KVM tuning and testing, and SMP enhancement

Eddie Dong, Yunfeng Zhao, Xin Li

Legal Disclaimer

- 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 PRODUCTS ARE NOT INTENDED FOR USE IN MEDICAL, LIFE SAVING, OR LIFE SUSTAINING APPLICATIONS.
- Intel may make changes to specifications and product descriptions at any time, without notice.
- All products, dates, and figures specified are preliminary based on current expectations, and are subject to change without notice.
- 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 and the Intel logo are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.
- *Other names and brands may be claimed as the property of others.
- Copyright © 2007 Intel Corporation.

Throughout this presentation:
VT-x *refers to Intel® VT for IA-32 and Intel® 64*
VT-i *refers to the Intel® VT for IA-64, and*
VT-d *refers to Intel® VT for Directed I/O*

Agenda

- Performance tuning
- Kernel interrupt controller status
- SMP support
- Testing

Back to KVM-18

- Kernel build performance was only 1/3 of Xen
  - We suspected shadow page table may not be optimized
- We used oprofile to analyze the overhead
  - Anthony & Avi started looking at performance issues at same time

Top 5 findings from oprofile

- Guest only get ~25% cycles
- Excessive MSR save/restore
  - Such as SYSCALL_MASK, LSTAR, CSTAR, KERNEL_GS_BASE, EFER, and K6_STAR
  - load_msrs costs ~7%
  - save_msrs costs ~3.7%
  - kvm_vmx_return costs ~6.1%
    - Hardware VM Exit does save/load for some of the MSRs
- vmx_vcpu_run costs ~3.2%
  - Most time is spent in HOST_FS/GS_BASE write and fx save/restore

Light-weight vs. heavy-weight VM Exit

- A light-weight VM Exit is handled in KVM and returned to guest directly, without host context switch
  - Mostly for shadow page fault
  - Cover 93% of all VM Exits in KVM-18
- A heavy-weight VM exit involves host context switch or transition to Qemu
  - Such as I/O or when signal is pending
  - Require save/restore of MSRs

