共生关系感知的虚拟机间通信优化

Co-location Aware Optimizations for Inter Virtual Machine Communication

任　怡
计算机学院

2016.01

國防科学技術大学
NATIONAL UNIVERSITY OF DEFENSE TECHNOLOGY

主要内容

研究背景

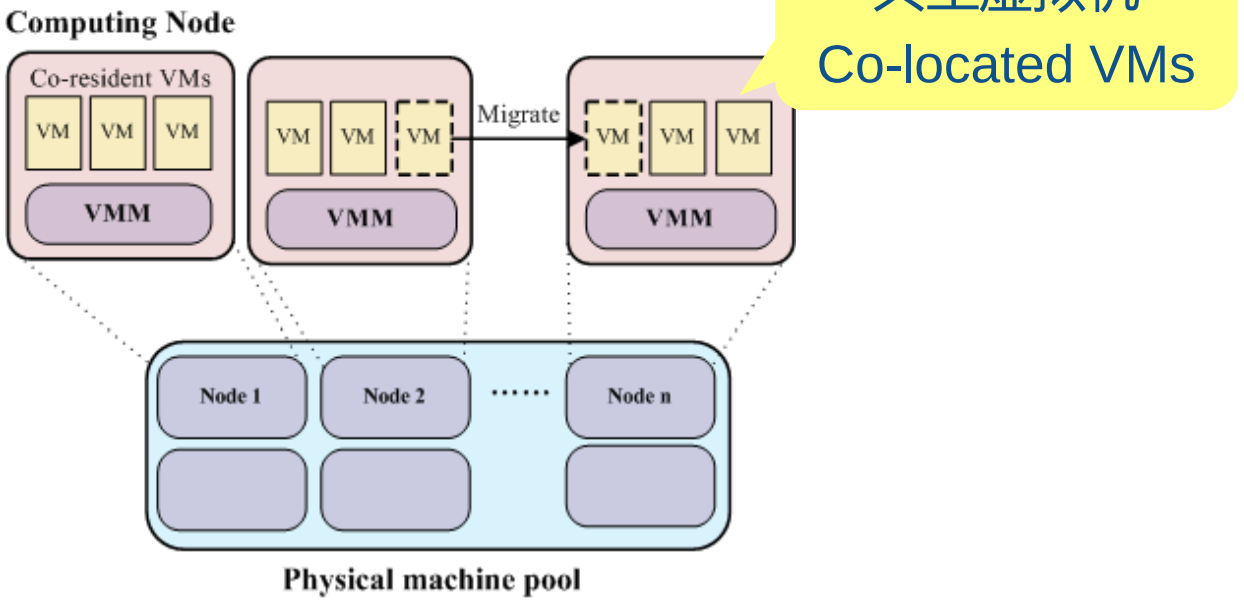
- 虚拟化技术被广泛用于云计算等平台的基础架构层的构建
  - 使用虚拟化技术管理和组织计算资源，将虚拟机作为资源封装的基本单位

提高资源利用率

降低投入成本

可靠性和隔离性

便捷的运维管理



研究背景

- 共生虚拟机间通信的性能损耗
  - 高负载情况下虚拟机之间资源的竞争带来性能折损
  - 某些情况下甚至与不同物理机间通信的性能相差无几
- 网络密集型应用的迫切需求
  - 高性能计算、大规模分布式计算、Web 事务处理…

