PRESENTED BY RED HAT

LEARN. NETWORK.
EXPERIENCE OPEN SOURCE.

www.theredhatsummit.com

Intel & Red Hat
Pushing the Scalability Envelope

Fal Diabate: Intel Strategic Relations Manager
Prarit Bhargava: Red Hat Principal Engineer

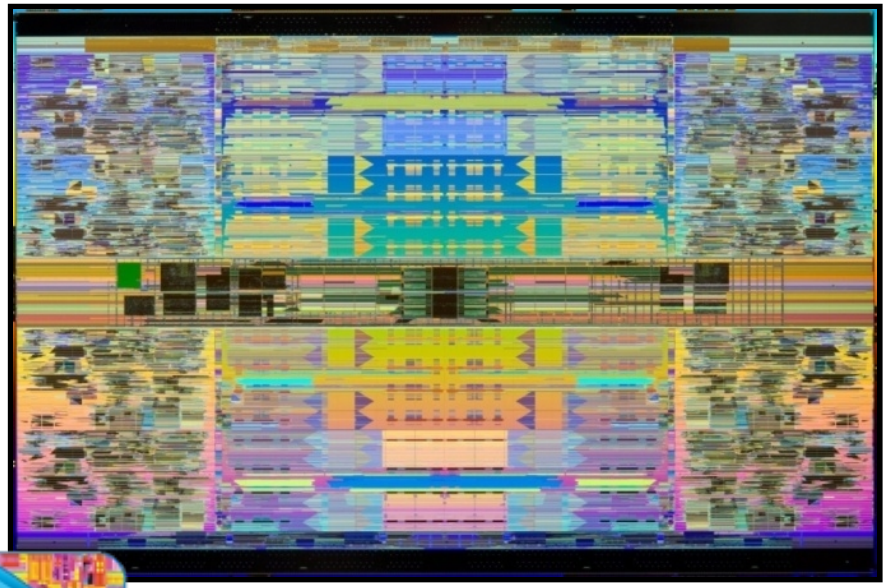
Friday June 25, 2010

Intel Legal Disclaimer

- Intel may make changes to specifications and product descriptions at any time, without notice.
- Performance tests and ratings are measured using specific computer systems and/or components and reflect the approximate performance of Intel products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance. Buyers should consult other sources of information to evaluate the performance of systems or components they are considering purchasing. For more information on performance tests and on the performance of Intel products, visit Intel Performance Benchmark Limitations
- Intel does not control or audit the design or implementation of third party benchmarks or Web sites referenced in this document. Intel encourages all of its customers to visit the referenced Web sites or others where similar performance benchmarks are reported and confirm whether the referenced benchmarks are accurate and reflect performance of systems available for purchase.
- Intel processor numbers are not a measure of performance. Processor numbers differentiate features within each processor family, not across different processor families. See www.intel.com/products/processor_number for details.
- Intel, processors, chipsets, and desktop boards may contain design defects or errors known as errata, which may cause the product to deviate from published specifications. Current characterized errata are available on request.
- Intel Virtualization Technology requires a computer system with a processor, chipset, BIOS, virtual machine monitor (VMM) and applications enabled for virtualization technology. Functionality, performance or other virtualization technology benefits will vary depending on hardware and software configurations. Virtualization technology-enabled BIOS and VMM applications are currently in development.
- 64-bit computing on Intel architecture requires a computer system with a processor, chipset, BIOS, operating system, device drivers and applications enabled for Intel® 64 architecture. Performance will vary depending on your hardware and software configurations. Consult with your system vendor for more information.
- Lead-free: 45nm product is manufactured on a lead-free process. Lead is below 1000 PPM per EU RoHS directive (2002/95/EC, Annex A). Some EU RoHS exemptions for lead may apply to other components used in the product package.
- Halogen-free: Applies only to halogenated flame retardants and PVC in components. Halogens are below 900 PPM bromine and 900 PPM chlorine.
- Intel, Intel Xeon, Intel Core microarchitecture, and the Intel logo are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.
- © 2009 Standard Performance Evaluation Corporation (SPEC) logo is reprinted with permission

Introducing the Intel® Xeon® Processor 7500 Series

Formerly known as Nehalem-EX



A New Generation of Intelligent Servers

Common IT Challenges

- Server Sprawl
- Power and Cooling
- Under-utilized Assets
- High Operating Costs
- Insufficient Space
- Capital Constraints