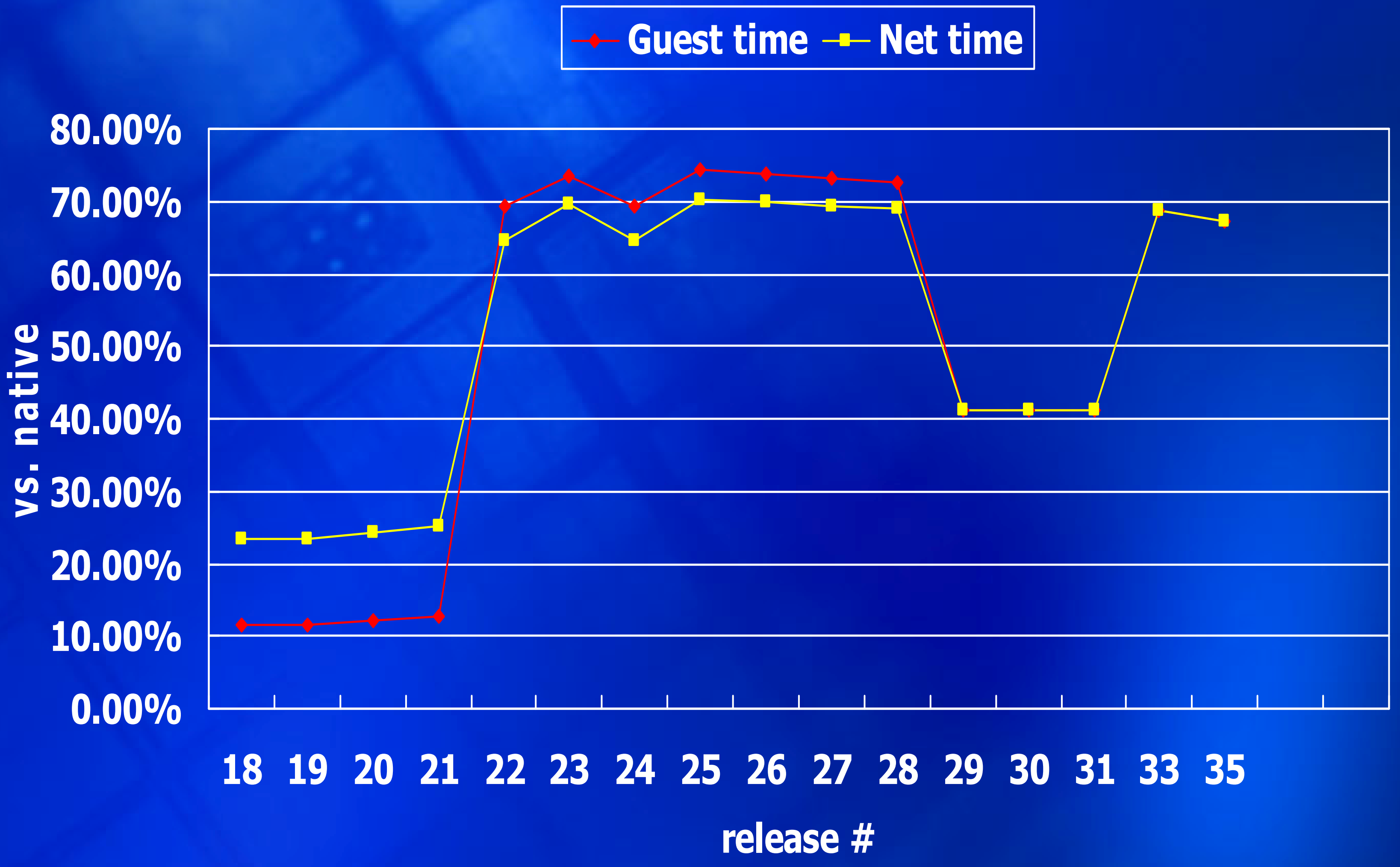
Reduce VM Exits

- Improve shadow page table code
  - Combine guest PTE update with shadow PTE update (Avi Kivity, Qumranet)
  - Increase shadow page table size (Avi Kivity, Qumranet)
- Misc.
  - Port 0x80 access goes to hardware directly (Qing He, Intel)

Shorten VM Exit handling

- Provide quick path for light-weight VM Exit
  - Minimize context save/restore for light-weight VM Exit (Eddie Dong, Intel)
  - Avoid hardware MSR save/restore (Eddie Dong, Intel)
  - Lazy MSR_EFER save/restore (Eddie Dong, Intel)
- Fine tune heavy-weight VM Exit path to save/restore necessary context only
  - Lazy FP (Anthony Liguori, IBM)
  - Some MSRs are not changed in certain environment (Anthony, Avi and etc.)
  - Unbundle fs from gs reload for better SMP support (Laurent Vivier, Bull)

