

News from the 2010 RedHat Summit

James Pryor, Jason Smith (BNL)

Date: July 12th, 2010

Time: 1:00 PM (ET)

Location: ITD Seminar Room

SUMMIT

**JBoss
WORLD**

PRESENTED BY RED HAT



SUMMIT JBoss WORLD
PRESENTED BY RED HAT
**LEARN.
NETWORK.**

CH BIN
JE SUIS
SOY RED HAT
AMEU SOU
OTO R
JSEM
SONO
CH BIN
JE SUIS
SOY RED HAT
AMEU SOU
OTO R
JSEM

SUMMIT JBoss WORLD
PRESENTED BY RED HAT
**EXPERIENCE
OPEN SOURCE.**

SUMMIT JBoss WORLD
PRESENTED BY RED HAT
**LEARN.
NETWORK.**

SUMMIT JBoss WORLD
PRESENTED BY RED HAT
**EXPERIENCE
OPEN SOURCE.**

SUMMIT JBoss WORLD
PRESENTED BY RED HAT
LEARN. NETWORK. EXPERIENCE OPEN SOURCE.

SUMMIT JBoss WORLD
PRESENTED BY RED HAT
**EXPERIENCE
OPEN SOURCE.**

SUMMIT JBoss WORLD
PRESENTED BY RED HAT
**LEARN.
NETWORK.**

SUMMIT JBoss WORLD
PRESENTED BY RED HAT
**EXPERIENCE
OPEN SOURCE.**

SUMMIT JBoss WORLD
PRESENTED BY RED HAT
**EXPERIENCE
OPEN SOURCE.**

CH BIN
JE SUIS
SOY RED HAT
AMEU SOU
OTO R
JSEM
SONO
CH BIN
JE SUIS
SOY RED HAT
AMEU SOU
OTO R
JSEM

SUMMIT JBoss WORLD
PRESENTED BY RED HAT
**LEARN.
NETWORK.**

SUMMIT

**JBoss
WORLD**

PRESENTED BY RED HAT

**LEARN. NETWORK.
EXPERIENCE OPEN SOURCE.**

www.theredhatsummit.com

RHEL 6 Schedule

- Beta 1 – public availability April 2010 via ftp and RHN for customers
- Beta 2 – SOON! shortly after RH Summit
 - Jump in! We'd love to get your feedback
- Release – later this year
 - Remember, RHEL 6 becomes available to all customers with active RHEL subscriptions.

SUMMIT

**JBoss
WORLD**

PRESENTED BY RED HAT



Major RHEL 6 release themes

Optimized foundation OS

For large-scale, centrally-managed enterprise deployments. Lower Total Cost of Ownership. Secure. Optimized for maximum efficiency of latest generation of high core-count systems – memory, scalability, RAS, power efficiency. Resource control.

Virtualization – optimize RHEL as host or guest

Deployment, provisioning and flexibility for dynamic workloads from datacenter to desktop. Emphasis on performance, storage flexibility, security, and guest isolation.

Green IT

Through power management and dynamic guest migration

Innovation and technology leadership

With the latest enterprise ready components in Storage and File Systems, Networking, Tools, Cluster, Desktop, Installer, and Services. Providing customer access to leadership technology throughout the RHEL product lifecycle.

SUMMIT

**JBoss
WORLD**

PRESENTED BY RED HAT



RHEL 6 foundation features

- **Virtualization – making RHEL an optimized host & guest**
 - KVM
 - Industry leading virt performance, flexibility, security for both host & guest environments
 - Device I/O optimization
- **Improved manageability**
 - For large scale virtualization deployments, server & desktop
 - RHEV-M (virt/cloud management) enablers
 - Samba enhancements for Windows active directory and file sharing
- **Power management**
 - Efficiency – lower deployment costs, reduced carbon footprint
 - For virt, bare metal, laptop
 - Hardware level as well as dynamic system service startup and suspend

SUMMIT

**JBoss
WORLD**

PRESENTED BY RED HAT



RHEL 6 foundation features (continued)

- **RAS (Reliability, Availability, Serviceability)**
 - Hotplug, memory error reporting, filesystem data integrity
 - Support tools – automated crash detection and bug reporting infrastructure
- **Hardware enablement and scalability**
 - Maximum efficiency with latest generation of highly scalable servers with headroom to grow
 - Large configurations (cpu, memory, busses, I/O), NUMA awareness
 - UEFI – new bios boot loader interface
 - Supported architectures: x86, x86-64, PPC64, s390x
- **Desktop**
 - VDI- virtualized thin client, SPICE integration
 - Mobility – dynamic network config
 - Display – external monitors, multihead, projectors, docking station

SUMMIT

**JBoss
WORLD**

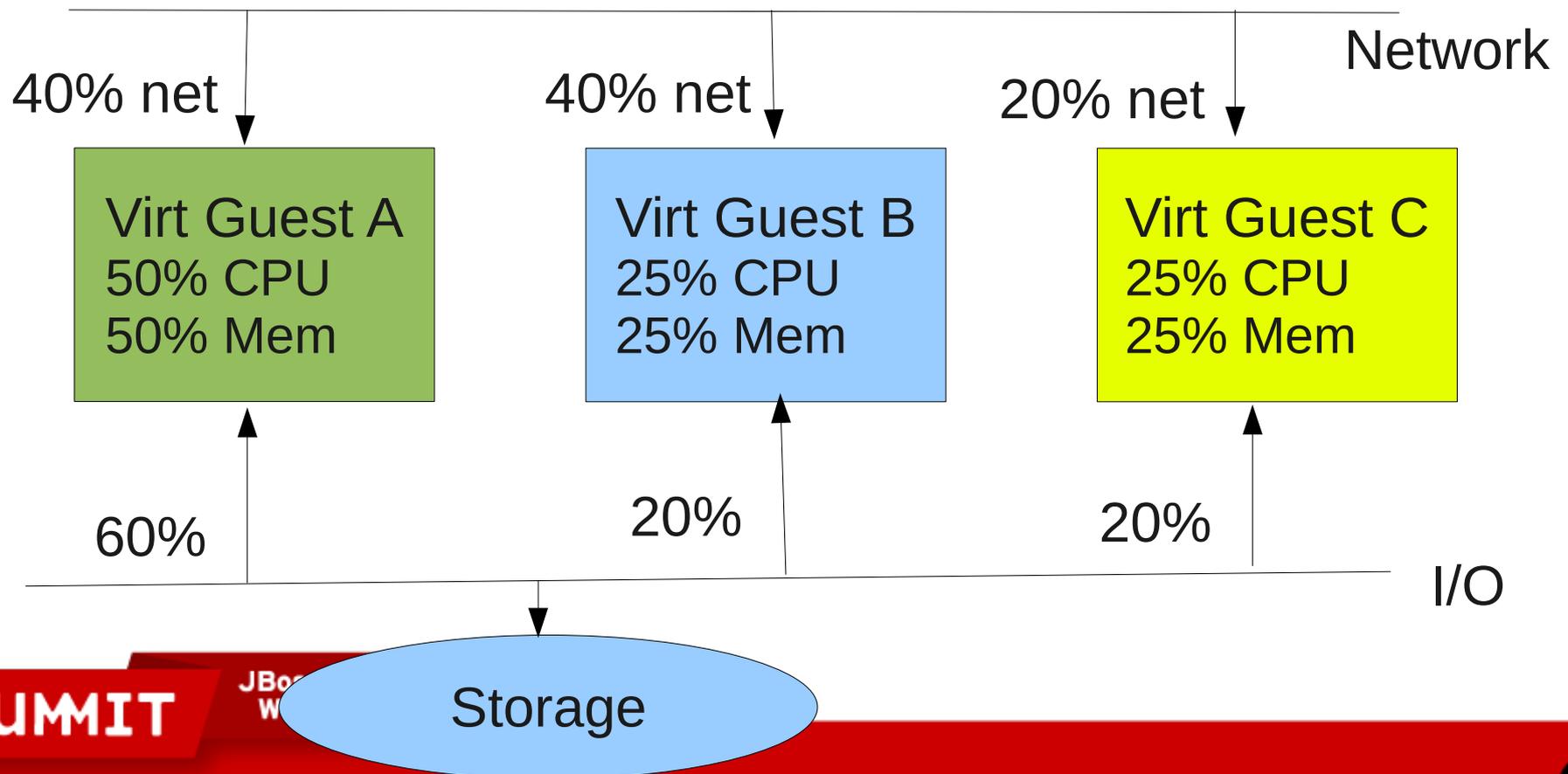
PRESENTED BY RED HAT



Kernel Resource Management

- **Illustrative cgroup use cases**

- Database workload dedicated 90%, background backup utility 10%
- Virtualization hosting provider – allows QoS (quality of service guarantees based on pricepoint)



SUMMIT

JBoss
W

Storage

PRESENTED BY RED HAT



Kernel resource management

- **Cgroup – Control group**

- A control group provides a generic framework where several “resource controllers” can plug in and manage different resources of the system such as process scheduling, memory allocation, network traffic, or IO bandwidth.
 - Can be tracked to monitor system resource usage
 - Sysadmin can use tools to allow or deny these groups access to resources

- **Memory resource controller**

- Isolates the memory behavior of a group of tasks – cgroup – from the rest of the system (including paging). It can be used to:
 - Isolate an application or a group of applications
 - Create a cgroup with limited amount of memory

- **Cgroup scheduler**

- CFS – Hierarchical proportional fair scheduler (SCHED_OTHER)
- Static priority scheduler with constant bandwidth limits (SCHED_FIFO)

SUMMIT

**JBoss
WORLD**

PRESENTED BY RED HAT



Kernel resource management (continued)

- **I/O controller**

- Designate portion of I/O bandwidth (based on controller queue depth)

- **Network controller**

- Define classes & queues between generic network layer and NIC. Tagging packets with class identifier with different priorities, placing outbound packets in different queues for traffic shaping.

- **Libcgroup**

- SELinux policy
- Cgroup creation, deletion, move and configuration management.
- Rules based automatic task placement, PAM module, daemon, uid/gid based rules

- **Illustrative cgroup use cases**

- Database workload dedicated 90%, background backup utility 10%
- Virtualized hosting provider – allows QoS (quality of service guarantees based on pricepoint)

SUMMIT

**JBoss
WORLD**

PRESENTED BY RED HAT



Power management

- **Objective** — reduced deployment costs through efficiency
- **Kernel**
 - Tickless kernel – fewer interrupts, more idle time to drop to lower power states – x86/x86-64 only
 - ASPM (Active State Power Management) – PCI Express reduced power states on inactivity
 - ALPM (Aggressive Link Power Management) – SATA links in low power mode when no I/O pending
 - Energy efficient turbo and deep C states
 - **Relatime** drive access optimization reducing filesystem metadata write overhead
 - Graphics power management
- **System services / daemons**
 - Intelligent drive spin down
 - Application audit and redesign where necessary to be event based rather than needless polling
 - TuneD – adaptive tuning daemon – powerdown idle peripherals & latency policy scripts. Providing a variety of power tuning pre-canned profiles
- **Virtualization management – RHEV-M integration**
 - Systems which are powered off are the most efficient. Workload consolidation with power management automatic migration policy

SUMMIT

**JBoss
WORLD**

PRESENTED BY RED HAT



Kernel scalability

- **Objective** – providing **scaling headroom** anticipating many years of upcoming hardware generations. Tested and supported limits will likely grow over the course of product lifespan.
- **Scalability features** – enhancements in algorithms. Applicable to bare-metal and virtualized guests
 - Split LRU VM – different eviction policies for file backed vs swap backed
 - Ticket spinlocks fixes NUMA starvation
 - CFS scheduler – better NUMA balancing
 - UEFI boot loader install & boot support on > 2TB disk partitions
- Virtualization scalability accommodates running older releases on newer hardware. ie, RHEL 4 guests on RHEL 6 host.