Characteristics of Data Demanding Processing

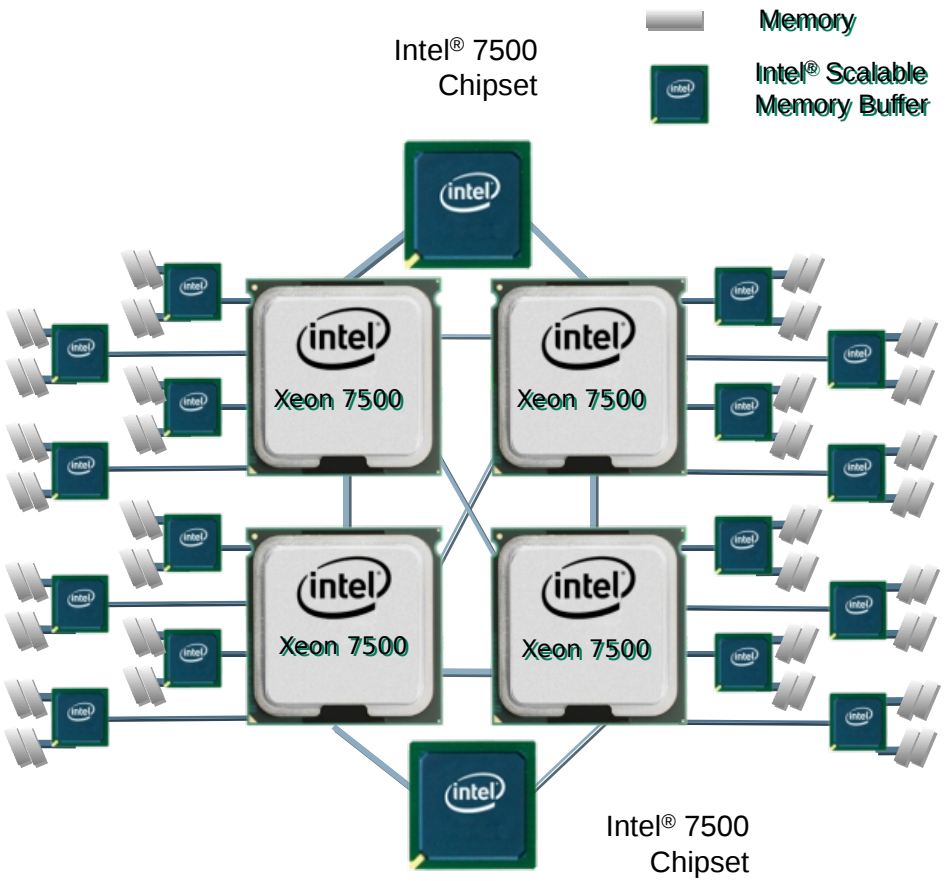
Maximum Performance	• Biggest Workloads • Unpredictable Workloads
Expandability	• Accommodate Growth
High Availability	• Mission Critical Usage • Lower Down Time
Best Performance/$ @ Maximum Capacity	• Economic Processing of Large Workloads

Data Demanding Workloads Requires Optimized Hardware

Intel® Xeon® Processor 7500 Series

- New processor architecture
- New platform architecture
- New memory subsystem
- New I/O subsystem
- New Mission Critical RAS
- New Levels of Scalability

The biggest performance jump ever in Xeon® history!



Scalable Performance

Flexible Virtualization

Advanced Reliability

Delivering Intelligent Performance

Nehalem Generation Intel® Microarchitecture

Threaded Applications

- 45nm – up to 8-core Intel® Xeon® Processors
- Intel® Hyper-threading Technology

