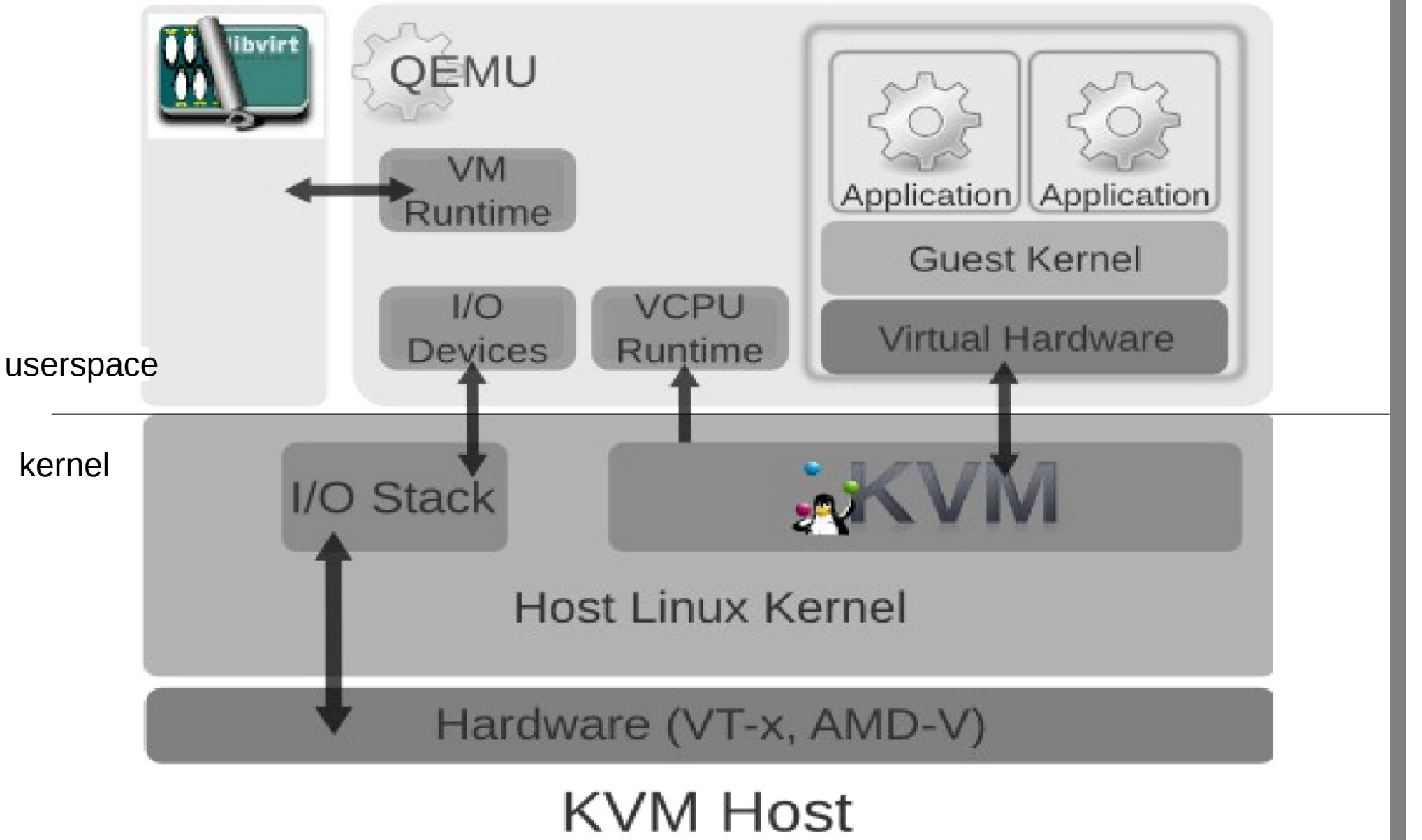
Weather Report

June 2012

Agenda

- What is KVM
- Performance
- Networking
- Block
- RAS
- Desktop
- Cloud

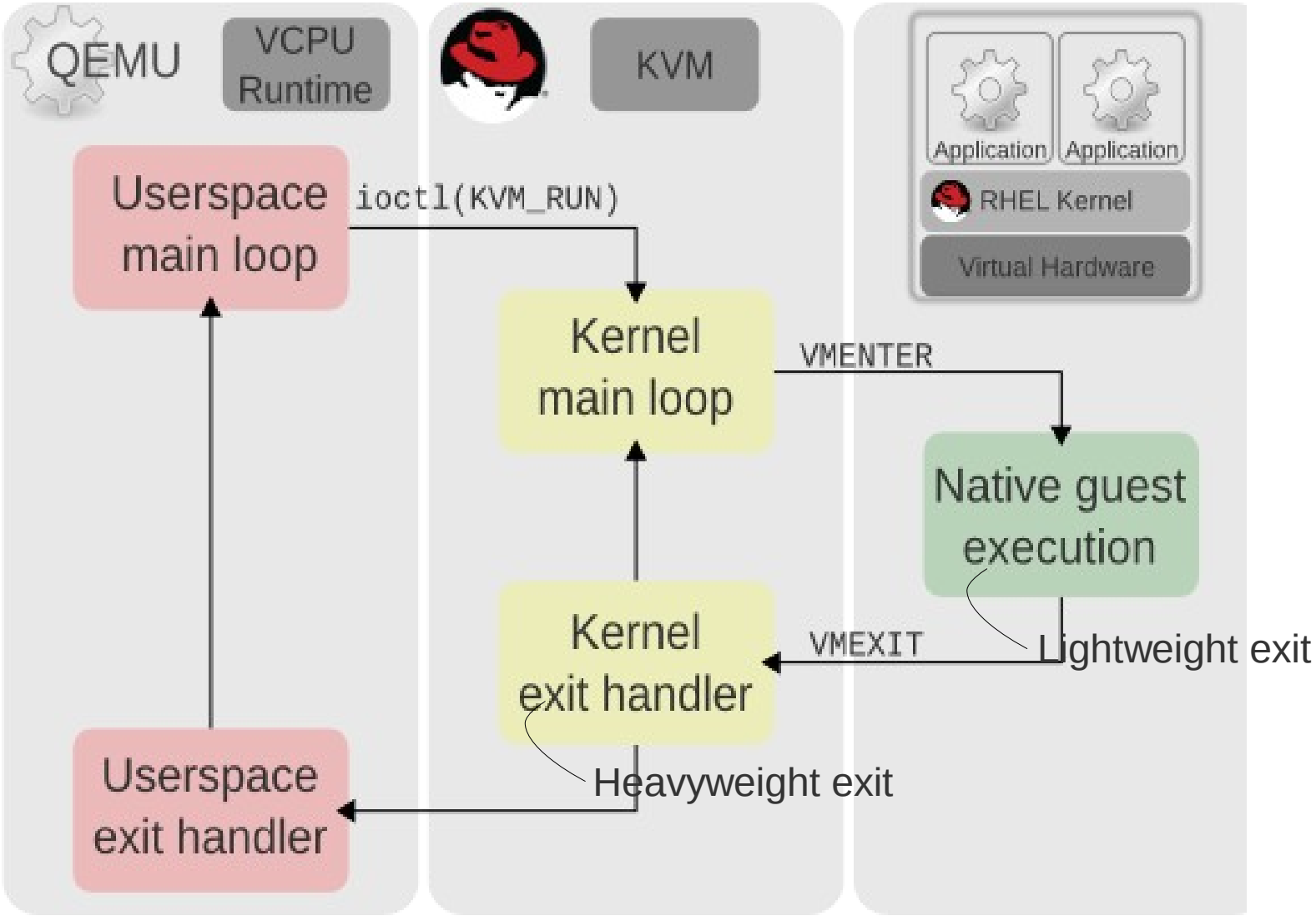
KVM Architecture



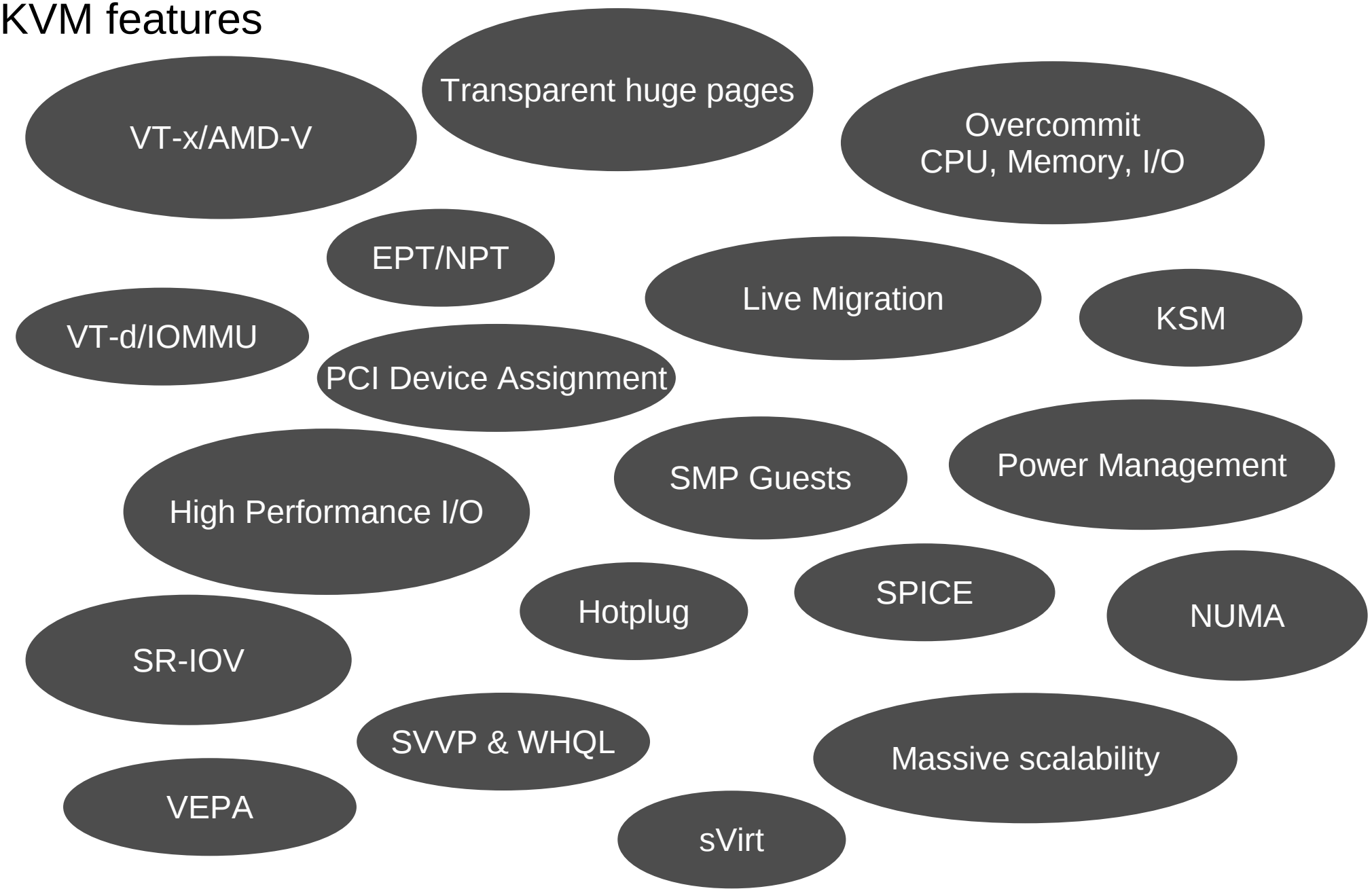
Why **reinvent** the wheel?

Focus on virtualization.

KVM Architecture



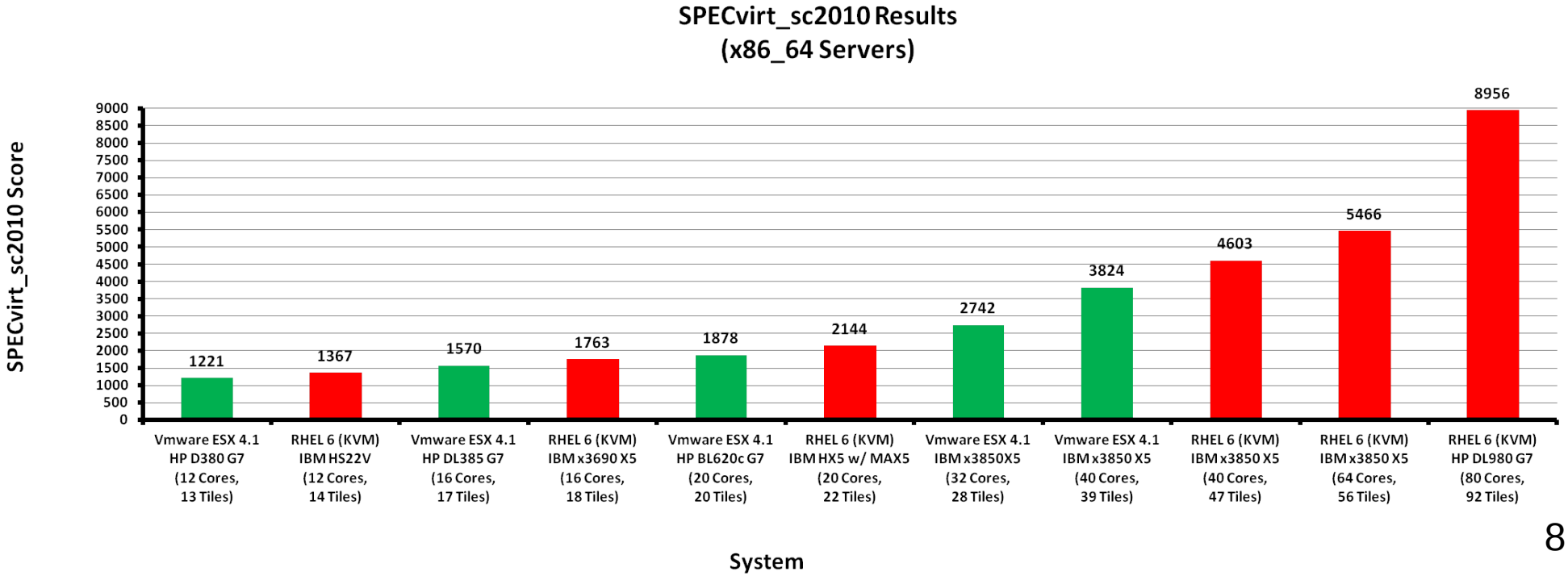
KVM features



What does it **add** up to?

Performance – SPECVirt (April 2012)

- Top RHEL/KVM score beats the top VMware ESX score by a factor of 2
- KVM bests VMware ESX wherever head-to-head comparison is possible
- Key enablers of KVM's leadership virtualization performance include: SR-IOV, Huge Pages, NUMA, Node Binding



Why KVM outperforms the rest?

- **Linux, Linux, Linux**
  - Hardware enablement : drivers, partners, ecosystem
  - Scheduler, MMU, IO stack
  - Hybrid mode
  - OSS – best minds in the world
  - More
- **We own the guest and the host**
  - Paravirt clock, steal time
  - Paravirt GPU (spice)
  - Paravirt interrupt controller (x2apic)
  - Paravirt page faults
  - Paravirt spinlocks
  - Vmchannel (virtio-serial)

