

# Virtualization at Scale

## in SUSE Linux Enterprise Server

**Jim Fehlig**

Software Engineer

[jfehlig@suse.com](mailto:jfehlig@suse.com)



# Agenda

- General guidelines
- Network guidelines
- Disk guidelines
- CPU and memory guidelines
- NUMA guidelines
- Scale-out considerations
- Scale-up considerations

# General Guidelines

- Minimize software installed on the host
  - Reduces resources
  - Reduces security risks/increases availability
- Synchronize time
  - Use NTP to synchronize time on the host AND guests
- Consider host resource requirements
  - Host uses resources too!
- Avoid over-allocating resources to guests
- Use paravirtual drivers for better performance

# General Guidelines

- Xen

- Disable autoballooning of domain0
  - Xen parameter 'dom0\_mem=xxG'
  - /etc/xen/xl.conf: autoballoon="off"
- Limit domain0 vcpus
  - Xen parameter 'dom0\_max\_vcpus=xx'
- Use tmpfs for xenstore db
  - Default configuration for SLES12 and newer

# Network Guidelines

- Use multiple networks to avoid congestion
  - admin, storage, live migration, ...
  - May require using `arp_filter` to prevent ARP flux

```
echo 1 > /proc/sys/net/ipv4/conf/arp_filter
```

- Same MTU in all devices to avoid fragmentation

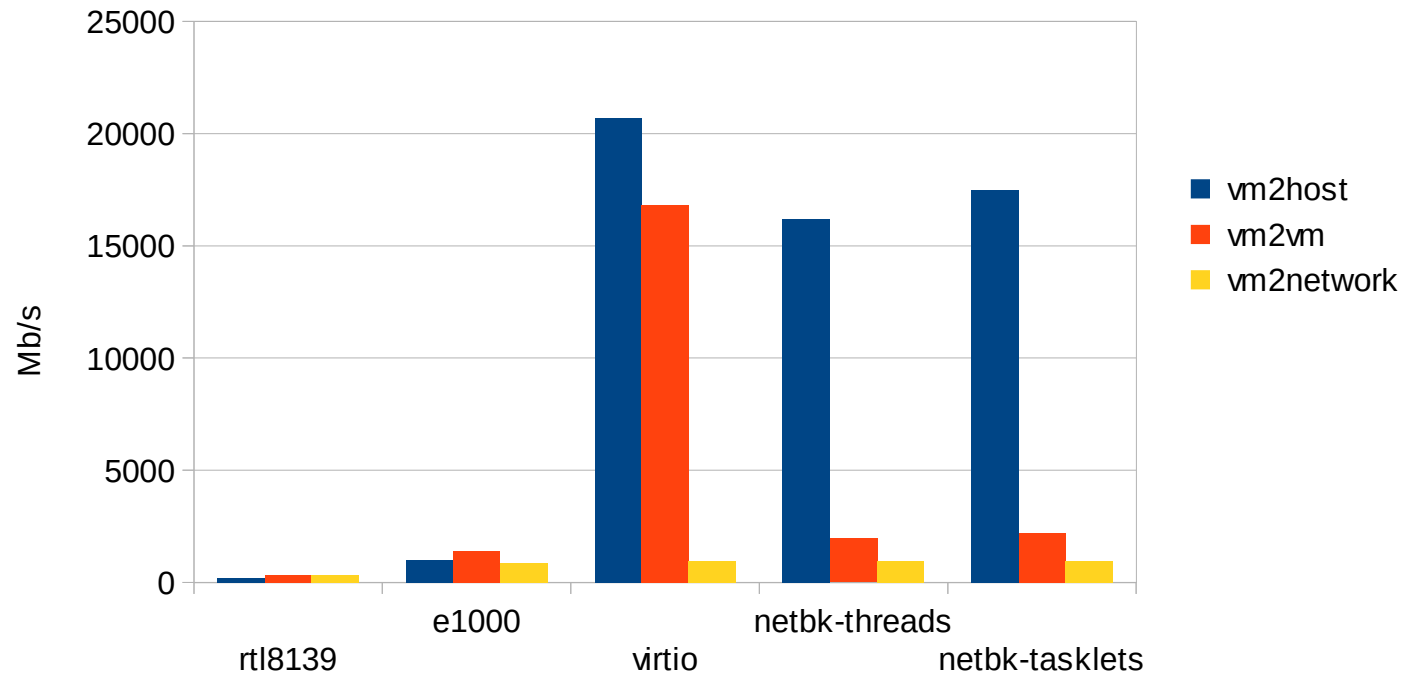
# Network Guidelines

- Virtual NICs
  - virtio-net (KVM)
    - multi-queue option
    - vhost-net
      - virtio-net accelerator (automatically loaded by libvirt, unless explicitly excluded)
  - netbk (Xen)
    - kernel threads vs tasklets
- Emulated NICs
  - e1000
    - Default and preferred emulated NIC
  - rtl8139

# Network Guidelines

## Bandwidth of vNICs

1Gb Network



# Network Guidelines

- Shared physical NICs
  - SR-IOV
  - macvtap
- Passthrough of physical NICs, aka PCI passthrough
  - Not supported by Intel due to security concerns
- Note: These approaches offer increased performance, but may complicate migration



# Disk Devices and Double Vision

- Two page caches
  - Two copies of data in memory
- Two IO schedulers
  - Guest and host both reordering and delaying IO
- Possibly two filesystems
  - Guest filesystem
  - Host filesystem containing the image
- Possibly two volume managers
  - Guest and host both using LVM
- Rx
  - Configure guest or host to bypass one of the redundant layers

# Disk Guidelines

## Block Devices -vs- Image Files?

- Block devices
  - Better performance
  - Use “standard” tools for administration/disk modification
  - Accessible from host (pro and con)
  - Eliminates one of the filesystems
- Image Files
  - Easier system management
    - Easier to move, clone, backup
  - Comprehensive toolkit (guestfs) for image manipulation
  - Fully allocated vs sparse
    - Performance vs resource consumption

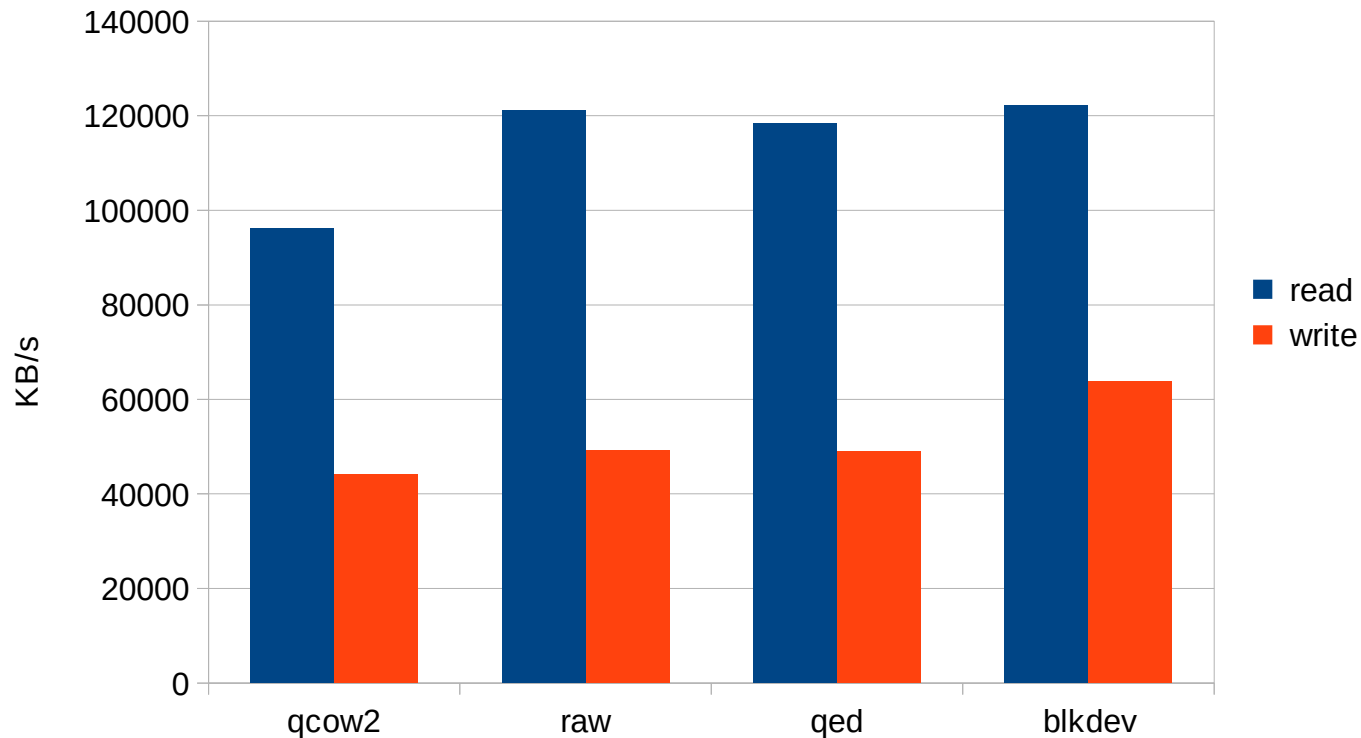
# Disk Guidelines

## Image Formats

- raw
  - Most common format
  - Historically, best performance
- qed
  - Next generation qcow
- qcow2
  - Required for snapshot support in libvirt + tools
  - Improved performance and stability
    - Compared to older versions of qcow2
- vhd/vhdx/vmdk/others
  - Suggest using for import/export only