Performance on Demand

- Intel® Turbo Boost Technology
- Intel® Intelligent Power Technology

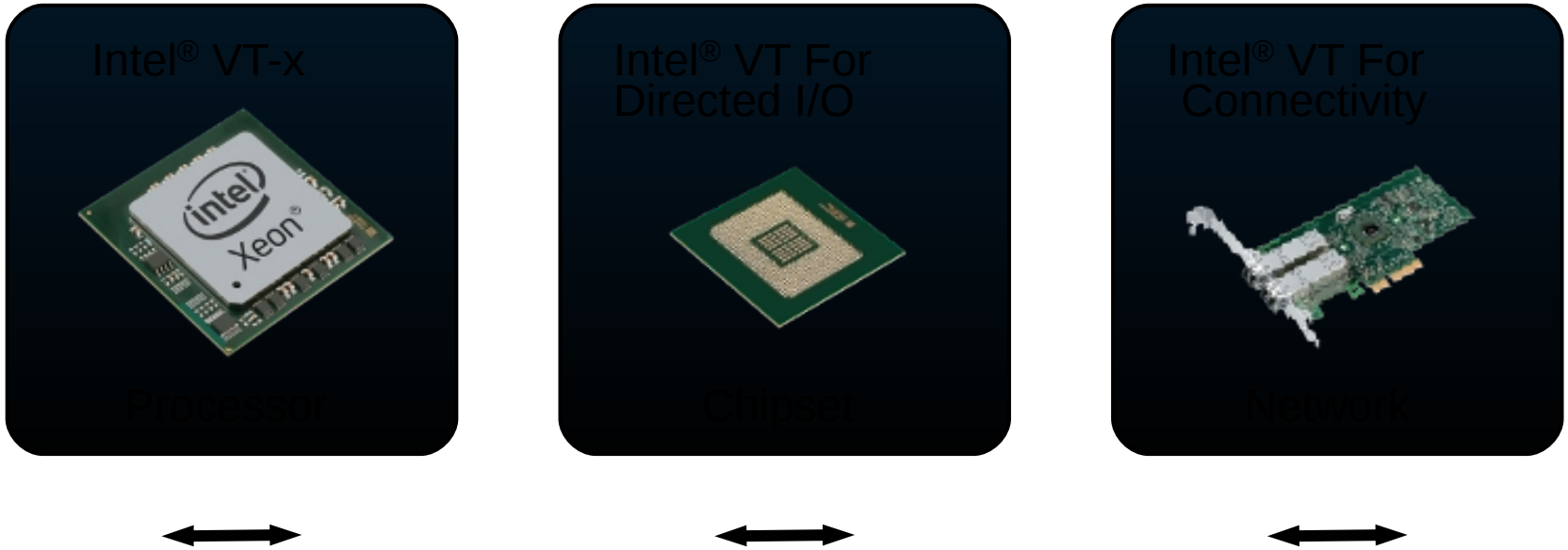
Bandwidth Intensive

- Intel® QuickPath Technology
- Integrated Memory Controller

Performance That Adapts to Your Software Environment

Intel® Xeon® 7500/6500 Series:

End-to-end Platform Virtualization



Extended Page Tables | Intel® VT-d | Intel® VT-c | Intel® VT Flex Migration

Intel Platform Virtualization Technologies

Intel® VT Flex Priority | Virtual Machine Device Queues (VMDq)

Advanced RAS* Delivers Value For IT

Protects Your Data

- Reduces circuit-level errors
- Detects data errors across the system
- Limits the impact of errors

Minimizes Planned Downtime

- Maintain partitions instead of systems
- Replace components before they fail

Increases Availability

- Heals failing data connections
- Migrates workloads from failing CPU & memory
- Helps predict failures before they happen
- Recovers from uncorrected data errors

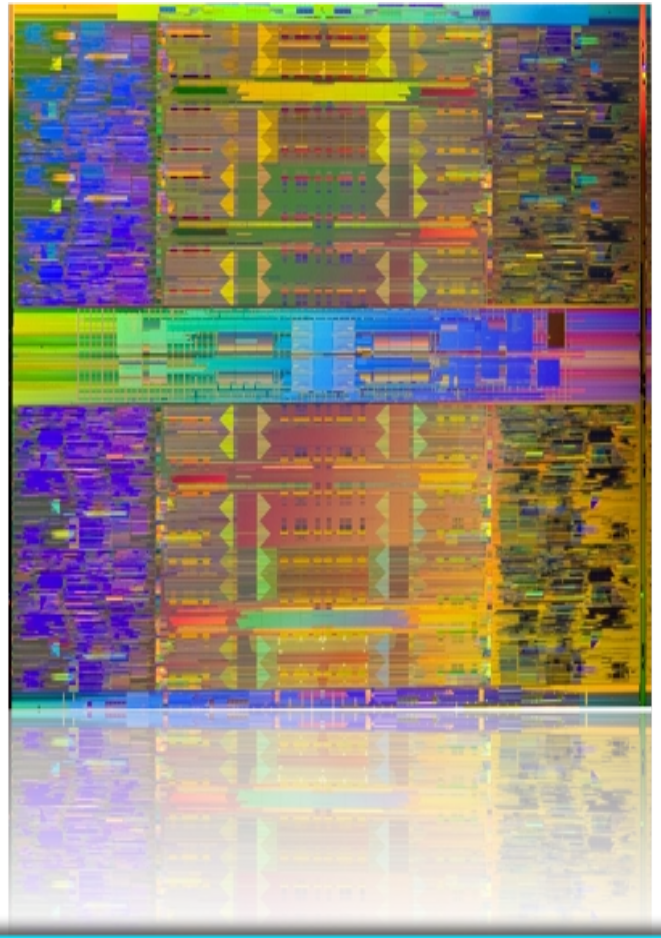
Nehalem-EX Solutions Span Silicon, OS, System