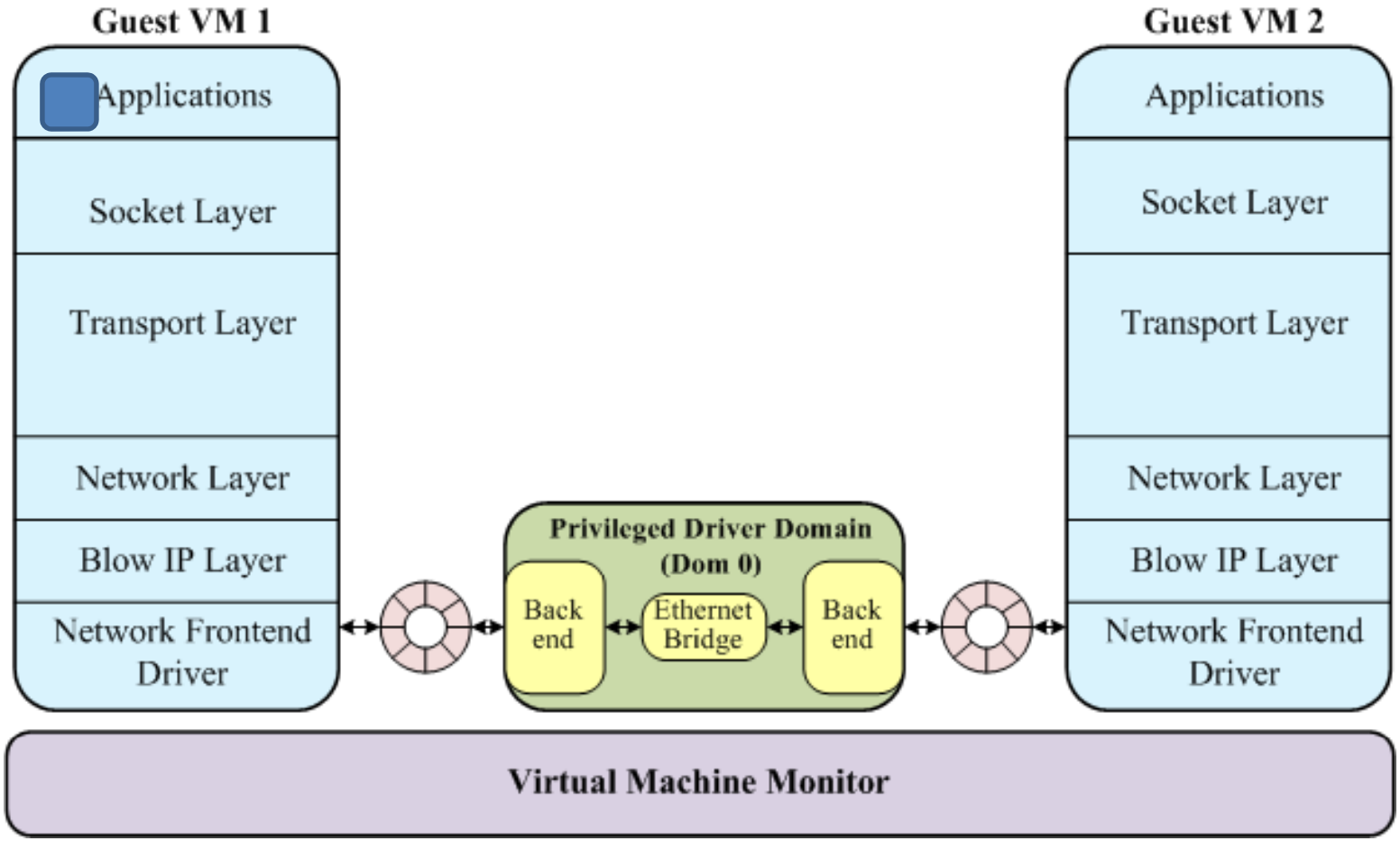
如何提高共生虚拟机间的通信效率？

研究背景

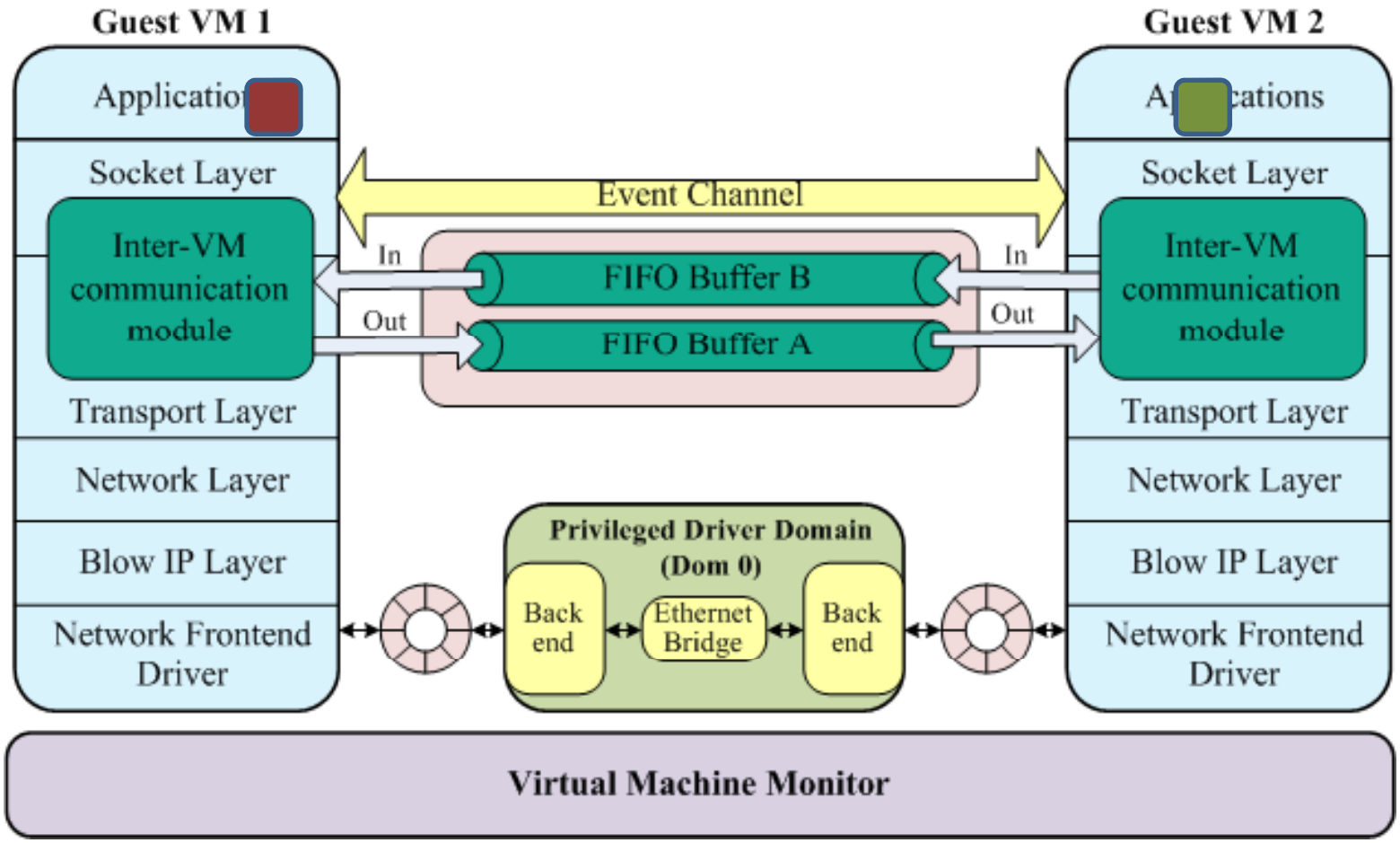
- 共生虚拟机间通信优化的两个层面
  - 较高层——减少资源竞争
    - » 运行不同类型应用的多虚拟机的部署策略
  - 较低层——旁路共生虚拟机间传统的TCP/IP通信路径，走捷径
    - » 共生虚拟机间采用高效透明的通信协议

研究背景

- 基于TCP/IP的虚拟机间通信(netfront/netback)