Kernel scalability limits - x86-64

Parameter	RHEL5 Support Limit	RHEL6 Support Limit	RHEL6 Theoretical Limit
CPUs	64 (192 – platform dependent)	4096	4096
Memory – Physical addressing	1T	8 T (pending testing)	64TB
Memory – process virtual address space (note – hardware dependent both RHEL5&6)	128T user 64T kernel	128T user 128T kernel	128TB
IRQs	239	33024	33024
# of processes	32000	32000 (larger pending testing)	4 million
KVM guest memory	512	Same as bare metal	Same as bare metal
KVM guest cpus	32	64 (pending testing)	64

SUMMIT

JBoss
WORLD

PRESENTED BY RED HAT



Determinism & realtime enhancements

- Some capabilities from Red Hat in MRG-realtime kernel (currently shipping as a RHEL 5 layered product) mainstreamed in RHEL 6
 - **Determinism** – Ability to schedule priority tasks predictability and consistently
 - **Priority** – Ensure highest priority applications are not blocked by low priority High Resolution Timers
 - **Timer** – Microsecond precision not timer interrupt ~millisecond precision
- **CFS scheduler** (completely fair scheduler)
 - Provides fair interactive response times by equally sharing available cycles rather than fixed quantum of timeslice
 - Includes modular scheduler framework – realtime task scheduler first
 - Priority inheritance algorithm prevents low priority processes from blocking higher priority by temporarily boosting priority to allow completion
- There will be a separate MRG-realtime offering for RHEL 6
 - Includes threaded interrupts and features not yet incorporated upstream
 - Allows rapid kernel innovation in supported product offering



RHEL 6 virtualization

- Advancements to make virtualization ubiquitous:
 - **Easier to deploy and manage**
 - Better control of resource allocation
 - Migration among non-identical hardware
 - **Performance close to bare metal**
 - Allowing all classes of workloads to benefit from virtualization flexibility – allowing a “run anywhere” deployment strategy.
 - Scalability - I/O, memory, CPU
 - More direct device access by guests, avoiding hypervisor overhead
 - Heterogeneous - Includes focus on Windows guests as well as Linux
 - **Secure**
 - Further guest isolation when cohabitating
 - **Compatibility of RHEL ecosystem**
 - Consistent application environment - Obviate need for applications to be aware of virt vs bare-metal, vs cloud deployment
 - KVM's tight kernel integration avails the majority of kernel features to virtualized guests. Examples: cgroups, CFS scheduler, timer precision. Additionally allows paravirtualization of clock, interrupt controllers, etc.

SUMMIT

JBoss

PRESENTED BY RED HAT



Virtualization – virtual memory enhancements

- **Transparent hugepages**

- Hugepages is a mechanism to efficiently manage large memory allocations (ie 2MB) as a unit rather than as small 4K chunks. Often 4X more efficient memory handling.
- Historically, hugepages suffered usability challenges as it required system startup time pre-allocation and is not swappable.
- Transparent hugepages obviates need to manually reserve memory. Dynamically allocates hugepage VM mappings for large allocation requests. Provides migration flexibility.
- Flexible policy controls, can enable per guest or process group.
- Beneficial to applications requesting large memory chunks.
- Highly beneficial to KVM hypervisor to more efficiently manage and allocate guest memory.

- **Extended Page Table (EPT) age bits**

- Enhancements in paging/swapping algorithm, nested page table support
- Allows host to make smarter choices in swapping when under memory pressure
- Allows swapping of transparently allocated hugepages by breaking up into smaller pages.

SUMMIT

**JBoss
WORLD**

PRESENTED BY RED HAT



Virtualization – performance enhancements

- **KSM – Kernel Shared Memory Swapping**

- Identifies duplicate pages, consolidating duplicates. Major example use case is Windows zero'ing all of memory at startup.
- Previously shared pages were not swappable
- New in RHEL6 is ability to swap KSM shared pages, beneficial to alleviate memory pressure in overcommit situations.

- **User return notifiers**

- Allows register caching and avoids needlessly preserving register states during context switching (expensive operations) when optional components like floating point are not currently utilized.



Virtualization – scalability enhancements

- **SMP** kernel synchronization enhancements (more fine grained)
 - Benefit: Scalable to 64 cpus per guest (vs 16 on RHEL5)
 - RCU kernel locking – utilizing a lock-free mechanism
 - Guest spin lock detector – causes guest to yield if spinning on same instruction too long (another guest not running may hold lock).
 - Intel & AMD hardware primitives “PLE exit” optimize
- **Guest hotplug** – CPU, disk & net
 - Allows virtual CPUs to be added/removed to running guests
 - Can also add/remove disk and network devices
 - Memory hotplug not currently supported
- **x2apic**
 - A virtual interrupt controller allowing direct guest access, obviating need for KVM emulation overhead.



Virtualization – network optimizations

- Vhost-net – a ring buffer abstraction between guest/host
- **(Performance)** Much of the network **implementation moved into kernel** (from Qemu user space) for optimization. Fewer context switches and vmexits. Increases multithreading.
- **Raw socket mode for SRIOV**
 - Previously networking interrupts handled through software bridging in “tap mode”
 - Bypasses bridge – allowing logical NICs assigned to guests direct PCI pass-through access. While optimized this ties guest to specific hardware and limits migration flexibility.
 - **(Migration Flexibility)** The vhost-net abstraction makes SRIOV allocation transparent to guest, allowing migration, even among non-identical systems.
- Network boot using **gpxe** – providing a more modern environment for pxe network booting



Virtualization – storage enhancements

- **AIO** – asynchronous IO – allows initiating large number of IO operations (ie database workloads)
 - RHEL5 provides AIO emulation – Qemu spawns individual threads per IO operation
 - Utilizing native AIO infrastructure yields 20% improvement in many IO intensive workloads
- **External ring buffers** – used in host/guest interfaces
 - Allows more concurrent IO operations to be in progress, not limited by finite buffer descriptors
 - Doesn't consume extra buffer space when not needed
- **Block alignment** storage topology awareness
 - Interrogates underlying storage hardware and pass through optimal alignment and physical sector size to guests. Ie, in support of 4K sectors. Requires storage device commands providing the info.
 - Allows optimal filesystem layout and application aware IO optimizations.

SUMMIT

**JBoss
WORLD**

PRESENTED BY RED HAT



Virtualization – migration enhancements

- **Static PCI slots**
 - Minor differences in device ordering preclude migration, especially PCI slot numbering among different versions of host/guest (ie RHEL4.7 on 5.5)
 - This capability enables logical assignment of PCI slots, preserving across migration – ensuring consistency of device namespace.
- **CPU capability enumeration**
 - Providing accurate physical CPU type to applications & libraries allows usage of optimization instructions, example SSE4 in recent instruction set.
 - Allows optimization of application performance with dynamic adaptation to match capabilities among migration domains.
- **Vhost over SRIOV**
 - Logically separate physical / virtual device assignment. Guest sees virtual device.
 - SRIOV optimizations previously were hardwired to specific units, precluding migration in many cases.
 - Host dynamically binds SRIOV resources to guests, allowing migration among non-identical systems.

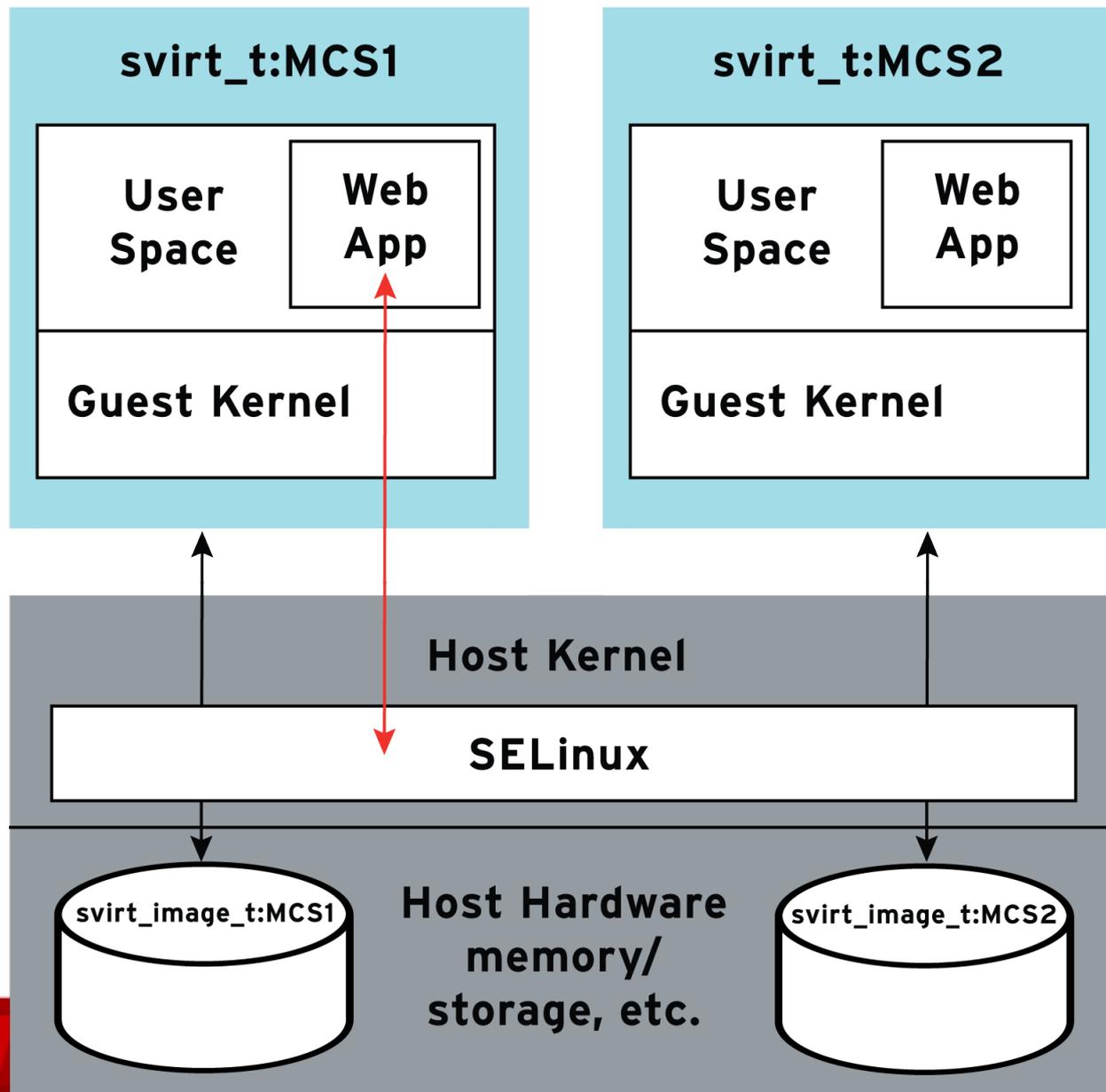
SUMMIT

**JBoss
WORLD**

PRESENTED BY RED HAT



Virtualization – svirt – SELinux virtual guest containment



SUMMIT

PRESENTED BY RED HAT



Virtualization – Xen interaction

- RHEL 5 includes both Xen & KVM hypervisors
 - RHEL 5 can accommodate RHEL 6 guests on either hypervisor
 - Xen accommodates PV (paravirt) & FV (fully virtualized) Linux guests
- RHEL 6 includes KVM hypervisor
- Migration tool provided to convert RHEL 5 Xen guests to KVM format to run on RHEL 6