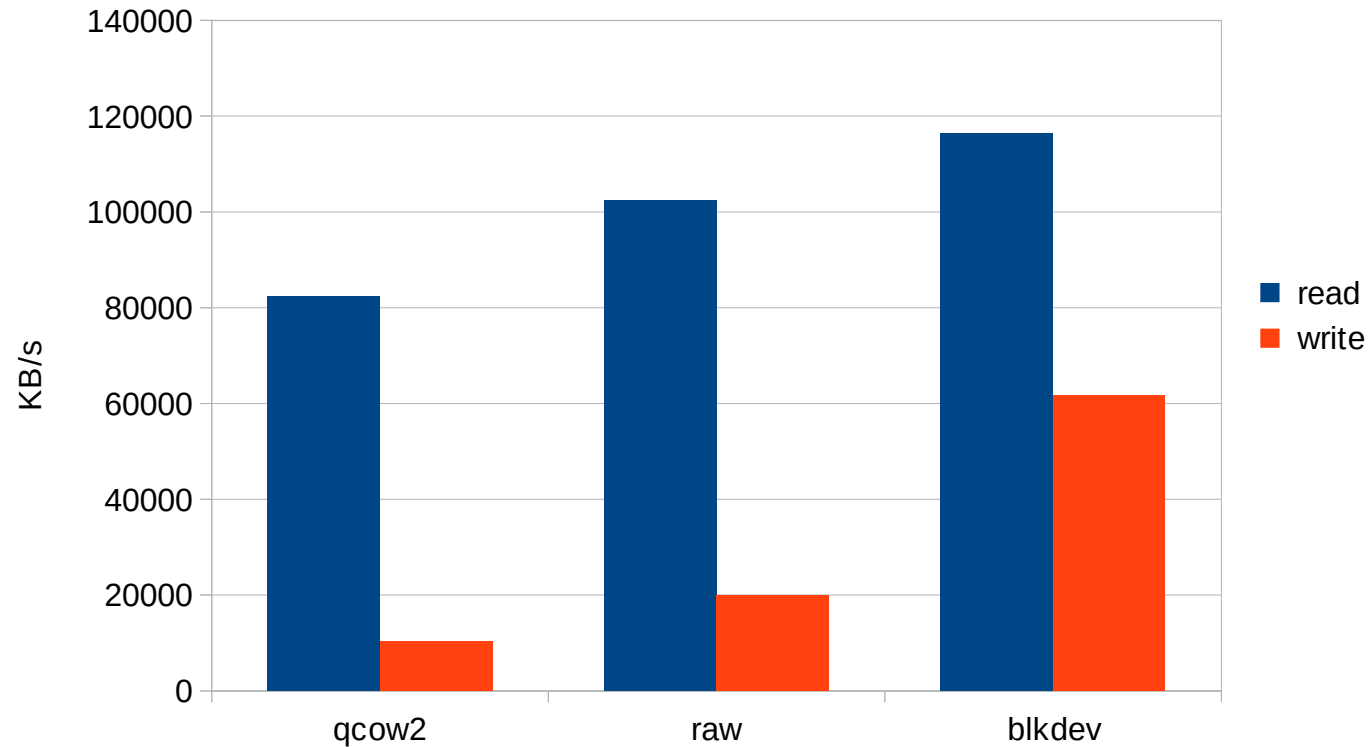
# Disk Guidelines

## Image File vs Block Device (KVM)



# Disk Guidelines

## Image File vs Block Device (Xen)



# Disk Guidelines

## Cache Modes

- writeback
  - Host page cache enabled
  - Disk write cache enabled
  - Guest flush commands honored
  - Default mode in KVM and Xen
- writethrough
  - Host page cache enabled
  - Disk write cache disabled
  - Guest informed no writeback cache

# Disk Guidelines

## Cache Modes

- `directsync`
  - Host page cache disabled
  - Disk write cache disabled
  - Writes reported completed only when committed to storage device
  - Useful for guests that don't send flush commands
- `None`
  - Host page cache disabled
  - Disk write cache enabled
  - `O_DIRECT` semantics
  - Guest informed of writeback cache

# Disk Guidelines

## Cache Modes

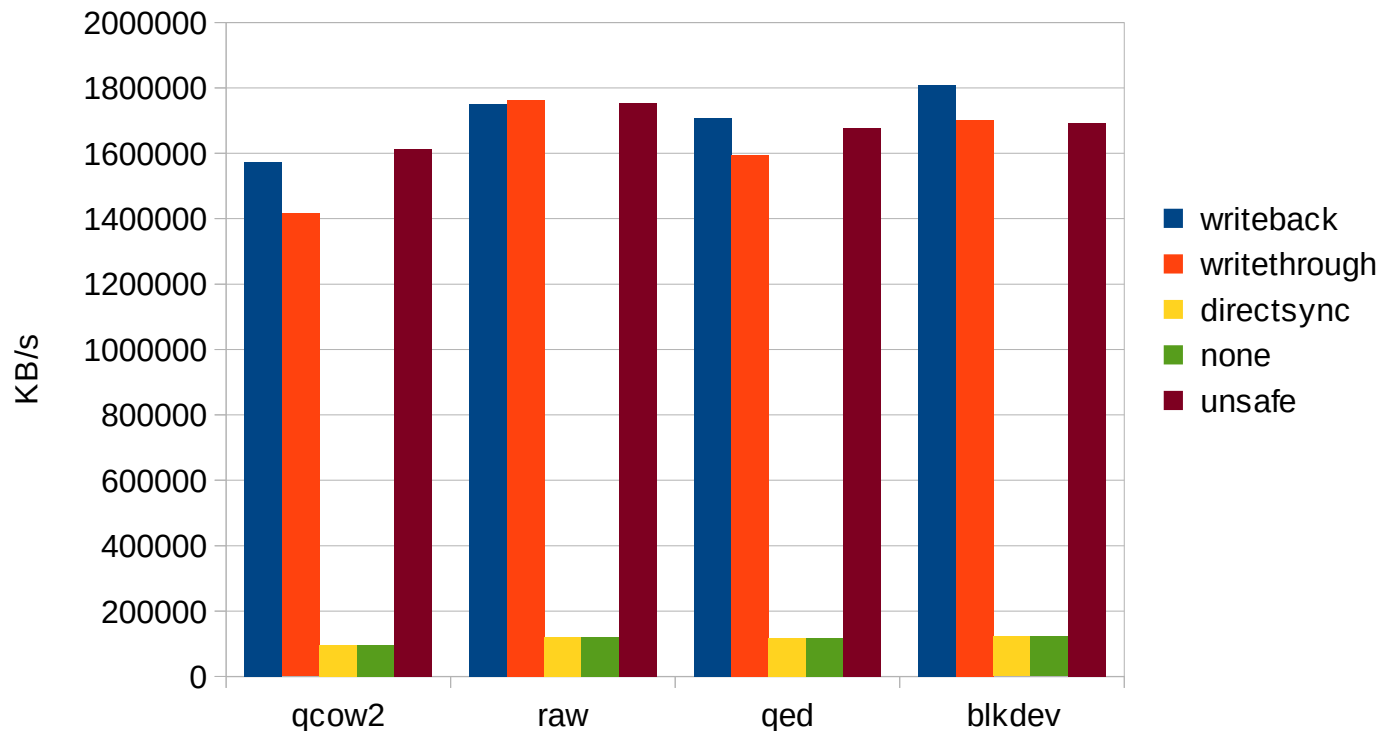
- Unsafe
  - Host page cache enabled
  - Disk write cache enabled
  - Guest flush commands ignored