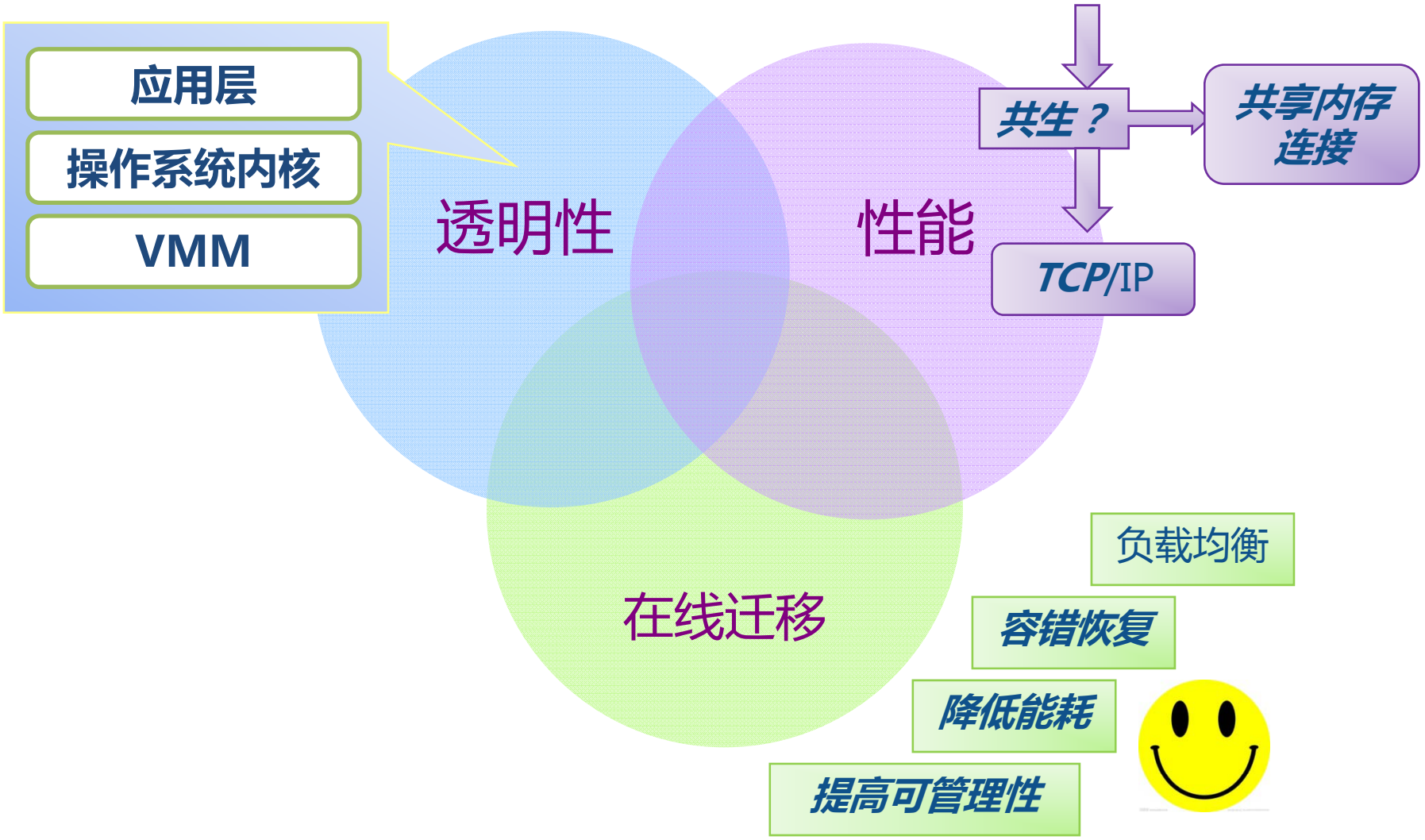
为什么采用基于共享内存的方法？



为什么采用基于共享内存的方法？

- 优点：
  - Less dependent on Dom0, less data copies
  - Bypass the default TCP/IP network path
  - Writes are visible immediately
  - Fewer hypercalls
- 以下机制可支撑虚拟机间共享内存的建立和维护
  - Xen Grant Table: 共生虚拟机之间内存页面的映射/拷贝支持
  - Xen Event Channel: 共生虚拟机之间的异步通知机制
  - XenStore: 辅助追踪虚拟机共生关系的变化

研究目标


应用层
操作系统内核
VMM
透明性
性能
在线迁移
共生？
共享内存连接
TCP/IP
负载均衡
容错恢复
降低能耗
提高可管理性

研究目标

高效	• 能够截获每个数据通信请求 • 能够区分通信双方是否是共生的
支持虚拟机在线迁移	• 适应共生虚拟机的动态加入和退出 • 支持在基于TCP/IP远程通信模式与 • 基于共享内存的本地通信模式之间的透明切换
保证透明性	• 不改变编程API，无需修改OS内核和VMM代码 • 无需重新编译和链接操作系统内核和VMM

國防科学技術大学
NATIONAL UNIVERSITY OF DEFENSE TECHNOLOGY

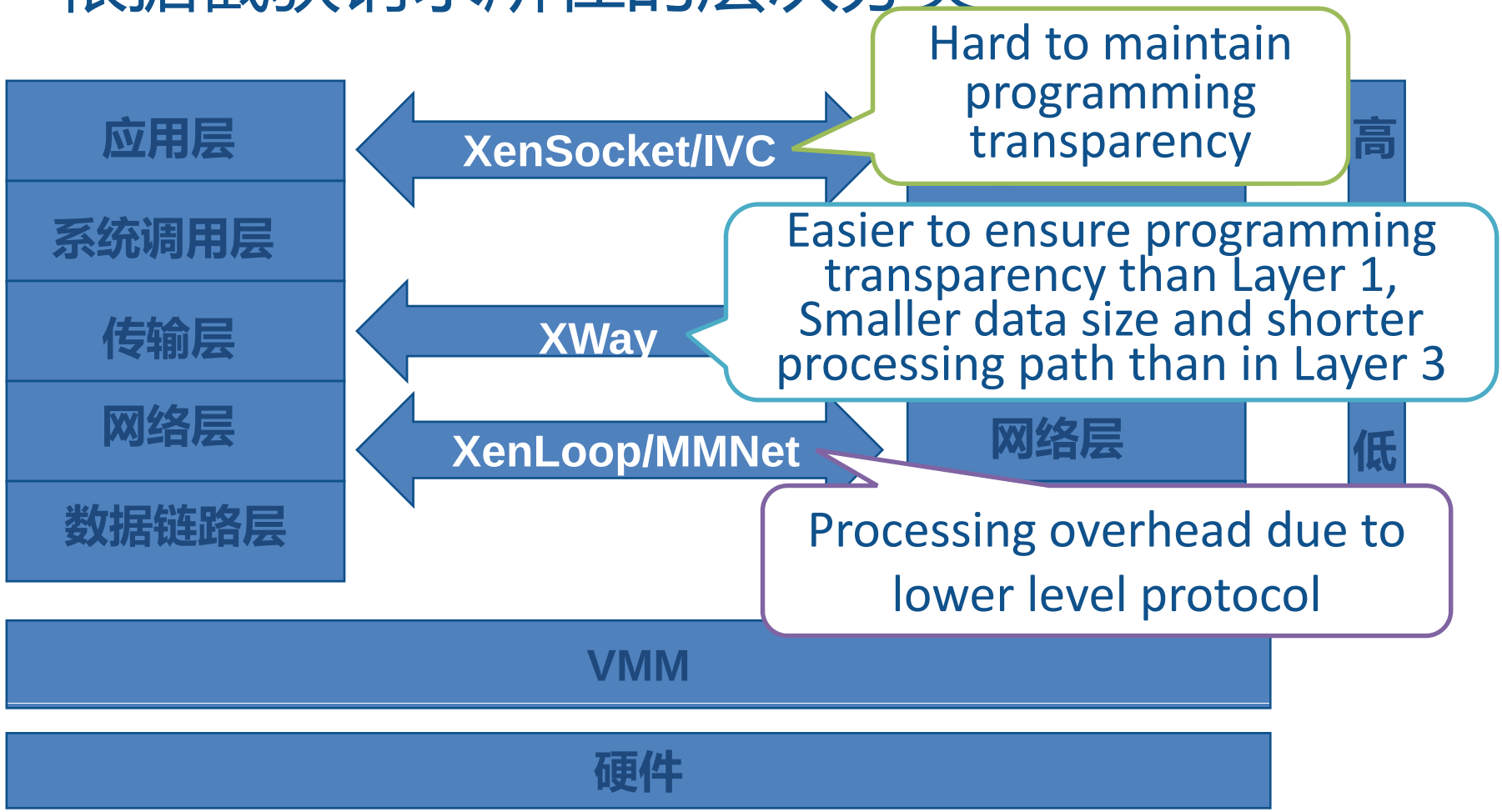
主要内容

相关工作

- 相似之处
  - 需求：更快
  - 基本思想：截获网络通信请求，旁路TCP/IP路径
  - 基本支撑机制：Xen Grant Table、Event Channel等
- 不同之处
  - 技术成熟度：Xen based vs. KVM based
  - 在软件栈中的实现层次
  - 透明性保证
  - 虚拟机在线迁移的支持

