

Towards Mission Critical Xen -- RAS Enabling Update

Jiang, Yunhong
<yunhong.jiang@intel.com>

Ke, Liping <liping.ke@intel.com>

Liu, Jinsong <jinsong.liu@intel.com>



Software and Solutions Group



Legal Information

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.

Intel may make changes to specifications and product descriptions at any time, without notice.

All products, computer systems, dates, and figures specified are preliminary based on current expectations, and are subject to change without notice.

Intel is a trademark of Intel Corporation in the U.S. and other countries.

*Other names and brands may be claimed as the property of others.

Copyright © 2009, Intel Corporation. All rights are protected.



Agenda

- **Overview**
- CPU/Memory error handling
- Host CPU hot-plug support
- Guest vCPU Hot-plug



Continuous Improvement on RAS Support

	February 2009 (Xen Summit North America 2009)	Now
CPU/Memory error handling	<ul style="list-style-type: none"> • Infrastructure proposed • Implementation WIP 	Checked-in
I/O error handling	<ul style="list-style-type: none"> • Infrastructure proposed • PV guest supported 	WIP for HVM guest support
Host CPU hot-add	Not Started Yet	checked-in
Host memory hot-add	Not started yet	WIP
Guest vCPU hot-plug	Not started yet	Ready for send out
Guest vMem hot-plug	Not started yet	May not support

Agenda

- Overview
- CPU/Memory error handling
- Host CPU hot-plug support
- Virtual CPU Hot-plug



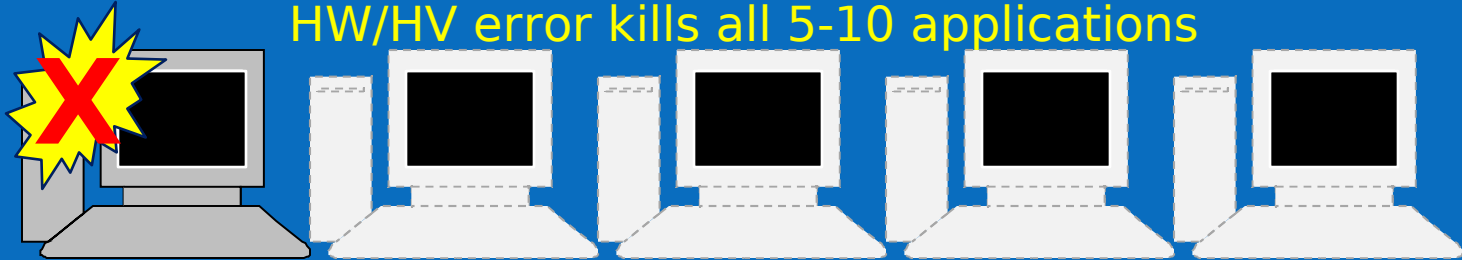
Why error handling enhancement – Motivation