Storage management

- **Topology awareness** – I/O (alignment and chunk size) based on info from the storage device. This is in dm, LVM, md, and utilities such as parted and mkfs standardized interfaces to obtain alignment and optimal I/O stripe width.
- DIF/DIX scsi **data integrity** commands (checksum) superior integrity
 - End-to-end data integrity check (SCSI DIF/DIX). Initially this extra checksum will be from the HBA to the storage. Added to applications and filesystems in the future. (requires storage hardware providing this capability)
 - Initially targeted at database use case on raw partition & HBA to storage in filesystem
- **FCoE** (fibre channel over ethernet) on specialized adapters (Emulex, QLogic, Cisco), and on standard NICs. FCoE install & boot support with DCB.
- **iSCSI** root/boot, including target
- **Thin provisioning** (virtual storage overcommit) via “discard” command – in LVM & filesystem. Requires storage device capability. Improves SSD wear leveling.
- **Block Discard**. Optimizes thin provisioning in the storage device, and improves wear leveling of SSDs. Currently used by XFS and ext4. Currently not usable with LVM/DM/multipath or md
- **SRIOV, NPIV** – driver virtualization IO accelerators – guest direct access
- **VSAN** – virt SAN fabric – based on NPIV, each guest has a separate WID, allows per-guest access control

SUMMIT

IBPS
WORLD

PRESENTED BY RED HAT



Storage management

- **LVM/DM** (Logical Volume Manager / Device Mapper)
 - LVM hot spare, a disk or group of disks used to replace a failing disk
 - Online resize of mirrored & multipath volumes
 - Snapshot scalability enhancements for virtualization
 - Multipath enhancements
 - Dynamic multipath load balancing. Path selection based on queue depth, or I/O service time
 - Mirroring enhancements
 - Snapshot of snapshot mirror
 - Mirrored mirror log. Improves mirror log availability, to avoid the need for a re-synch after a failure
 - Cluster mirror
 - Snapshot merge. Provides the ability to "rollback" changes that were made since the snapshot was taken. (Additional work to integrate this with Anaconda and yum will be post 6.0).
 - dm-crypt enhancements. Selectable hash algorithm for LUKS header, new cryptsetup commands, new libcryptsetup with versioned API.
 - Remote Replication tech. preview.

SUMMIT

**JBoss
WORLD**

PRESENTED BY RED HAT



Filesystem – larger & faster

- Ext4 - will be the default file system and scale to 16TB
- XFS - optional offering to support extremely large file systems > 16TB, up to 100TB. Tuned for larger servers & high end arrays.
- NFS
 - NFS4.0 – clients will default to use NFS4.0 (tunable via mount or config file)
 - NFS4.1 – enhanced support for referrals, delegation & failover.
 - Support for enhanced encryption types for kerberized NFS
 - Added IPv6 support
- GFS2 - optional
 - Targeting high availability clusters of 2-16 nodes.
 - Clustered samba (CTDB) – parallel (concurrent) servers for scalability & availability
- Filesystem utilities enhancements:
 - Creation tools warn about unaligned partitions, and new partitions are created on aligned boundaries with preferred block sizes. (hardware dependent)
 - Enhanced write barriers for increased data reliability – for ext3, ext4, GFS2, xfs



Networking improvements

- **10GbE** Driver support – on card switch and 8-16 pci devices.
 - Virtual guest can access the full NIC directly – SR-IOV enhancement – ie a single virt guest can saturate a 10GbE link.
 - **Data Center Bridging** (DCB) support – in ixgbe driver
 - Uses 802.1p VLAN priority tags to schedule and control traffic rates
 - Uses 802.1Qaz (priority grouping) and 802.1Qbb (priority flow control) to physically separate traffic flows that coexists on the same physical link
- **FCoE** (Fibre Channel over Ethernet)
 - Working on performance improvements throughout the storage stack (locking changes in the block and SCSI layers, improved interrupt handling).
- **RDMA** support – over 10GbE & Infiniband
 - Add IPv6 support
 - NFSoRDMA

SUMMIT

JBoss
WORLD

PRESENTED BY RED HAT



Networking improvements

- Major new features post RHEL5
 - Ipv6 Mobility support, RFC 3775
 - UDP-lite support, benefits multimedia protocols such as Voice Over IP, RFC 3828
 - Add Mutiqueue hardware support API
 - Large Receive Offload in network devices
 - Network controller for cgroup
 - Add multi-queue, DDR scheduler
 - Add TCP Illinois and YeAH-TCP congestion control algorithms
- General Networking Stack Performance improvement
 - RCU (read copy update) SMP locking optimization adoption in networking stack
 - Use RCU for the UDP hash lock
 - Convert TCP & DCCP has tables to use RCU.
 - RCU handling for Unicast packets
 - Multi-CPU rx to pull in from the wire faster
 - Multi-queue xmit networking for multiple transmit queues devices
 - New monitor tools for dropped packets, tc and dropwatch

SUMMIT

**JBoss
WORLD**

PRESENTED BY RED HAT



System services enhancements

- Dracut replacement for initramfs, mkinitrd
 - Better long-term supportability of storage configurations
 - Can automatically add raid members via udev rules
 - Allows change in hardware setup without needing to recreate initramfs
- NetworkManager - iSCSI, FCoE config, IPv6, Bridging
- Upstart flexible system service startup infrastructure
- CUPS printing enhancements
 - SNMP-based monitoring of ink/toner/supply levels and printer status
 - Device discovery speed improvements (backends now run in parallel)
 - Automatic PPD configuration for PostScript printers (PPD options values queried from printer) -- available in CUPS web interface
- Portreserve
 - Avoids network port allocation failures for network services

SUMMIT

**JBoss
WORLD**

PRESENTED BY RED HAT



SUMMIT

**JBoss
WORLD**

PRESENTED BY RED HAT

**LEARN. NETWORK.
EXPERIENCE OPEN SOURCE.**

www.theredhatsummit.com

EXT4 Pros & Cons

- **Ext4 has many compelling new features**
 - Extent based allocation
 - Faster fsck time (up to 10x over ext3)
 - Delayed allocation
 - Higher bandwidth
 - Should be relatively familiar for existing ext3 users
- **Ext4 challenges**
 - Large device support not finished in its user space tools
 - Limits supported maximum file system size to 16TB
 - Has different behavior over system failure



XFS Pros and Cons

- XFS is very robust and scalable
 - Very good performance for large storage configurations and large servers
 - Many years of use on large (> 16TB) storage
 - Red Hat tests & supports up to 100TB
- XFS challenges
 - Not as well known by many customers and field support people
 - Performance issues with meta-data intensive (small file creation) workloads



BTRFS

- Btrfs is the newest local file system
 - Has its own internal RAID and snapshot support
 - Does full data integrity checks for metadata and user data
 - Can dynamically grow and shrink
- Supported in RHEL6 as a tech preview item
 - Developers very interested in feedback and testing
 - Not meant for production use!



RHEL5 Local File Systems

- ext3 is our default file system for RHEL5
 - ext4 is supported as a tech preview in (5.4)
- xfs offered as a layered product (5.5+)



RHEL6 Local FS Summary

- FS write barrier enabled for ext3, ext4, gfs2 and xfs
- FS tools warn about unaligned partitions
 - parted/anaconda responsible for alignment
- Size Limitations
 - XFS for any single node & GFS2 for clusters up to 100TB
 - Ext3 & ext4 supported < 16TB



RHEL6 Support for Alignment

- New standards allow storage to inform OS of preferred alignment and IO sizes
 - Few storage devices currently export the information
- Partitions must be aligned using the new alignment variables
 - fdisk, parted, etc snap to proper alignment
 - FS tools warn of misaligned partitions
- Red Hat engineering is actively working with partners to verify and enhance this for our customers



RHEL6 Support for Discard

- File system level feature that informs storage of regions no longer in active use
 - SSD devices see this as a TRIM command and use it to do wear leveling, etc
 - Arrays see this as a SCSI UNMAP command and can enhance thin lun support
- Discard support is off by default
 - Some devices handle TRIM poorly
 - Might have performance impact
 - Test carefully and consult with your storage provider!



RHEL6 NFS Features

- NFS version 4 is the default
 - Per client configuration file can override version 4
 - Negotiates downwards to V3, V2, etc
- Support for industry standard encryption types
- IPV6 Support added for NFS and CIFS
 - NFS clients fully supported in 6.0
 - NFS server support for IPV6 aimed at 6.1

SUMMIT

**JBoss
WORLD**

PRESENTED BY RED HAT



Upcoming Local File System Features

- Union mounts
 - Allow a read-write overlay on top of a read-only base file system
 - Useful for virt guests storage, thin clients, etc
- Continuing to help lead btrfs development towards an enterprise ready state
- Support for ext4 on larger storage
- Enhanced XFS performance for meta-data intensive workloads



Upcoming NFS Features

- PNFS support
 - pNFS and more 4.1 features aimed at a minor 6.x release
 - No commercial arrays support pNFS yet
 - Ongoing work on open source (GFS2, object, etc) pNFS servers
- Working with standards body to add support for passing extended attributes over NFS
 - Goal is to enable SELinux over NFS



SUMMIT

**JBoss
WORLD**

PRESENTED BY RED HAT

**LEARN. NETWORK.
EXPERIENCE OPEN SOURCE.**

www.theredhatsummit.com

Minimal Platform Install

- Goals
 - Reduce Attack Surface
 - Minimize package count
 - Add back things needed for secure operation
 - Need to be able to disable services
 - Cron jobs for maintenance
 - Mail delivery for cron jobs
 - Update packages
 - Iptables, audit, and sshd



Minimal Platform Install

**RED HAT
ENTERPRISE LINUX 5**

The default installation of Red Hat Enterprise Linux Server includes a set of software applicable for general internet usage. What additional tasks would you like your system to include support for?