# Disk Guidelines

## Cache Modes

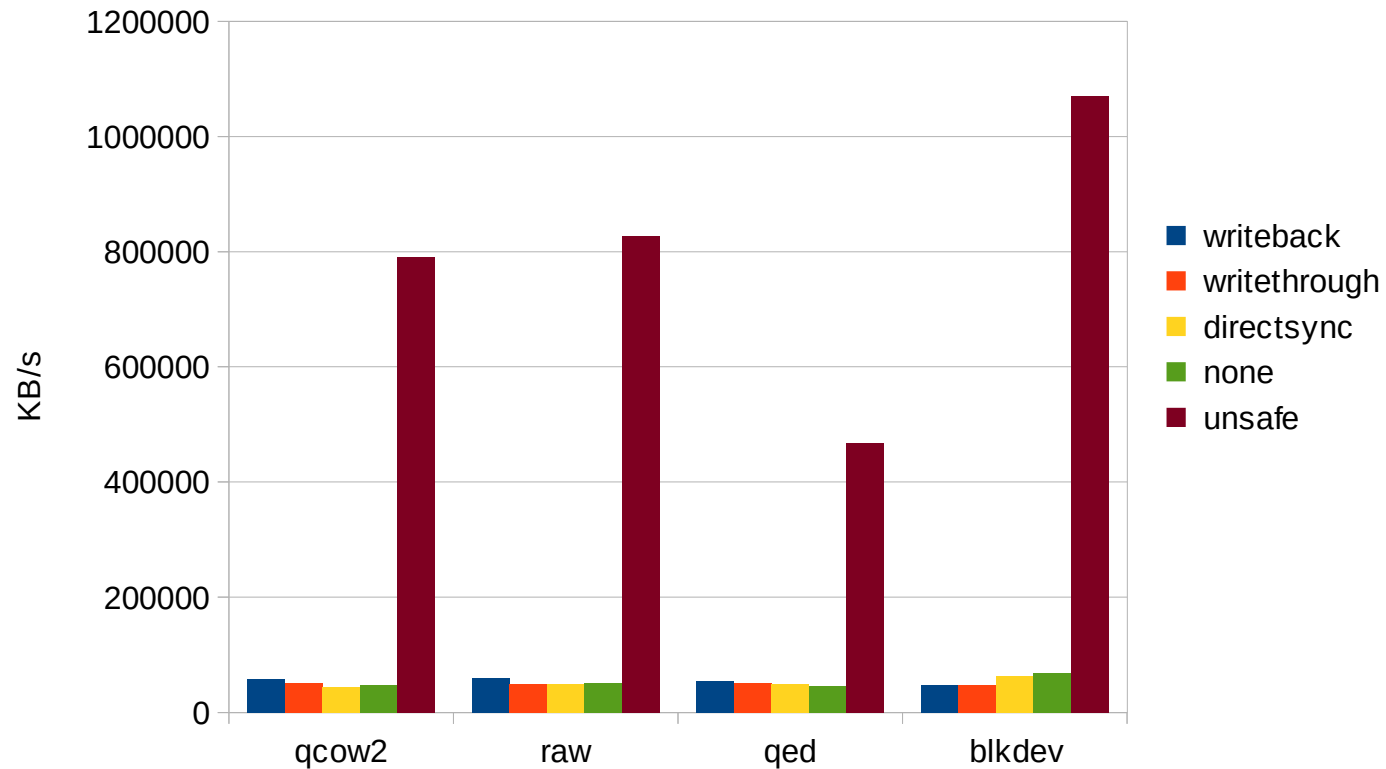
Cache Modes and Read Bandwidth (KVM)



# Disk Guidelines

## Cache Modes

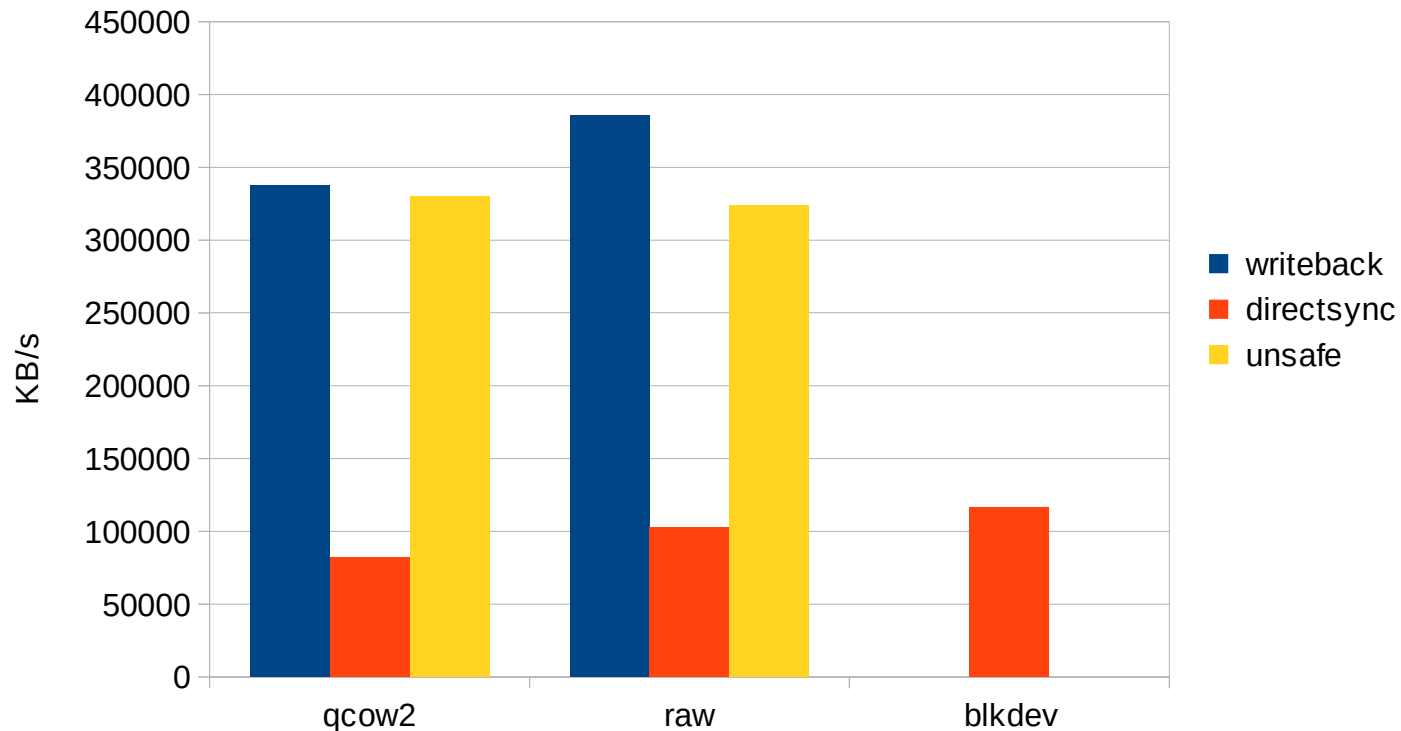
Cache Modes and Write Bandwidth (KVM)