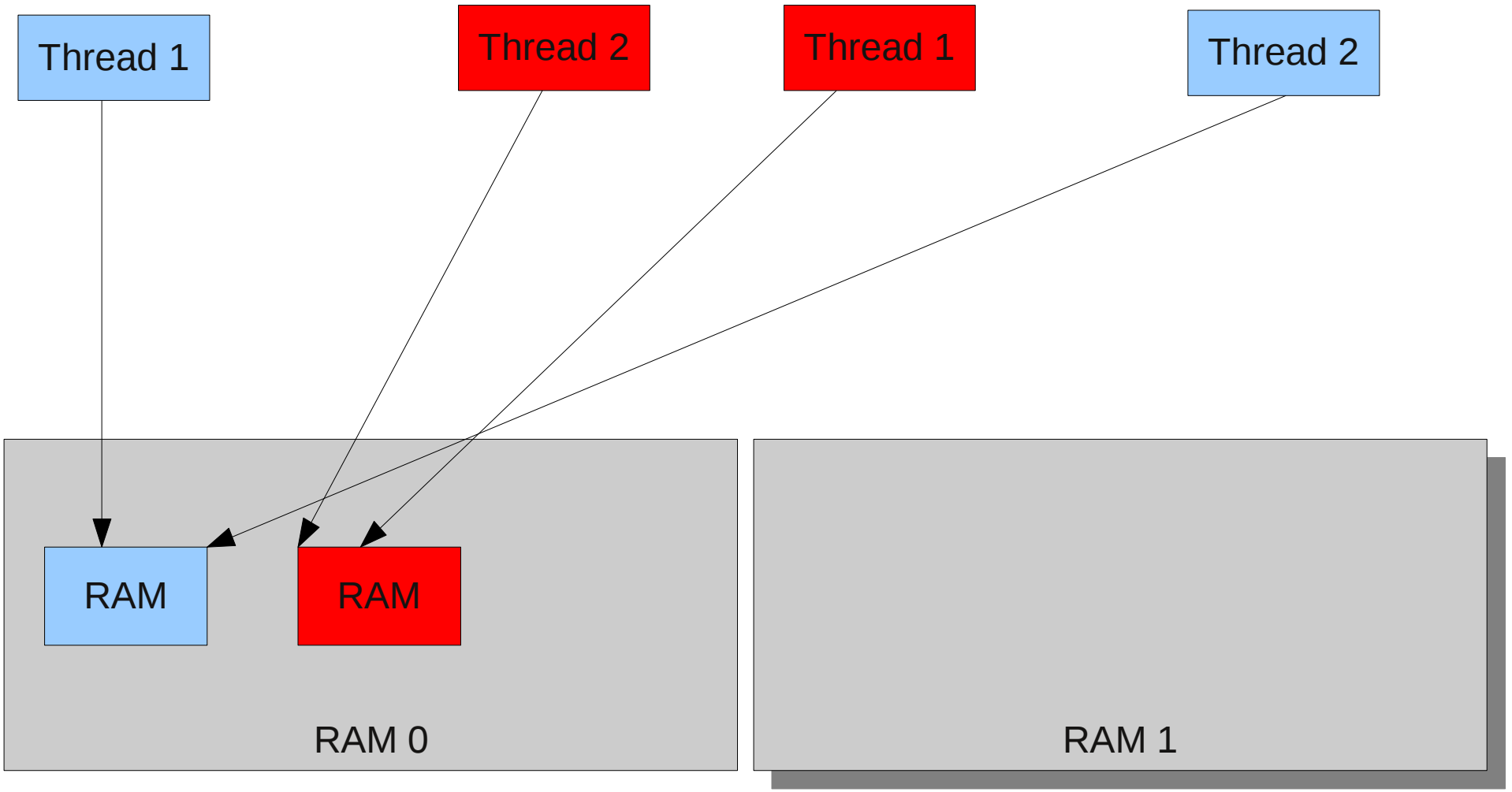
Performance

- Up to 160 virtual CPUs
- Up to 2TB Ram
- MMU and guest page fault handling performance improvements
- Dirty logging performance improvements
- PCID/INVPCID for guests with EPT – tlb tagging for reducing the need for tlb flush
- Paravirt spinlocks (ticketlock)

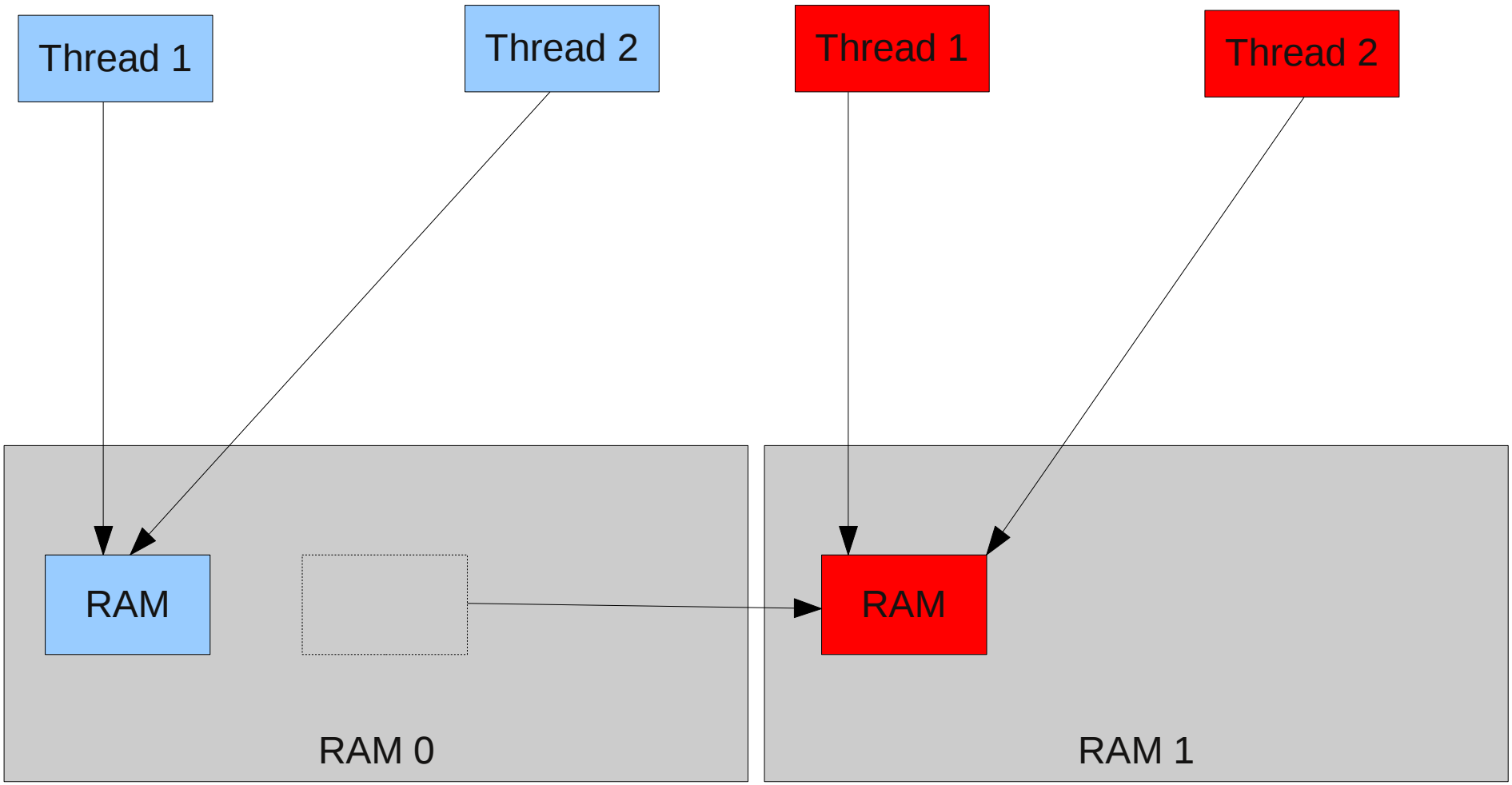
Performance

- Autonuma/schednuma - Automatic page migration
  for better NUMA localization
  - Kernel implementation
- Numad- Non-Uniform Memory Access Daemon
  - User space implementation
  - http://fedoraproject.org/wiki/Features/numad