- Software Development
- Virtualization
- Web server

You can further customize the software selection now, or after install via the software management application.

Customize later Customize now

[Release Notes](#) [Back](#) [Next](#)

SUMMIT

**JBoss
WORLD**

PRESENTED BY RED HAT



Minimal Platform Install



SUMMIT

**JBoss
WORLD**

PRESENTED BY RED HAT



Minimal Platform Install

RHEL5 (5.5 used for testing)

- Packages - 879
- Setuid - 33
- Setgid - 11
- Daemons - 44
- Networked services - 18
- Space – 2.2 Gb
- Notes: Boots into X even though no packages checked

SUMMIT

JBoss
WORLD

PRESENTED BY RED HAT



Minimal Platform Install

**RED HAT
ENTERPRISE LINUX 5**

Desktop Environments
Applications
Development
Servers
Base System
Virtualization
Languages

Administration Tools
 Base
 Dialup Networking Support
 Java
 Legacy Software Support
 OpenFabrics Enterprise Distribut
 System Tools

This group is a collection of graphical administration tools for the system, such as for managing user accounts and configuring system hardware.

[Optional packages](#)

[Release Notes](#) [Back](#) [Next](#)

SUMMIT

**JBoss
WORLD**

PRESENTED BY RED HAT



Minimal Platform Install

RHEL5 (5.5 used for testing)

- Packages - 437
- Setuid - 29
- Setgid - 9
- Daemons - 39
- Networked services – 16
- Space – 1006 Mb
- Notes: Boots to runlevel 3

SUMMIT

JBoss
WORLD

PRESENTED BY RED HAT



Minimal Platform Install

**RED HAT®
ENTERPRISE LINUX® 6** 

The default installation of Red Hat Enterprise Linux is a basic server install. You can optionally select a different set of software now.

virtualhost
 Desktop
 Software Development Workstation
 Minimal

Please select any additional repositories that you want to use for software installation.

HighAvailability
 LargeFileSystem
 LoadBalance
 Red Hat Enterprise Linux

You can further customize the software selection now, or after install via the software management application.

Customize later Customize now

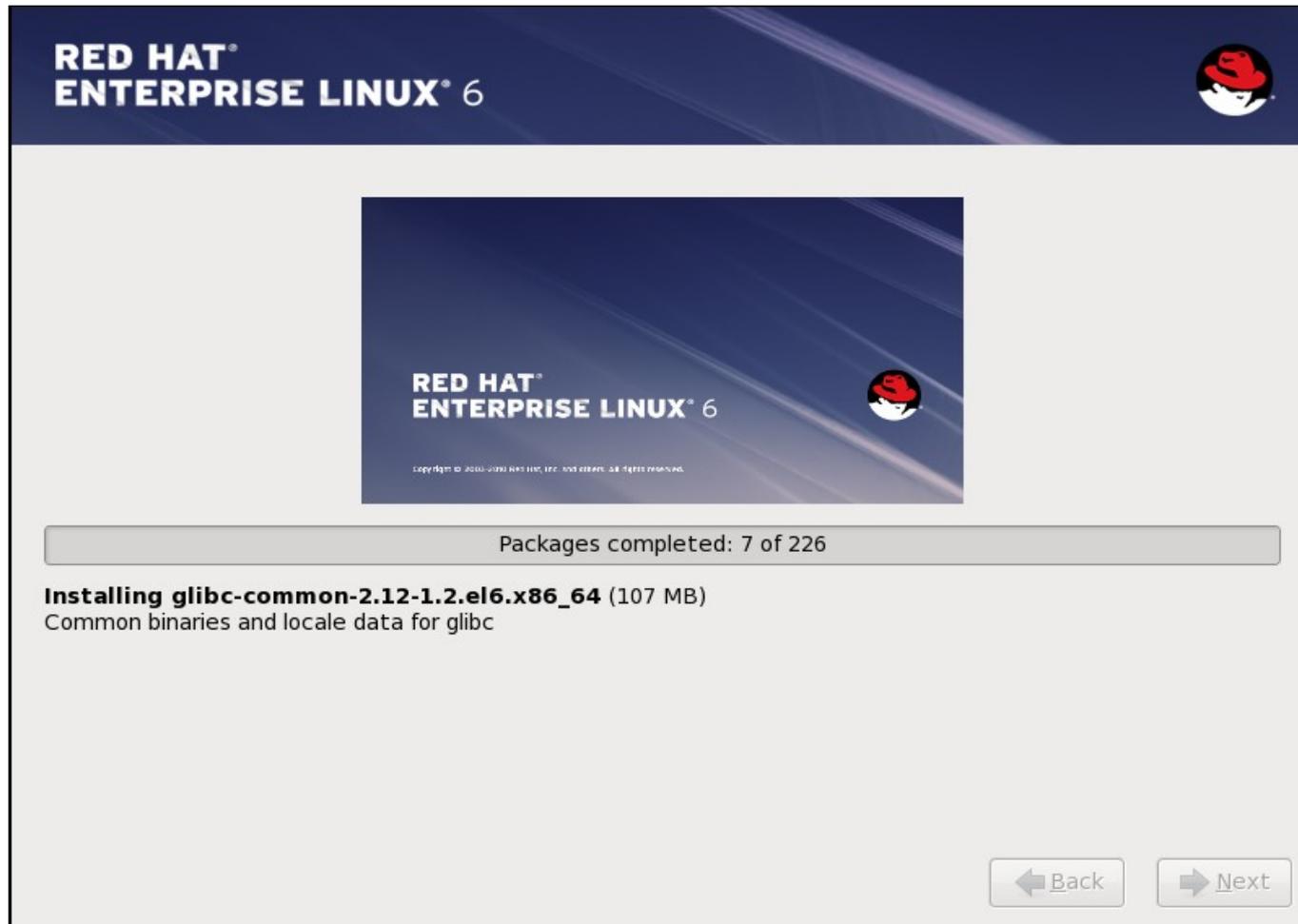
SUMMIT

**JBoss
WORLD**

PRESENTED BY RED HAT



Minimal Platform Install



SUMMIT

JBoss
WORLD

PRESENTED BY RED HAT



Minimal Platform Install

RHEL6 (pre-beta2)

- Packages - 226
- Setuid - 20
- Setgid - 7
- Daemons - 13
- Networked services – 5
- Space – 565 Mb
- Notes: Boots to runlevel 3 very quickly



Minimal Platform Install - Summary

	Packages	Setuid	Setgid	Daemons	Network Services	Space
RHEL5	879	33	11	44	18	2200
RHEL5 base	437	29	9	39	16	1006
RHEL6	226	20	7	13	5	565

SUMMIT

**JBoss
WORLD**

PRESENTED BY RED HAT



Stronger Hashes

- MD5 was being used in many places for integrity or password hashes
- Attacks against MD5 have been getting better
- NIST's Policy on Hash Functions:
 - Federal agencies should stop using SHA-1 for digital signatures, digital time stamping and other applications that require collision resistance as soon as practical, and must use the SHA-2 family of hash functions for these applications after 2010.
- Needed to adjust all tools that touch software from source code to system verification.



Stronger Hashes

- Shadow-utils, glibc, pam, authconfig were done during RHEL5
- Started Project for Fedora 11. Changed:
 - Rpm, koji, spacewalk/satellite, yum, createrepo, punji, RHN, yaboot
- To do:
 - Changes for grub password hash expected in 6.1



SUMMIT

**JBoss
WORLD**

PRESENTED BY RED HAT

**LEARN. NETWORK.
EXPERIENCE OPEN SOURCE.**

www.theredhatsummit.com

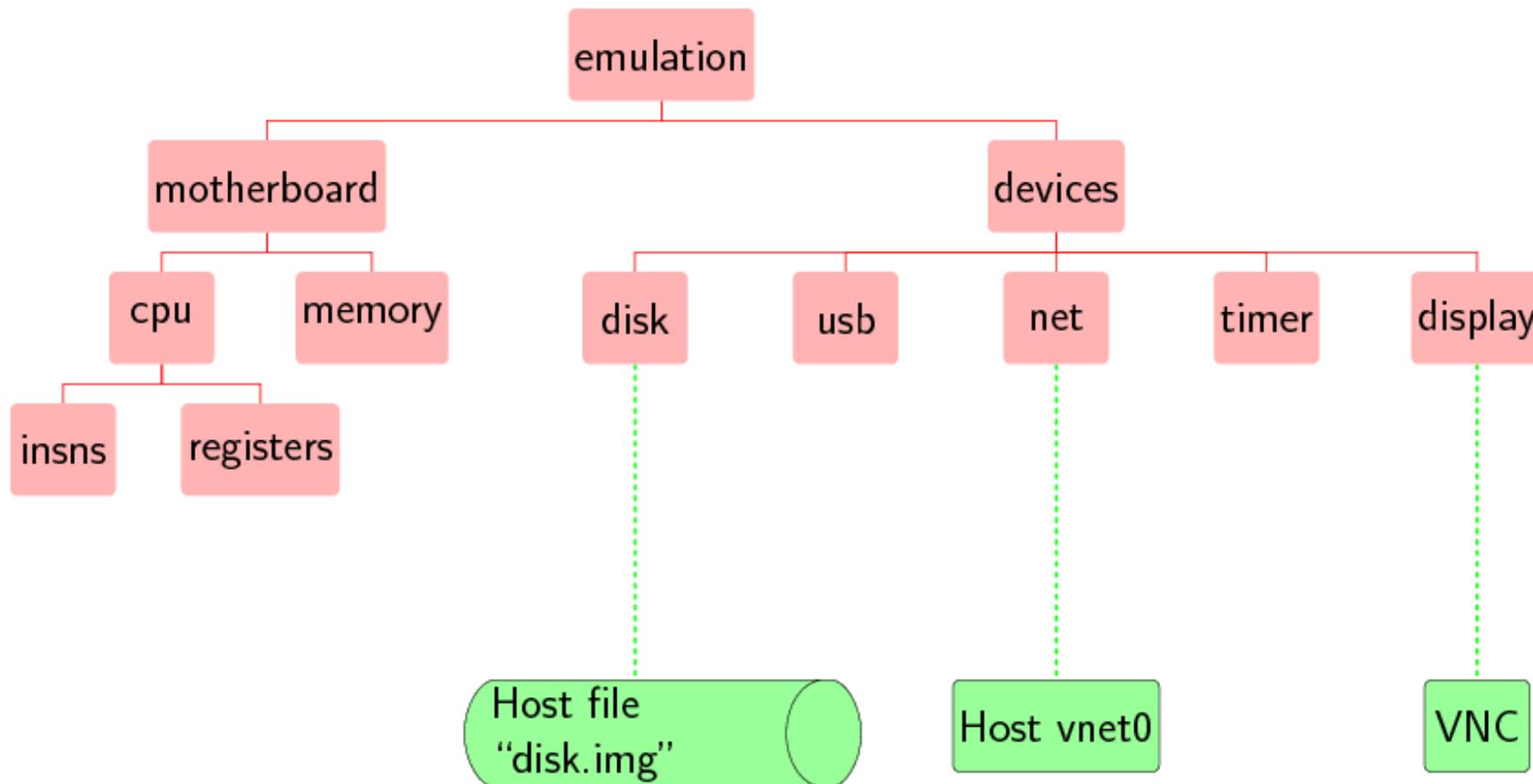
libguestfs

a library for accessing and modifying
virtual machine disk images

Richard W.M. Jones — people.redhat.com/~rjones
sponsored by Red Hat Inc.