Kernel Build



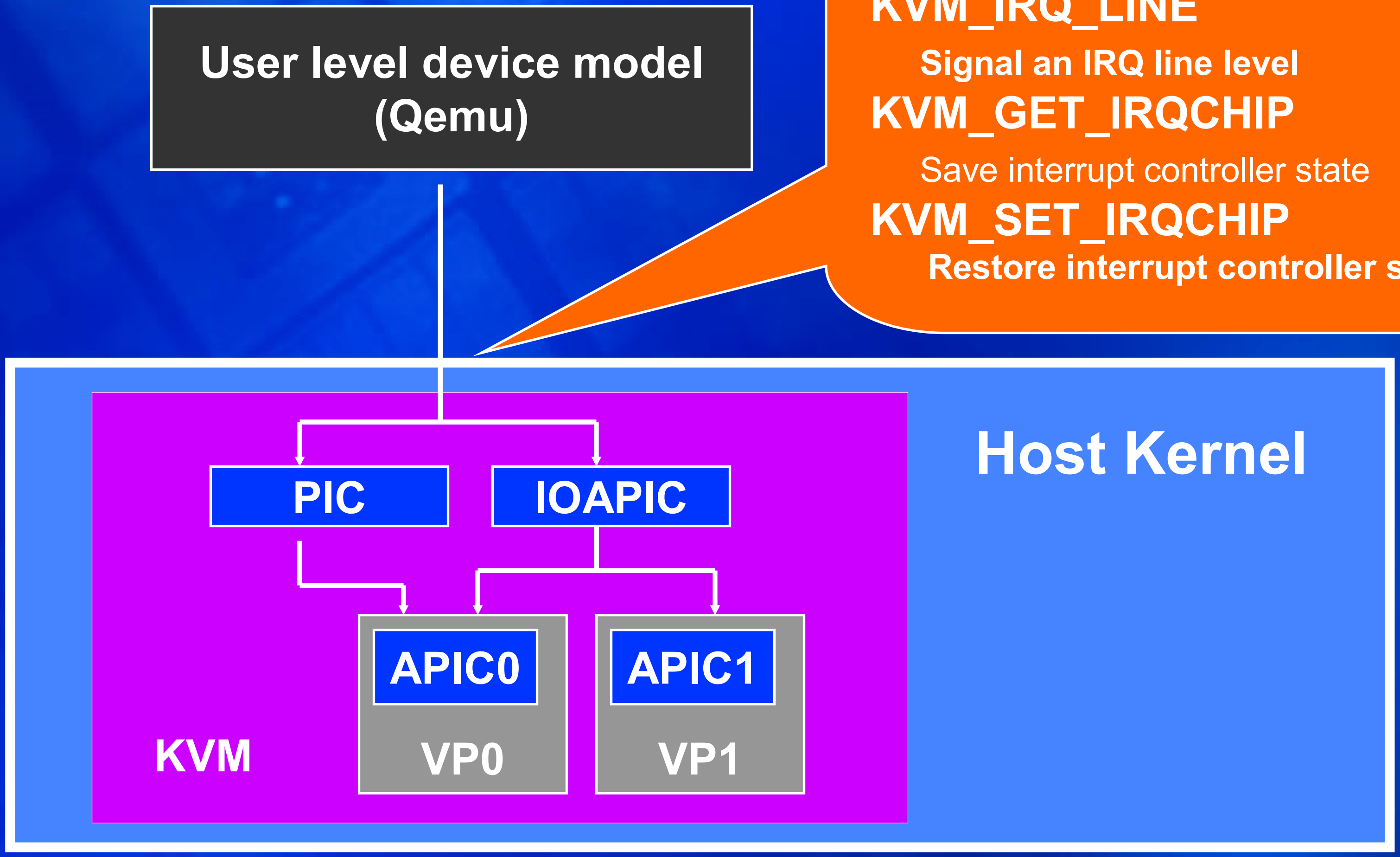
Agenda

- Performance tuning
- Kernel interrupt controller status
- SMP support
- Testing

Where to virtualize interrupt controller ?

- User level
  - Pro: Can reuse Qemu device model
  - Con: Performance concern for kernel devices
- Mixed mode (APIC in kernel, PIC/IOAPIC in user level initially)
  - Pros
    - Flexible code structure
    - Better SMP support
  - Cons
    - Complexity
    - Performance concern if IOAPIC is in user level
- Kernel level (lapic5 branch)
  - Pro: Better SMP support, better performance
  - Con: Kernel is subject to device model failure

Kernel interrupt controller I/Fs



KVM_IRQ_LINE

Signal an IRQ line level

KVM_GET_IRQCHIP

Save interrupt controller state

KVM_SET_IRQCHIP

Restore interrupt controller state

Convert vector based VCPU ops to gsi based VM ops

Lapic5 status

- Current Status
  - PIC/IOAPIC/APIC are implemented
  - Live migration is supported
  - SMP Windows/Linux works
- TODO
  - Merge with master branch
  - Stabilize
  - Guest MSI

Lapic5 Quality Status

🟩	Pass
🟧	Can boot, but has issues
🟥	Fail
⬜	N/A

Guest OS	Guest/Host			
	32/32p	32p/32p	32p/64	64/64
Linux 2.6.9 UP	🟩	🟩	🟩	🟩
Linux 2.6.9 SMP	🟩	🟩	🟩	🟩
Linux 2.6.18 UP	🟩	🟩	🟩	🟩
Linux 2.6.18 SMP	🟩	🟩	🟩	🟩
Linux 2.6.22 UP	🟧	🟧	🟧	🟧
Linux 2.6.22 SMP	🟧	🟧	🟧	🟧
Win2k3 R2 No-ACPI HAL	🟩	🟩	🟩	⬜
Win2k3 R2 UP ACPI HAL	🟩	🟩	🟩	🟩
Win2k3 R2 MP ACPI HAL	🟩	🟩	🟩	🟩
Win2k Srv No-ACPI HAL	🟩	🟩	🟩	⬜
Win2k Srv UP ACPI HAL	🟧	🟧	🟧	⬜
Win2k Srv MP ACPI HAL	🟧	🟧	🟧	⬜
WinXP No-ACPI HAL	🟩	🟩	🟩	⬜
WinXP UP ACPI HAL	🟩	🟩	🟩	🟩
Vista UP ACPI HAL	⬜	🟥	🟥	🟥
Vista SMP ACPI HAL	⬜	🟥	🟥	🟥

