Current & Future State of Red Hat Enterprise Linux

Brad Hinson
Email: bhinson@redhat.com
Worldwide Lead, Linux on System z

redhat

Agenda

- **Overview**
- **Current technology through RHEL 5.5 relating to** System z
- **Roadmap moving forward** (**RHEL 6)**
  - New file systems
  - Kernel patches & improvements
  - Performance
- **Open Dialog**

RHEL 5 on System z, Current Technology

Through RHEL 5.4

RHEL 5 Summary: Highlights

- **New Package: ktune [5.3]**
  - Service that sets several kernel tuning parameters to values suitable for specific system profiles.
  - Provides a profile for large-memory systems running disk-intensive and network-intensive applications.
- **NETIUCV Driver Removed, Deprecated [5.0]**
  - Network driver only, socket API still supported
- **SELinux Per-Packet Access Controls [5.3]**
  - Add Secmark support to core networking
    - Allows security subsystems to place security markings on network packets (2.6.18)

RHEL 5 Summary: Highlights

- **3592 Tape Encryption Support [5.1]**
  - Use IBM 3592 encryption feature to securely store data
- **Samba: Rebased from 3.0.28 to 3.0.32 for Bug Fixes**
  - Now supports Windows Vista and 2008
  - Fixes for DC functionality (interoperability with Citrix and Domain trusts)

RHEL 5 Summary: File Systems/Storage

- **Block device encryption support, including support for /root partition, including configuration in anaconda installer.**
- **ext4 file system tech preview [5.3, with updates in 5.4]**
    - Support for delayed allocation
- **Ecryptfs support [5.2, with updates in 5.3]**
- **Improved multipath over PAV support [5.2]**
- **Add integrity check to cryptsetup-luks, in order to meet FIPS-140 requirements [5.4]**
- **File system freeze/quiesce interface added to support hardware snapshots for file systems.**
- Full support for FUSE and libfuse
    - Allow end users to more easily install and use their own user space FUSE file systems.

RHEL 5 Summary: File Systems/Storage

- **xDR/GDPS System Initialization [5.3]**
  - Disaster recovery solution for DASD and FCP.
  - Multipath access to PPRC (primary/mirror) volume pair.
  - Hyperswap support in DASD driver.
- **DASD Fast Fail Support**
  - Increase failover speed in multipath setup with DASD and PAV setup.
- **Multipath Support in Installer [5.2]**
  - Automatically detect multipathed DASD and SCSI/FCP disks. Support for multipathed root file system.

s390-tools Package Rebased to v1.8.1

- **This package provides *the* essential tool chain for Linux on System z. It contains everything from the boot loader to dump related tools for a system crash analysis.**
- New Features
  - DASD related tools: Add Large Volume Support for ECKD DASDs
  - IPL tools: Can be used to change the reipl & shutdown behavior
  - ziomon tools: Set of tools to collect data for zfcp performance analysis.
  - lsluns: List available SCSI LUNs depending on adapter or port.
  - lszcrypt/chzcrypt: Show/modify information about zcrypt devices and configuration.

s390-tools Package Rebased to v1.8.1

- New Features (continued)
  - cpuplugd: Daemon that manages CPU- and memory-resources based on a set of rules. Depending on the workload CPUs can be enabled or disabled. The amount of memory can be increased or decreased exploiting the Cooperative Memory Management (CMM1) feature.
  - lschp/chchp: Tool to show/modify channel-path states and available channel paths.
  - mon_procd: Daemon that writes process information data to the z/VM monitor stream.
  - vmur: Tool to work with z/VM spool file queues (reader, punch, printer).
  - zfcpdump_v2: Version 2 of the zfcpdump tool. Now based on the upstream Linux kernel 2.6.23.
- **Plus various bug fixes**