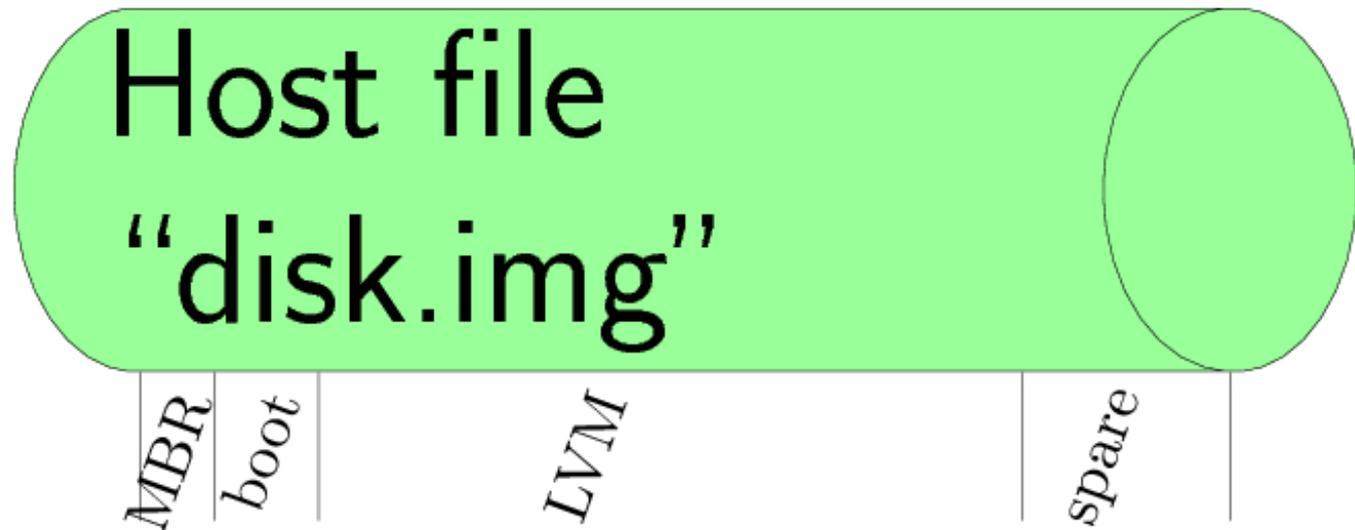
libguestfs.org

What *is* a virtual machine?



Disk image ops

- . Clone the machine and change the hostname &c
- . What licensed software is installed?
(*auditing*)
- . Is it running out of disk space? (*monitoring*)
- . Resize it (*admin*)
- . Take a back-up, or copy bits in or out
- . VM doesn't boot → edit `grub.conf`
- . Make a new one from scratch



```
# kpartx -a /dev/vg_trick/F13Rawhide64
# ls -l /dev/mapper/vg_trick-F13Rawhide64*
/dev/mapper/vg_trick-F13Rawhide64
/dev/mapper/vg_trick-F13Rawhide64p1
/dev/mapper/vg_trick-F13Rawhide64p2
# vgscan
Reading all physical volumes. This may
  take a while...
Found volume group "vg_f13rawhide" [..]
Found volume group "vg_trick" [..]
# lvs
LV          VG          Attr      LSize [..]
lv_root    vg_f13rawhide -wi---    8.83g
lv_swap    vg_f13rawhide -wi---   992.00m
```

API

Perl

OCaml

Ruby

Python

Java

Shell

Haskell

C/C++

C#

guestfish

virt-ls

virt-rescue

virt-cat

virt-inspector

virt-tar

virt-win-reg

virt-df

virt-resize

virt-v2v

virt-edit

Since the link is quite interesting, I'll go over [the Augeas configuration API](#) page. So I won't go through that here.

[Demonstration of using the API from Perl and Python]

This is the Perl example. Notice the use of [the Augeas configuration API](#) to pull out the list of NTP servers:

```
#!/usr/bin/perl -w

use strict;

use Sys::Guestfs;

my $g = Sys::Guestfs->new ();
$g->add_drive_ro ("disk.img");
$g->launch ();

my @logvols = $g->lvs ();
print "logical volumes: ", join (" ", @logvols), "\n\n";

$g->mount_ro ("/dev/vg_f12x32/lv_root", "/");
print "----- ISSUE file: -----\n";
print ($g->cat ("/etc/issue"));
print "----- end of ISSUE file -----\n\n";

# Use Augeas to list the NTP servers.
$g->aug_init ("/", 16);
my @nodes = $g->aug_match ("/files/etc/ntp.conf/server");
my @ntp_servers = map { $g->aug_get ($) } @nodes;
print "NTP servers: ", join (" ", @ntp_servers), "\n\n";
```

Virt-df

Virt-df is df for virtual guests. Run the program on the host / dom0 to display disk space used and available on all partitions on all guests.

```
# virt-df -h
Filesystem                Size      Used    Available   Use%
Ubuntu904x64:/dev/sda1    9.4G      2.1G      6.8G    27.7%
Debian5x64:/dev/debian5x64/home 3.4G      761.9M    2.5G    27.0%
Debian5x64:/dev/debian5x64/root 321.5M    111.1M    193.8M   39.7%
Debian5x64:/dev/debian5x64/tmp 302.1M     10.0M    276.5M    8.5%
Debian5x64:/dev/debian5x64/usr 3.4G      1.1G      2.1G    38.3%
Debian5x64:/dev/debian5x64/var 1.7G      612.6M    1001.9M  41.1%
Debian5x64:/dev/sda1     227.9M     18.6M    197.1M   13.5%
F10x32:/dev/VolGroup00/LogVol00 8.8G      3.1G      5.2G    40.3%
F10x32:/dev/sda1        189.9M     20.2M    159.9M   15.8%
CentOS5x32:/dev/VolGroup00/LogVol00 8.6G      3.9G      4.2G    50.6%
CentOS5x32:/dev/sda1    98.7M     23.5M     70.1M   29.0%
Win2003x32:/dev/sda1    20.0G      2.1G     17.9G   10.4%
```

SUMMIT

**JBoss
WORLD**

PRESENTED BY RED HAT

**LEARN. NETWORK.
EXPERIENCE OPEN SOURCE.**

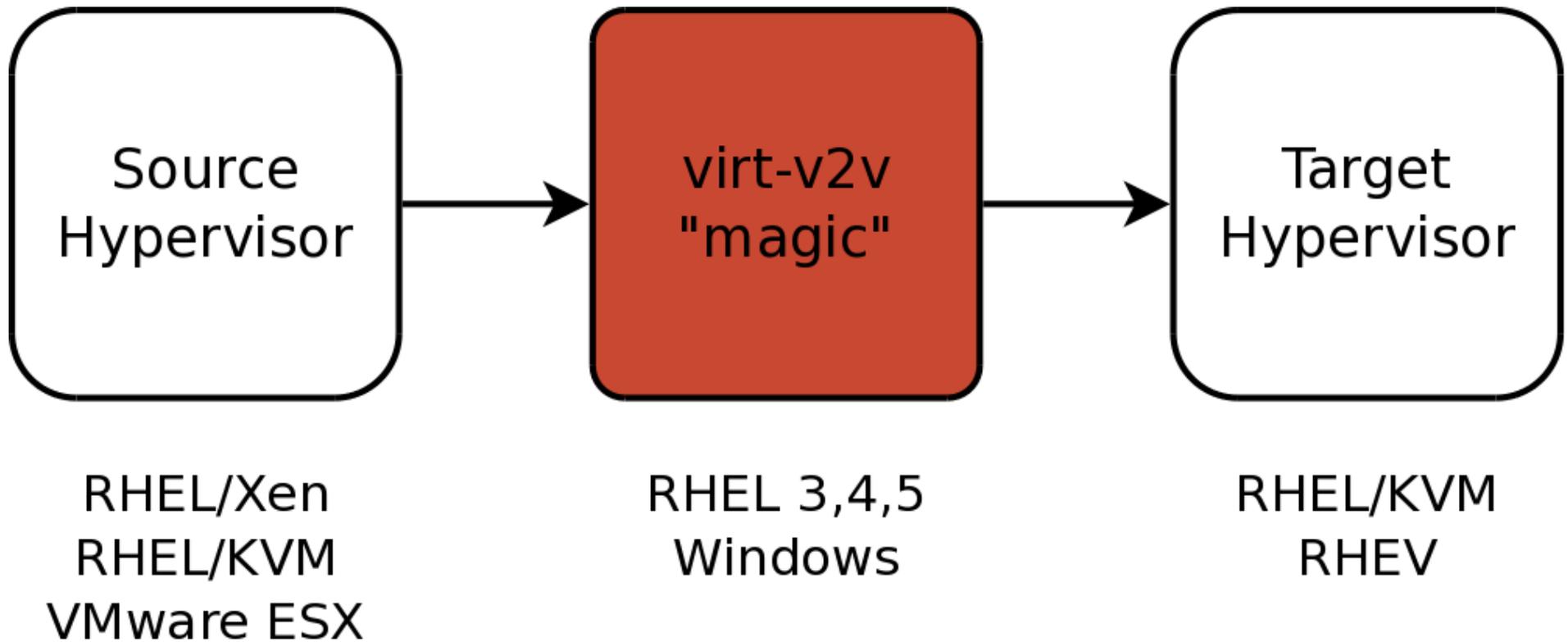
www.theredhatsummit.com

What is virt-v2v?

A command-line tool to manage virtual machine conversions.



What does virt-v2v support?