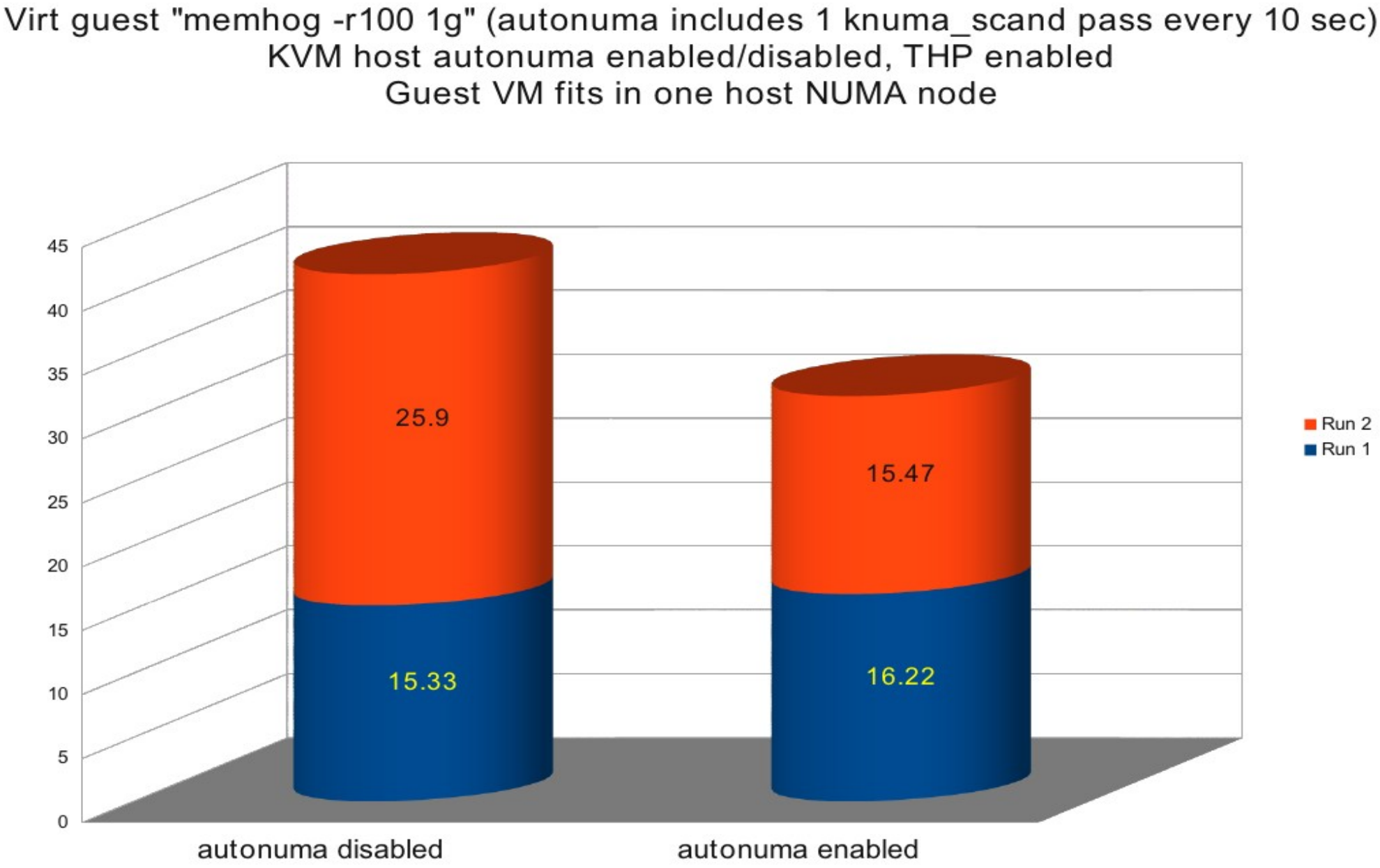
Performance – before autonuma



Performance – after autonuma



autonuma benchmark

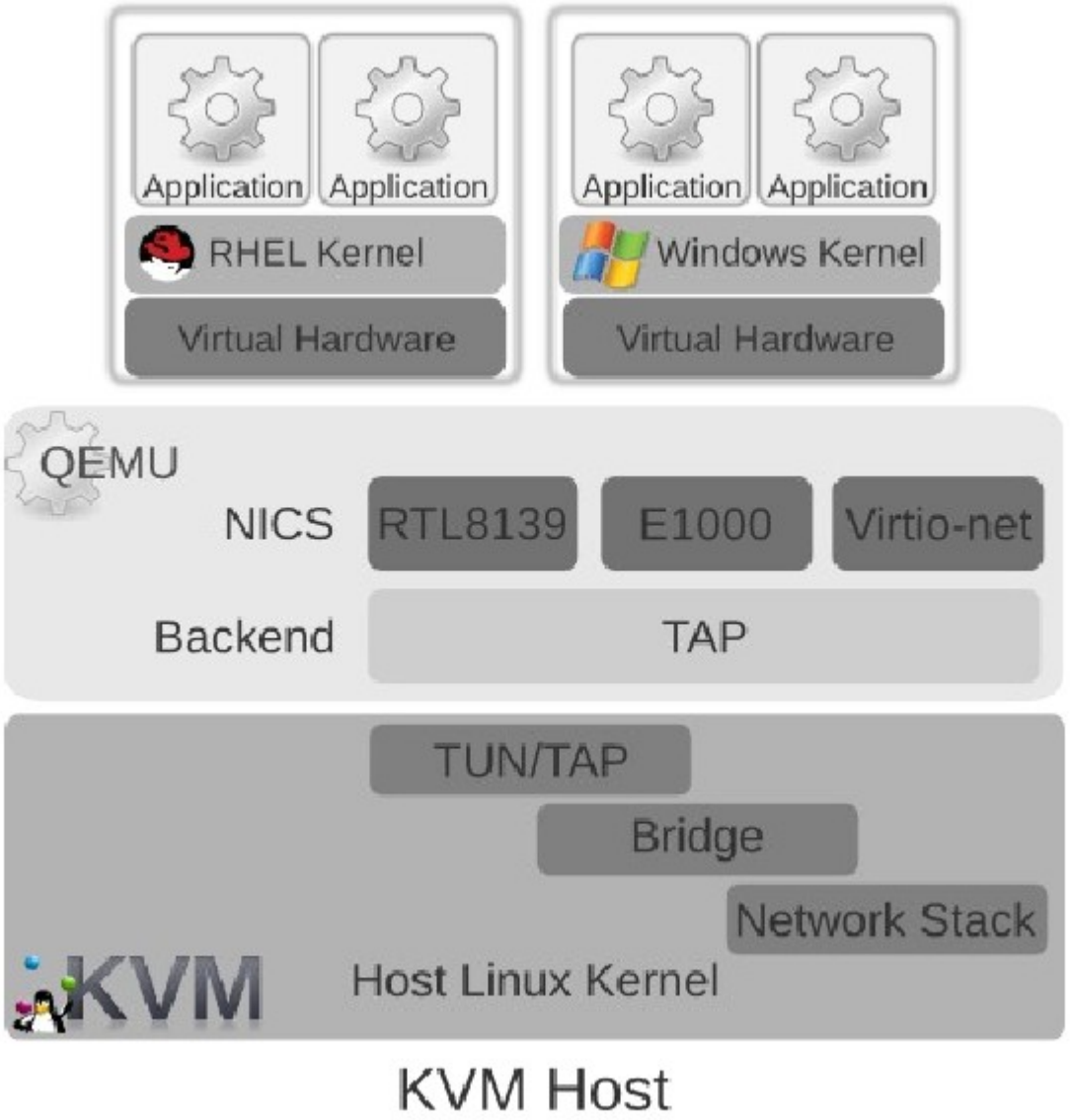


Performance

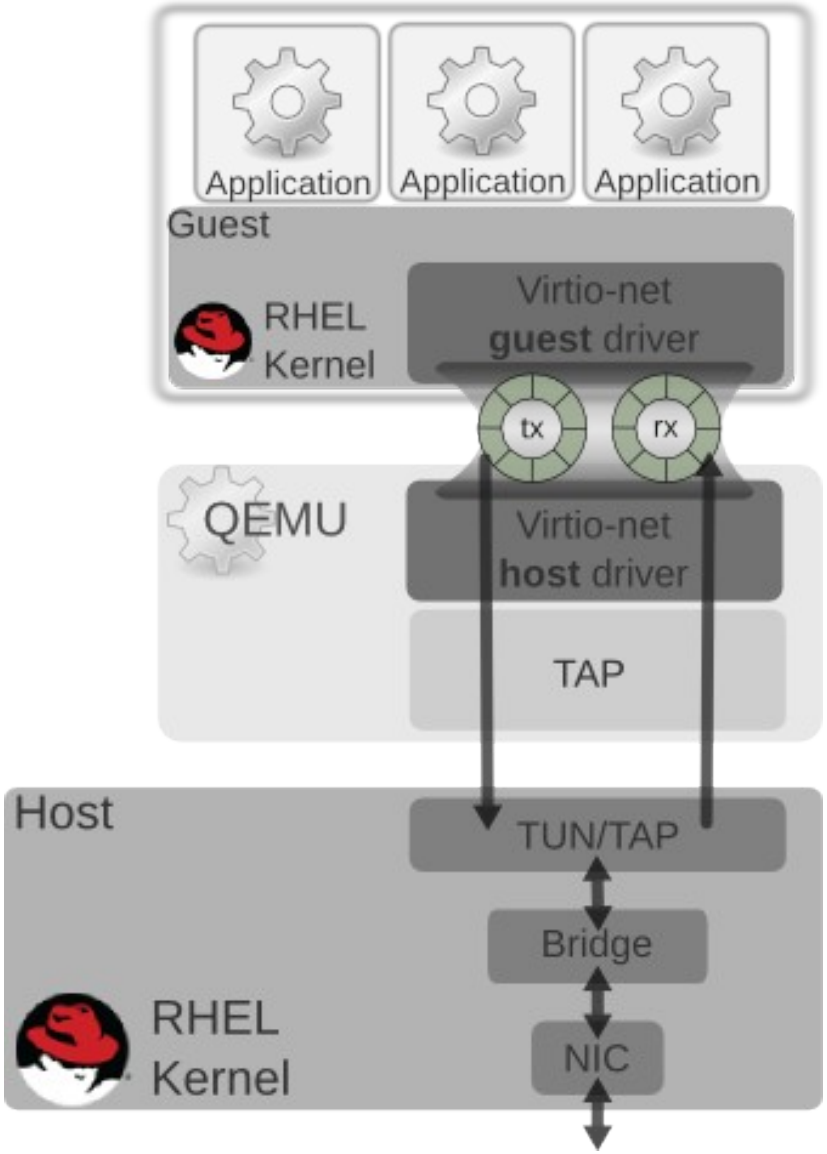
- Hyper-V enlightened guest interface support in KVM:
  - increases performance of MS guests on KVM. (per Microsoft Hypervisor Functional Specification)
  - Feature and interface discovery
  - Scheduling/spinlocks
  - Virtual APIC & Others