Pre-Virtualization – 1 App/svr, error kills 1 application



Post-Virtualization – ~5-10 App/svr

HW/HV error kills all 5-10 applications



Error handling enhancements

HW/HV error kills 1 virtual application that error localized



Retrospect: CPU/Memory Error Handling

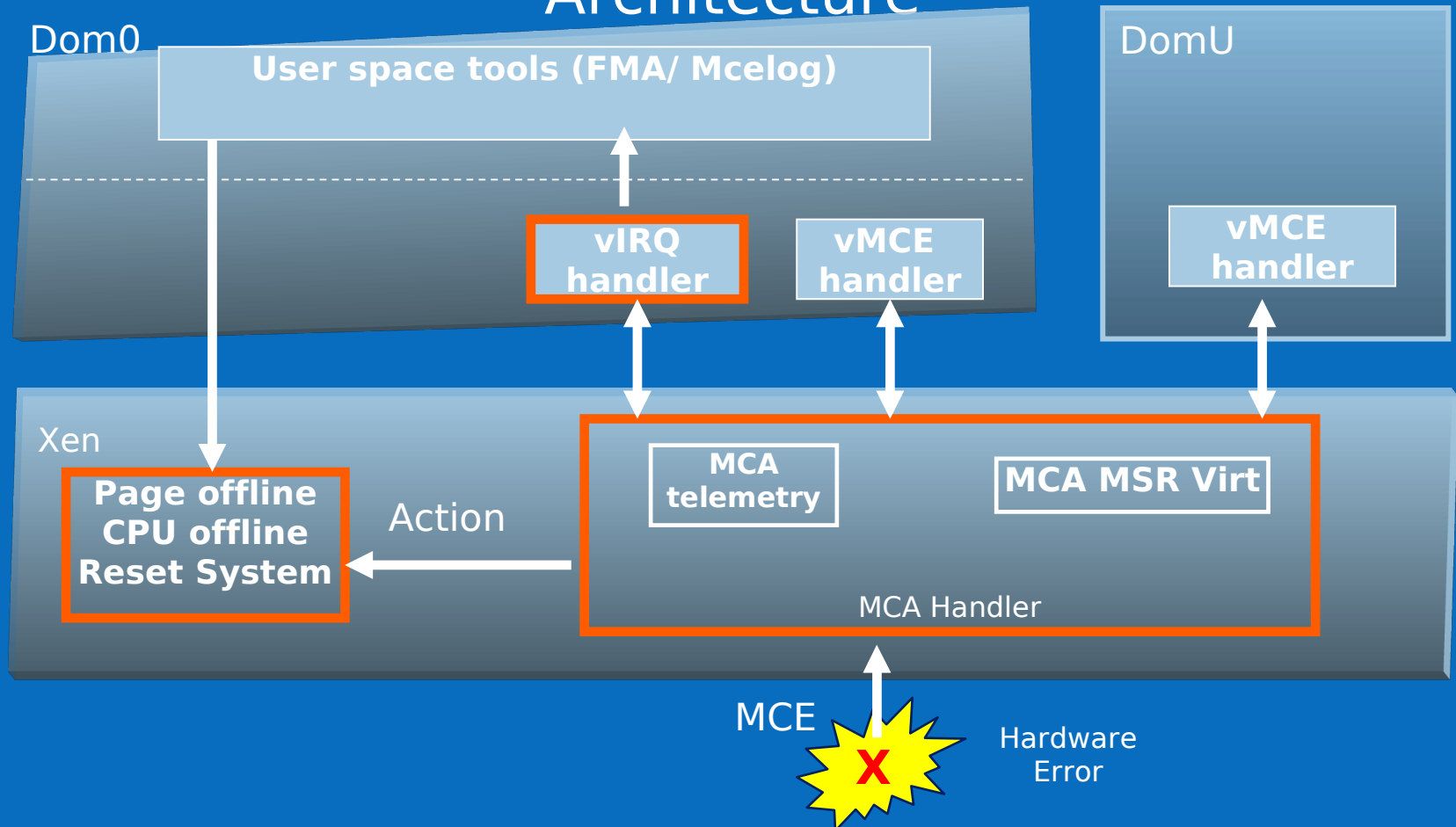
Flow

- Error happens to CPU or memory and is detected by hardware
 - E.g. ECC error in L3 Cache , ECC error in memory cell
- A MCE (Machine Check Exception) is raised to Xen hypervisor
- Xen hypervisor MCE handler will parse error information
 - Information from: MCE error code, MCE MSR's etc
- Xen hypervisor take action to recover the error, e.g.
 - Offline a page for memory error
 - Killing impacted guest
- System continues if error is recovered successfully
 - Log the error information in dom0

- Otherwise reset the system



Retrospect: CPU/Memory Error Handling Architecture



Infrastructure is implemented Already by Ke, Liping/Frank van der Linden/ Christoph Egger

Next Step for CPU/Memory Error Handling

- Add xen awareness to Linux MCA tools
- Support for more MCA error code
 - Now Supports two software recoverable Errors
 - UCR Errors detected by memory controller scrubbing
 - UCR Errors detected during L3 cache explicit write-backs
- Enhancement to memory error handling (see next pages)



Memory Error Handling– Current Solution

- Reading access to broken memory affects data integrity
 - Whole system may even crash
- Recover action from xen hypervisor

Type	Probability	Action
Free memory	Depends on workload	Offline the page
Guest memory	Large	<ul style="list-style-type: none">• Off line the page when it is freed• A virtual MCE is sent to guest
Critical Memory	Small	Reset the system
<ul style="list-style-type: none">• Xen's private data/heap• Shadow/HAP/IOMMU/P2M page tables• The granted memory used by backend service		



Memory Error Handling Enhancement (Cont'd)

- Issue: Xen may access broken guest memory
 - Xen scans guest's page tables when killing shadow mode guest
 - Xen hypervisor crashes if one page table page is broken
 - Xen access guest's memory for instruction emulation
 - KExec access the broken pages
 -
 - Proposal: Avoid high-possibility access
- Issues: guest's access to broken memory is not prevented
 - Malicious guest can trigger system crash
 - Proposal: Detecting the access in Xen hypervisor in advance