RAS* = Reliability, Availability, and Serviceability

Advanced Reliability Starts With Silicon….
Nehalem-EX Reliability Features

Memory

- Inter-socket Memory Mirroring
- Intel® Scalable Memory Interconnect (Intel® SMI) Lane Failover
- Intel® SMI Clock Fail Over
- Intel® SMI Packet Retry
- Address Parity via Memory Lockstep Operation
- Failed DIMM Isolation
- Physical Memory Board Hot Add/remove
- Dynamic/OS Assisted Memory Migration
- Dynamic/OS Memory On-lining (capacity change)
- Demand and Patrol scrubbing
- Fail Over from Single DRAM Device Failure (SDDC)
- DIMM and Rank Sparing
- Intra-socket Memory Mirroring
- Mirrored Memory Board Hot Add/Remove



I/O Hub

- Physical IOH Hot Add
- Dynamic/OS IOH On-lining (capacity change)
- PCI-E Hot Plug

CPU/Socket

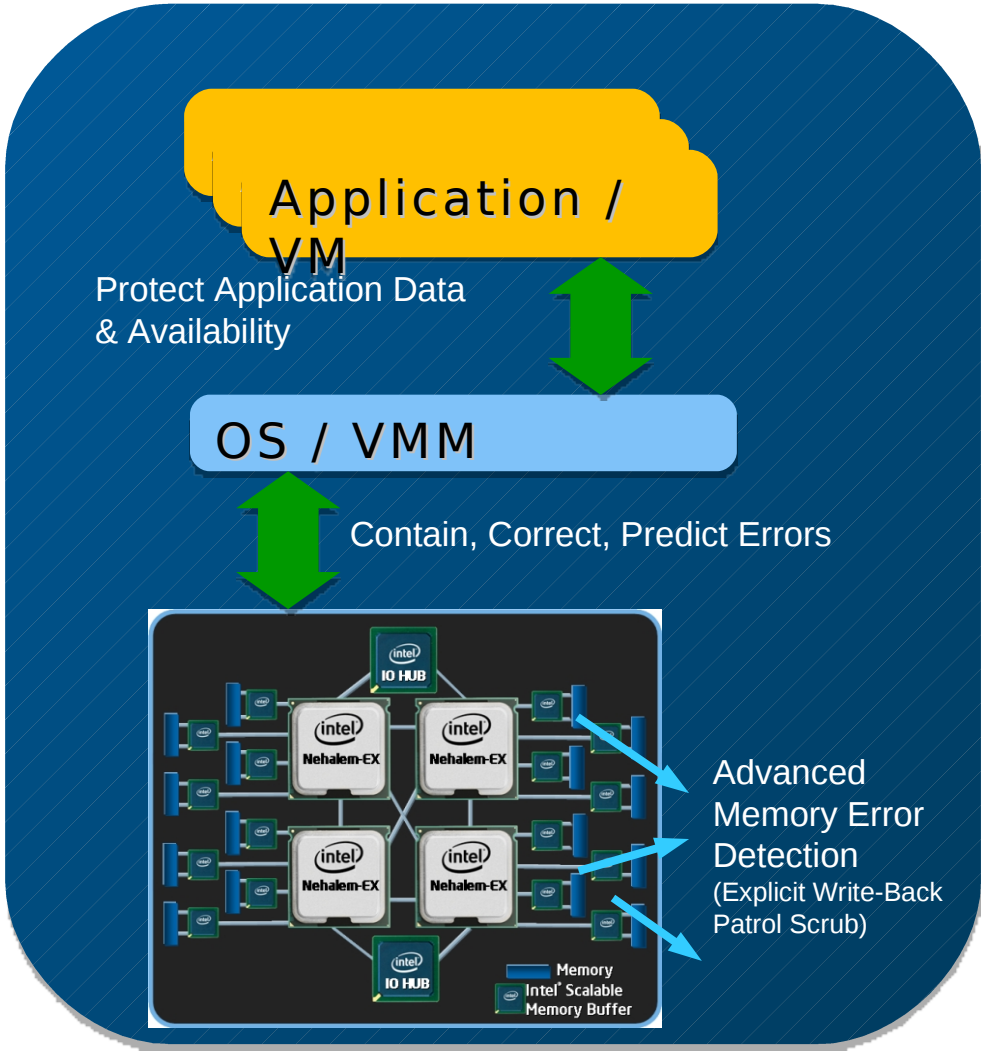
- MCA-recovery
- CMCI
- Data Poisoning/ and Viral Mode
- Dynamic Processor Sparing and Migration
- Static Hard Partitioning
- On-Die Error Protection