# Disk Guidelines

## Cache Modes

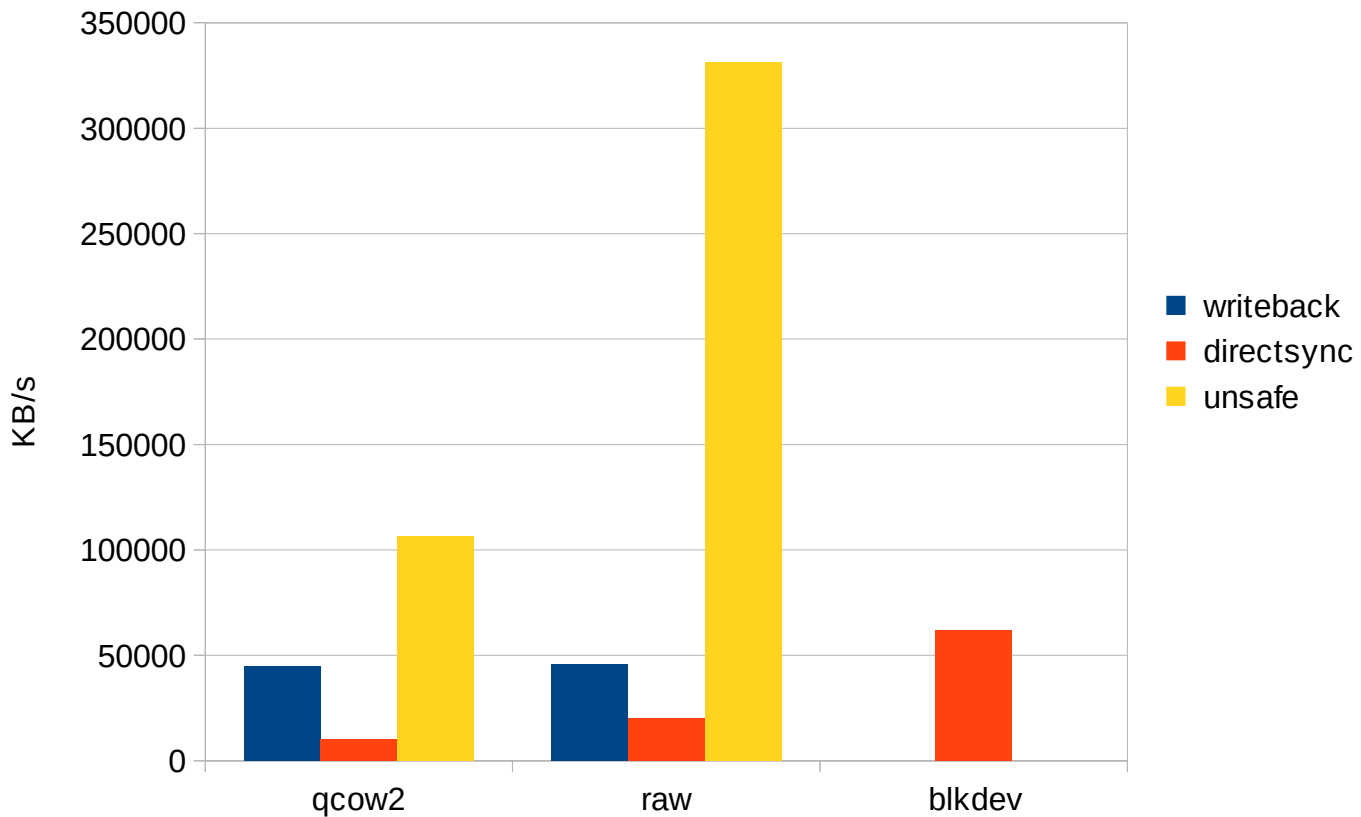
Cache Modes and Read Bandwidth (Xen)



# Disk Guidelines

## Cache Modes

Cache Modes and Write Bandwidth (Xen)



# Disk Guidelines

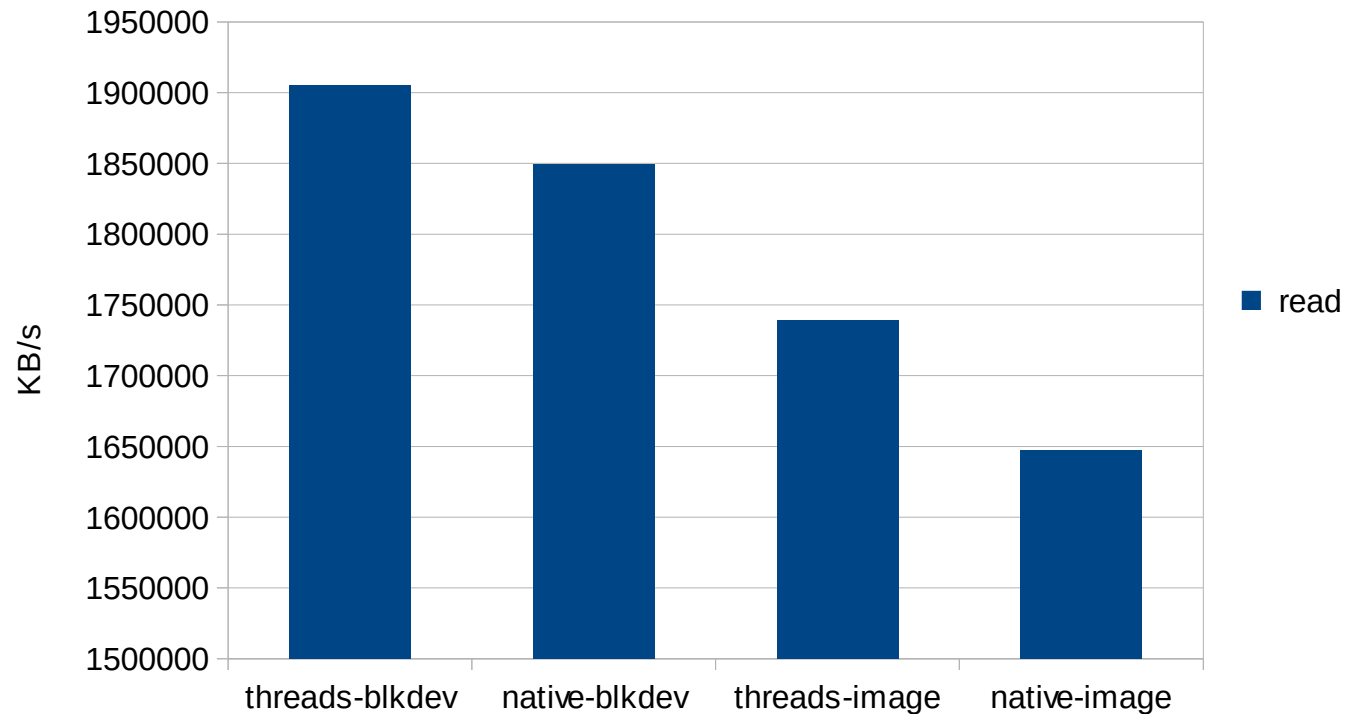
## IO Modes (KVM Only)

- native
  - Kernel asynchronous IO
  - Lower CPU overhead
- threads
  - Host user-mode based threads
  - Better throughput
  - Default mode in SLES

# Disk Guidelines

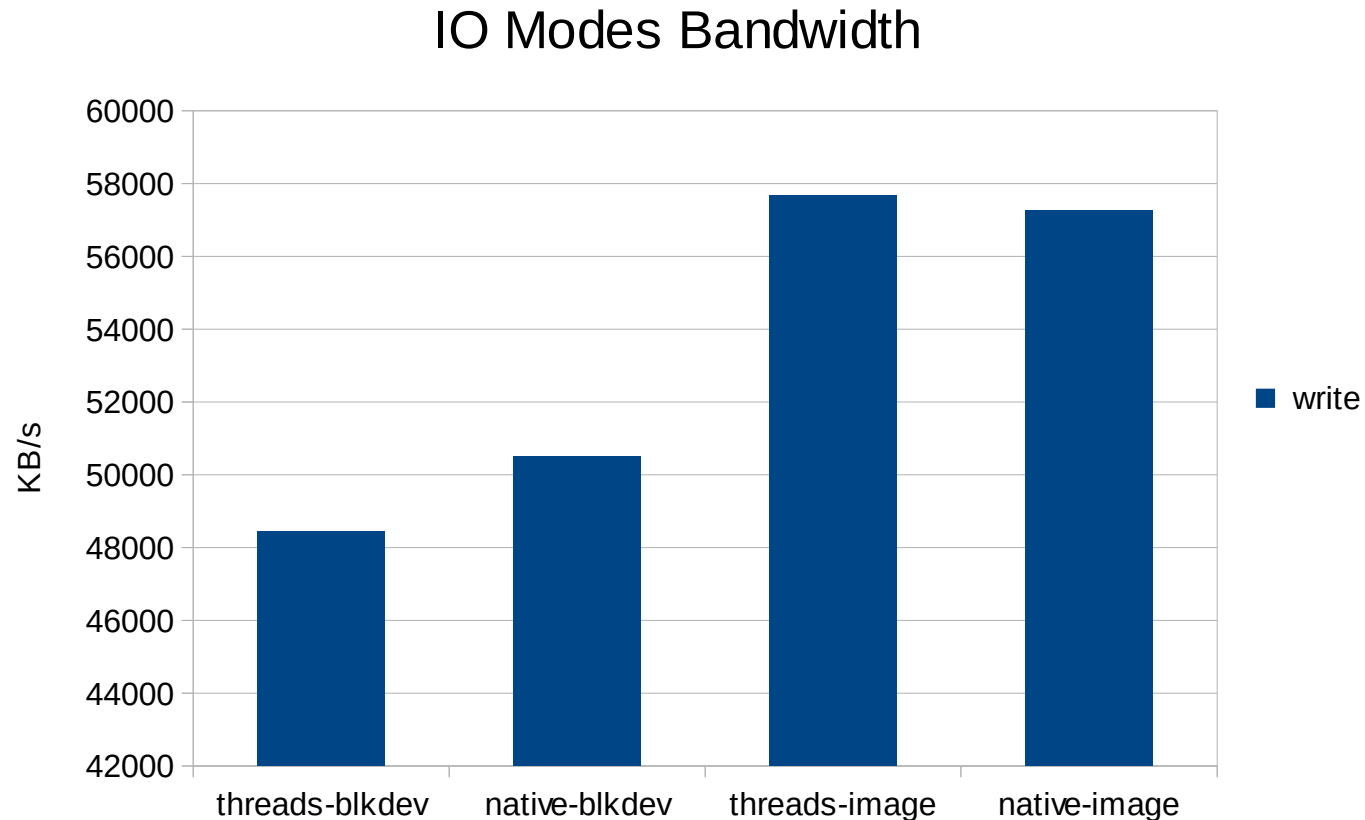
## IO Modes (KVM Only)

