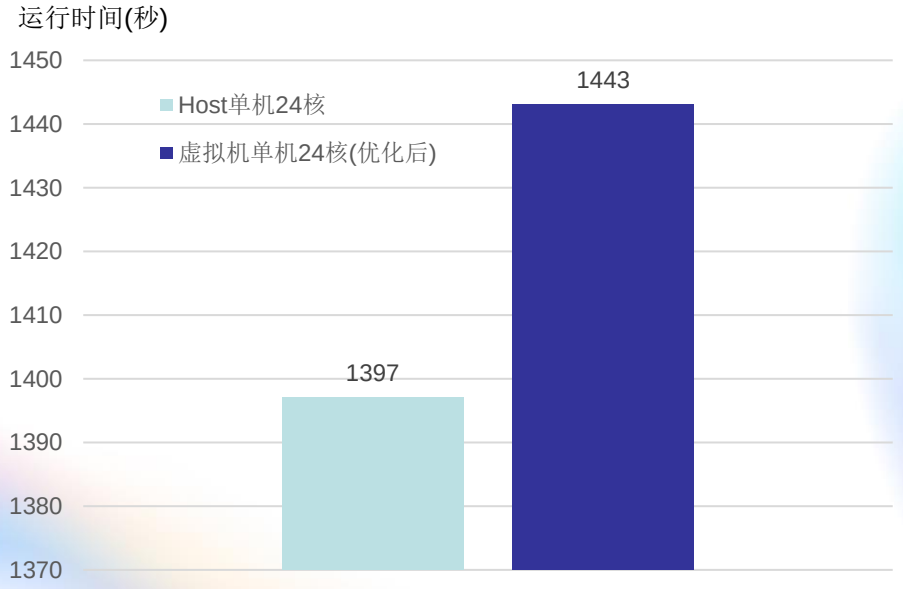
24VCPU/120GB单节点应用性能

Dyna



在虚拟机中运行时间相比于宿主机增加1%

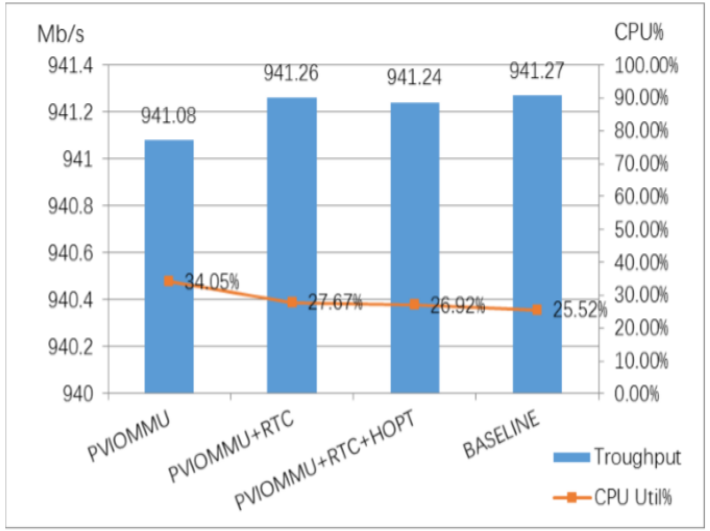
Gaussian09
(C6H6单点能量计算)