Intel® QuickPath Interconnect

- Intel QPI Packet Retry
- Intel QPI Protocol Protection via CRC (8bit or 16bit rolling)
- QPI Clock Fail Over
- QPI Self-Healing

Over 20 New RAS features at all levels!

Advanced RAS Example:
Machine Check Architecture (MCA) Recovery



Increases system availability

- OS can terminate/restart an application
- VMM can terminate a single guest OS and keep the server running
- More benefits in virtualized environments

Protects application data

- Supports error containment to isolate error location before it affects other data

Can reduce service costs

- Allow for failure prediction and corrective action through Corrected Error Signaling
- Allows failing components to be identified and replaced during planned maintenance cycles

First Machine Check Recovery in Xeon®-based Systems

Intel® Xeon® Processor 7500 Series Benefits

Scalable Performance	Flexible Virtualization	Advanced Reliability
Up to 3.8X performance boost over Xeon 7400	I/O Virtualization	Over 20 new RAS features
1 terabyte of memory *(4S)*	Lower cost/VM vs 2skt EP	Machine Check Architecture-recovery • Recover from fatal errors • 1st time on X86 architecture •
Scaling: 2-8+ sockets	Intel VT Flex-Migration Assist for live migration across multi-generations of Xeon servers	Broad software and server support on a full range of server designs
9X memory bandwidth boost		
Over 20x performance vs. single-core servers (4S)		
Biggest Performance Leap Ever for Xeon	**Top VM Capability & Investment Protection**	**Mission Critical Reliability**

Back to Prarit

Intel & Red Hat

Pushing The Scalability Envelope

Fal Diabate: Intel Strategic Relations Manager
Prarit Bhargava: Red Hat Principal Software Engineer

Friday June 25, 2010

RHEL 6 and Intel Xeon 7500 Processors

- RAS (Reliability Availability Serviceability) Features
- Virtualization Features

RHEL 6.0 RAS Features

- MCA/MCE
- PCIe AER
- CPU Physical Add
- CPU Logical Add/Remove
- Hardware Poisoning

RHEL 6.0 RAS Features

- MCA/MCE
  - Machine Check Architecture
  - Machine Check Error/Machine Check Exception
- PCIe AER
- CPU Physical Add
- CPU Logical Add/Remove
- Hardware Poisoning

RHEL6 RAS: Intel Xeon 7500 Processors & MCE

- On console you would see

```
Machine check events logged
Machine check poll done on CPU 23
Starting machine check poll CPU 23
Machine check events logged
Machine check poll done on CPU 23
```

RHEL6 RAS: Intel Xeon 7500 Processors & MCE

- Run ‘mcelog’ utility to determine what happened

```
[root@intel-75xx mce-inject]# mcelog
HARDWARE ERROR. This is *NOT* a software problem!
Please contact your hardware vendor
MCE 0
CPU 23 BANK 1
ADDR abcd
TIME 1273773919 Thu May 13 14:05:19 2010
MCG status:
MCi status:
Error enabled
MCi_ADDR register valid
MCA: No Error
STATUS 9400000000000000 MCGSTATUS 0
MCGCAP 1000c16 APICID 72 SOCKETID 3
CPUID Vendor Intel Family 6 Model 46
```