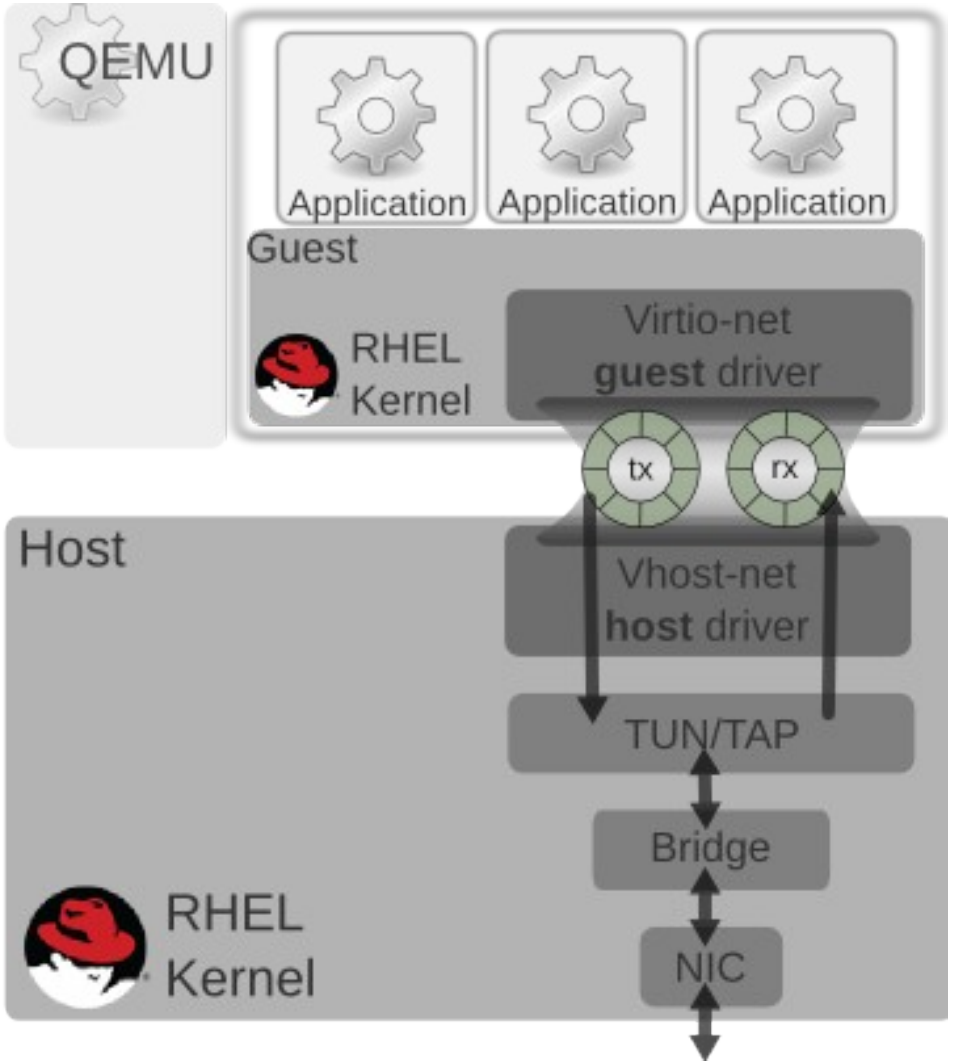

networking
success
informal
planning
mutual
meeting
internet
relationships
jobs
support
opportunities
community
professional
colleagues
efficiency
common
online
interests
interaction
social
connect
marketing
business
contacts
people