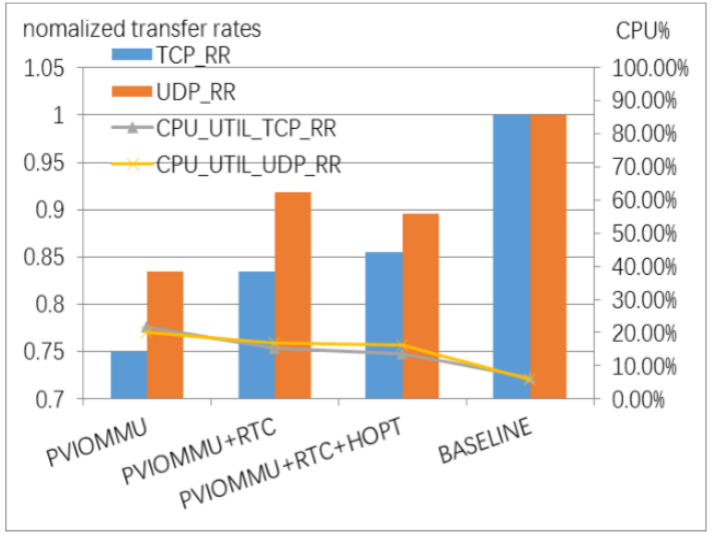


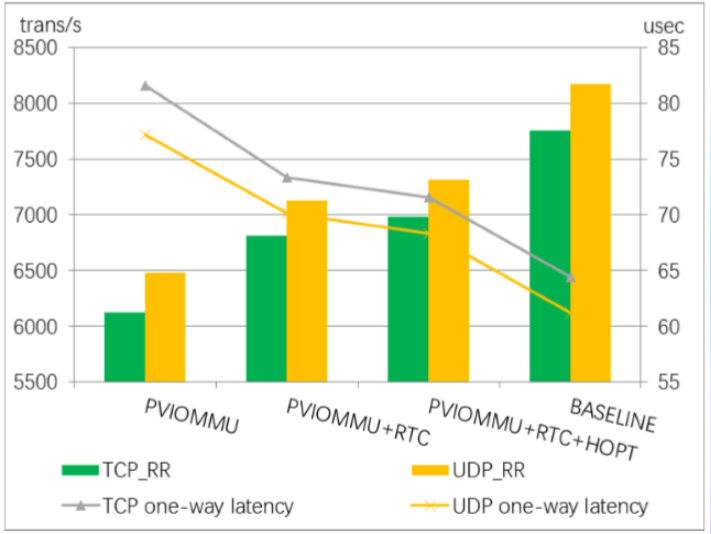
在虚拟机中运行时间相比于宿主机增加3%

性能评测

Pass-Through Intel I210千兆网卡性能测试



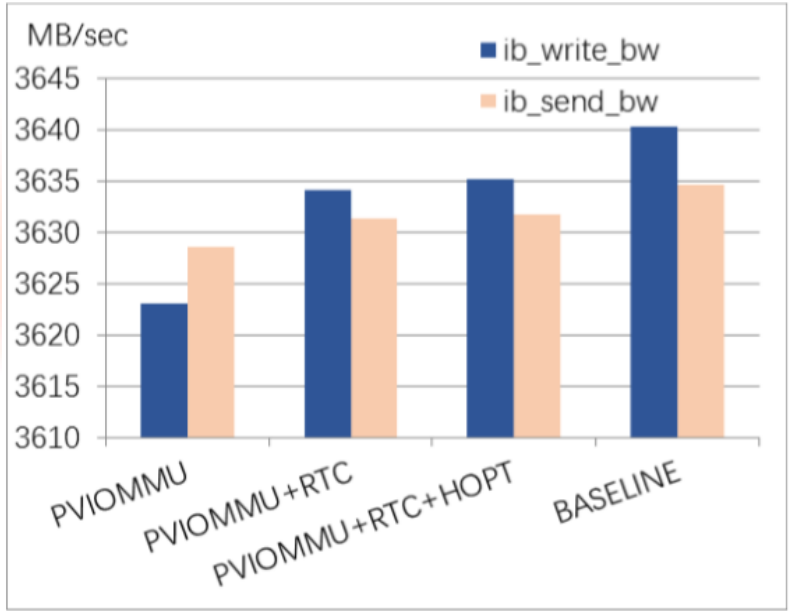




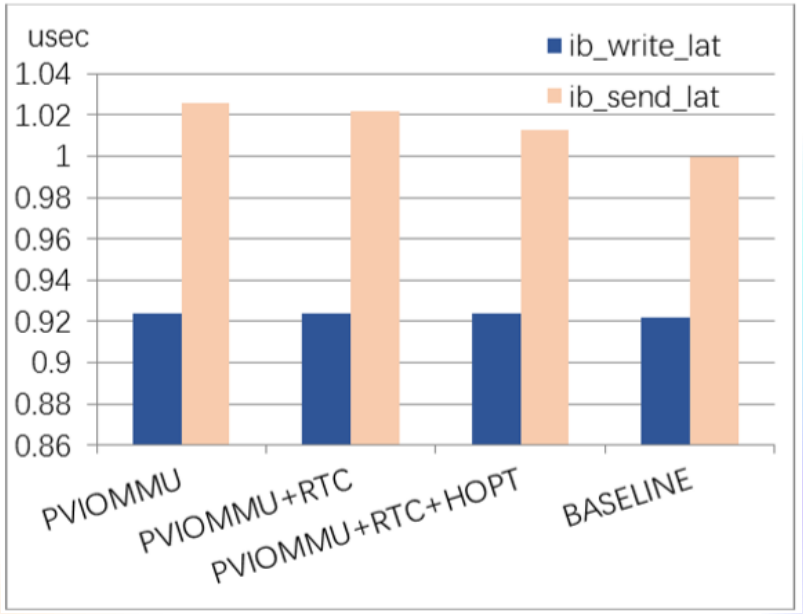
IOMMU半虚拟化性能开销：通信带宽几乎没有影响，TCP 通信延迟增加了 13 微妙，UDP 通信延迟增加了 4 微妙