基于Xen的相关工作

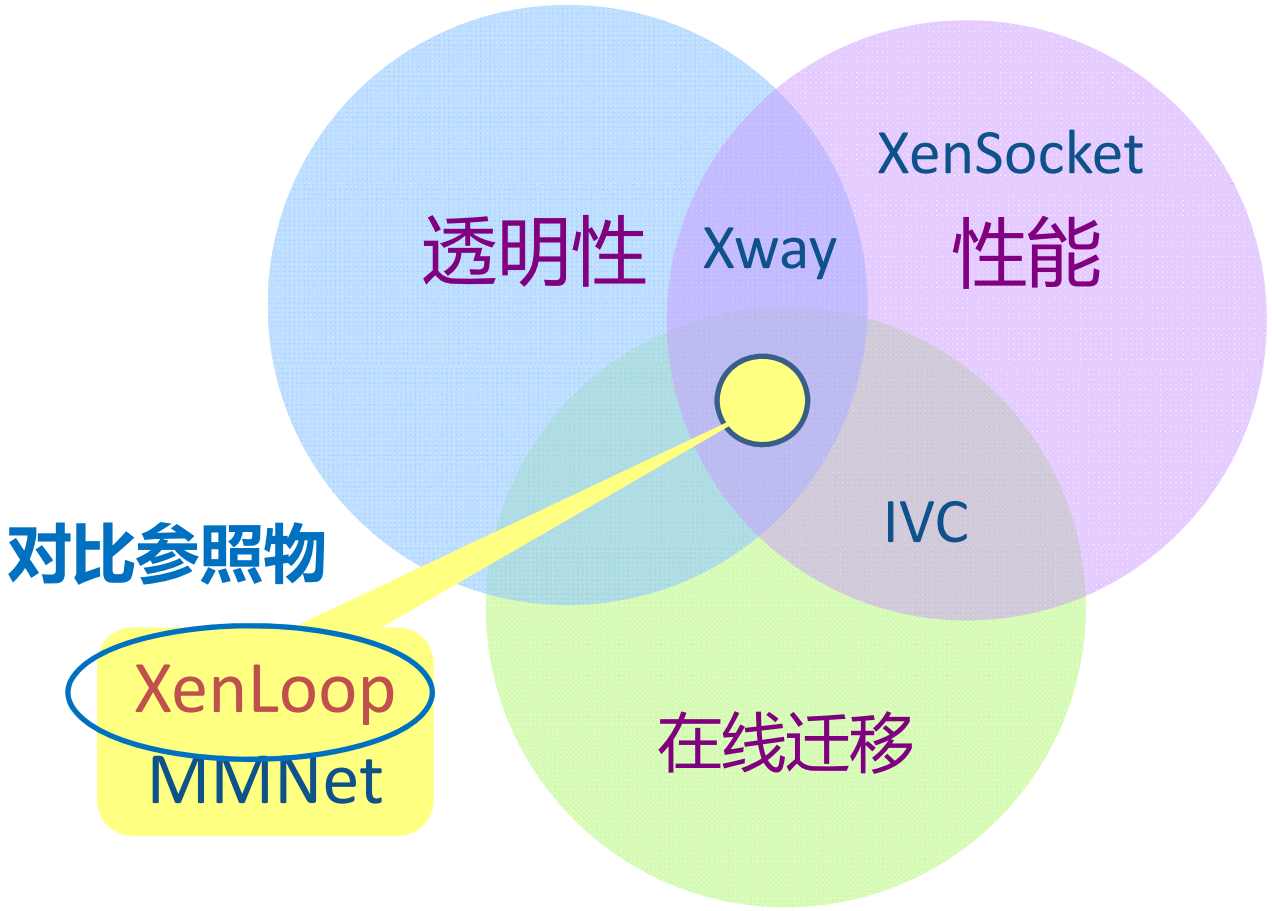
- 根据截获请求所在的层次分类



相关工作对比

	实现层次	是否支持在线迁移	应用层透明性	内核层透明性	VMM透明性	支持协议	开源否
XenSocket	应用层	否	否	是	是	无	是
IVC	应用层	是	否	NA	NA	无	否
Xway	传输层	否	是	否	否	TCP	是
XenLoop	IP层下	是	是	是	是	TCP/UDP	是
MMNet	IP层下	是	是	是	是	TCP/UDP	否

相关工作



主要内容

Xen相关机制研究

- 授权表(Grant Table)
  - 虚拟机之间对读写页面读写权限的授予表
  - 两种授权方式：内存映射和内存传递
- 事件通道(Event Channel)
  - 虚拟机间异步通知机制
- Xenstore：层次的目录结构
  - 存储虚拟机配置信息：可用于虚拟机状态监控
- 虚拟机在线迁移机制
  - 活动状态→暂停状态(内存拷贝)→活动状态

國防科学技術大学
NATIONAL UNIVERSITY OF DEFENSE TECHNOLOGY

Linux相关机制研究

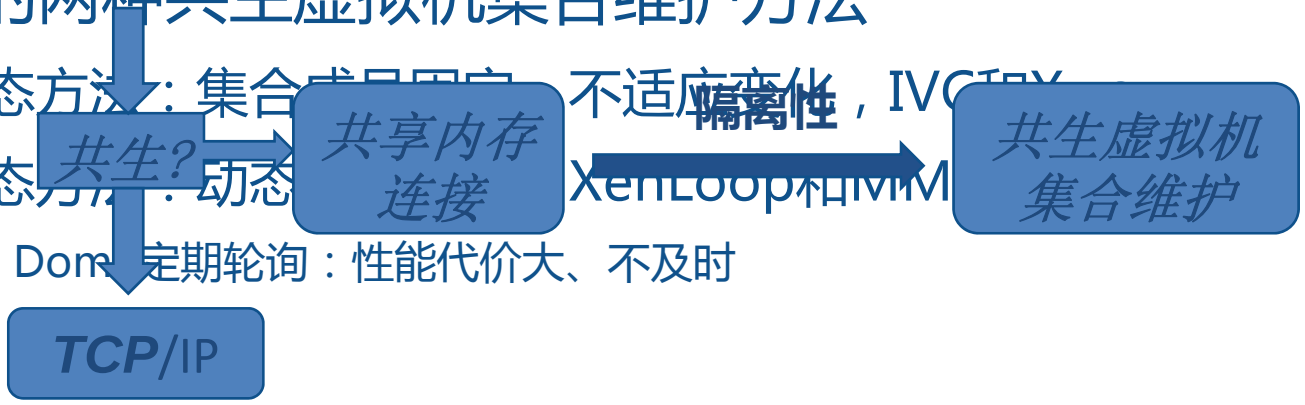
- 系统调用截获
  - 通信数据截获与数据旁路
- 锁机制
- 数据链路层协议
  - 消息机制-虚拟机间信息传递
- sk_buff数据结构
  - 消息发送
  - 构造UDP数据包
- Linux网络设备接收机制
- socket、sock、tcp_sock数据结构
  - 通信旁路、遗留数据处理