KVM Network Architecture

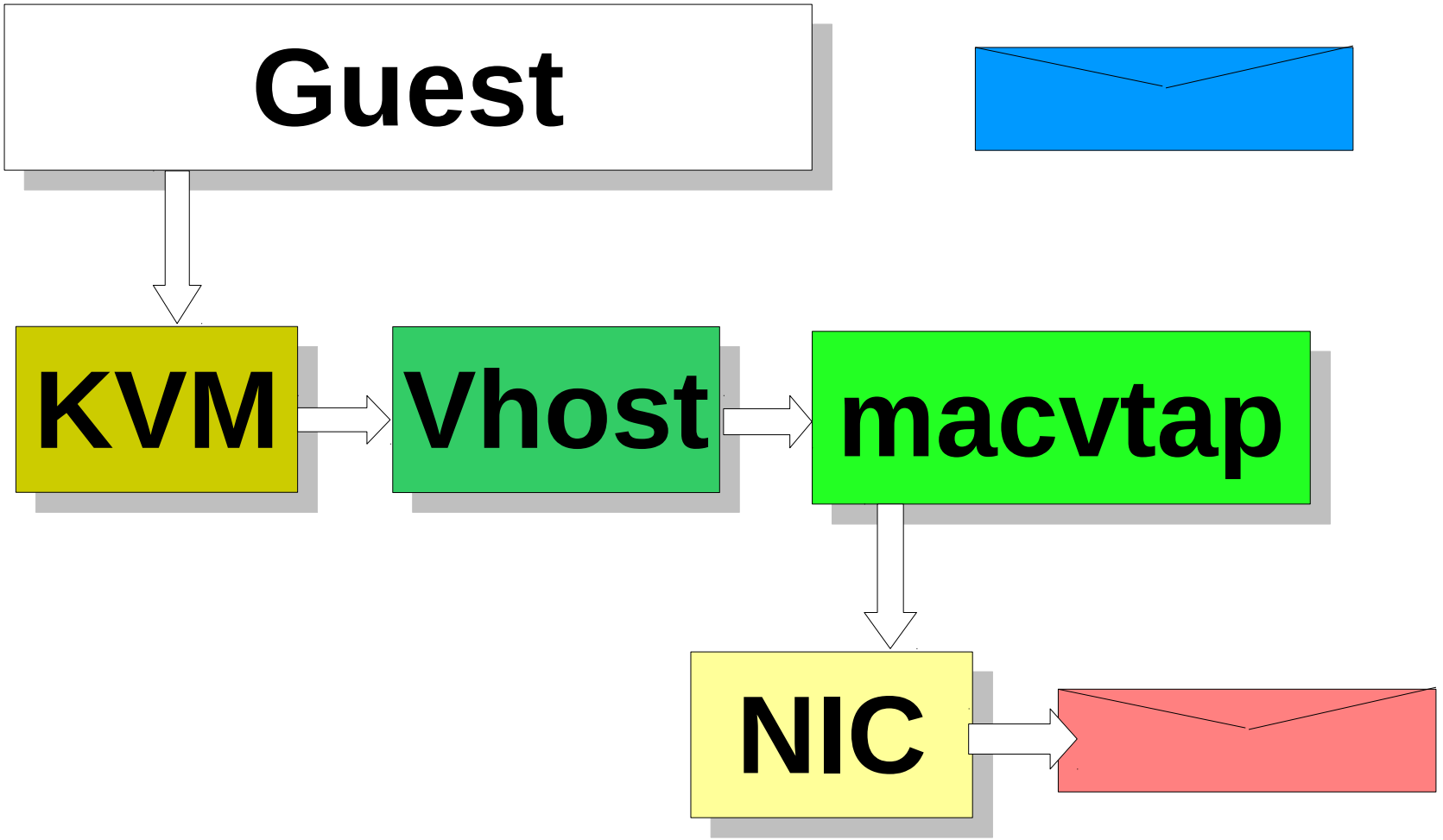


KVM virtio network architecture

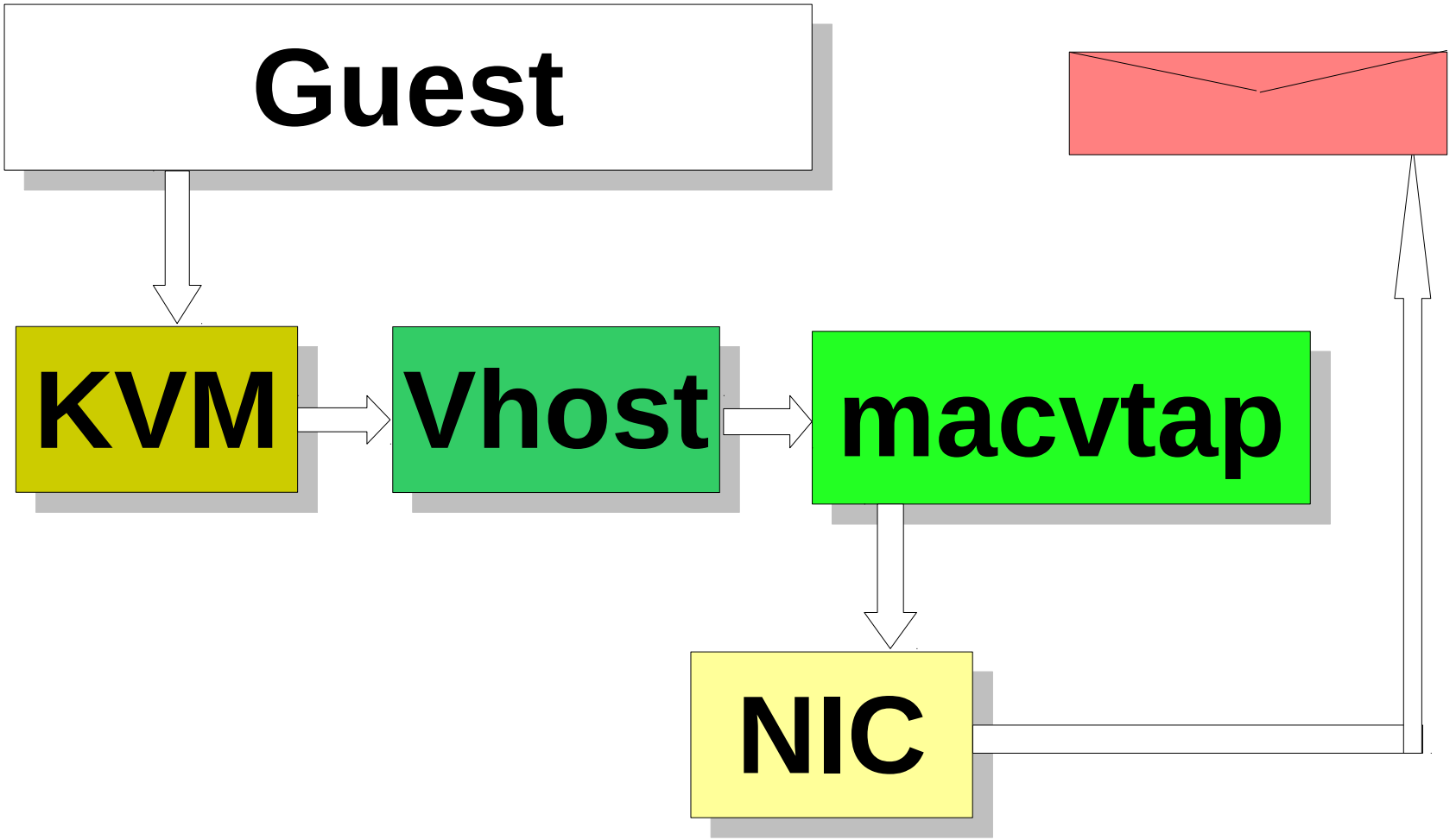




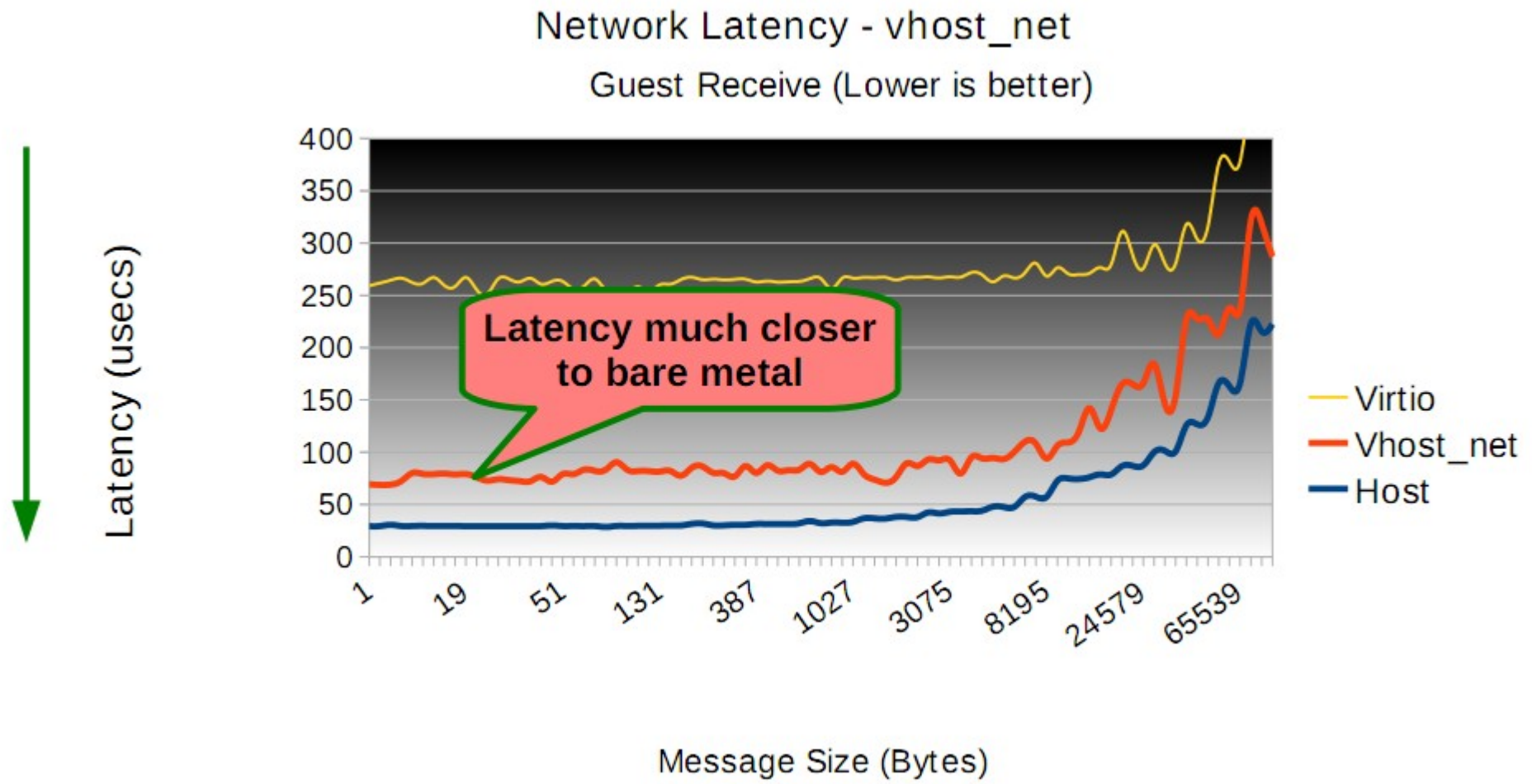
virtio-net TX



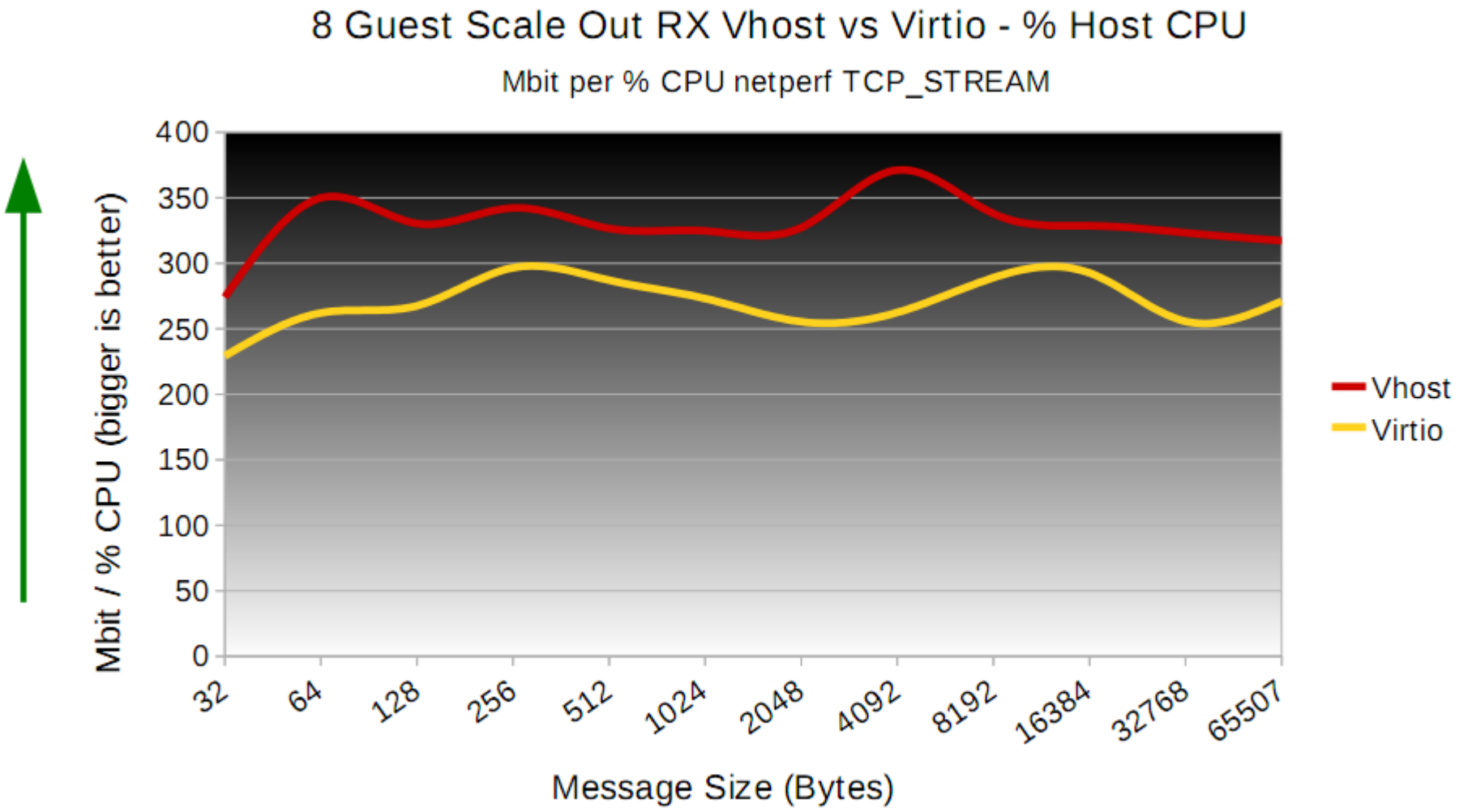
virtio-net TX w/ zero copy macvtap



vhost_net performance

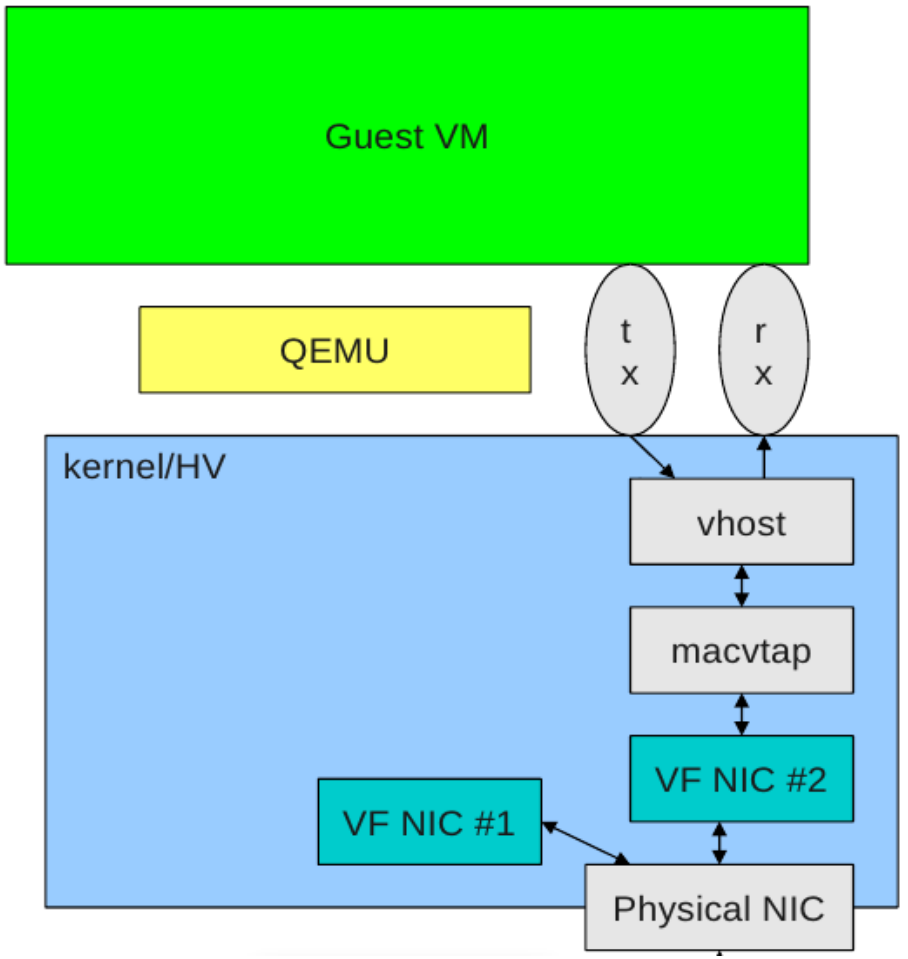


vhost_net performance



Virtio over macvtap with SR-IOV

- Guest only knows virtio
- Migration friendly
- Good performance
- Zero copy

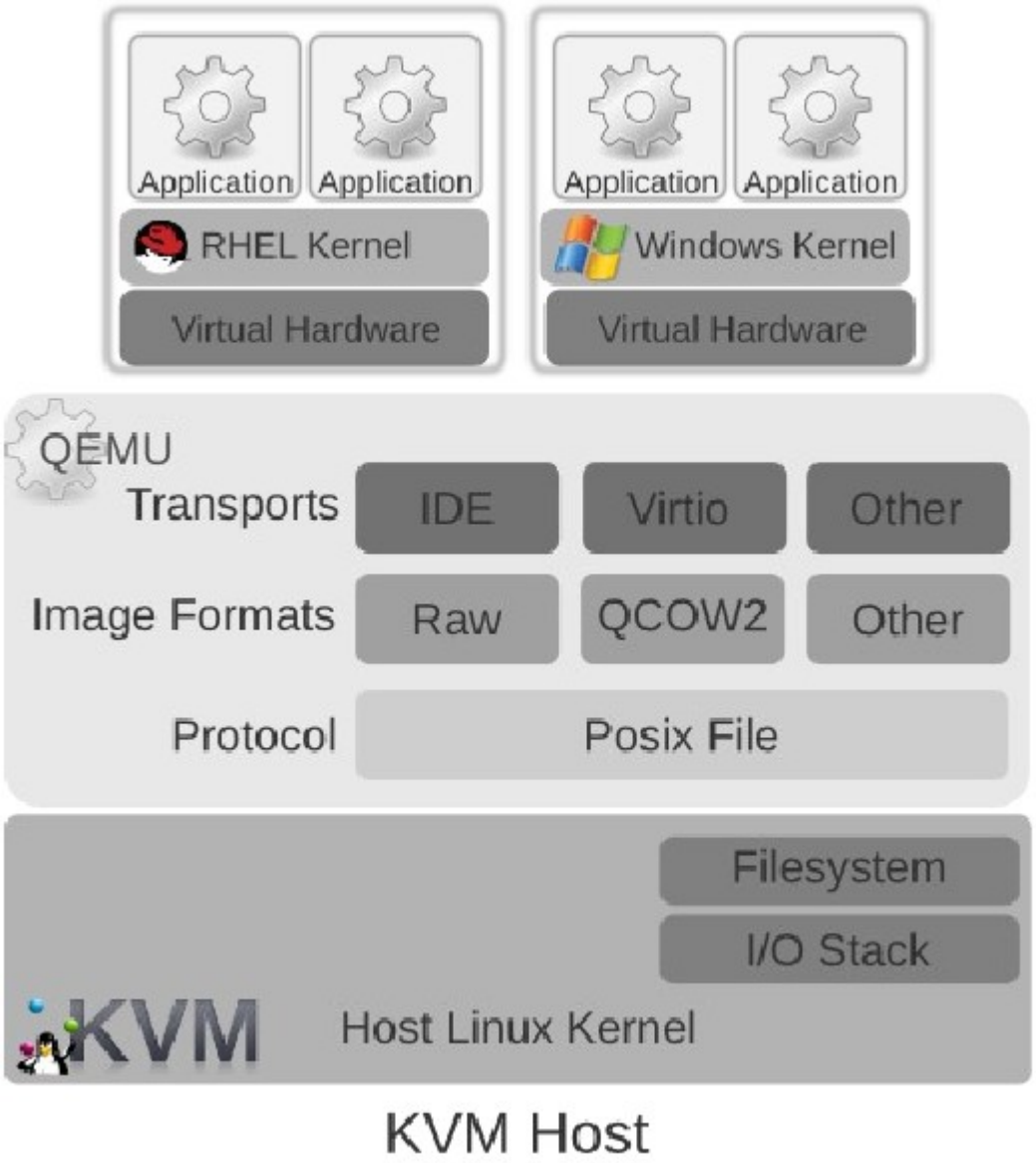


Networking

- Paravirt EOI (End of Interrupt) – reduces the number of EOI exits
- Zero copy tx bridge support – zero copy without macvtap

block

KVM Block Architecture



Virtio-scsi

- New KVM storage architecture based on SCSI
- Allows the usage of hundreds of devices per guest
- Supports SCSI pass-through and SCSI reservations
- Rich features - Feature set depends on the target, not on virtio-scsi
- Multipath: one virtio-scsi device = one SCSI host
- Multiple target choices: QEMU, lio
- Drop-in physical disk replacement
- True SCSI devices, good p2v/v2v migration