RHEL6 RAS: Intel Xeon 7500 Processors & MCE

- What about a fatal error and a system crash?
- Console displays

```
intel-75xx.lab.bos.redhat.com login:
HARDWARE ERROR
CPU 0: Machine Check Exception:                0 Bank 0: b200000000000000
TSC 14ad69e42e3
PROCESSOR 0:206e6 TIME 1273774963 SOCKET 0 APIC 0
No human readable MCE decoding support on this CPU type.
Run the message through 'mcelog --ascii' to decode.
This is not a software problem!
```

RHEL6 RAS: Intel Xeon 7500 Processors & MCE

- Again, start mcelog but with –ascii option

```
[root@intel-75xx ~]# mcelog –ascii < /tmp/console.out

CPU 0: Machine Check Exception:                 0 Bank 0: b200000000000000
TSC 37374fa664a
CPU 0 BANK 0 TSC 37374fa664a
TIME 1273774218 Thu May 13 14:10:18 2010
STATUS b200000000000000 MCGSTATUS 0
PROCESSOR 0:206e6 TIME 1273774218 SOCKET 0 APIC 0
```

RHEL 6.0 RAS Features

- MCA/MCE
- PCIe AER
  - PCIe Advanced Error Reporting
  - PCIe Advanced Error Recovery
  - (PCIe AERR) PCI Advanced Error Reporting and Recovery
- CPU Physical Add
- CPU Logical Add/Remove
- Hardware Poisoning

RHEL 6 RAS: Intel Xeon 7500 Processors & PCIe AER

- Specialized PCIe hardware required
- Detects correctable and uncorrectable errors on PCIe devices
- PCIe AER recovery requires driver modifications
- e1000e, igb, ixgb, netxen, arcmsr, lpfc, qla2xxx, etc.

RHEL 6 RAS: Intel Xeon 7500 Processors & PCIe AER

- PCIe AER errors report against specific device
- PCI address shown in error message

```
PCIeport 0000:80:00.0: AER: Corrected error received: id=8000
```

RHEL 6 RAS: Intel Xeon 7500 Processors & PCIe AER

- PCIe AER errors report against specific device
- PCI address shown in error message

PCIeport 0000:80:00.0: AER: Corrected error received: id=8000

[root@intel-s3e36-03 rhel6]# lspci | grep '80:00.0'
80:00.0 PCI bridge: Intel Corporation 5500 Non-Legacy I/O Hub PCI Express Root Port 0 (rev 22)

RHEL 6 RAS: Intel Xeon 7500 Processors & PCIe AER

```
PCIeport 0000:80:00.0: AER: Uncorrected (Non-Fatal) error received: id=8000
PCIeport 0000:80:00.0: PCIe Bus Error: severity=Uncorrected (Non-Fatal), type=Data Link
Layer, id=8000(Completer ID)
PCIeport 0000:80:00.0:   device [8086:3420] error status/mask=001ff011/00100000
PCIeport 0000:80:00.0:    [ 0] Unknown Error Bit (First)
PCIeport 0000:80:00.0:    [ 4] Data Link Protocol
PCIeport 0000:80:00.0:    [12] Poisoned TLP
PCIeport 0000:80:00.0:    [13] Flow q      Control Protocol
PCIeport 0000:80:00.0:    [14] Completion Timeout
PCIeport 0000:80:00.0:    [15] Completer Abort
PCIeport 0000:80:00.0:    [16] Unexpected Completion
PCIeport 0000:80:00.0:    [17] Receiver Overflow
PCIeport 0000:80:00.0:    [18] Malformed TLP
PCIeport 0000:80:00.0:    [19] ECRC
PCIeport 0000:80:00.0:   TLP Header: 00000000 00000001 00000002 00000003
PCIeport 0000:80:00.0: broadcast error_detected message
PCIeport 0000:80:00.0: broadcast mmio_enabled message
PCIeport 0000:80:00.0: broadcast resume message
PCIeport 0000:80:00.0: AER driver successfully recovered
```

RHEL 6.0 RAS Features

- MCA/MCE
- PCIe AER
- CPU Physical Add
  - CPU “Hot” Add
  - Socket Add
- CPU Logical Add/Remove
- Hardware Poisoning

RHEL 6 RAS: Intel Xeon 7500 Procs & CPU Hot Add

- Memory controller on die
- Memory “behind” processor comes and goes with processor
- Automatic memory online
- udev brings CPUs online