RHEL 5.4: Kernel

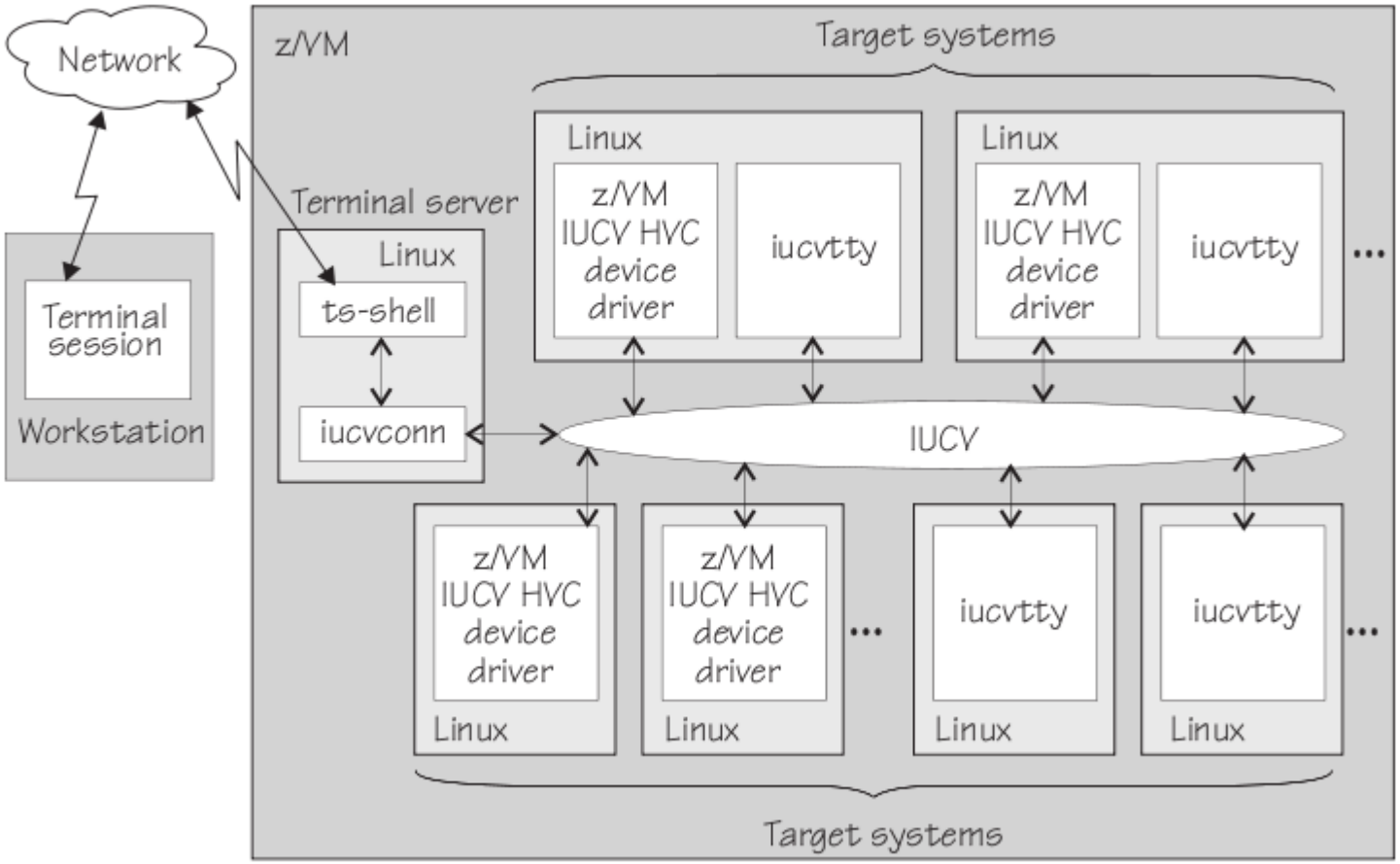
- **Control Program Identification (CPI)**
  - If your RHEL5.4 Linux instance runs in LPAR mode, you can now use the extended control program identification (CPI) module, sclp_cpi and the sysfs interface */sys/firmware/cp*i to assign names to your Linux instance
  - The names are used, for example, to identify the Linux instance on the HMC.
  - *This feature is only available while running in LPAR*
- **Extra Kernel Parameter via VMPARM**
  - Modify the IPL records to append extra parameters specified with the z/VM VMPARM option to the kernel command line.
- **Support for Processor Degradation**
  - Adds support for processor degradation, which allows processor speed to be reduced in some circumstances (i.e. system overheating). This new feature allows automation software to observe the machine state.

RHEL 5.4: Kernel

- **TTY Terminal Server Over IUCV**
  - Provide central access to the Linux console for the different guests of a z/VM.
  - The terminal server connects to the different guests over IUCV.
  - The IUCV based console is ASCII based.
  - Fullscreen applications like *vi* are usable on the console.

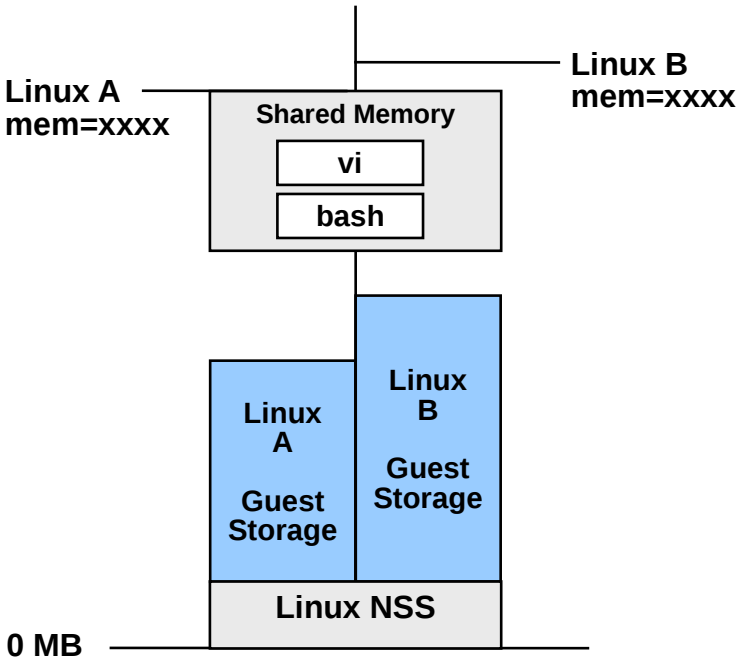
redhat.

TTY Terminal Server Over IUCV



Shared Kernel

- **Named Saved Segments (NSS)**
  - Using NSS the z/VM hypervisior makes operating system code in shared real memory pages available to z/VM guest virtual machines.
  - With this update, Linux guest operating systems using z/VM can boot from the NSS and be run from a single copy of the Linux kernel in memory.



RHEL 5.4: Networking

- **HiperSockets Layer3 Support for IPv6**
  - How IPv6 support for HiperSockets devices running in layer 3 mode is available
  - IPv6 is supported on:
    - Ethernet interfaces of the OSA-Express adapter running in QDIO mode.
    - HiperSockets layer 2 and layer 3 interfaces
    - z/VM guest LAN interfaces running in QDIO mode.
  - IPv6 is not supported on the OSA-Express Token Ring and ATM features.