IO Modes Bandwidth



# Disk Guidelines

## IO Modes (KVM Only)



# Disk Guidelines

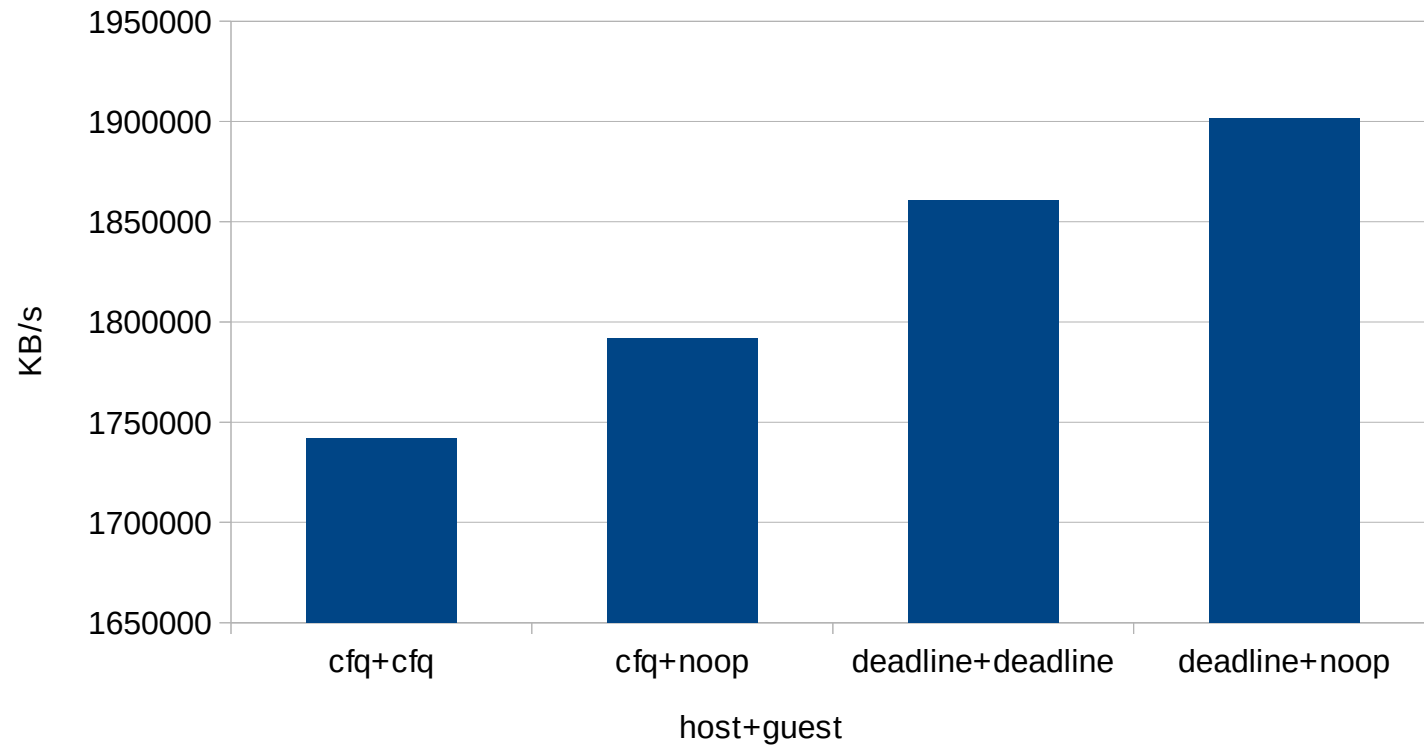
- I/O scheduler
  - Completely Fair Queuing (CFQ), deadline, noop
  - CFQ is default
  - Tunable per device
    - `echo noop > /sys/block/<device>/queue/scheduler`
  - Disable one of the schedulers
    - noop in the VM, deadline in the host
    - noop in the VM, CFQ in the host



# Disk Guidelines

## I/O Scheduler

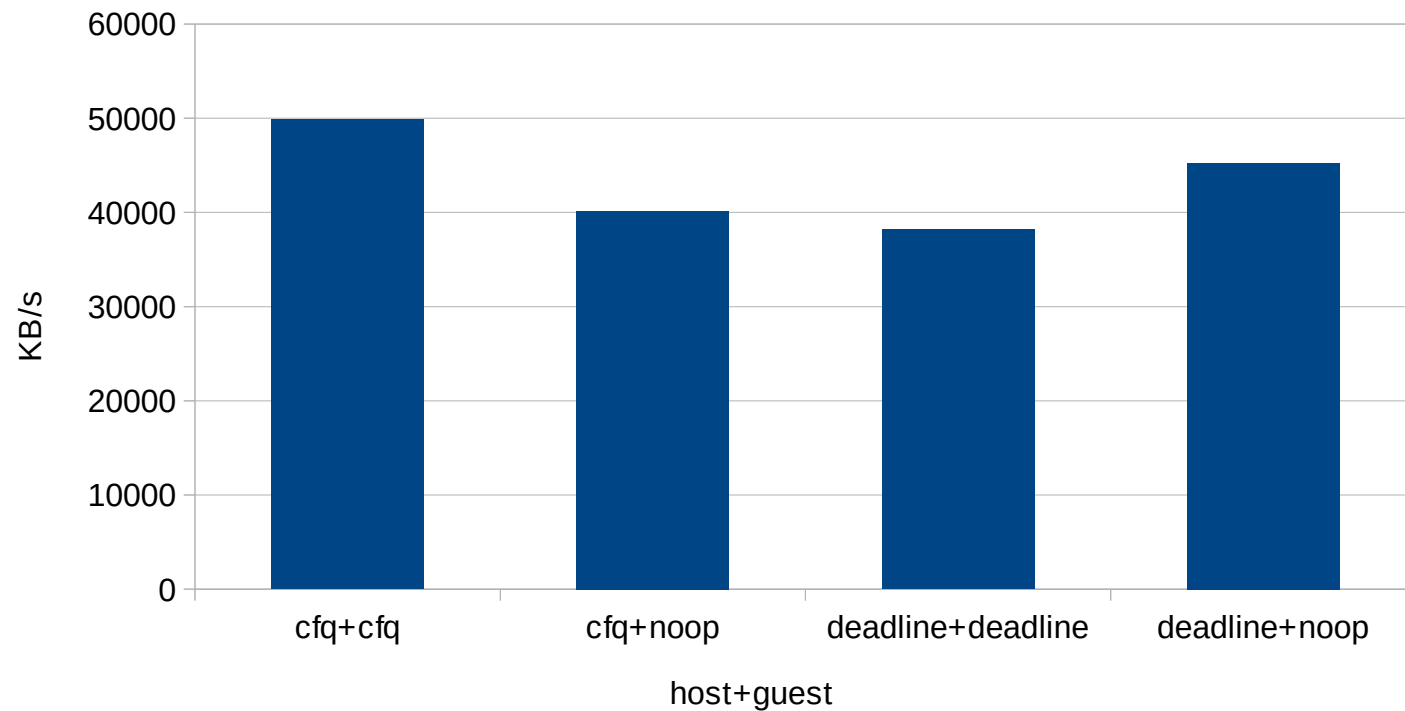
I/O Scheduler and Read Bandwidth



# Disk Guidelines

## I/O Scheduler

I/O Scheduler and Write Bandwidth



# CPU Guidelines

- Avoid CPU contention
  - Due to excessive CPU overcommit
- Scheduler
  - CFS tuned with kernel.sched\_\* parameters
- vCPU model and features
  - Normalize to allow migration among heterogeneous hosts
    - virsh capabilities | virsh cpu-baseline /dev/stdin >> all-hosts-cpu-caps.xml
    - virsh cpu-baseline all-hosts-cpu-caps.xml
- SLES12 Tuning Guide