Agenda

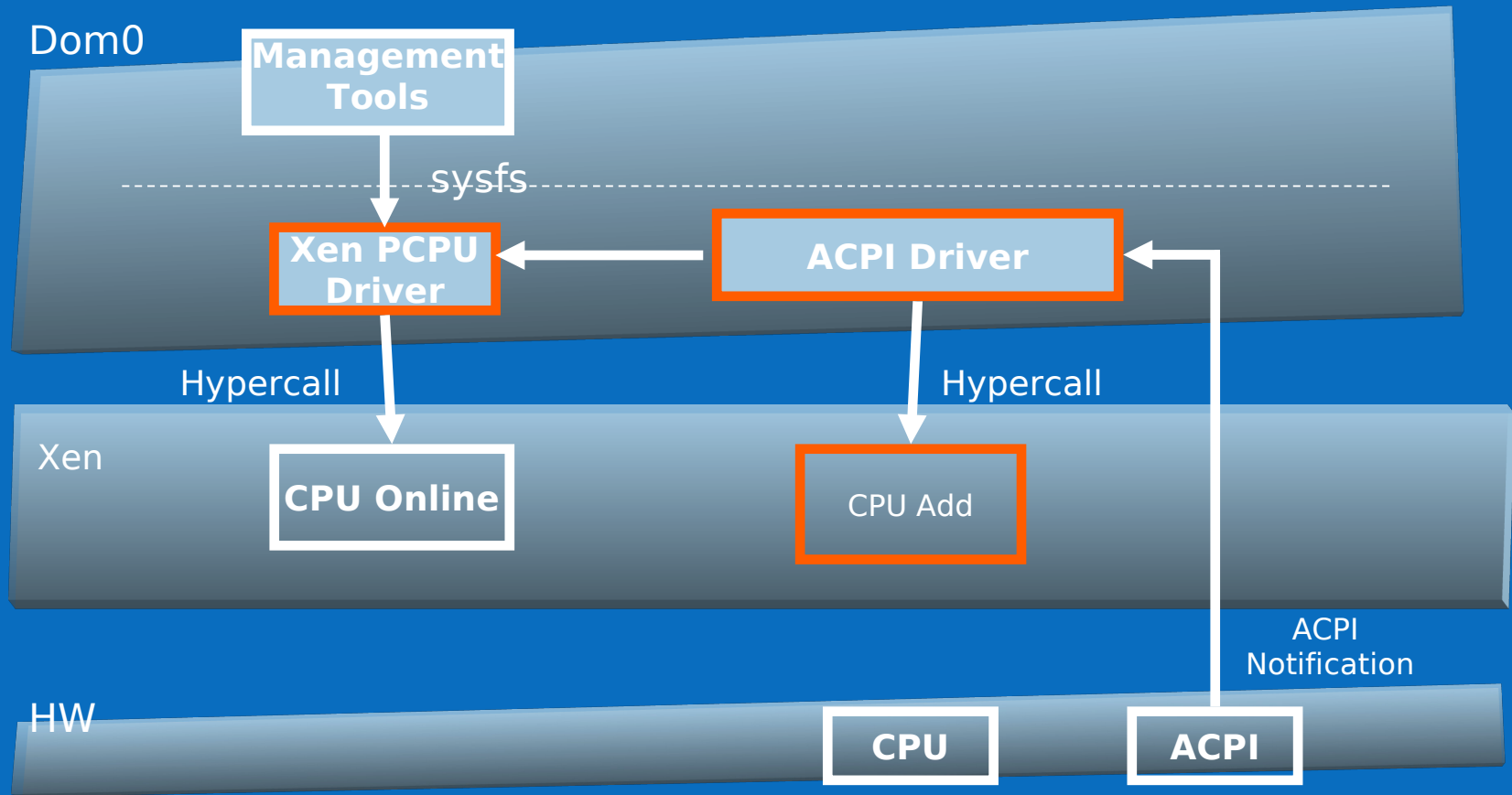
- Overview
- CPU/Memory error handling
- Host CPU hot-plug support
- Guest vCPU Hot-plug



Host CPU Hot-add Support

- Host CPU hot-add works in Xen environment through 2 steps
- Step 1: CPU is marked present
 - A CPU is hot-added to physical platform
 - Platform raise a interrupt (SCI) to dom0
 - Dom0's ACPI driver parses the ACPI table to get CPU information
 - Dom0's ACPI driver notify Xen hypervisor of a new CPU added
 - The new CPU is marked present in xen hypervisor, but will not be scheduled
- Step 2: management tools notify Xen hypervisor to bring the CPU online