RHEL 5.4: Reliability, Availability, & Serviceability

- **Multi Volume Dump Support for DASDs**
  - Added the ability to dump on multiple ECKD DASD devices, which can be necessary, if the system memory size is larger than the size of a single DASD device.
- **Service Levels of Hardware & Hypervisor**
  - A new Interface which provides service levels of hardware and z/VM service-levels to the Linux userspace. Interface: */proc/service_levels*

```
# cat /proc/service_levels
VM: z/VM Version 6 Release 1.0, service level 0901 (64-bit)
zfcp: 0.0.010a microcode level b02
qeth: 0.0.0800 firmware level V611
```

RHEL 5.4: Reliability, Availability, & Serviceability

- **Shutdown Actions Interface**
  - The new shutdown actions interface allows to specify for each shutdown trigger (halt, power off, reboot, panic) one of the five available shutdown actions (stop, ipl, reipl, dump, vmcmd).
  - A sysfs interface under /sys/firmware is provided for that purpose.
  - Possible use cases are e.g. to specify that a vmdump should be automatically triggered in case of a kernel panic or the z/VM logoff command should be executed on halt.
- **Automatic IPL After Dump**
  - The new shutdown action dump_reipl is introduced. It combines the actions dump and re-ipl, first a dump is taken, then a re-ipl of the system is triggered

RHEL 5.4: Storage

- **FCP Performance Data Collection & Reports:**
  - Fibre Channel Protocol (FCP) performance data can now be measured. Metrics that are collected and reported on include:
  - Performance relevant data on stack components such as Linux devices, Small Computer System Interface (SCSI) Logical Unit Numbers (LUNs) and Host Bus Adapter (HBA) storage controller information.
  - Per stack component: current values of relevant measurements such as throughput, utilization and other applicable measurements.
  - Statistical aggregations (minimum, maximum, averages and histogram) of data associated with I/O requests including size, latency per component and totals.

RHEL 5.4: Storage

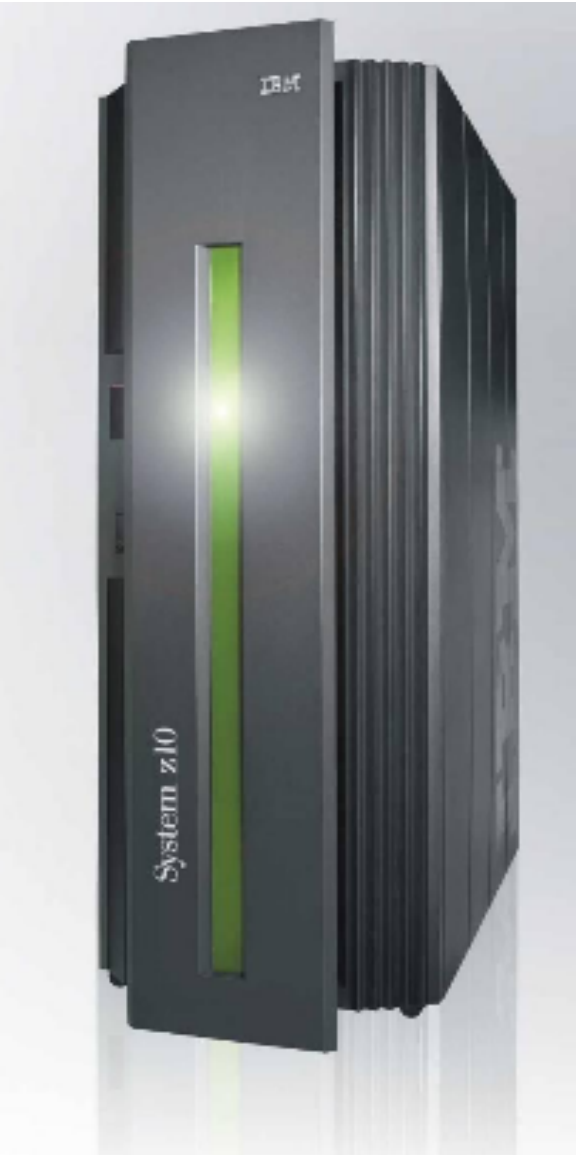
- **DS8K Encryption Support**
  - This feature enhances s390-tools to be able to display if the Storage has its disk encrypted or not.
- **Kernel Support to Issue Control I/O to DASD on EMC Symmetrix Arrays**
  - Support has been added to the kernel to issue EMC Symmetrix Control I/O. This update provides the ability to manage EMC Symmetrix storage arrays.
- **lstape Support for SCSI Tapes**
  - With this feature it is now possible to list installed FCP-attached tape devices (SCSI tapes) besides channel attached tapes using the *lstape* command

RHEL 5.4: Security

- **Long Random Numbers Generation [5.3]**
  - Provide access to the random number generator feature on the Crypto card. High volume random number generation compared to a CPU based solution.
- **Crypto Device Driver Use of Thin Interrupts [5.4]**
  - Performance update: Use interrupts instead of polling
  - Eliminates wasted CPU cycles