实现层次的权衡

- 相关因素
  - 通信请求的截获方式
  - 性能
  - 透明性
  - 可靠性保证

共生虚拟机集合维护

- 每个物理机上的共生虚拟机组成一个集合
- 为什么需要维护共生虚拟机集合？
  - 通信时要判断采用哪种通道：共享内存 or TCP/IP
  - 集合的成员是不可预测的：动态加入/退出
  - 客户虚拟机不能感知到彼此的存在
- 目前的两种共生虚拟机集合维护方法
  - 静态方法：集合成员固定，不适应变化，IVC和XWay
  - 动态方法：动态发现共生虚拟机，XenLoop和MMNet
    - Dom0定期轮询：性能代价大、不及时



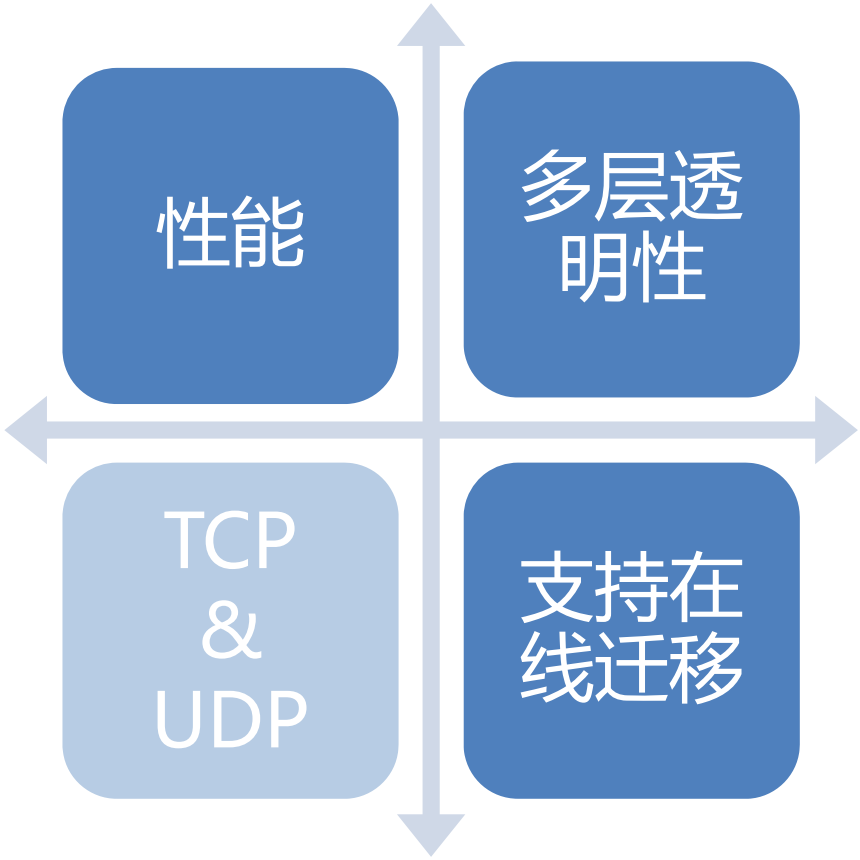
通信模式的透明切换

- 两种通信模式
  - 本地模式：基于共享内存通道
  - 远程模式：基于TCP/IP通道
- 为什么要切换？而且要透明？
  - 在线迁移引起虚拟机的迁出、迁入
  - 正在通信中的虚拟机间的通信模式
    - 本地→远程；远程→本地
  - 透明：用户不感知，平滑地迁移
- 关键和难点
  - 通道的管理和切换
  - 遗留数据的处理

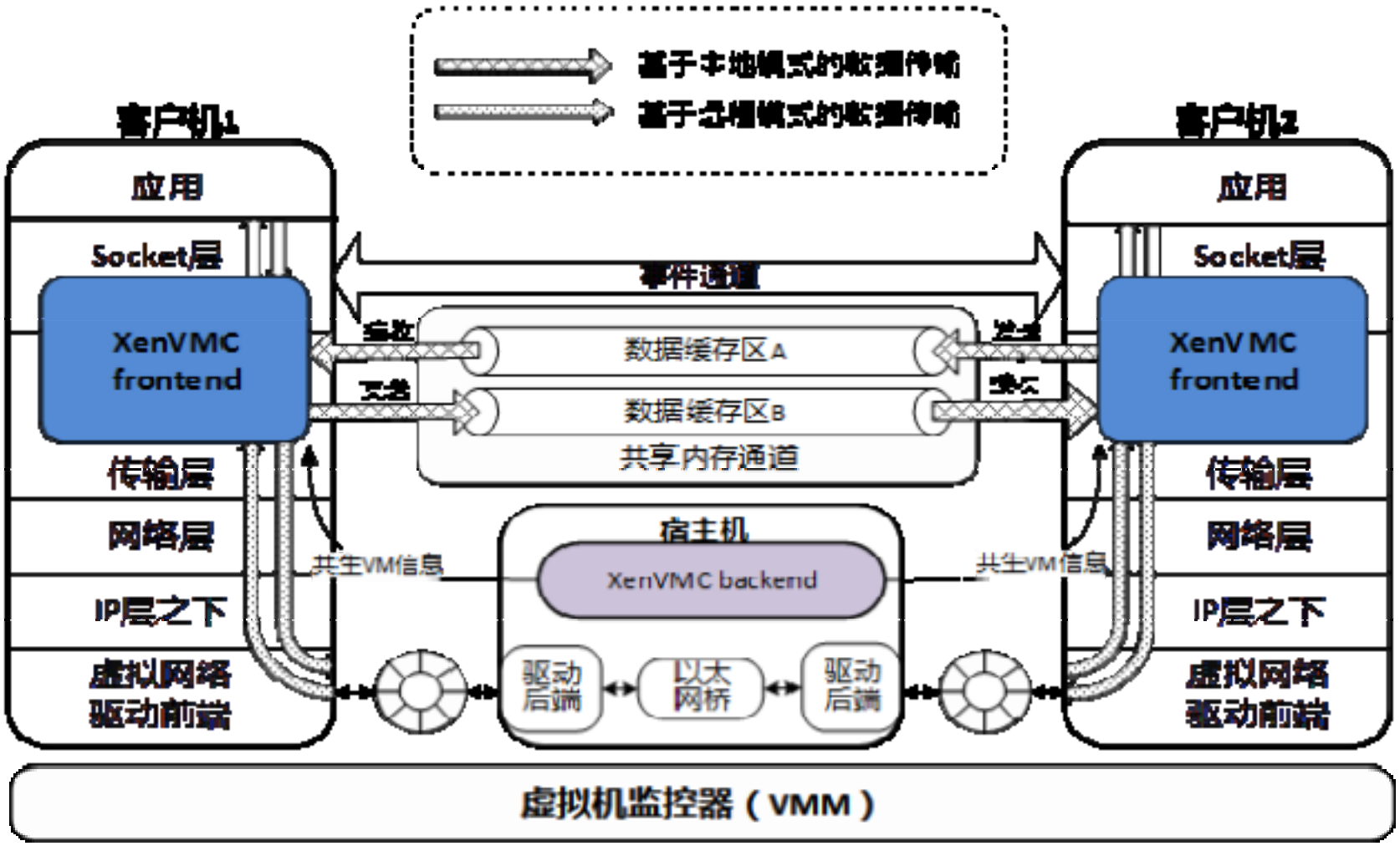
國防科学技術大学
NATIONAL UNIVERSITY OF DEFENSE TECHNOLOGY