Lapic5 has same functionality with master branch

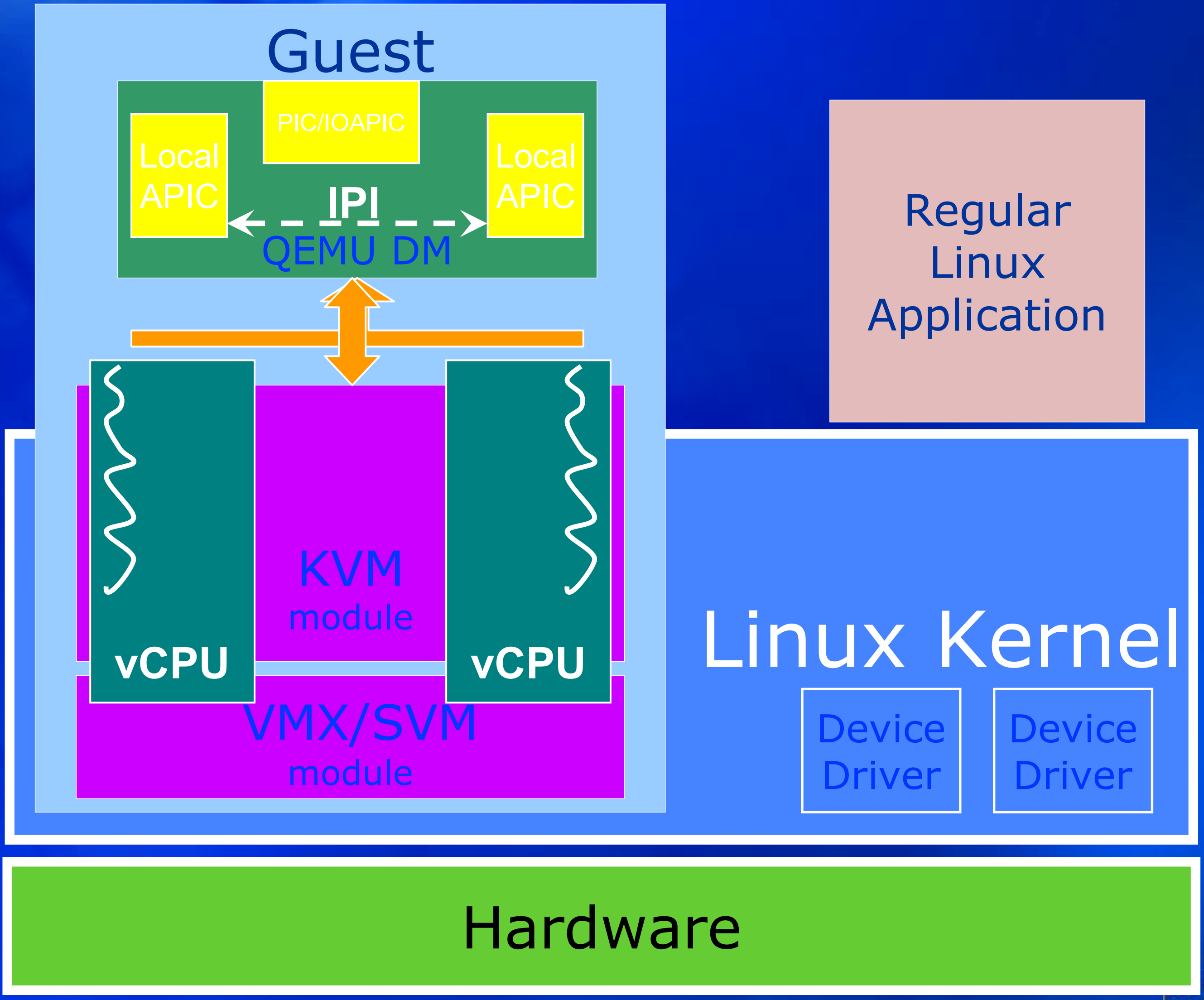
Agenda

- Performance tuning
- Kernel interrupt controller status
- SMP support
- Testing

KVM SMP development

- Originally enabled in June based on Greg’s in kernel APIC V09 (Xin Li, Intel)
  - Each vCPU has a dedicated thread
- User level SMP is enabled in KVM-29 (Avi Kivity, Qumranet)
- In kernel interrupt controller (lapic5 branch) based SMP is enabled (Xin Li, Intel)

KVM SMP


Guest
PIC/IOAPIC
Local APIC
Local APIC
IPI
QEMU DM
KVM module
vCPU
vCPU
VMX/SVM module
Regular Linux Application
Linux Kernel
Device Driver
Device Driver
Hardware