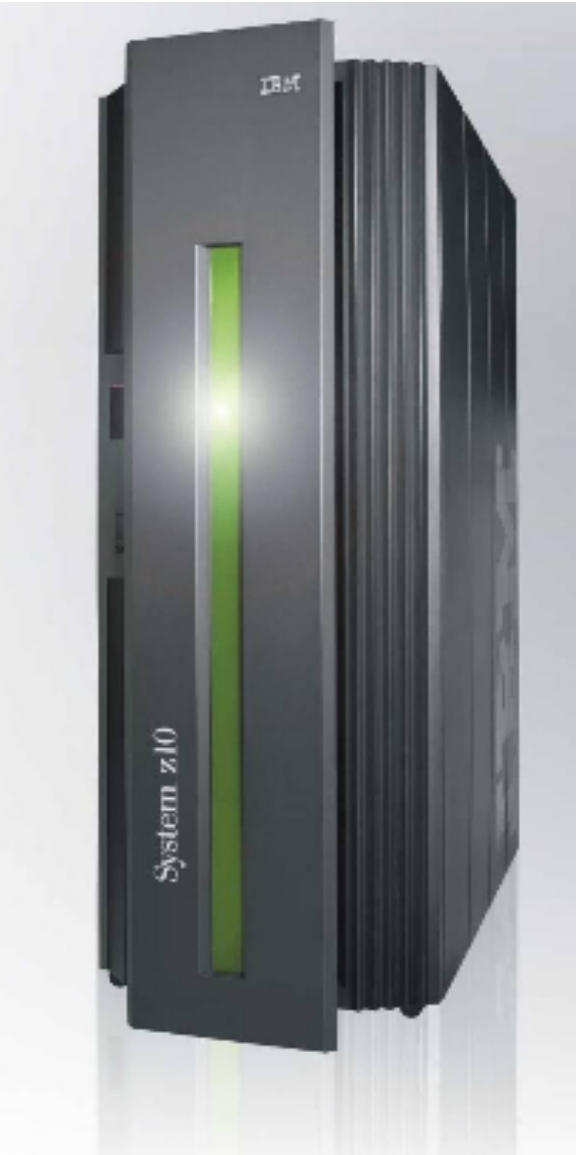
z10 Hardware-specific Updates

- **New Capabilities Through New Instruction Support**
  - Compiler supports *Decimal Floating Point* (DFP)
  - Kernel and user space libraries support *new CPU crypto algorithms* like SHA-512/384 and AES192/256
- **z10 Exploitation for Improved LPAR Performance**
  - *Node affinity* aligns process scheduling to book boundaries
  - *Vertical CPU Management* concentrates workload on fewer physical CPUs



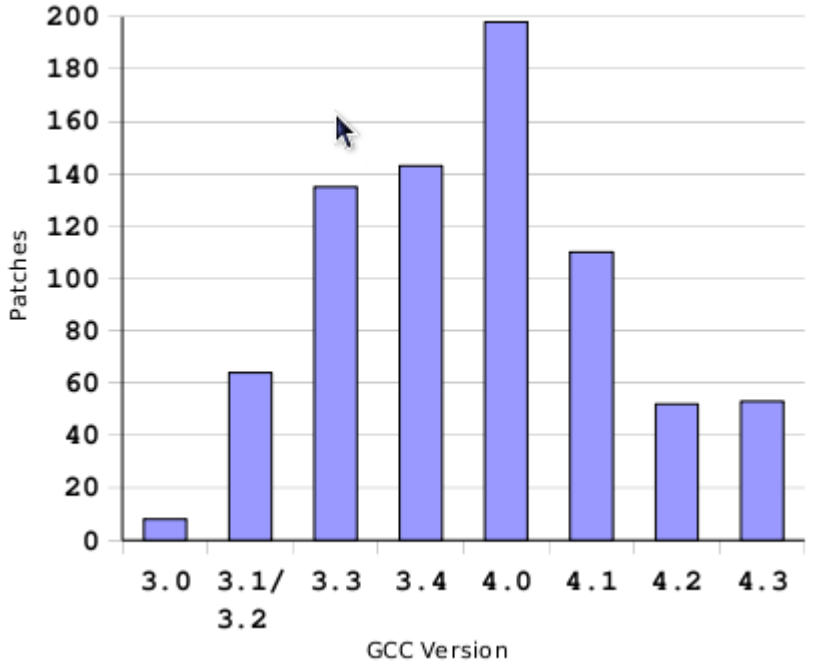
z10 Hardware-specific Updates

- **New Functions with z10**
  - *Large Page Support* minimizes lookup overhead into areas of large memory, with large page emulation on older hardware
  - *HiperSockets Layer 2* support for simplification of Linux environments
  - *Crypto Express 3* Enablement and *performance improvements*



redhat

GCC Compiler Updates



- **General Optimizer Improvements**
  - New data flow analyzer framework (GCC 4.3)
- **System z Machine Support**
  - System z10 processor support (GCC 4.4)
    - Exploit instruction new to z10
    - Selected via -march=z10 / -mtune=z10
  - Decimal floating point support (GCC 4.3)
    - For newer machines with hardware DFP support
  - 64 bit registers for 31 bit applications (> GCC 4.4)
    - Work in progress, harder than it looks
- **System z Compiler Performance**
  - Overall enhancement > 10% on z9 with industry-standard integer benchmark

RHEL 5.5

GA March, 2010

RHEL 5.5: New Features: Kernel

- **Feature: DS8000 Large Volume Support (kernel 2.6.30)**
- **Summary: Implement support for DS8k volumes, which have more than 64k cylinders. This allows DASD/ECKD larger than 50GB, reducing the need for volume management of small disks.**
- **AF_IUCV SOCK_SEQPACKET Support (kernel 2.6.31)**
- **Summary: Provides a sequenced, reliable, two-way connection-based data transmission path for datagrams of fixed maximum length. Introduces AF_IUCV sockets of type SOCK_SEQPACKET that map read/write operations to a single IUCV operation. The socket data is not fragmented. The intention is to help application developers who write applications using the native IUCV interface, e.g. Linux to z/VSE.**
- **Feature: Digitally Sign s390x Kernel**
- **Summary: Digitally sign all modules in the kernel, so that authenticity can be checked immediately by the “tainted” flag.**