SUMMIT

**JBoss
WORLD**

PRESENTED BY RED HAT



```
[root@t500 /]# virt-v2v -ic esx://yellow.rhev.marston/?no_verify=1 -o rhev -osd blue.rhev.marston:/nfs/export --network rhevm RHEL5-64
** HEAD https://yellow.rhev.marston/folder/RHEL5-64/RHEL5-64_1-flat.vmdk?dcPath=ha-datacenter&dsName=yellow%3Astorage1 ==> 401 Unauthorized
** HEAD https://yellow.rhev.marston/folder/RHEL5-64/RHEL5-64_1-flat.vmdk?dcPath=ha-datacenter&dsName=yellow%3Astorage1 ==> 200 OK
** GET https://yellow.rhev.marston/folder/RHEL5-64/RHEL5-64_1-flat.vmdk?dcPath=ha-datacenter&dsName=yellow%3Astorage1 ==> 200 OK (330s)
virt-v2v: RHEL5-64 configured with virtio drivers
[root@t500 /]# █
```

I

SUMMIT

**JBoss
WORLD**

PRESENTED BY RED HAT



Conclusion

- Manage the whole conversion process
- Convert RHEL/Windows virtual machines
- Convert Xen/KVM/ESX to RHEL or RHEV
- Available in a child channel of RHEL 5 with a RHEV subscription
- Will be available in RHEL 6
- Open Source: available in Fedora



SUMMIT

**JBoss
WORLD**

PRESENTED BY RED HAT

**LEARN. NETWORK.
EXPERIENCE OPEN SOURCE.**

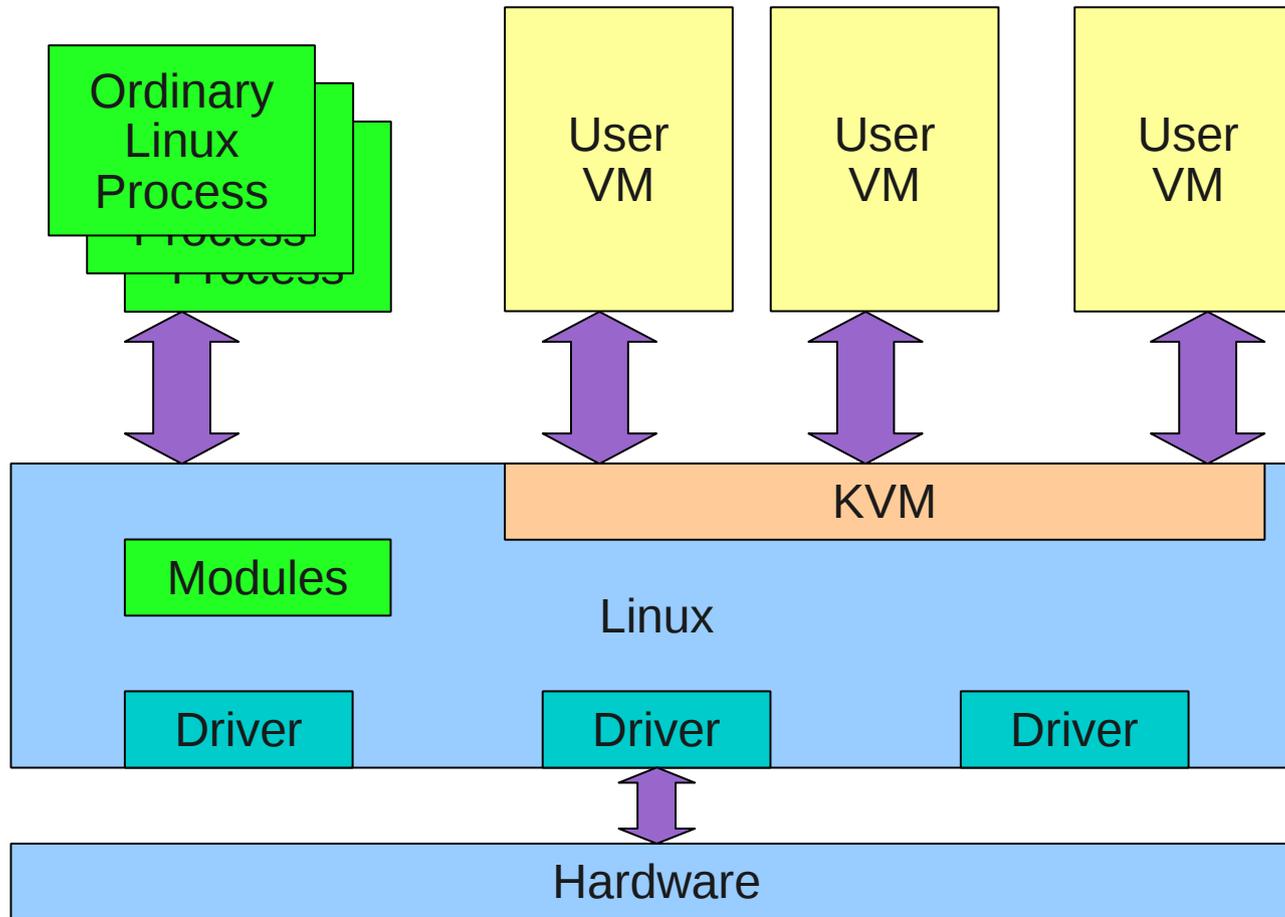
www.theredhatsummit.com

KVM

- Kernel-based virtual machine
- Linux Kernel modules turn Linux into a hypervisor
 - In Linux since 2.6.20
 - User space component is qemu-kvm
- Leverages Linux kernel features and support
- Leverages hardware virtualization support



General Architecture (KVM)



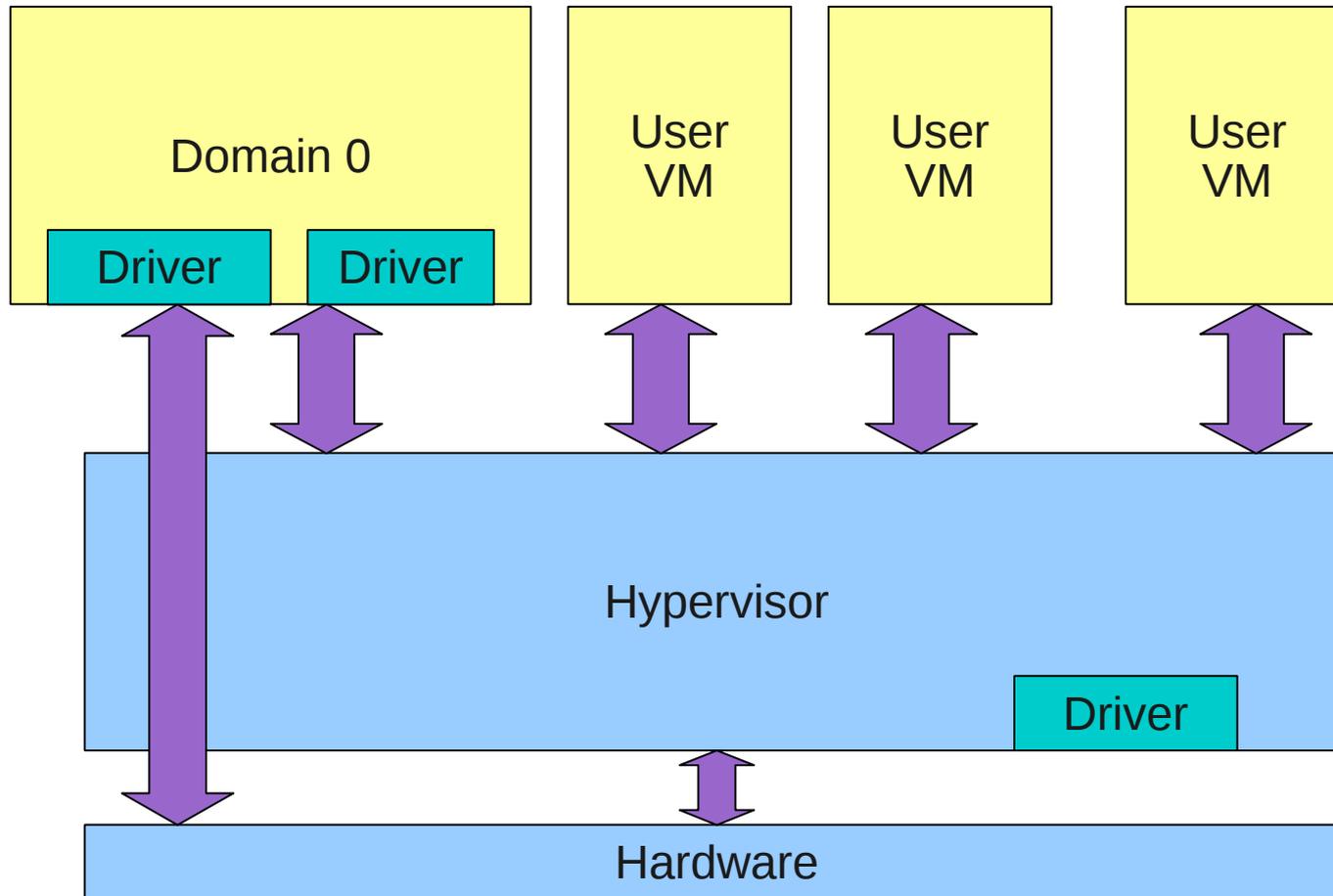
SUMMIT

JBoss
WORLD

PRESENTED BY RED HAT



General Architecture (Xen)



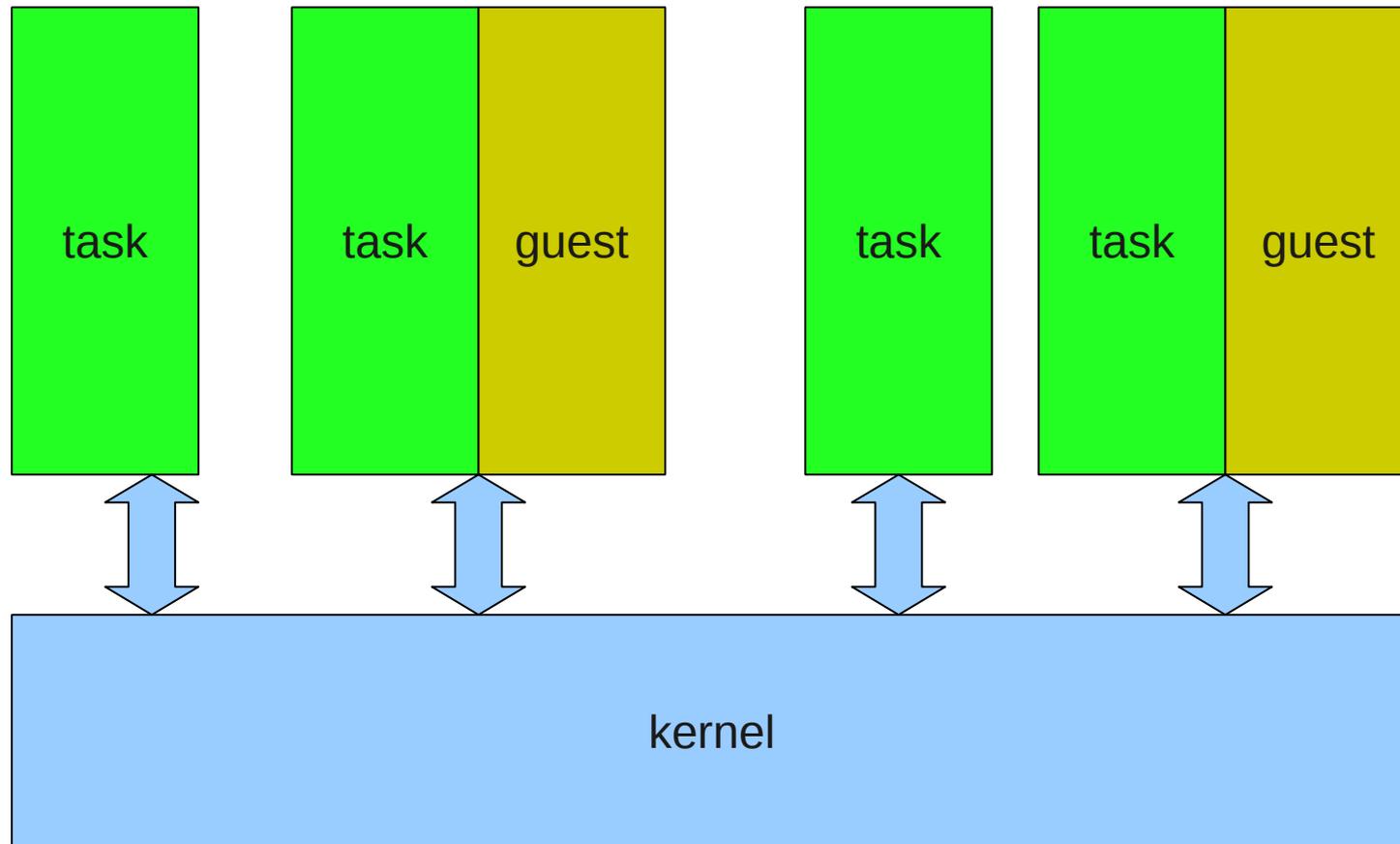
SUMMIT

JBoss
WORLD

PRESENTED BY RED HAT



General Architecture (process model)



