

## Xen Summit Asia 2009

### Abstracts

Nov 19-20, 2009



**Speaker:** Jun Kamada , Fujitsu

**Author:** Jun Kamada, Fujitsu & Simon Horman, VA Linux Systems Japan

**Title:** Evaluation and improvement of I/O scalability for Xen

**Abstract:**

For cloud computing, it is required to consolidate many physical servers into one virtualized environment. At that time, I/O scalability is very important element to be considered. This presentation will report evaluation result of the I/O scalability on Xen.

And also, I/O Bandwidth isolation is important for ensuring fair access to I/O resources by guests. This presentation will look at how bandwidth isolation works in the context of Xen and how hardware capabilities of SR-IOV NICs may fit into this framework.

**Speaker:** Isaku Yamahata, VA Linux Systems Japan

**Author:** Isaku Yamahata, VA Linux Systems Japan & Akio Takebe, Fujitsu

**Title:** PCI Express Support in qemu

**Abstract:**

In this presentation, the on-going development for PCI express support in qemu and its status will be shown. Currently passing through of PCI is supported and PCI express device can be passed through as PCI device. However it can't as PCI express natively. PCI express has more features than PCI like MMCONFIG, native hot plug (not ACPI based), ARI(Alternative Route ID), AER(Advanced Error Reporting) and stuff. It requires to enhance QEMU and BIOS. The issues for PCI express will be discussed in this session.

**Speaker:** Ben Lin & Weidong Han, Intel

**Title:** Graphics Passthru with VT-d

**Abstract:**

Graphics passthrough lets guest direct access graphics cards and gets good performance of graphics processing in guest. But graphic passthrough is more complicated than general PCI passthrough. This presentation will introduce the current status of graphic passthrough in xen upstream, then will introduce the

necessary steps to basic graphics passthrough for virtualization friendly graphics cards, additional changes needed for passthrough Intel integrated graphics, and hacks needed for passthrough some non virtualization friendly graphics cards. We will then summarize learnings from our graphics passthrough project.

**Speaker:** Zhigang Wang, Oracle

**Title:** Xen Debugging

**Abstract:**

In this presentation, I summarized the various Xen debugging techniques and tools, including guest core dumps, Xen debug keys, XenTrace, gdbSX, kdb for xen, Xen crash dump, Windows debug and Xend debug etc.. I will also update the status of gdbSX/kdb development.

**Speaker:** George Dunlap, Citrix

**Title:** Xen Scheduler

**Abstract:**

The Credit scheduler has served Xen well for many years now, but it is showing its age. It doesn't handle latency-sensitive workloads gracefully, such as audio or video pass-through. It also doesn't handle large numbers of processors well. When it was written, a 4-way system was at the high end. We may soon have four sockets with 8 cores each with 2 threads, giving a total of 64 schedulable entities. This talk will discuss problems with the Credit scheduler, goals for the new scheduler, and new designs and plans.

Complete paper on this topic at

[http://www.xen.org/files/xensummit\\_intel09/George\\_Dunlap.pdf](http://www.xen.org/files/xensummit_intel09/George_Dunlap.pdf)

**Speaker:** Haicheng Li, Intel

**Title:** Practical Xen Testing at Intel

**Abstract:**

Xen is evolving rapidly with hundreds of patches monthly going to upstream. Xen testing and quality assurances have become more complicated than ever. It requires an efficient test activity and powerful test infrastructure to expose defects timely and locate culprit commit quickly.

Intel VMM QA team keeps working on Xen validation to ensure Xen quality from nightly build to each major release. This presentation will show our test process, test infrastructure, as well as test methodology. We will also introduce our recent test efforts for new Intel VT features of Xen, such as RAS, VT-d, SR-IOV, XenPM and etc.

**Speaker:** Tao Ma & Joel Becker, Oracle

**Title:** Reblink Update

**Abstract:**

The REFLINK operation creates a new inode that shares the data extents of a source inode in a Copy-on-Write (CoW) fashion. It is an important feature for image snapshot and quick deployment of images . Joel Becker has talked about the prototype in Xen Summit Feb.2009, so this talk will show some new update. And the good thing is that it is now integrated into Linux mainline kernel 2.6.32.

**Speaker:** He Qing, Intel

**Title:** Nested Virtualization

**Abstract:**

Hardware virtualization technologies like VMX have made the implementation of a full virtualized environment much simpler than before. However, there is no direct support of nested virtualization, i.e. exposing VMX to VMX non-root environment. This idea is interesting not only that it completes a missing part of hardware virtualization, but also it's becoming more useful as the application of hardware VT widens.