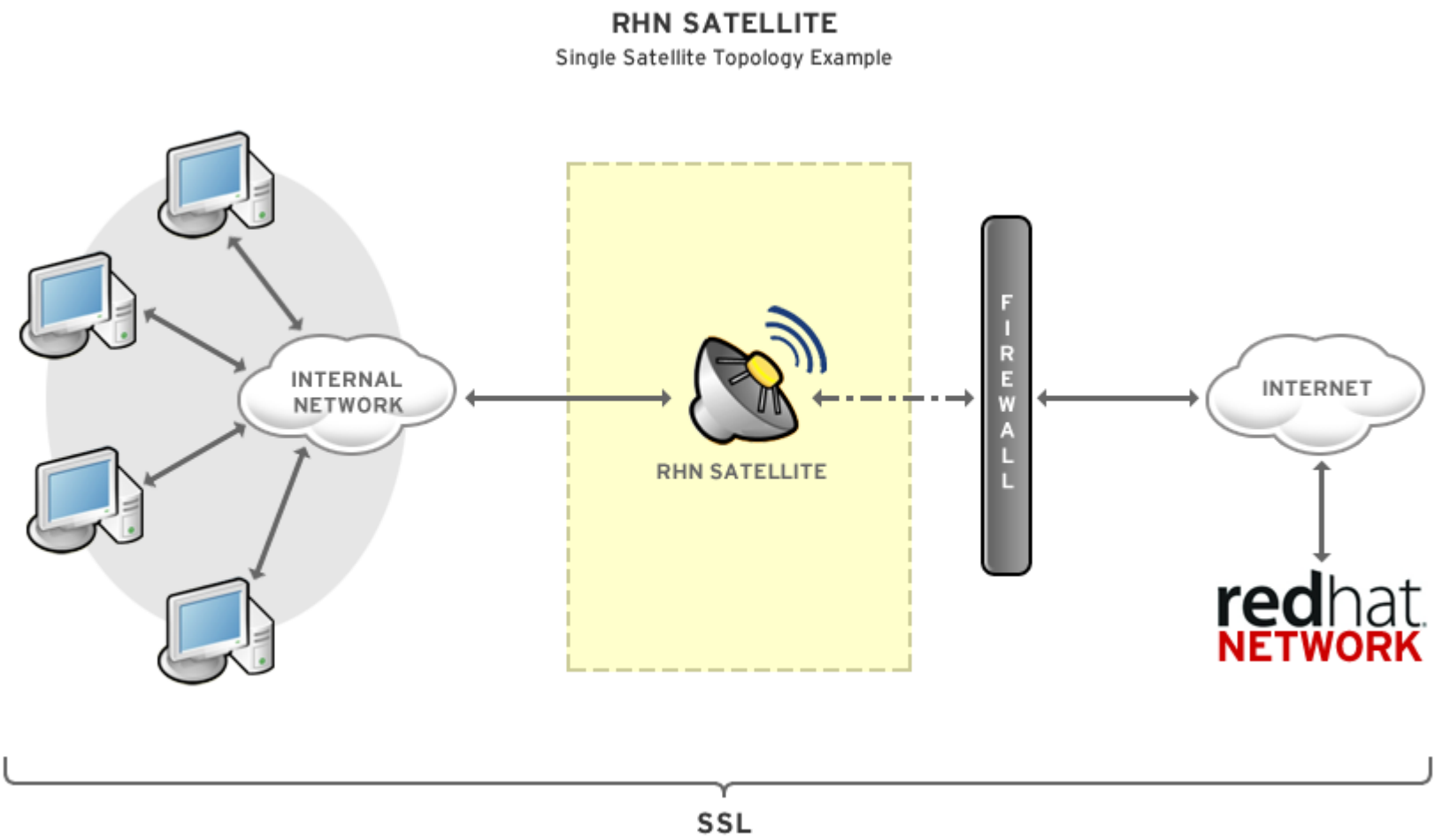
RHEL 5.5: New Features: Installer

- **Feature: Provide CMS Script for Initial IPL under z/VM**
- **Summary: In order to begin the install process, the user must create a REXX EXEC to IPL the kernel, initial RAMdisk, and PARM file. A sample EXEC file is now provided on the installation media, so this no longer needs to be manually created with CMS XEDIT.**
- **Feature: Enhanced ReIPL Support from Alternate Location**
- **Summary**: Take advantage of the ReIPL capabilities in the kernel, enabling the installer to boot into a non-default location. Works in all cases, including:
  - LPAR on machine older than z9
  - LPAR on z9 or newer (required to reIPL from FCP)
  - z/VM 5.2 or older
  - z/VM 5.4 or later (required to reIPL from FCP)