KVM SMP

- N model: each vCPU has a dedicated thread
  - Global lock to DM
    - May use device locks instead
  - Each vCPU thread need to handle signals
    - Asynchronous events may be delivered to any thread
- N+1 Model: to add a dedicated thread to handle asynchronous events
  - Such as DMA/AIO
  - Simplify vCPU thread logic

Agenda

- Performance tuning
- Kernel interrupt controller status
- SMP support
- Testing

KVM Test

Goals

- Ensure KVM works well on all Intel Platforms
  - functionality and performance

Activities

- Test KVM master branch daily
- Report issues and regressions ASAP
- Track issues and help community developers to fix issues
- Develop test cases for new KVM features

KVM Test Suites

Test Suite	Test Scope
VM Management	Create/destroy different guest configurations (memory, #VCPU, ACPI, 32/64), Save & Restore, Live Migration etc.
Device Model	Disk, NIC, VGA, Timer, Keyboard, Mouse
Guest OS	LTP, kernel parameters, Windows (HCT,DTM), Guest OS installation (RHEL5, FC6, RHEL4U3, SLES10, OpenSuse10, SLES9, Windows XP/2k/2k3/Vista)
Regression tests	Specific tests for previous failures
Stress	Linux: LTP stress, Crashme, misc workloads Windows: HCT Stress
Performance	Linux: CPU2K, Kernel build, Lmbench, Iometer, SpecJBB, Sysbench, Byte, NetPerf Windows: Sysmark, CPU2k, SpecJBB, PCmark
Nightly Test	Basic test cases for KVM main features, like Save/Restore, SMP Windows/Linux, live migration, and basic virtual devices

Test Frequency

🟩	Daily
⬜	Monthly
🟨	On demand

Guest/Host	32/32p	32p/32p	32/64	32p/64	64/64
Nightly Test	🟩	🟩	🟩	🟩	🟩
Device Model	⬜	⬜	⬜	⬜	⬜
Regression	⬜	⬜	⬜	⬜	⬜
Guest/Guest Installation	⬜	⬜	⬜	⬜	⬜
Performance	🟨	⬜	🟨	🟨	⬜
Stress	🟨	🟨	🟨	🟨	⬜

Test Infrastructure

- Common interface for easy test case development
- Run tests according to predefined configuration and scenario
- Can handle host hang/crash/reboot situations
- Automatic report generation
  - Outputs a journal file in a standard well-defined format

Sample Test Result

```
Issue list
========================================================
1. Could not create kvm guest with memory >=2040

Details
========================================================
PAE:
1. boot guest with 256M memory                      PASS
2. boot two windows xp guest                        PASS
…
IA32e:
1. boot 4 32-bits guest in parallel                PASS
2. boot 4 64-bits guest in parallel                 FAIL
…
Test Log
```

Current Status

🟩	Pass
🟧	Can boot, but has issues
🟥	Fail
⬜	N/A

Guest OS	Guest/Host			
	32/32p	32p/32p	32p/64	64/64
Linux 2.6.9 UP	🟩	🟩	🟩	🟩
Linux 2.6.9 SMP	🟩	🟩	🟩	🟩
Linux 2.6.18 UP	🟩	🟩	🟩	🟩
Linux 2.6.18 SMP	🟩	🟩	🟩	🟩
Linux 2.6.22 UP	🟧	🟧	🟧	🟧
Linux 2.6.22 SMP	🟧	🟧	🟧	🟧
Win2k3 R2 No-ACPI HAL	🟩	🟩	🟩	⬜
Win2k3 R2 UP ACPI HAL	🟩	🟩	🟩	🟩
Win2k3 R2 MP ACPI HAL	🟩	🟩	🟩	🟩
Win2k Srv No-ACPI HAL	🟩	🟩	🟩	⬜
Win2k Srv UP ACPI HAL	🟧	🟧	🟧	⬜
Win2k Srv MP ACPI HAL	🟧	🟧	🟧	⬜
WinXP No-ACPI HAL	🟩	🟩	🟩	⬜
WinXP UP ACPI HAL	🟩	🟩	🟩	🟩
Vista UP ACPI HAL	⬜	🟥	🟥	🟥
Vista SMP ACPI HAL	⬜	🟥	🟥	🟥

We need your help

- **Use Bug Tracker instead of email to track issues**
  - submit issue
  - assign owner
  - update bug status
- **Give us feedback on our test and its results**
- **What more can we do for KVM?**

intel®

Leap ahead™