主要内容

XenVMC设计目标

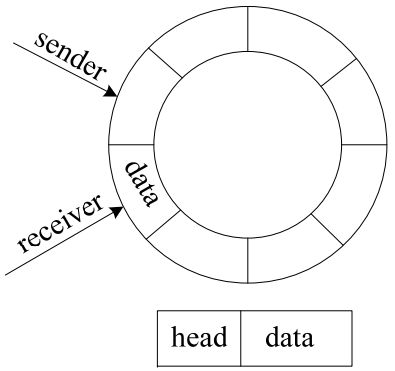


XenVMC总体框架


基于本地模式的数据传输
基于总相模式的数据传输
客户机1
应用
Socket层
XenVMC
frontend
传输层
网络层
IP层之下
虚拟网络
驱动前端
事件通道
数据缓存区A
数据缓存区B
共享内存通道
宿主机
XenVMC backend
共生VM信息
共生VM信息
驱动
后端
以太
网桥
驱动
后端
客户机2
应用
Socket层
XenVMC
frontend
传输层
网络层
IP层之下
虚拟网络
驱动前端
虚拟机监控器（VMM）

XenVMC实现

- 对用户透明的系统调用截获
  - Hooked into system call processing path
  - Replacing existed system call handlers with self defined ones
- 共享内存连接的建立与释放
- 数据发送与接收
  - FIFO buffer: Data flow
  - Event Channel: Control flow



XenVMC实现

- 事件驱动的动态共生虚拟机集合维护算法
  - 发生事件的虚拟机触发集合更新
  - 事件归类：虚拟机加入、删除、准备迁移
  - 基于的观察
    - 虚拟机迁移等事件发生的频率并不高
    - 定期轮询的方法：性能代价大、不及时(X)
  - 特点：
    - 开销小
    - 响应速度快
    - 辅助支持在前迁移：迁移完成前感知迁移

XenVMC实现

- 通信模式透明切换
  - 迁入：远程模式→本地模式
  - 迁出：本地模式→远程模式
  - 切换过程中的遗留数据处理问题
    - 切换为本地模式时，原socket中可能有遗留数据
    - 切换为远程模式时，vmc_tcp_sock中可能有遗留数据
    - tcp_sock结构体：copied_seq和write_seq

实验结果及分析

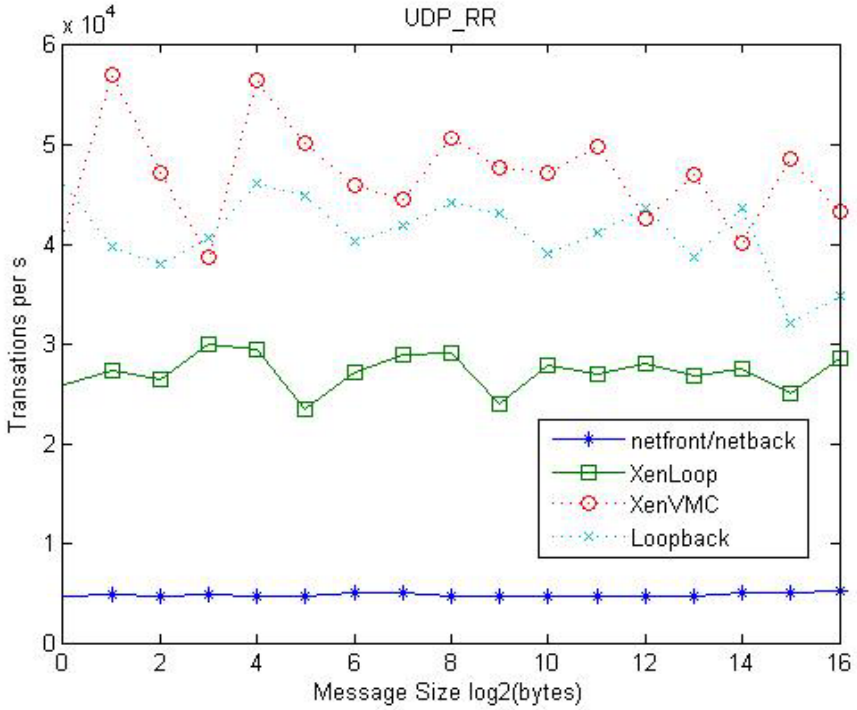
- 概况
  - Xen4.5&Linux3.13.0-rc8
  - XenVMC
  - 完成对比对象XenLoop的移植
- 性能测试
- 在线迁移支持测试

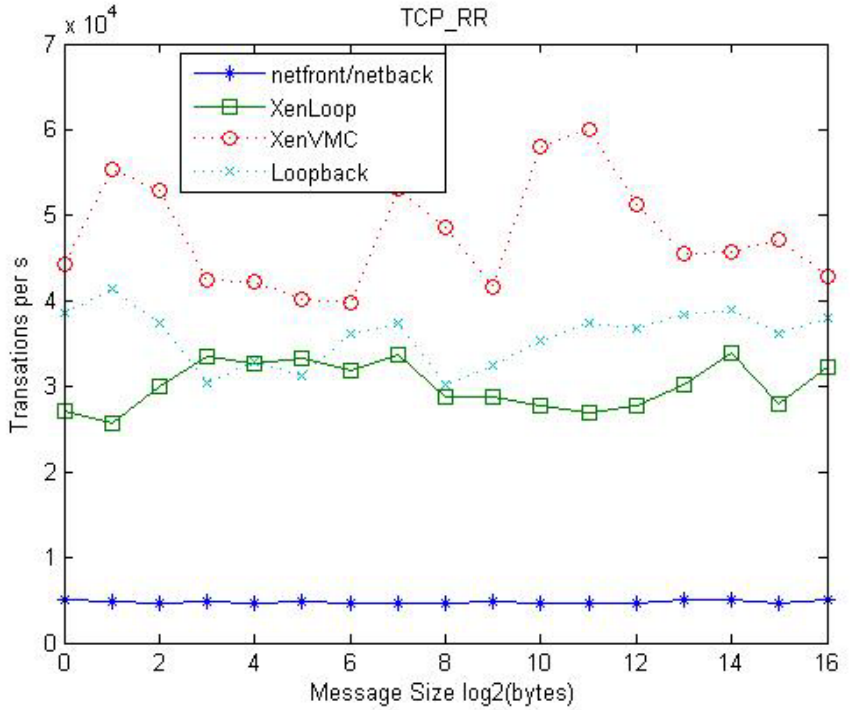
实验结果及分析

- 性能测试
  - benchmark: netperf
  - 测试对象
    - netfront/netback ：传统基于TCP/IP的通信机制
    - Loopback
    - XenLoop: IP层之下，支持在线迁移，多层透明
    - XenVMC
  - 测试内容
    - UDP_RR
    - TCP_RR
    - UDP_STREAM
    - TCP_STREAM