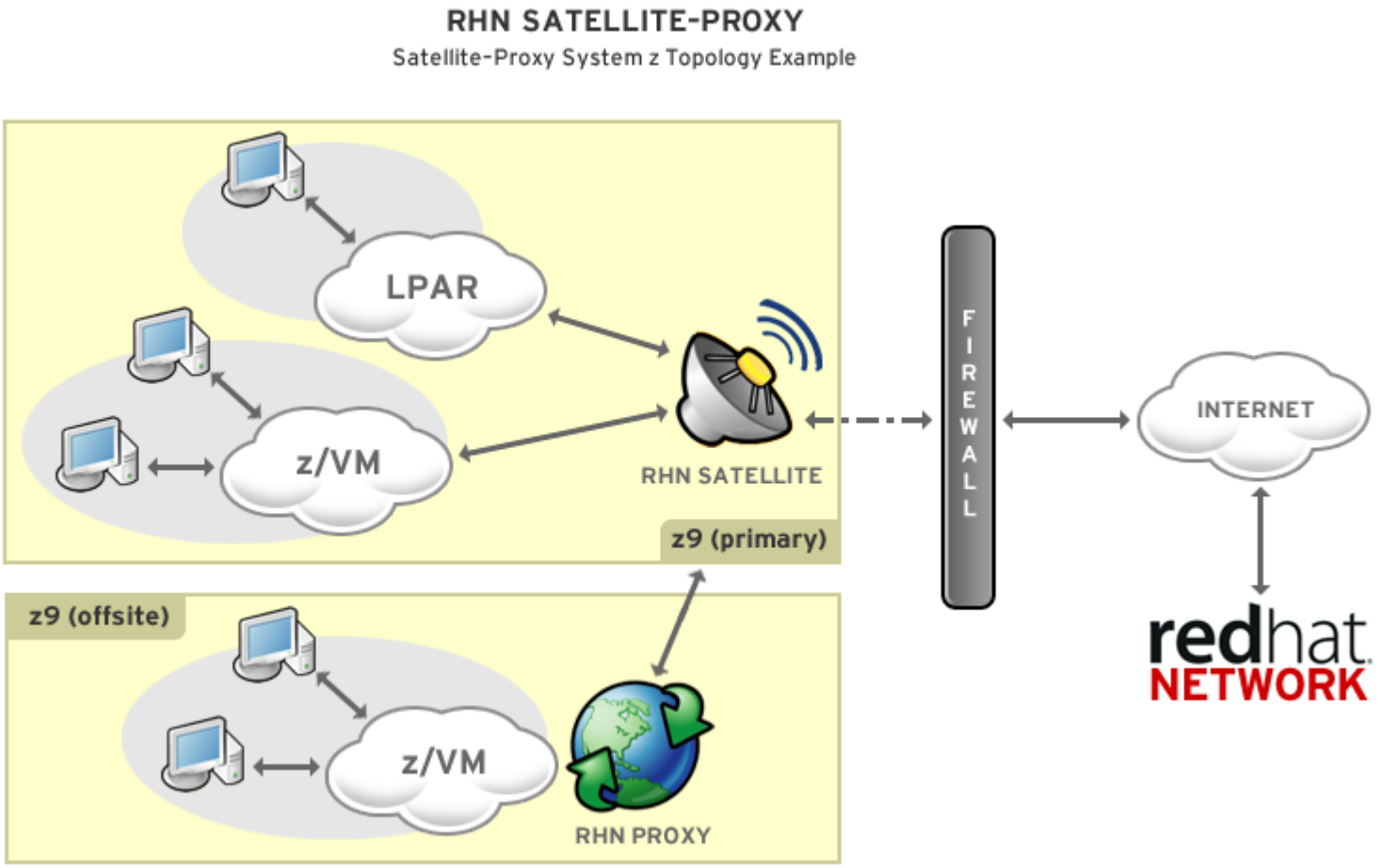
RHEL 5.5: New Features: Software

- **Update Java to Version 1.5 to SR11-FP1**
  - Critical Java updates
- **SBLIM Updates**
  - Standards Based Linux Instrumentation for Manageability
  - Add SFCB/SFCC to allow for common CIM/SEBM infrastructure across various Linux guests
  - Update Java CIM Client to version 1.3.9.1
  - Update versions of packages from sblim.org (27 updated packages total)
- **Rebase Systemtap to Version 1.1**
  - Instrumentation system for the 2.6 kernel used to collect data on the operation of the Linux guest

RHN Satellite



RHN Satellite on System z



PXE Deployment on System z

zPXE

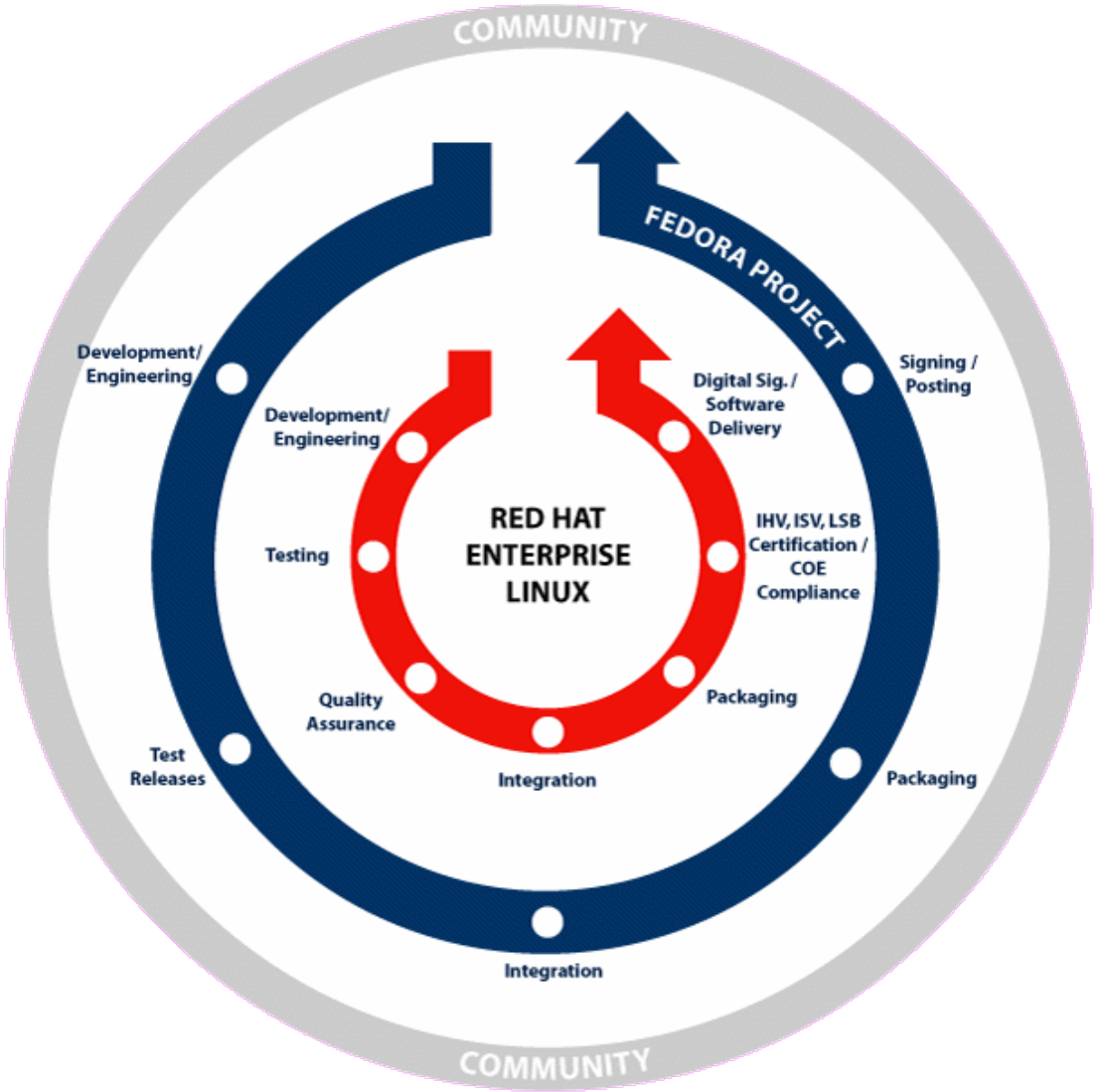
- Same configuration/profile on all guests
- Read-write 191 disk not required for each guest
- All changes kept on management server
- Flexibility of kickstart
- Same principles as traditional PXE, adapted to System z
- Fits with configuration management tools (cobbler)
- Easy to update