RHEL 6 RAS: Intel Xeon 7500 Procs & CPU Hot Add

- Flip a physical or remote switch to bring processor into service, trigger ACPI events
- Memory added first ....

```
Container driver received ACPI_NOTIFY_BUS_CHECK event
Hotplug Mem Device
On node 3 totalpages: 0
init_memory_mapping: 000000cd00000000-000000d100000000
 cd00000000 - d100000000 page 2M
[ffffea02cd9c0000-ffffea02cd9fffff] potential offnode page_structs
 [ffffea02cd800000-ffffea02cd9fffff] PMD -> [ffff880465600000-ffff8804657fffff] on node 3
[ffffea02cdb80000-ffffea02cdbfffff] potential offnode page_structs
[ffffea02cdd40000-ffffea02cddfffff] potential offnode page_structs
 [ffffea02cda00000-ffffea02cddfffff] PMD -> [ffff880461800000-ffff880461bfffff] on node 3
<snip>
```

RHEL 6 RAS: Intel Xeon 7500 Procs & CPU Hot Add

- ... then CPU components

```
ACPI: HARDWARE addr space,NOT supported yet
processor LNXCPU:30: registered as cooling_device48
Built 4 zonelists in Zone order, mobility grouping on.  Total pages: 12331603
Policy zone: Normal
processor LNXCPU:35: registered as cooling_device53
processor LNXCPU:36: registered as cooling_device54
processor LNXCPU:37: registered as cooling_device55
processor LNXCPU:38: registered as cooling_device56
<snip>
```

RHEL 6 RAS: Intel Xeon 7500 Procs & CPU Hot Add

- ... finally CPUs brought online

```
Booting Node 3 Processor 51 APIC 0x63
Booting Node 3 Processor 52 APIC 0x64
Booting Node 3 Processor 53 APIC 0x65
Booting Node 3 Processor 54 APIC 0x66
Booting Node 3 Processor 55 APIC 0x67
Booting Node 3 Processor 56 APIC 0x70
Booting Node 3 Processor 57 APIC 0x71
Booting Node 3 Processor 58 APIC 0x72
Booting Node 3 Processor 60 APIC 0x74
Booting Node 3 Processor 59 APIC 0x73
Booting Node 3 Processor 61 APIC 0x75
Booting Node 3 Processor 62 APIC 0x76
Booting Node 3 Processor 63 APIC 0x77
```

RHEL 6.0 RAS Features

- MCA/MCE
- PCIe AER
- CPU Physical Add
- CPU Logical Add/Remove
  - CPU Soft Add/Remove
- Hardware Poisoning

RHEL 6 RAS: Intel Xeon 7500 Procs & Soft Add/Remove

- This is the standard well-known procedure of taking a CPU offline
- Useful for serviceability

```
[root@intel-75xx ~]# echo 0 > /sys/devices/system/cpu/cpu23/online
CPU 23 is now offline
[root@intel-75xx ~]# echo 1 > /sys/devices/system/cpu/cpu23/online
Booting Node 2 Processor 23 APIC 0x36
```

RHEL 6.0 RAS Features

- MCA/MCE
- PCIe AER
- CPU Physical Add
- CPU Logical Add/Remove
- Hardware Poisoning

RHEL 6 RAS: Intel Xeon 7500 Procs & Hardware Poisoning

- RHEL5 – uncorrectable memory errors lead to panic
- RHEL6 – new feature, allows system recovery but not necessarily application recovery
- On RHEL6 ...

May 17 20:05:33 intel-75xx kernel: MCE 0x170f73: dirty LRU page recovery: Recovered
May 17 20:05:33 intel-75xx kernel: MCE: Killing firefox :30510 due to hardware memory corruption fault at 7f7189333000

RHEL 6: Intel 7500 Series & Virtualization

- Partitioning via virtualization (a.k.a. virtual partitioning)
  - VT-x, VT-d (IOMMU), VT-c and VMDq (SR-I/OV)
- CPU Migration (capacity change in guest)
  - Virtual CPU Soft Plug
- OS CPU Onlining (capacity change in host)
  - Physical CPU Hotplug

Questions & Answers (hopefully)

- Prarit Bhargava, prarit@redhat.com
- Fal Diabate, Fal.Diabate@intel.com

FOLLOW US ON TWITTER

www.twitter.com/redhatsummit

TWEET ABOUT IT

#summitjbw

READ THE BLOG

http://summitblog.redhat.com/