SUMMIT

JBoss
WORLD

PRESENTED BY RED HAT



General Architecture (process model cont.)

- Guests are ordinary processes
 - Each virtual cpu is thread
- Like a new operating mode: kernel, user and “guest”
- “Guest” mode can hypercall, but not syscall
- Leverages Linux kernel features like
 - Scheduling, Accounting, cgroups
 - KSM (Kernel Samepage Merging)
 - Power management



Brief virtualization history

- Mid 1960s IBM developed mainframe virtualization
- 2001 VmWare 1st x86 virtualization, binary translation
- 2003 Xen, para-virtualization
- 2005/6 Intel/AMD x86 hardware virtualization
- 2007 Linux 2.6.20 includes KVM
- 2009 Red Hat supports KVM in Red Hat Enterprise Linux 5.4
- 2010 RHEL 6

SUMMIT

**JBoss
WORLD**

PRESENTED BY RED HAT



Hardware features

- CPU support (Intel VMX, AMD SVM)
 - EPT/NPT
- IOMMU/VT-d
 - Protection for devices that are passed through to guests
- SR-IOV
 - Safe sharing of real hardware
 - Getting real traction with NICs
- NPIV
 - Allows sharing storage



RHEL 6 CPU Enhancements

- 64 Virtual CPUs per guest
- Minimized CPU overhead
 - RCU kernel “locking” improves large SMP performance
 - User space notifiers
 - X2apic, a virtual interrupt controller

SUMMIT

**JBoss
WORLD**

PRESENTED BY RED HAT

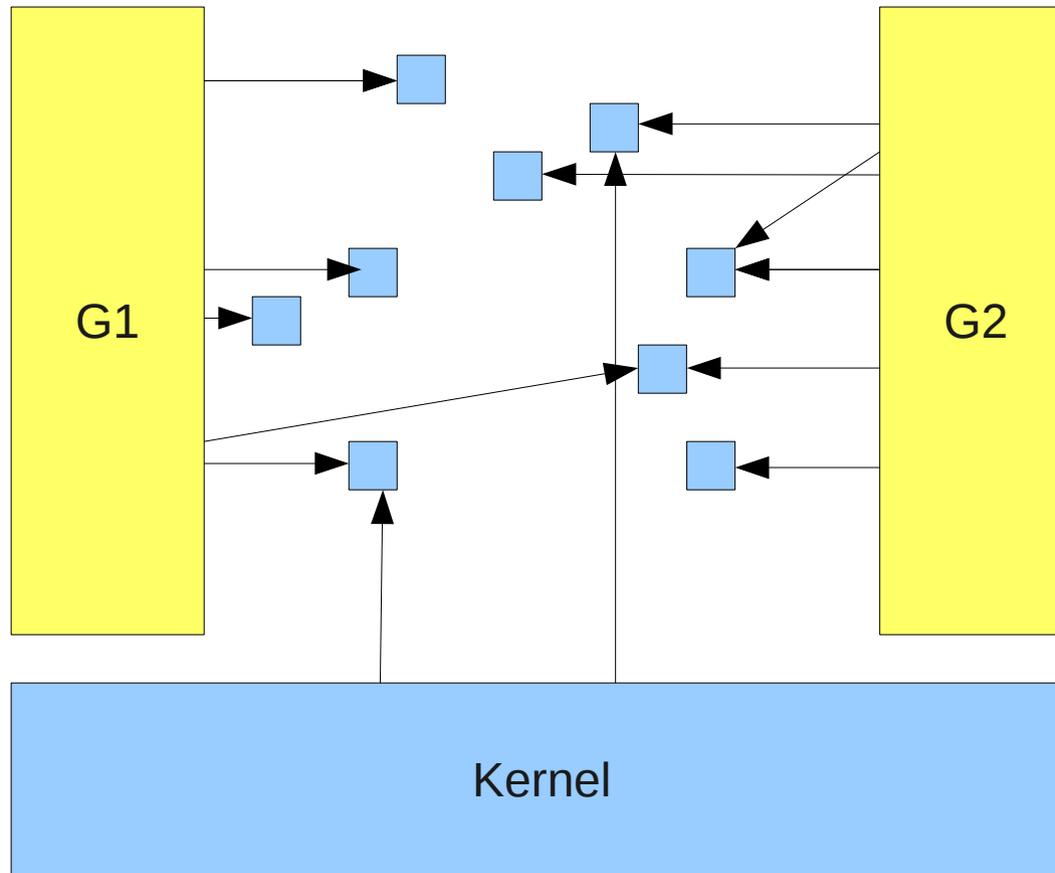


RHEL 6 Memory Enhancements

- Transparent Hugepages
 - Now dynamic, no boot time preallocation required
 - Can be broken down and swapped
- Extended/Nested page table aging improves swap choices
- Linux feature KSM (Kernel Samepage Merging) coalesces common pages.
 - A real win for Windows guests, where they zero all pages on boot



KSM example



SUMMIT

**JBoss
WORLD**

PRESENTED BY RED HAT



RHEL 6 Block I/O Enhancements

- Native AIO, and preadv/writev
- External ring buffers in guest/host interface
- Virtio barrier support
- MSI interrupt support
- Intelligent block alignment changes, better default
- Near native performance

SUMMIT

**JBoss
WORLD**

PRESENTED BY RED HAT

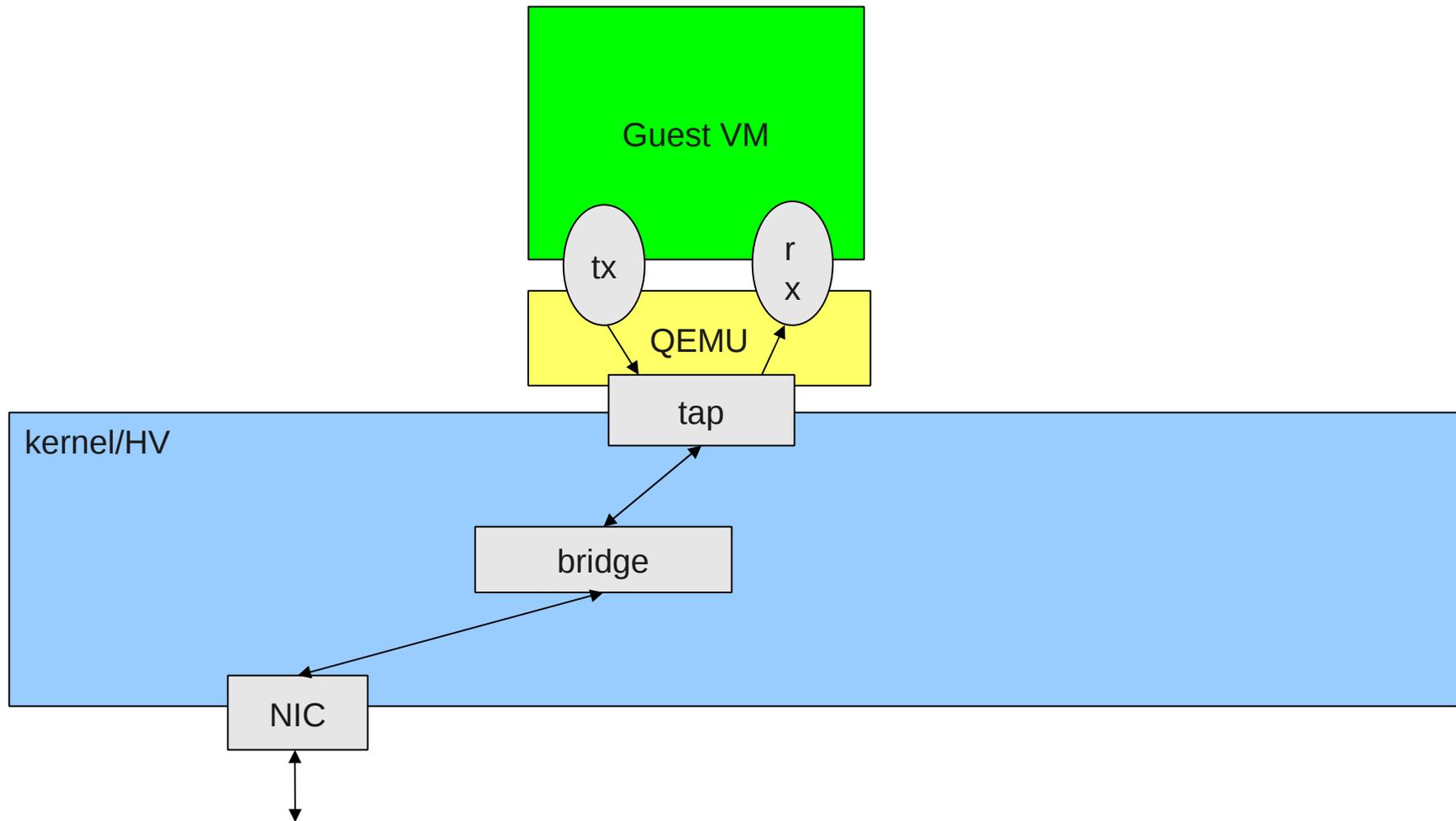


RHEL 6 Network I/O Enhancements

- Networking
 - Vhost-net, moves a portion of networking from user space to kernel.
 - More migratable than pass through
 - Can be used on top of SR-IOV devices, while preserving migratability
 - GPXE network boot supported.