https://fedorahosted.org/cobbler/wiki/SssThreeNinety

redhat

Linux *continues to enable new ways of doing business*

Red Hat Development Model



Advanced Virtualization

- **Dynamic Memory Add/Remove** (kernel 2.6.27)
  - Attach and use standby memory that is configured for a logical partition or z/VM guest.
  - Memory Attach & Detach requires running Linux on System z as a VM-guest requires z/VM 5.4 plus the PTF for APAR VM64524.
- **Standby CPU Activation/Deactivation (kernel 2.6.25)**
  - Allow standby CPUs to be activated / deactivated
- **Extra Kernel Parameter for SCSI IPL** (kernel 2.6.32)
  - Modify the SCSI loader to append append extra parameters specified with the z/VM VMPARM option to the kernel command line.

Advanced Virtualization

- **hvc_iucv: Provide IUCV z/VM User ID Filtering** (kernel 2.6.29)
  - Introduces the kernel parameter "hvc_iucv_allow=" that specifies a comma-separated list of z/VM user IDs.
  - If specified, the z/VM IUCV hypervisor console device driver accepts IUCV connections from listed z/VM user IDs only.
- **Suspend / Resume** (kernel 2.6.31)
  - With suspend and resume support, you can stop a running Linux on System z instance and later continue operations.
  - When Linux is suspended, data is written to a swap partition. The resume process uses this data to make Linux continue from where it left off when it was suspended.
  - A suspended Linux instance does not require memory or processor cycles.