实验结果及分析

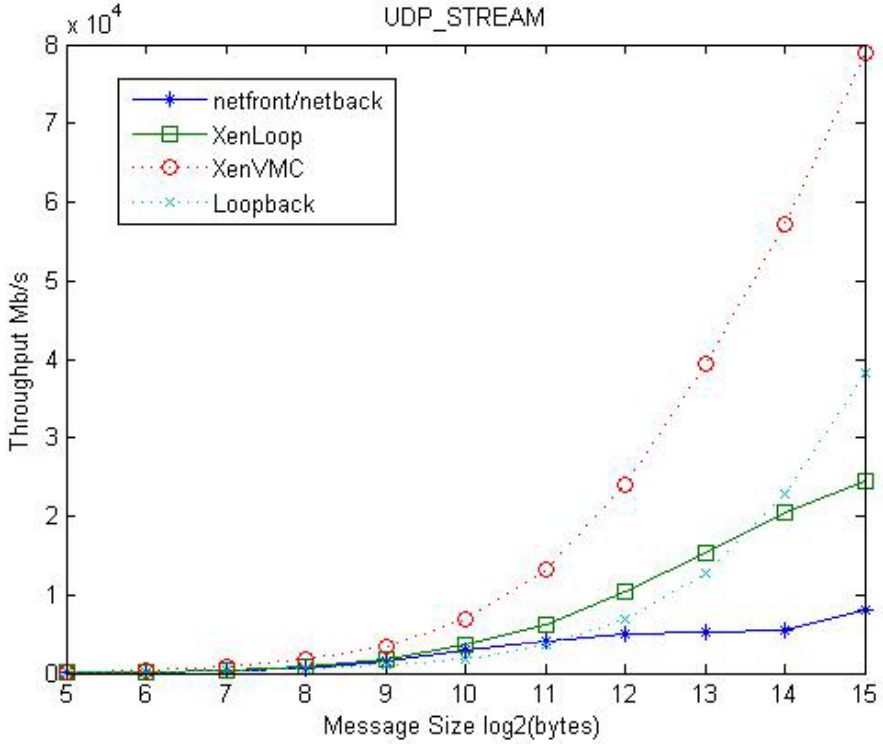
- 性能测试
  - UDP_RR和TCP_RR

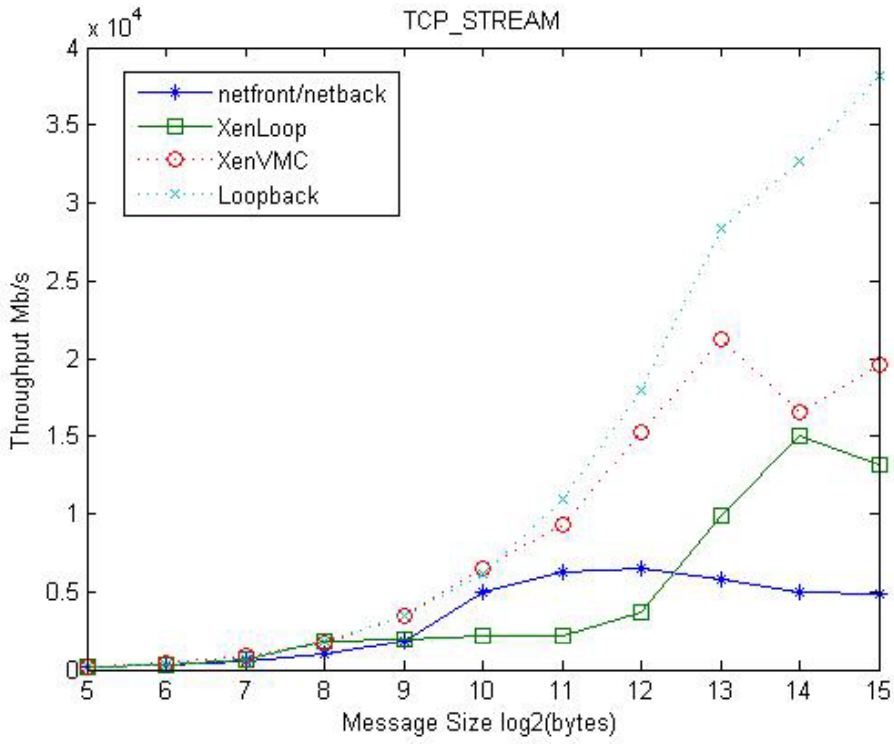




实验结果及分析

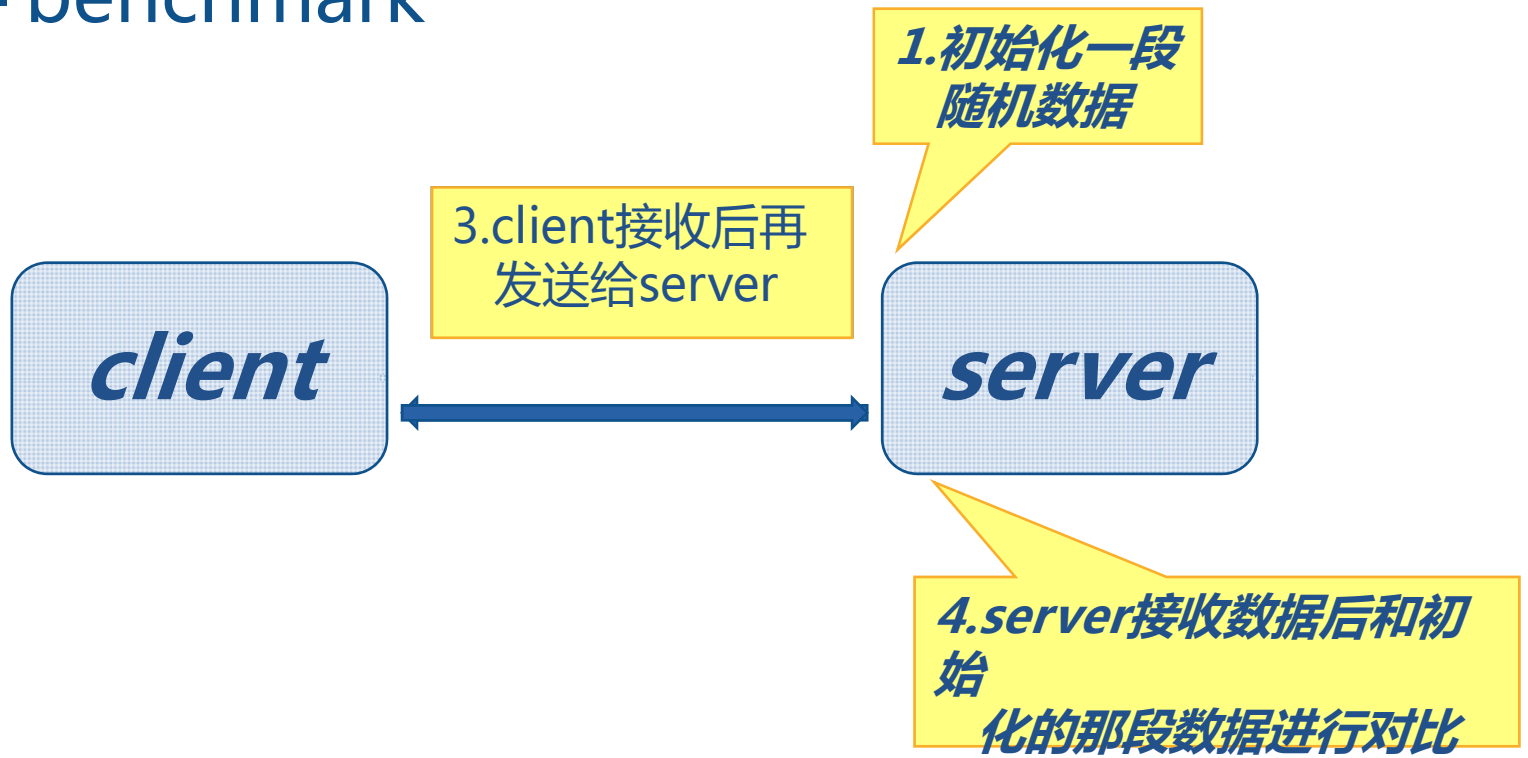
- 性能测试
  - UDP_STREAM和TCP_STREAM





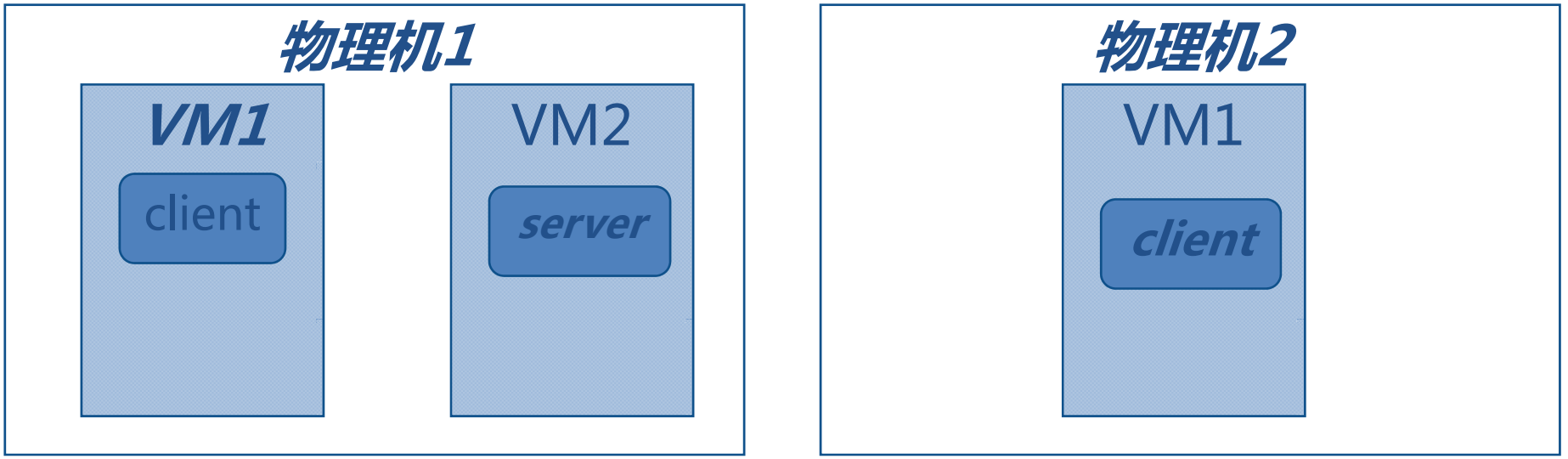
实验结果及分析

- 虚拟机在线迁移支持测试
  - benchmark



实验结果及分析

- 虚拟机在线迁移支持测试
  - 测试过程