Live Block Copy

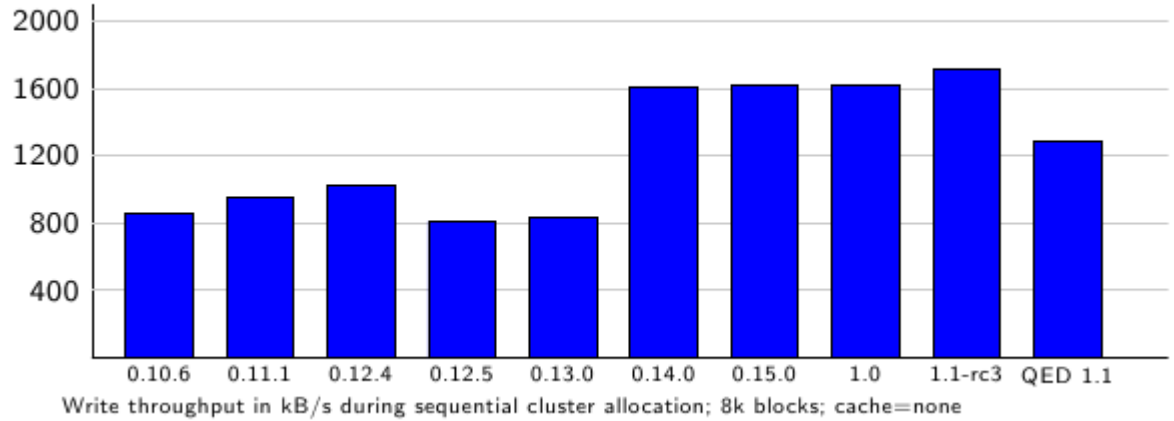
- Live block copy - copies guest image while the guest is running . You can use it to move a guest image to another location online.
- Image streaming – starts running the guest on a new location while the image is being copied to it.
- Live storage migration – migrates a guest with its image, a new implementation based on live block copy
- More today at 16:30 https://events.linuxfoundation.org/events/linuxcon-japan/bonzini

Block

- Coroutines – makes synchronous code asynchronous
- Qcow2 performance improvements
  - Zero/copy read/write
  - Introduces writeback meta data cache
  - Improves cluster allocation with writeback cache

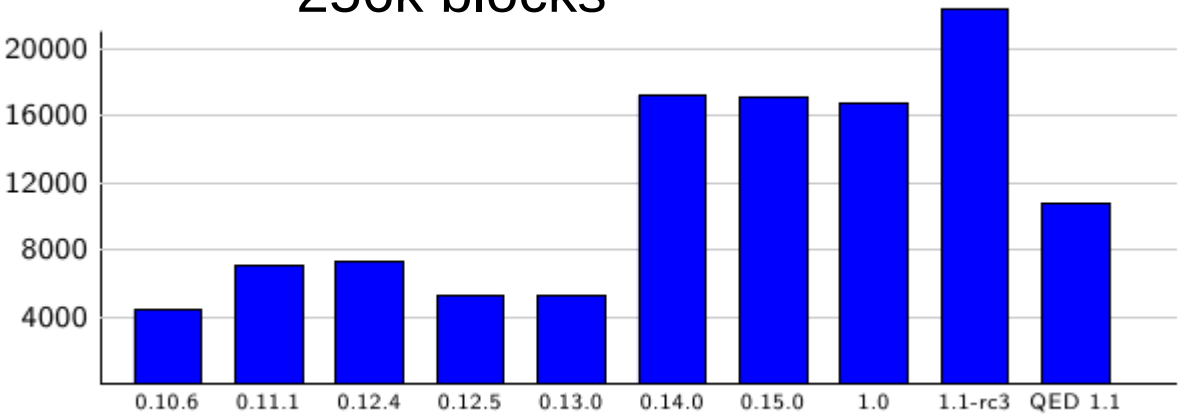
QCOW2 performance

8k blocks



Write throughput in kB/s during sequential cluster allocation; 8k blocks; cache=none

256k blocks



Write throughput in kB/s during sequential cluster allocation; 256k blocks; cache=none

RAS

RAS

- Power management for guests
  - Suspend to RAM (S3) and suspend to Disk (hibernate/S4)
- USB 2.0 support and SPICE improvements,
  - Remote wake up support which allows a suspended guest to resume from USB 2.0 devices
- Live migration improvements for boosting live migration convergence
  - Page delta compression
  - Migration thread
  - Post Copy
- New CPU models (Sandy/IvyBridge)

RAS

- Virtual CPU hot-plug
  - Host admin can dynamically adjust resources in the guests
- vPMU
  - Enable PMU on the guest for better guest profiling
  - Secure
  - Shareable
  - Model independent
- I/O throttling - Either through QEMU or cgroup.

VFIO – Virtual Function I/O for improved pci device assignment

- **VFIO – Virtual Function I/O**
  - Enhanced interrupt support
  - Virtualized PCI config space
  - Supports virtualization and userspace
  - VFIO is a device driver in the host
- **KVM device assignment (existing)**
  - PCI stub, PCI sysfs
  - Security
  - Depends on KVM
  - X86 only KVM is not a device driver (and should not be)

PCI Bus Enhancements

- New virtual platform chipset – q35
  - PCI-express bus support

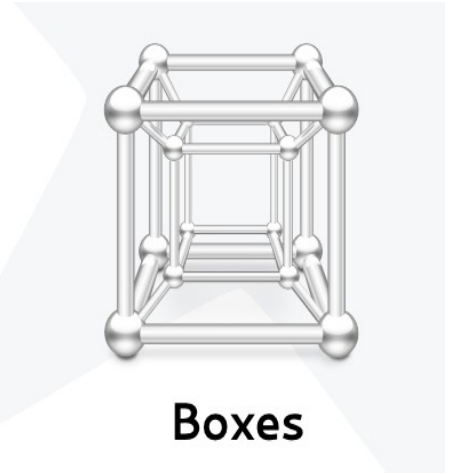
- PCI Bridge Support
  - Allows more than 32 PCI devices, each hot-pluggable

Security

- Sandbox virtual machines
  - Use new “seccomp” library to only allow certain syscalls to be executed

Desktop

- Boxes – application for managing virtual machines targeted towards typical desktop end-users https://live.gnome.org/Boxes



- Spice (Virtual Desktop Interface protocol)
  - New spice agent using GTK called spice-gtk.
  - Usbredir – Protocol for sending usb device traffic over a network connection

Cloud

Cloud

- Nested virtualization on Intel nVMX
- Nested TDP (Two Dimensional Page table) on AMD
- Open stack supports KVM

oVirt

- Open source Linux-based KVM virtualization project
- Provides a feature-rich server virtualization management system and advanced capabilities for hosts and guests.
- Includes high availability, live migration, storage management, system scheduler, and more.
- Come to the oVirt workshop on June 8th

Coming **soon**

- Virtio-net multiqueue (queue per guest virtual CPU)
- Guest memory hot plug
- EPT Access and Dirty bit
  - Important for KSM scanning mechanism
  - Needed to choose which guest pages are candidate for swap out

Coming **soon**

- QCOW2 format extensions
  - Qemu 1.1 has some basic support
  - zero clusters for keeping images sparse with copy-on-read/image streaming
  - It must be enabled explicitly during image creation (-o compat=1.1)
  - Images that use this new version cannot be read by older Qemu versions.

Weather is
Cloudy
with a chance of
total world domination


THANK YOU!