This presentation will describe the concept of the nested virtualization, along with the necessary changes and tricks brought to Xen in order to implment it with VMX from software approach. The presentation will also discuss memory management in nested virtualization when it's using shadow memory and EPT. Finally it will discuss the current implementation status, some performance analysis, and possible future improvements.

**Speaker:** Thomas Goirand & Cao Wei, GPLHost

**Title:** Running a VPS Hosting Business with Xen

**Abstract:**

Thomas Goirand, the CEO and main author of GPLHost web hosting tools. DTC is not only the only open source control panel that has reached major distributions like Debian (Lenny) and Ubuntu (Jaunty). It's also the only open source solution aiming at running a Xen VPS hosting business, and maybe the most achieved solution for commercial Xen VPS at all. My intention is to introduce our work in a short briefing, showing what can be done with DTC-Xen and DTC, and why it's the perfect solution for marketing Xen VPS.

**Speaker:** Zhiteng Huang, Intel

**Title:** I/O Virtualization Performance

**Abstract:**

Many heavy, complex enterprise workloads in real life requires high throughput disk & network IO, for example Database and Web server. IO performance has

always been challenging topic in virtualization, especially when high throughput storage and network are at the door. How IO virtualization IO performance on XEN? In this session, we'd show the performance of XEN software and hardware IO virtualization solution in various IO intensive workloads. Also we discussed the software overhead of using Direct IO to run IO workloads.

**Speaker:** Wang Chao, Red Flag

**Title:** Xen virtualization solution and application of case

**Abstract:**

Redflag using Gfs2 and Heartbeat3 with the Xen implements a integrated solution, it's based on fewer physical servers to build a high-availability system. Xen virtual machines deployed on host servers, running applications. Using Xen to guarantee the resource of host server could be more fully utilized. Gfs2 provides a shared storage which could be used by all the virtual machines. Heartbeat3 in a virtual environment, real-time monitoring of individual virtual machine and host server health, Once a virtual machine problems, Heartbeat3 will isolate the wrong virtual machine from virtual environments, and launch a new virtual machine to replace it. If a host server cracks, Heartbeat3 will switch all the virtual machines on that host server to another backup host server, to ensure the whole system could provide the continuous service.

**Speaker:** Xiaowei Yang, Intel

**Topic:** Extending IO scalability in Xen

**Agenda:**

Scalability is an important topic in server virtualization. In Xen Summit North America 2009, we presented our first round scalability evaluation and brought up some scheduler enhancements.

Since last summit we have done more scalability investigation and enhancements, especially at IO side. As SR-IOV devices require much more interrupt vectors, which can't be satisfied by global interrupt vector allocation policy obviously, we extended it to per-cpu. In high frequent interrupt (e.g. 10G NIC) environment, interrupt handling need special care. We optimized VT-d interrupt delivery code to avoid unnecessary IPIs, and shortened critical virtual interrupt handling path. In the meantime, we also addressed several VNIF scalability issues by introducing interrupt coalescing and multiple tasklets. In this summit, we would like to talk about the design, implementation and some performance data.

**Speaker:** Noboru Iwamatsu, Fujitsu

**Title:** Status Update of PVUSB

**Abstract:**

PVUSB is a driver that provides high-performance and flexible device assignment

of USB devices to guest domains with paravirtualized method. In the previous Xen Summit (North America 2009 & Tokyo 2008), we introduced the developing PVUSB drivers. PVUSB is now into Xen 3.4, but some management features were lacked and we should consider the API is not matured.

In this talk, we will report the latest PVUSB status, along with new device management mechanism and challenges to get beyond USB 2.0.

**Speaker:** Sang-bum Suh, Samsung

**Title:** Xen ARM Update

**Abstract:**

The talk will include the performance benchmark of Xen ARM compared with L4 solution, and contributions to Xen ARM from open community for H/W emulators like Andorid and RTOS. Xen ARM is the best in terms of performance.

**Speaker:** Yunhong Jiang, Intel

**Title:** Update on Mission-critical Xen

**Abstract:**

In this presentation, we will share our continuous effort to mission-critical Xen.

We will give update on error handling in Xen since last Xen summit presentation, for both CPU MCA support and PCI-E AER support.