Virtio drivers



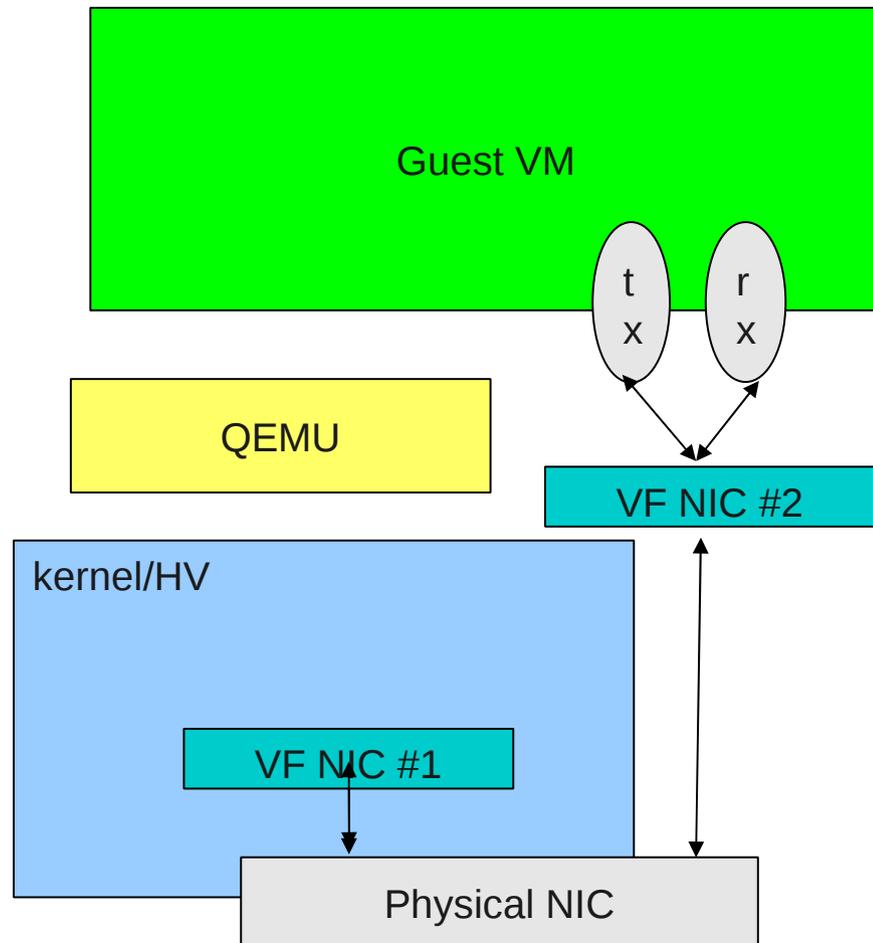
SUMMIT

**JBoss
WORLD**

PRESENTED BY RED HAT



Device assignment – SR-IOV, VT-d/IOMMU



- Low overhead
 - Best throughput
 - Lowest latency
- Complicates migration

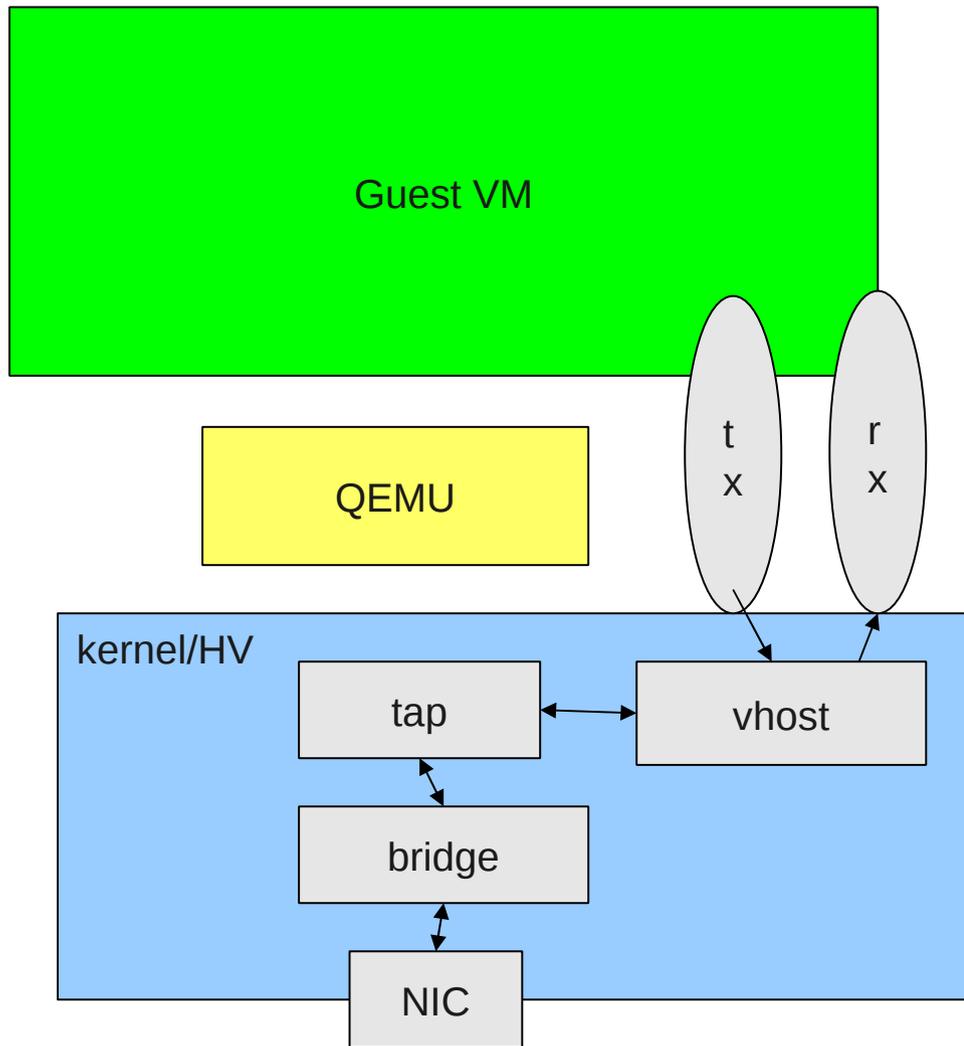
SUMMIT

**JBoss
WORLD**

PRESENTED BY RED HAT



In-Kernel Vhost-net



- Less context switching
- Low latency
- MSI
- One less copy

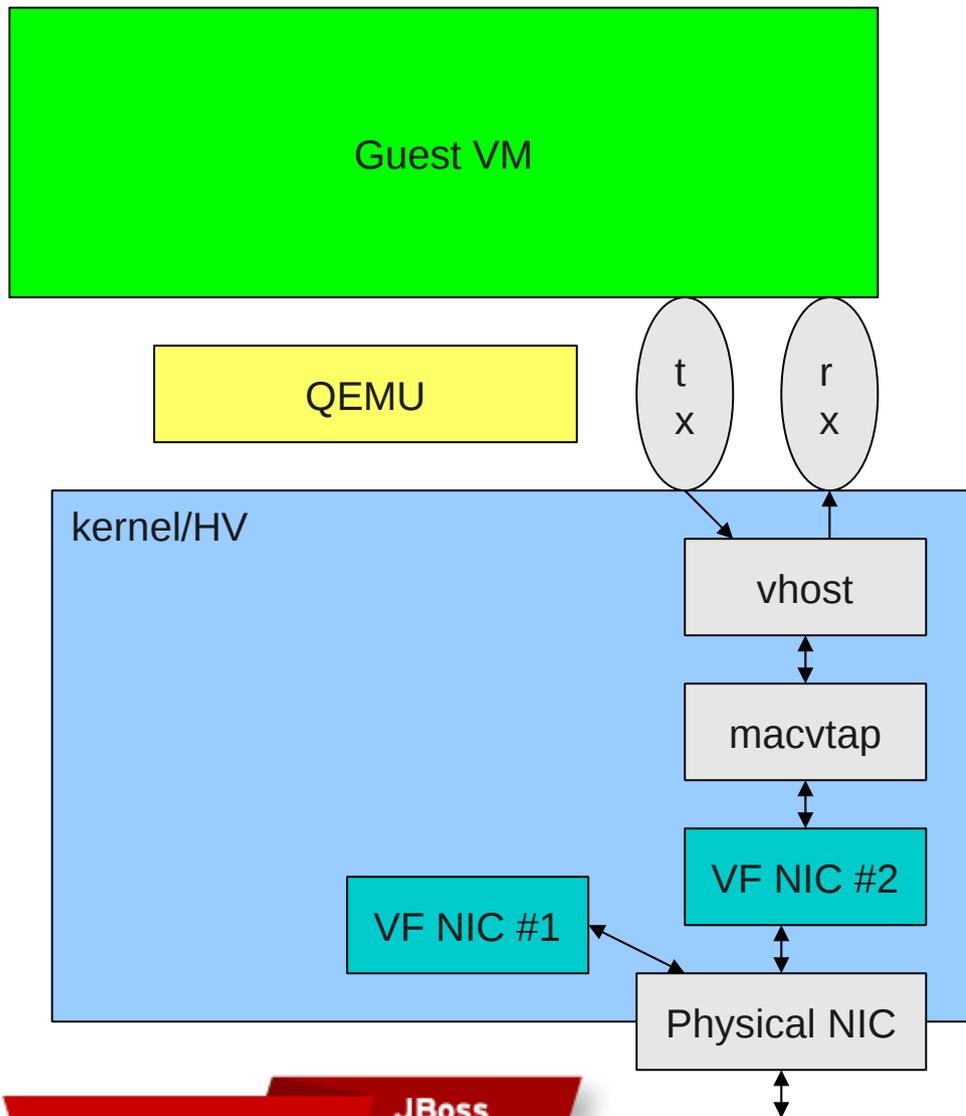
SUMMIT

**JBoss
WORLD**

PRESENTED BY RED HAT



Vhost over SR-IOV using macvtap



- Guest only knows virtio
- Migration friendly
- Excellent performance
- Future zero copy

SUMMIT

JBoss
WORLD

PRESENTED BY RED HAT



RHEL 6 RAS Enhancements

- QMP – QEMU Monitor protocol
- Virtio serial (vmchannel)
- Improved migration protocol
- Kvm-clock
- Cgroups
- sVirt
- Power management – tickless kernel
- Static PCI slots to allow easier migration

SUMMIT

**JBoss
WORLD**

PRESENTED BY RED HAT



Supported Guests

- RHEL 3/4/5/6
- Windows 2003, 2008, XP and 7

SUMMIT

**JBoss
WORLD**

PRESENTED BY RED HAT



Future (beyond 6.0)

- PV spin locks
- Vhost zero copy
- Vswitch, VEPA
- Nested VMX
- UIO PCI device assignment
- Deep C-state power management

SUMMIT

**JBoss
WORLD**

PRESENTED BY RED HAT



SUMMIT

**JBoss
WORLD**

PRESENTED BY RED HAT

**LEARN. NETWORK.
EXPERIENCE OPEN SOURCE.**

www.theredhatsummit.com