性能评测

IB通信性能测试



bandwidth

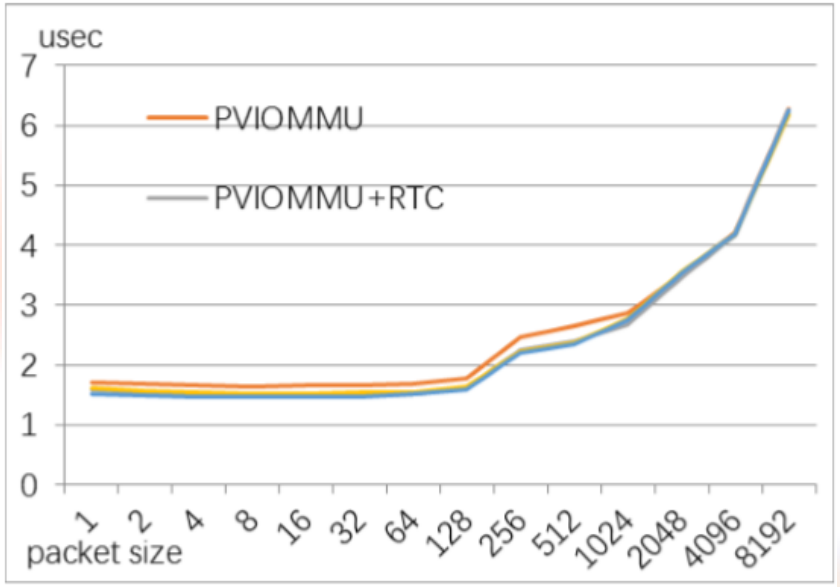


latency

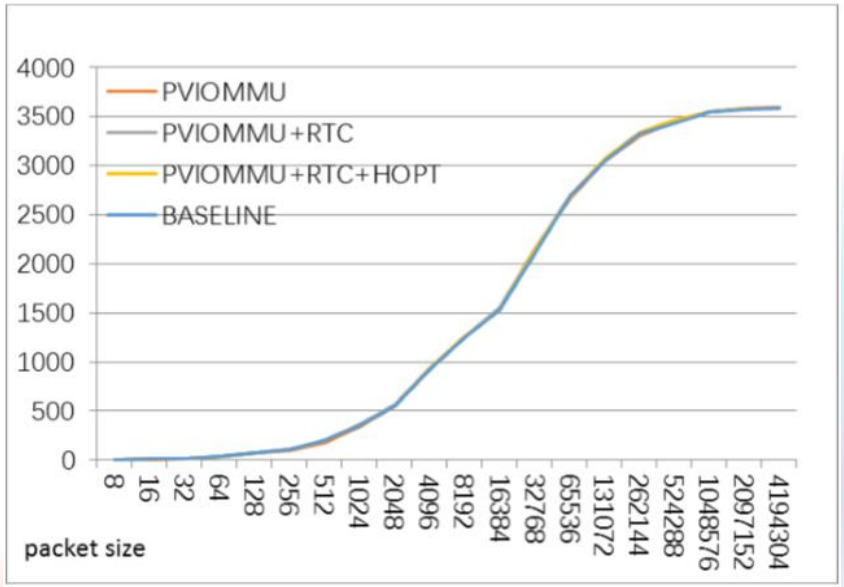
- IOMMU半虚拟化性能开销：延迟增加不到0.1us，带宽减少5MB/sec

性能评测

IB通信性能测试