Host CPU Hot-add



Patch is on the way to upstream

Agenda

- Overview
- CPU/Memory error handling
- Host CPU hot-plug support
- Guest vCPU Hot-plug

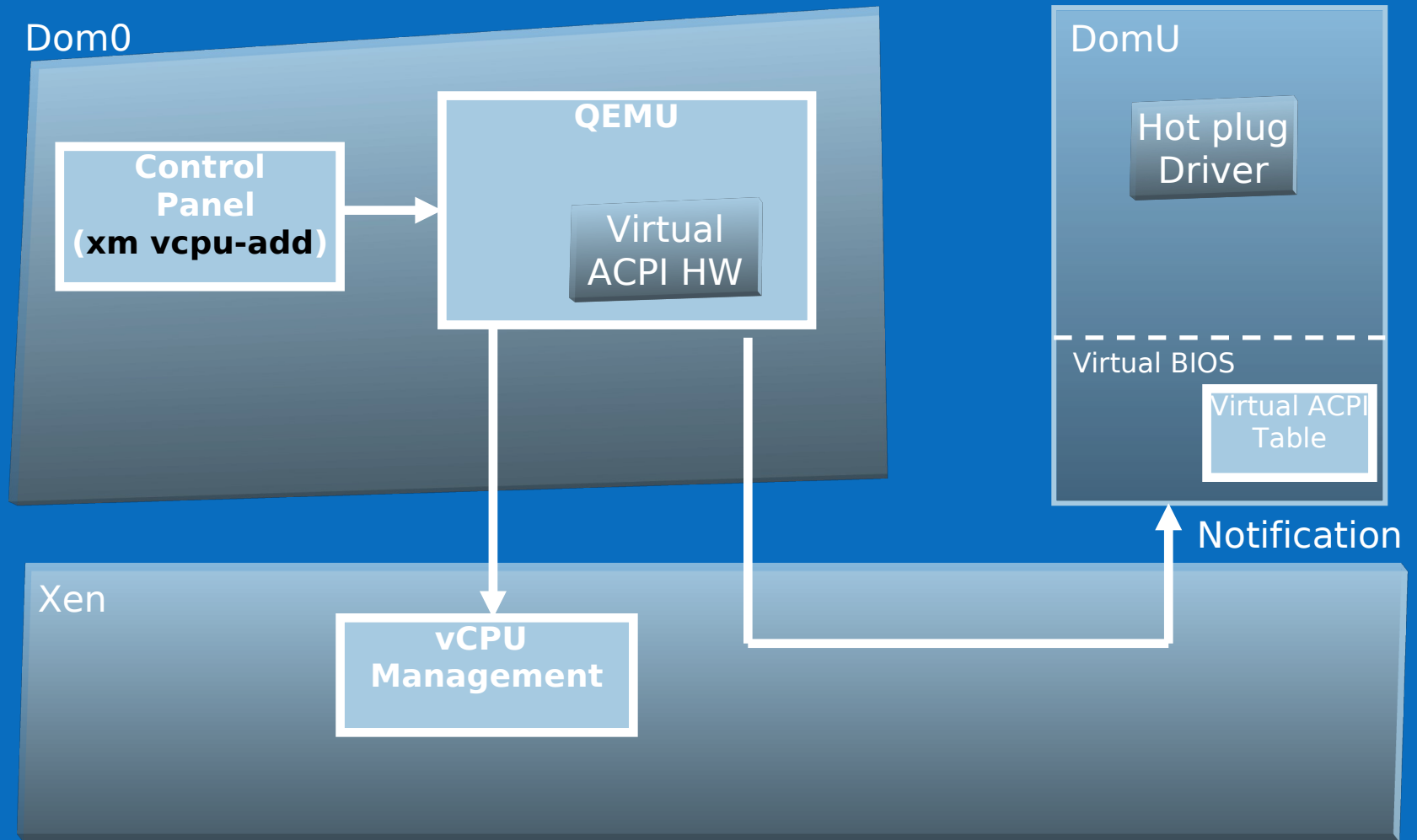


HVM Guest vCPU Hotplug

- Hot-add/remove HVM guest's vCPU dynamically
 - PV guest has vCPU hot-plug for a long time
- Code is almost ready



HVM Guest vCPU Hotplug



Next Step for RAS Effort

- CPU/Memory Enhancement
- PCI AER support for device assigned to HVM guest
 - Based on PCI-Express support in Qemu
- Host memory hot-add support