[https://www.suse.com/documentation/sles-12/book\\_sle\\_tuning/data/book\\_sle\\_tuning.html](https://www.suse.com/documentation/sles-12/book_sle_tuning/data/book_sle_tuning.html)



# CPU Guidelines

- vCPU topology
  - Multiple sockets with a single core and thread, on the same NUMA node, generally give best performance

- vCPU Pinning

- Constrain vCPU threads to a NUMA node

```
<cputune>
```

```
<vcpupin vcpu='0' cpuset='0-15'/>
```

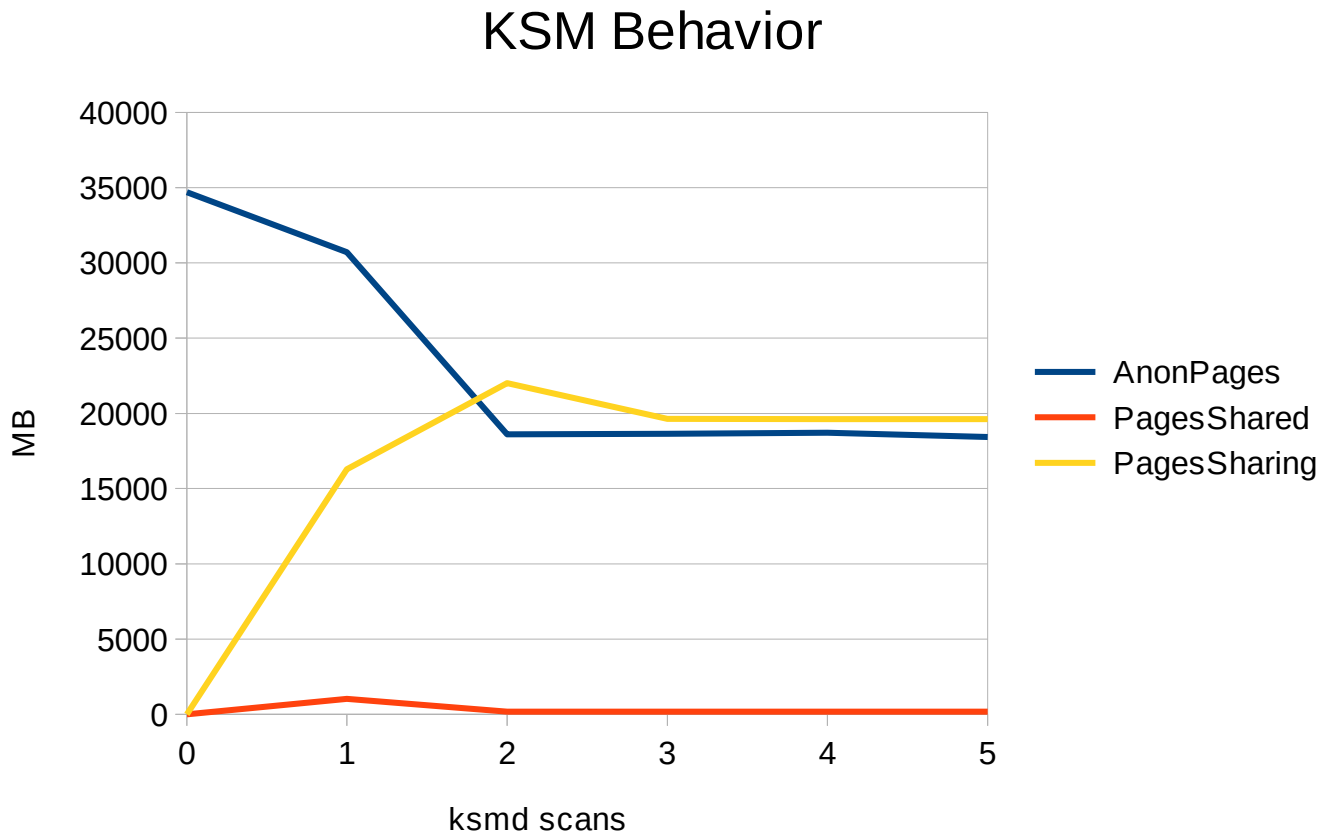
```
...
```

```
</cputune>
```

# Memory Guidelines

- Kernel Samepage Merging (KSM)
  - Memory-overcommit technique
  - Best results when running multiple instances of same image
  - ksmd thread consumes 5-10% of one core with default settings
  - `echo 1 > /sys/kernel/mm/ksm/run`
- Warning: By default, pages common across NUMA nodes are merged
  - Increased memory access latencies may be observed in VM
  - `echo 0 > /sys/kernel/mm/ksm/merge_across_nodes`

# Memory Guidelines - KSM



# Memory Guidelines

- Transparent Huge Pages (THP)
  - Enabled by default
  - Reduce Page Faults
    - 4K pages: 320343631
    - THP: 297214
  - Warning: Reduced performance for workloads with sparse access patterns (databases)
- Hugepages and hugeTLB
  - Manually manage hugepage allocation and use
- SLES12 Tuning Guide

[https://www.suse.com/documentation/sles-12/book\\_sle\\_tuning/data/book\\_sle\\_tuning.html](https://www.suse.com/documentation/sles-12/book_sle_tuning/data/book_sle_tuning.html)



# NUMA Guidelines

- Potentially huge impact on performance
- Consider host topology when sizing guests
  - `virsh {nodeinfo, capabilities, freecell}`
- Prevent vCPUs from floating across NUMA nodes
  - vCPU pinning
- Avoid allocating VM memory across NUMA nodes

```
<numatune>
```

```
  <memory mode='strict' nodeset='1' />
```