We will share our latest progress in mission-critical Xen. We will cover enabling physical RAS feature to Xen hypervisor, like CPU and Memory hot-plug support and socket-migration support. We will also talk the virtual CPU and memory hot-plug support to HVM guest

**Speaker:** Ying Song, Chinese Academy of Sciences

**Authors:** Ying Song, Chinese Academy of Sciences & Yuzhong Sun, Chinese Academy of Sciences

**Title:** Utility Analysis for Internet-Oriented Server Consolidation in VM-Based Data Centers

**Abstract:**

Server consolidation based on virtualization technology will simplify system administration, reduce the cost of power and physical infrastructure, and improve utilization in today's Internet-service-oriented enterprise data centers. How much power and how many servers for the underlying physical infrastructure are saved via server consolidation in VM-based data centers is of great interest to administrators and designers of those data centers. Various workload consolidations differ in saving power and physical servers for the infrastructure. The impacts caused by virtualization to those concurrent services are fluctuating

considerably which may have a great effect on server consolidation. This paper proposes a utility analytic model for Internet-oriented server consolidation in VM-based data centers, modelling the interaction between server arrival requests with several QoS requirements, and capability flowing amongst concurrent services, based on the queuing theory. According to features of those services' workloads, this model can provide the upper bound of consolidated physical servers needed to guarantee QoS with the same loss probability of requests as in dedicated servers. At the same time, it can also evaluate the server consolidation in terms of power and utility of physical servers. Finally, we verify the model via a case study comprised of one e-book database service and one e-commerce Web service, simulated respectively by TPC-W and SPECweb2005 benchmarks. Our experiments show that the model is simple but accurate enough. The Xen-based server consolidation saves up to 50% physical infrastructure, up to 53% power, and improves 1.7 times in CPU resource utilization, without any degradation of concurrent services' performance, running on Rainbow - our virtual computing platform, compared to the traditional dedicated servers. Our experimental results also show that the power consumed by the same workloads hosted on consolidated Xen-based servers is 30% less than that hosted on dedicated Linux servers, at the same time, the power consumed by the idle Xen platform is 9% less than that consumed by the same number of idle Linux platform.

**Speaker:** Jun Nakajima, Intel

**Authors:** Jun Nakajima, Intel, Sheng Yang, Intel, and Eddie Dong, Intel

**Title:** Optimizing and Enhancing VM for the Cloud Computing Era

**Abstract:**

We propose that we use and enhance para-virtualization (simply PV, hereafter) technologies more aggressively so that HVM guests can perform better in new environments, including cloud computing. As virtualization becomes the foundation of cloud computing, we need to further lower virtualization overheads as much as possible and to implement virtualization-specific features/optimizations. Because of the advanced and efficient H/W-based virtualization features, HVM guests outperform 64-bit PV guests in most benchmarks on new machines in the market today. However, it does not mean that the PV technologies in Xen are no longer required. Rather, we should use and enhance such technologies more aggressively so that HVM guests can perform better in such new environments. We compare performance of the PV guest, unmodified HVM guest, and the one with optimizations using Xen API (i.e. enlightenments), which we prototyped. Then we discuss two major architectures: 1) hybrid para-virtualization - PV guests to run either in software-based PV or in an HVM container depending on the H/W virtualization features available. 2) Enlightenments -- optimizations for HVM guests available only in virtualization. We also discuss the future of the dom0 architecture, especially in terms of driver support, as PV dom0 becomes less ideal for device drivers support because such performance, maintenance, and certification issues. We also believe that such PV techniques in HVM guests will be the key when implementing virtualization-

specific features, such as “notification upon live migration”, which never happens on bare-metal machines. By sending such notification to the kernel and performance-sensitive apps, the guest can use more optimized features upon live migration.

**Speaker:** Jun Kamada, Fujitsu

Authors: Jun Kamada, Fujitsu, Simon Horman, VA Linux Systems Japan

**Title:** Network Bandwidth Isolation and SR-IOV

**Abstract:**

Network Bandwidth isolation is useful for ensuring fair access to network resources by guests. This presentation will look at how bandwidth isolation works in the context of Xen and how hardware capabilities of SR-IOV NICs may fit into this framework.

**Speaker:** Prof. Chen Haibo, Fudan University

**Title:** Optimizing Crash Dump in Virtualized Environments

**Abstract:**

Crash dump, or core dump is the typical way to save memory image during system crash for future offline debugging. However, for typical server machines with likely abundant memory, the time of core dump could significantly increase the mean time to repair (MTTR) by delaying the reboot-based recovery. In this paper, we propose several optimization techniques to improve the crash dump in virtualized environments, to shorten the downtime of consolidated virtual machines facing a crash. First, we parallelize the process of crash dump and reboot of the crashed VM, by dynamically releasing and allocating resources between the crashed VM and the newly spawned VM. Second, we use the VMM to scan the critical data structures of the crashed VM to filter out the dump of unused memory. Finally, we implement a disk I/O rate control between the crash dump VM and other production VMs (including the recovery VM), to dynamically balance the disk bandwidth, thus ensure the QoS of the production VMs. We have implemented a working prototype, Vicover, that optimizes core dump and recovery on system crash of a virtual machine in Xen. Experimental results on TPC-W show that Vicover can shorten the downtime caused by crash dump by more than 2X, yet causes little impact on the performance and QoS on the normal execution of the recovering VM and other VMs.