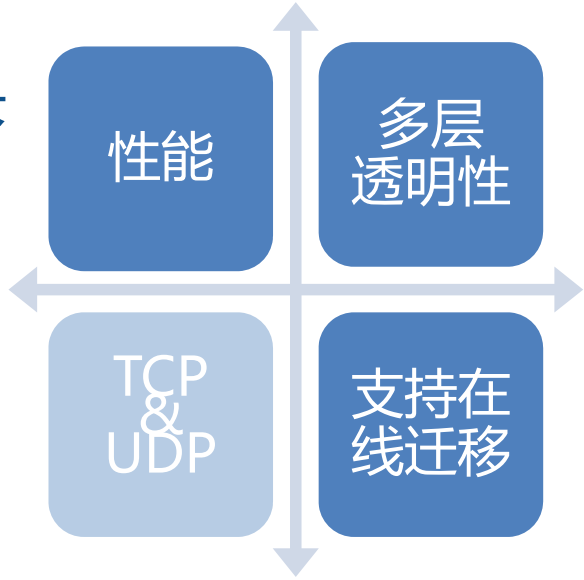


國防科学技術大学
NATIONAL UNIVERSITY OF DEFENSE TECHNOLOGY

主要内容

小结及展望

- 首次对基于共享内存的虚拟机间通信优化的相关工作进行了深入详尽的综述和分析
- XenVMC：达到设计目标
  - 共生虚拟机间采用共享内存通道加速通信
  - 支持虚拟机在线迁移
  - 应用编程透明；可加载的内核模块
  - 支持TCP/UDP语义
- 研究意义和价值



小结及展望

- 展望
  - 其它网络通信相关系统调用的截获
  - 降低空间复杂度
  - 大小数据分离处理
  - 进一步优化性能
  - 针对多处理机模型的通信优化
  - 考虑KVM平台上的可行性

谢 谢

Q&A