IMB PingPong延迟

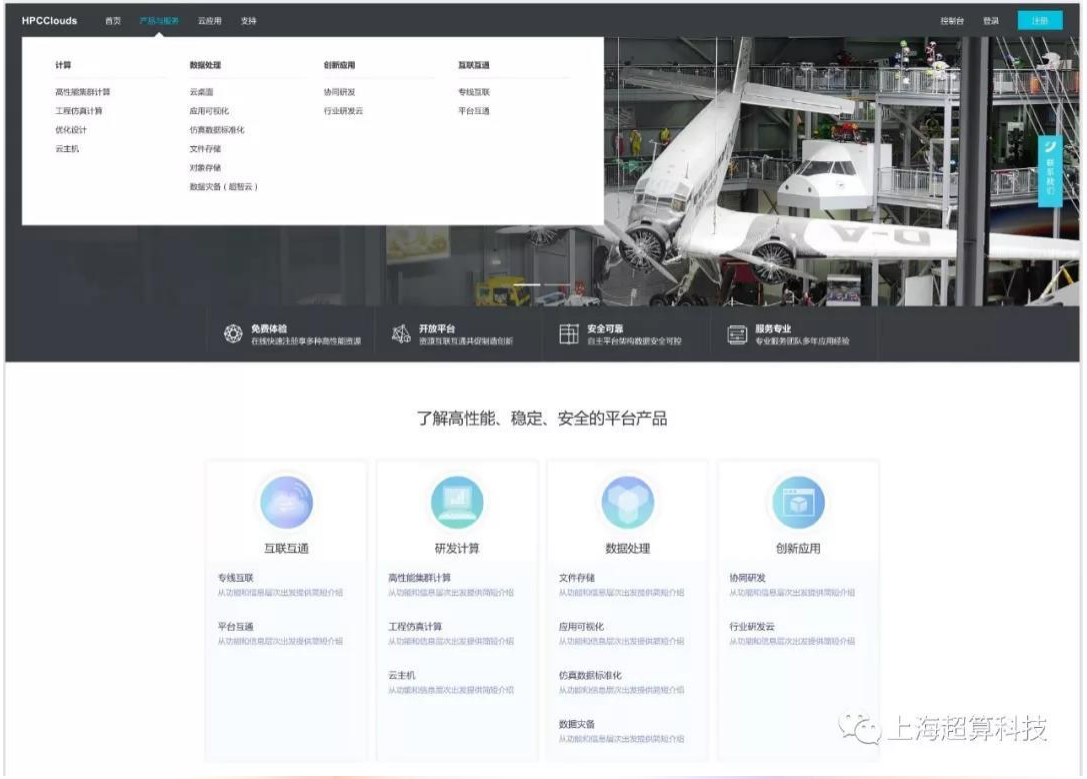


IMB PingPong带宽

- IOMMU半虚拟化性能开销：在MPI通信层面性能开销可以忽略不计

应用案例

上海超算科技HPCClouds



THANKS!

2021
TRUSTED CLOUD
SUMMIT