Suspend/Resume Support

- Ability to stop a running Linux on System z instance and later continue operations
- Memory image is stored on the swap device specified with a kernel parameter: `resume=/dev/dasd<x>`
- ```
  # grep swap /etc/fstab
  /dev/dasdd1 swap swap pri=-1 0 0
  /dev/dasde1 swap swap pri=-2 0 0
  ```
- Suspend operation is started with “echo” command to sysfs:

  ```
  # echo disk > /sys/power/state
  ```
- Resume is done automatically on IPL
- Use signal (Ctrl+Alt+Del) to automatically suspend a guest:

  ```
  ca::ctrlaltdel:/bin/sh -c "/bin/echo disk > \
  /sys/power/state || /sbin/shutdown -t3 -h now"
  ```
  ```

Storage Support

- **FCP Automated Port Discovery (kernel 2.6.25)**
  - Scan the connected fiber channel SAN and automatically activate all available and accessible target ports. This requires a proper SAN setup with zoning.
- **FCP SCSI Error Recovery Hardening (kernel 2.6.30)**
  - Avoid SCSI error recovery escalation in case of concurrent zfcp and SCSI error recovery.
- **FCP Adjustable Queue Depth (kernel 2.6.31)**
  - Customizable queue depth for SCSI commands in zfcp.

Storage Support

- **HyperPav** (kernel 2.6.25)
  - HyperPav is addressing the need to access more data with good performance and high availability!
  - This feature, which required a IBM DS8000™ disk storage system in average leads to a higher utilization, resulting in I/O transfer rates.
  - Activated automatically when the necessary prerequisites are there (DS8000 with HyperPAV LIC, z/VM 5.3). Transparent for the Linux on System z guest
- **I/O Configuration Support (kernel 2.6.27)**
  - Adds the infrastructure to allow Linux system to change the I/O configuration of a System z system.
  - Operations are addition, removal and reconfiguration/reassignment of I/O channels, control units and subchannels.
  - This support is available only when running Linux on System z in an LPAR.

Storage Support

- **High Performance FICON (HPF)** (kernel 2.6.29)
  - Added HPF support to the DASD Device Driver
  - HPF is an extension to the FICON architecture and is designed to improve the execution of small block I/O requests.
  - HPF streamlines the FICON architecture and reduces the overhead on the channel processors, control unit ports, switch ports, and links by improving the way channel programs are written and processed.
- **FCP SCSI Error Recovery Hardening** (kernel 2.6.30)
  - Avoid SCSI error recovery escalation in case of concurrent zfcp and SCSI error recovery.
- **DASD Large Volume Support** (> kernel 2.6.29)
  - Large Volume Support is a feature that allows to use ECKD devices with more than 65520 cylinders. This features is available with DS8000 R4.0

Networking

- **HiperSockets Network Traffic Analyzer (> kernel 2.6.33)**
  - Trace HiperSockets network traffic for problem isolation and resolution. Supported for layer 2 and layer 3
- **Crypto Express 3 (kernel 2.6.33)**
  - Support for Crypto Express3 Accelerator (CEX3A) and Crypto Express3 Coprocessor (CEX3C)
  - z/VM 6.1, 5.4, or 5.3 with the PTF for APAR VM64656 is required for z/VM guest-support.
- **Secondary Unicast Addresses** (kernel 2.6.27)
  - Allow secondary unicast MAC addresses to support MAC address based VLANs
  - This only works with an OSA interface running in layer 2 mode.

Networking

- **OSA QDIO Data Connection Isolation (> kernel 2.6.33)**
  - Feature (available for OSA-Express2 cards since early 2009) allows an operating system to configure the OSA adapter to prevent any direct package exchange between itself and other operating system instances that share the same OSA adapter.
  - The configuration of the isolation level is done through sysfs.
  - Starting with s390-tools 1.8.4 lsqeth indicates connection isolation in qeth attribute 'isolation'.
- **QETH Componentization** (kernel 2.6.25)
  - The qeth driver module is split into a core module and layer2-/layer3-specific modules. The default operation mode for OSA-devices is changed to layer2; for HiperSockets devices the layer3 default-mode is kept.
  - For layer3 mode devices the existence of (possibly faked) ethernet-headers is guaranteed to enable smooth integration of qeth devices into Linux.

RAS: Reliability, Availability, Serviceability

- **Automatic IPL After Dump** (kernel 2.6.30)
  - Extension to the shutdown action interface which combines the actions dump and re-ipl, first a dump is taken, then a re-ipl of the system is triggered

- **Compiler Improvements** (gcc 4.3/4.4)
  - The latest compiler enhancements allow a customer to recompile existing applications which can be optimized for the latest hardware generation without any changes to the source code.
  - This can lead up to a > 10 % performance improvement.

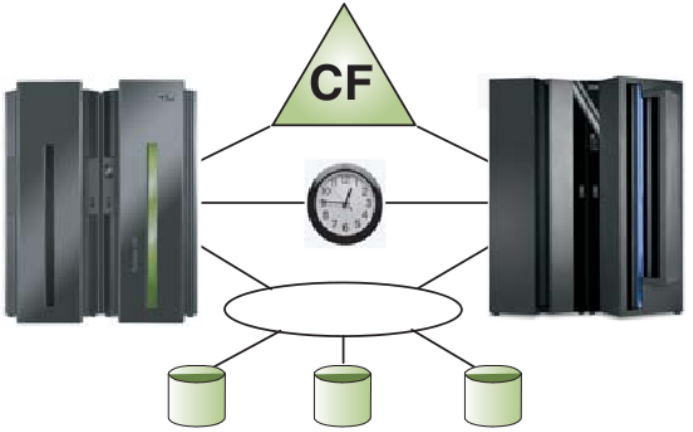
- **Large Page Support (**kernel 2.6.25)
  - Support for a new access method to allocate larger chunks of memory (1 MB page size), resulting in performance improvements, especially in Java based environments
  - This feature exploits z10 hardware features and provides a software emulation for older systems.

RAS: Reliability, Availability, Serviceability

- **Add Call Home Data on Halt and Panic if Running in LPAR (kernel 2.6.32)**
  - Report system failures (kernel panic) via the service element to the IBM service organization.
  - Improves service for customers with a corresponding service contract. (by default this features is deactivated)
- **CIO, DASD: Improved DASD Error Recovery (2.6.33)**
  - Improved the DASD error recovery procedures used in the early phases of IPL and DASD device initialization.

Miscellaneous

- **Kernel VDSO Support** (kernel 2.6.29)
  - Kernel provided shared library to speed up a few system calls (gettimeofday, clock_getres, clock_gettime)
- **Kernel Image Compression (> kernel 2.6.33)**
  - The kernel image size can be reduced by using one of three compression algorithms: gzip, bzip2 or lzma.
- **STP/ETR Support** (kernel 2.6.27)
  - Support for clock synchronization using the server time protocol (STP) or an external



Miscellaneous

- **KULI (2009-06-24)**
  - kuli is experimental userspace sample to demonstrate that KVM can be used to run virtual machines on Linux on System z.
  - This experimental proof of concept is unsupported and should not be used for any production purposes.
- **Oprofile**
  - Starting with version 0.9.4, oprofile supports sampling of Java byte code applications for Linux on System z.
- **Eclipse 3.3**
  - Starting with Eclipse 3.3, Linux on System z is officially supported.

Filesystems

- **Ext4 – main, default filesystem. Incremental scalability (ie 100TB files) & performance enhancements over ext3.**
- **NFS**
  - NFS4.0 – as default
  - NFS4.1 – referrals, delegation & failover.
  - Secure NFS, selinux labels over the wire – enabler for secure virt – timeframe likely post-GA update
- **FUSE kernel portion as enabler.**
- **BTRFS – next generation, enhanced data integrity, ease of use, & scalability. Likely tech preview at GA, non root/boot.**

Kernel Scalibility

- **Scalability Limits – Maximum Values**
  - Max CPUs – dynamic allocation of CPU structs (2.6.29) allows supported limit of 4096 and theoretical limit of 64K.
  - Max IRQs – dynamic allocation (2.6.28) allows limit of 256
  - Max memory 48-bit addressing (256TB) requires pending patch incorporation
  - Max # processes – 4 million on 64 bit kernels
  - Max threads per process remains at 32000 (same as RHEL5)
  - Filesystem limits for ext4 – 100T (practical target – bounded by fsck time)

- **Scalability Features**
  - Split LRU VM – different eviction policies for file backed vs swap backed

IPA Client (Identity, Policy, Audit)

- **IPA Client in Core RHEL for Centralized Security Management**
  - Kerberos authentication with host based access control
  - Provides central storage of extended user attributes
  - Enables centralized policy for applications, including SELinux policy
  - Audit log aggregation services & search capabilities

Security enhancements

- **Virtualization Isolation in Conjunction with SELinux**
  - May be post-GA update
  - Labeled NFS for filesystem isolation
  - Cut-and-paste window controls per authorization level
  - Guest confinement via SELinux policy enhancements
- **NSS Crypto**
  - Broaden the core services which utilize NSS crypto libraries
  - Allows cheaper implementation of new features, ie TPM & centralized key store
  - Incremental targeted conversion of: Openswan, openldap, glibc
  - Add new crypto GUI for key import & establishment of trust

Security enhancements

- **Volume Encryption**
  - Basic operation already present in RHEL5 – incremental centralized key management for RHEL6
- **Sectool**
  - Compliance checking / intrusion detection utility – validates system admin config: file permissions, valid UIDs, reasonable passwords, etc

Documentation Links

- **Documentation/Getting Started**
  - **Redbook, “z/VM and Linux on IBM System z: The Virtualization Cookbook for Red Hat Enterprise Linux 5.2”**
    - http://www.redbooks.ibm.com/abstracts/sg247492.html
    - Coming soon: RHEL 6 and z/VM 6.1
  - **DeveloperWorks**:
    - http://www.ibm.com/developerworks/linux/linux390/documentation_dev.html
  - **Knowledgebase**:
    - http://kbase.redhat.com/
    - Search “s390”
  - **http://www.redhat.com/z**

Questions and Open Dialog