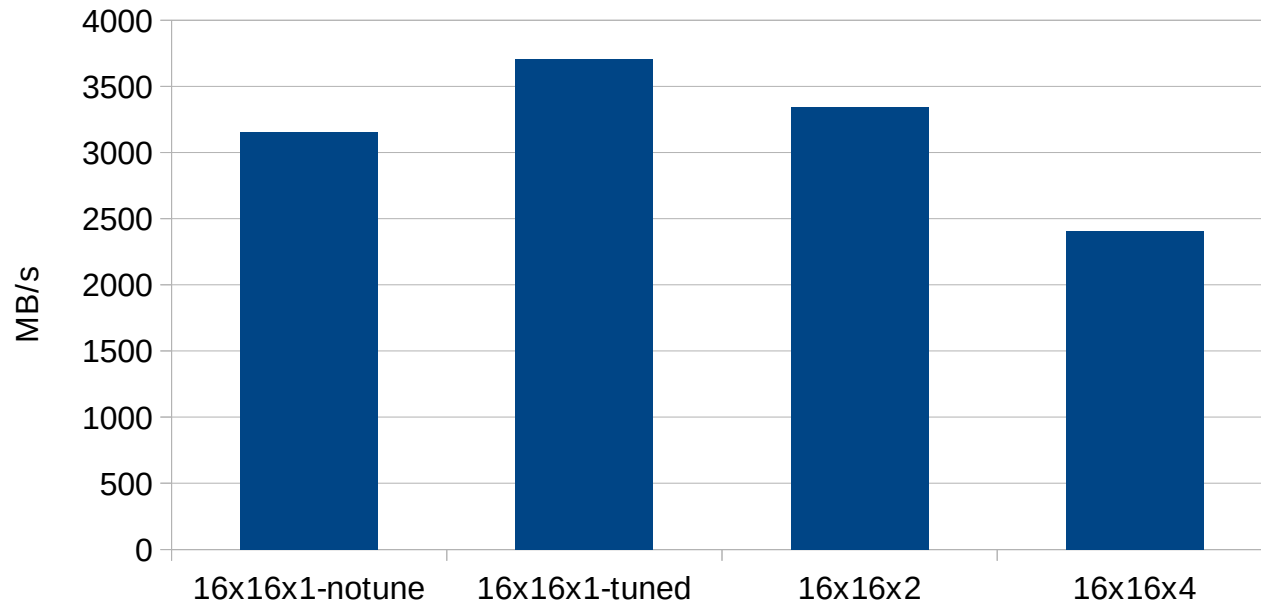
```
</numatune>
```



# NUMA Guidelines

## KVM Memory Bandwidth

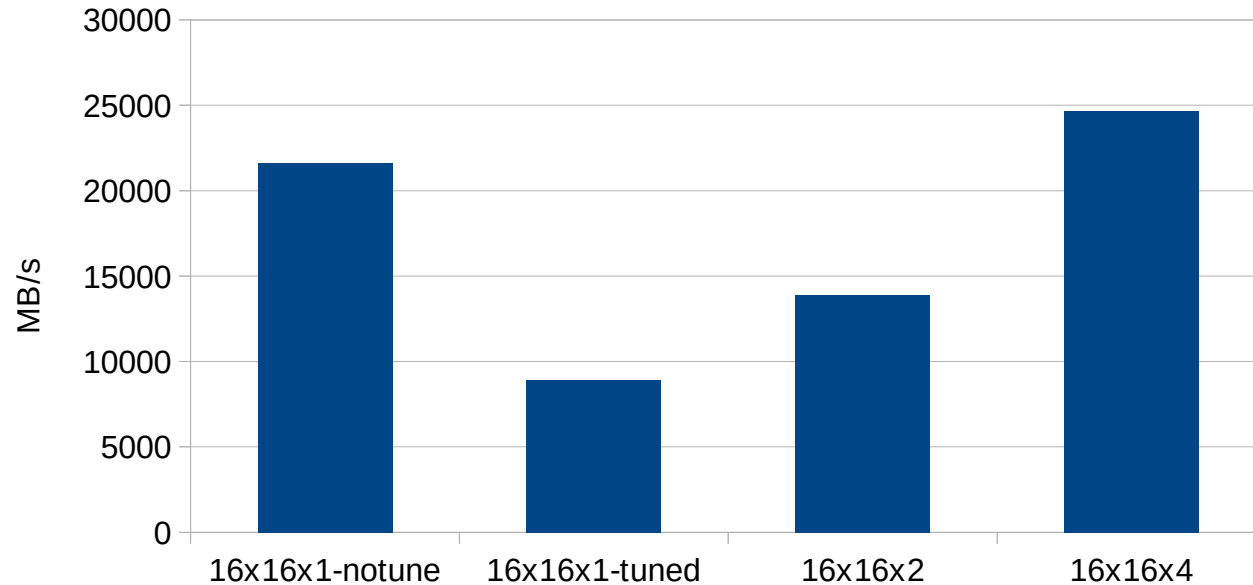
1 process, 12G



# NUMA Guidelines

## KVM Memory Bandwidth

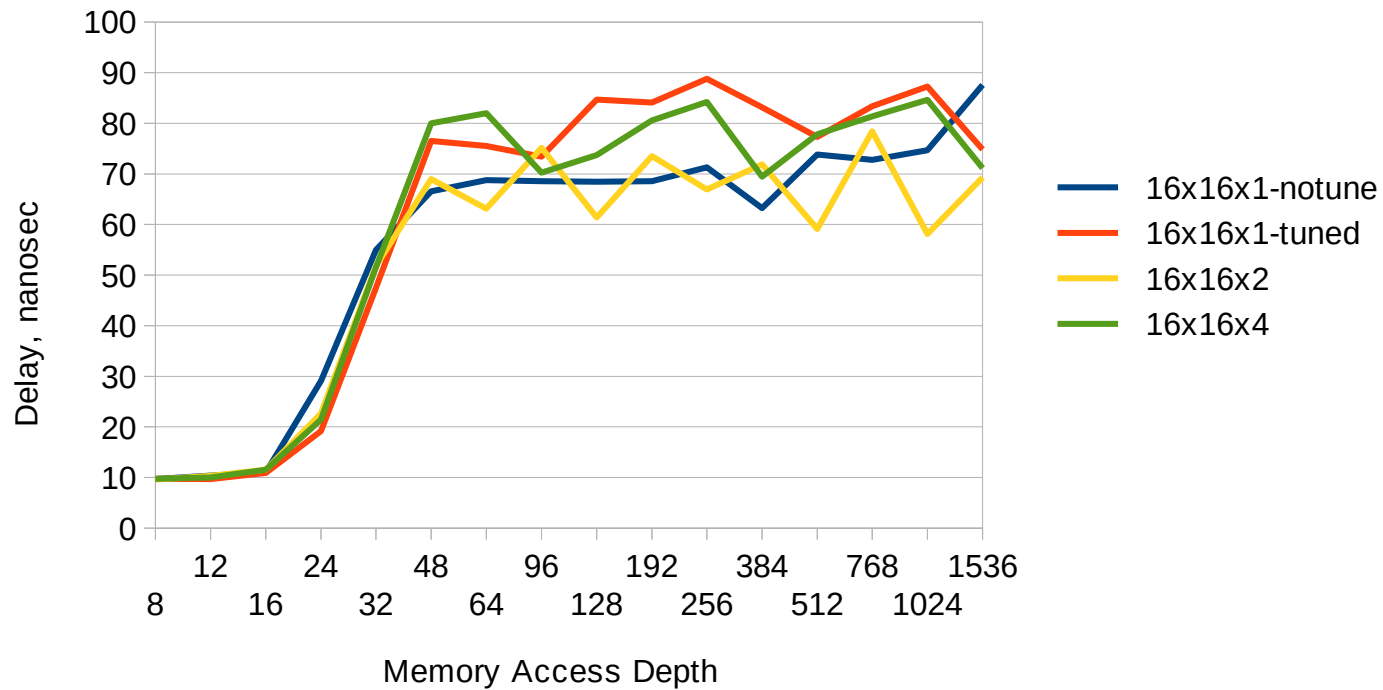
16 processes, 1G each



# NUMA Guidelines

## KVM Memory Latency

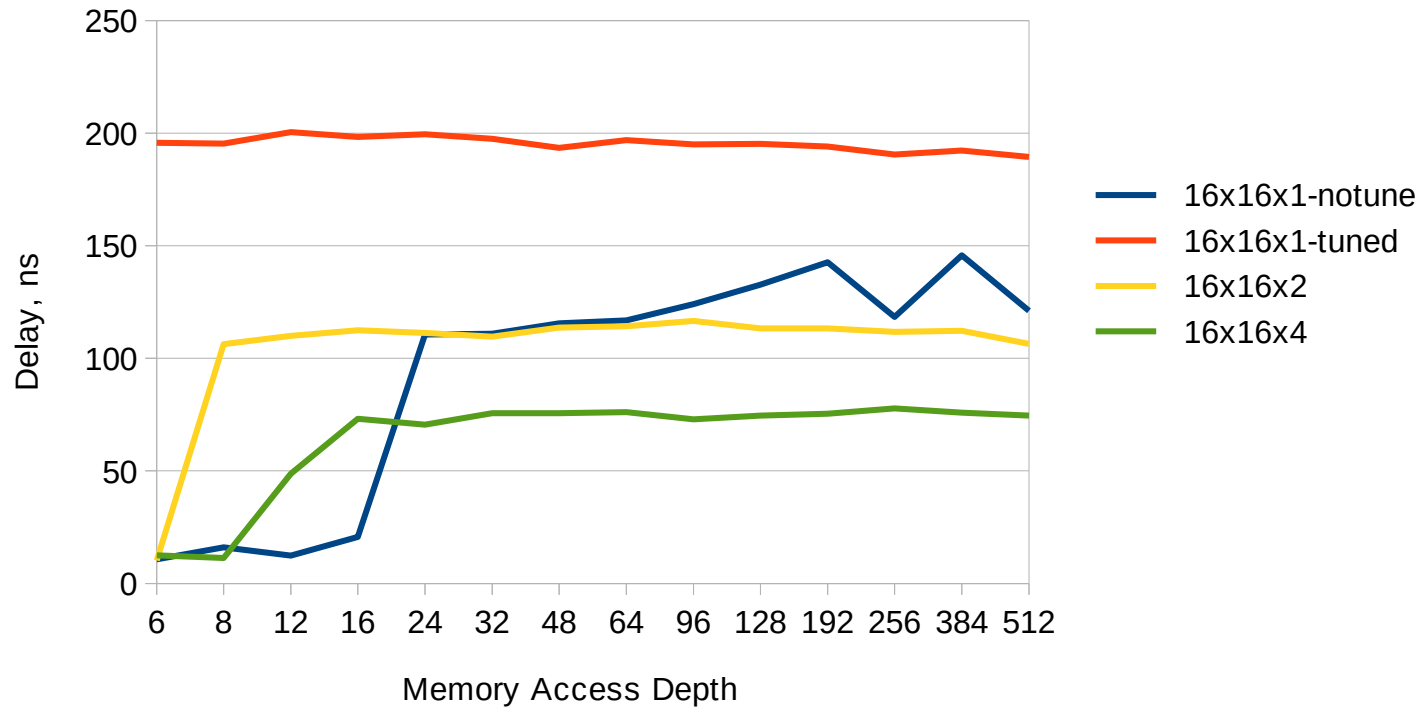
1 process



# NUMA Guidelines

## KVM Memory Latency

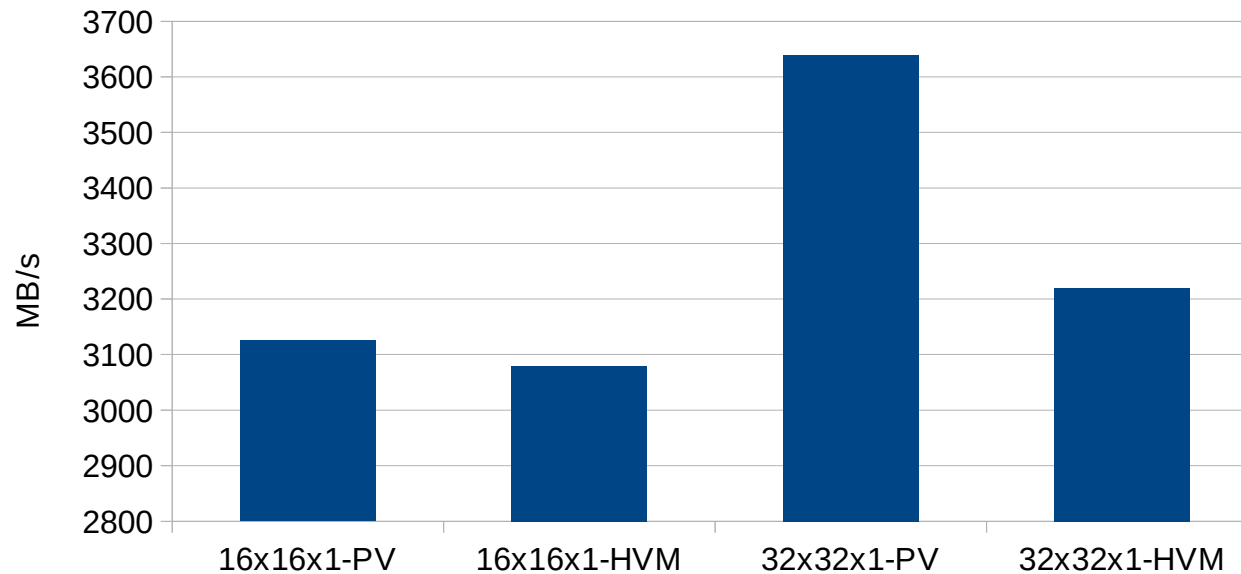
16 processes



# NUMA Guidelines

## Xen Memory Bandwidth

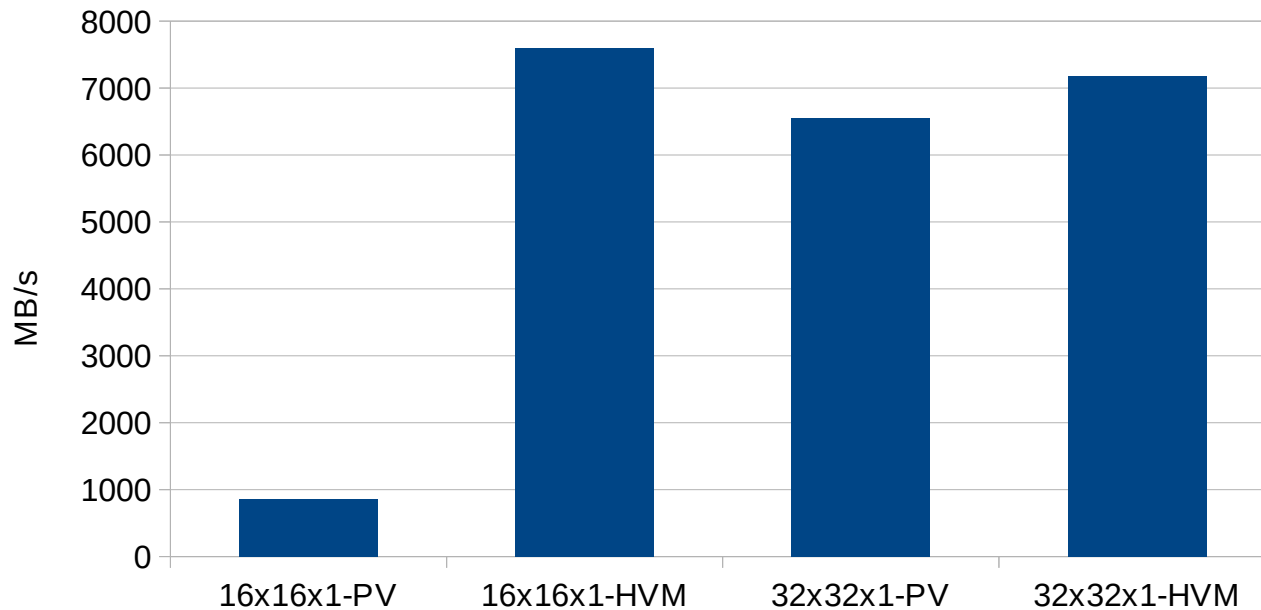
1 process, 12G working set



# NUMA Guidelines

## Xen Memory Bandwidth

16 processes, 1G working set



# Scale-out Considerations

- Size VMs to fit on a single NUMA node
- Confine VMs to a single NUMA node
  - Don't pin vCPU to a single pCPU
  - Do pin vCPU to all pCPUs on a NUMA node
  - Do use memory from same NUMA node where vCPUs run
- Use block devices with 'native' IO mode
- Use vhost-net for network devices
- Control VM resource consumption
  - vCPU shares and periods, IO tuning, etc...

# Scale-out Considerations

- Don't resource-starve host
  - Host services I/O from all VMs
  - Limits may be exceeded, e.g. loop devices
- Host virtualization toolstack less responsive
  - Busy gathering stats from all block device in use by all VMs



# Scale-up Considerations

- vNUMA
  - Expose host NUMA to VM
  - Virtual NUMA node, CPUs, and memory logically mapped to corresponding host resources
- virtio-net with multiqueue (+vhost-net)
  - Possible to have an rx/tx queue for each vCPU
- Use block devices with 'threads' IO mode
- Increased time to create/destroy/migrate large VMs
  - Xen scrubs all VM memory on destroy

Have a lot of fun!  
[www.suse.com/virtualization](http://www.suse.com/virtualization)

Thank you.





## **Unpublished Work of SUSE LLC. All Rights Reserved.**

This work is an unpublished work and contains confidential, proprietary and trade secret information of SUSE LLC. Access to this work is restricted to SUSE employees who have a need to know to perform tasks within the scope of their assignments. No part of this work may be practiced, performed, copied, distributed, revised, modified, translated, abridged, condensed, expanded, collected, or adapted without the prior written consent of SUSE. Any use or exploitation of this work without authorization could subject the perpetrator to criminal and civil liability.

## **General Disclaimer**

This document is not to be construed as a promise by any participating company to develop, deliver, or market a product. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. SUSE makes no representations or warranties with respect to the contents of this document, and specifically disclaims any express or implied warranties of merchantability or fitness for any particular purpose. The development, release, and timing of features or functionality described for SUSE products remains at the sole discretion of SUSE. Further, SUSE reserves the right to revise this document and to make changes to its content, at any time, without obligation to notify any person or entity of such revisions or changes. All SUSE marks referenced in this presentation are trademarks or registered trademarks of Novell, Inc. in the United States and other countries. All third-party trademarks are the property of their respective owners.

