

**Xen Summit North America**  
**April 28 – 29, 2010**  
**Topic Abstracts**



April 28-29, 2010  
Sunnyvale, CA

Sponsored By:



**Topic: Xenalyze: Analyzing Xen Traces**

**Speaker: George Dunlap**

**Abstract:**

Xen's trace infrastructure can produce a wealth of information about the execution of a running Xen system, useful for profiling, debugging, or just figuring out what's going on. However, sorting through that data and making sense of it is a much more difficult matter.

Xenalyze is a tool I've been developing over the last three years to make sense out of the data. Its first big feature is its attempt to reconstruct the order that traces occurred originally across multiple processors, even in the face of clock skew and lost records. The second is to track individual vcpus as they migrate across physical cpus, collecting statistical information about them. Finally, it can collect statistical information and display it in a "summary" form (across the whole run), various graphs, or a record-by-record exposition.

The talk will briefly describe Xen tracing infrastructure, the xenalyze tool, and various uses to which it can be put.

**Topic: Graphics Passthrough Challenges**

**Speaker: Allen Kay**

**Abstract:**

I will start this talk with a description of the fundamentals of Xen HVM PCI passthrough as it works today in upstream Xen for devices such as NIC and USB controllers. I will then provide details on special challenges and enhancements necessary for bringing up discrete graphics controllers and various generations of Intel integrated graphics devices in the guest environment.

**Topic: Status Update on PCI Express Support in QEMU**

**Speaker: Isaku Yamahata**

**Abstract:**

This presentation is update of my last presentation which I gave at the summit in Shanghai. 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.

**Topic:** Case Study: IaaS using XCP and XAPI

**Speaker:** Marco Sinhoreli

**Abstract:**

Globo.com is the Internet branch of Organizações Globo, the biggest media conglomerate in Latin America, with offices in Rio de Janeiro and São Paulo. It aggregates the greatest web vertical portals in Brazil, spreading from News (G1), Sports (globoesporte.com), Videos (Globo Videos) to Celebrities (Ego). The company is also responsible for providing SaaS, strategy and technology support on Internet for all other businesses of the organization. It is considered benchmark on high volume web distribution and is responsible for the highest simultaneous video streaming audiences in the country. Today it has more than one thousand servers in the datacenter running services from web applications to databases. The main website receives more than 150 million page view per day and average of 27 million unique visitors in February from all around the country through over 60 Gbit internet links. Xen utilization started 2 years ago on development and Quality Assurance environments. Due to it's complex datacenter architecture, in the beging of 2009, we decided start a research for better resource utilization and manageability. Xen Cloud Platform was chosen as a virtualization component for the Globo.com IaaS as it has paravirtualization, a complete API for storage, pooling, network management and community support as well. The XCP is integrated with other internal APIs to make a complete IaaS solution, including switches and balancer management, deployment, monitoring and ticket system. Given company's size and quantity of teams, the APIs were all integrated in a single ajax web interface and a workflow to manage the IaaS. Only open source components were used. The web interface was written in django/python and it supports a horizontal growth. The Globo.com IaaS shows the traditional virtual machine views and a enterprise view grouping owners > products > environments > farms > virtual machines and a search engine to seek all objects related between thousands of existing virtual resources.

**Topic:** Application of Fuzzy Control Theory to Resource Management in a Virtualized System

**Speaker:** Sho Niboshi, Hitoshi Oi

**Abstract:**

Nowadays, virtualization technology is widely used to reorganize data centers and server systems with a small number of computers by incorporating multiple systems into a single physical computer. On the shared system, the resource controller controls resource assignment to virtual machines (VMs) and it plays important role determining the virtualized system performance. However, the resource controller in current systems does not have any guarantees for application performance because the allocation only utilizes the information from VM instead of application itself. This paper demonstrates the resource controller that takes application state, e.g. Quality of Service (QoS), into account to identify resource demand. We applied fuzzy control theory for the resource allocation to model the complex relationship between QoS and the demand. Also, we report evaluations of the fuzzy rule-based controller on a Xen-based system with two guest VMs running mail and Java application servers, and its advantages over Xen default scheduler are shown.

**Topic:** Supporting Soft Real-Time Tasks in the Xen Hypervisor

**Speaker:** Min Lee, A.S. Krishnakumar, P. Krishna, Navjot Singh, Shalini Yajnik

**Abstract:**

Note: This work was presented at VEE 2010. Virtualization technology enables server consolidation and has given an impetus to low-cost green data centers. However, current hypervisors do not provide adequate support for real-time applications, and this has limited the adoption of virtualization in some domains. Soft real-time applications, such as media-based ones, are impeded by components of virtualization including low-performance virtualization I/O, increased scheduling latency, and shared-cache contention. The virtual machine scheduler is central to all these issues. The goal in this paper is to adapt the virtual machine scheduler to be more soft-real-time friendly. We improve two aspects of the VMM scheduler – managing scheduling latency as a first-class resource and managing shared caches. We use enterprise IP telephony as an illustrative soft real-time workload and design a scheduler S that incorporates the knowledge of soft real-time applications in all aspects of the scheduler to support responsiveness. For this we first define a laxity value that can be interpreted as the target scheduling latency that the workload desires. The load balancer is also designed to minimize the latency for real-time tasks. For cache management, we take cache-affinity into account for real time tasks and load-balance accordingly to prevent cache thrashing. We measured cache misses and demonstrated that cache management is essential for soft real time tasks. Although our scheduler S employs a different design philosophy, interestingly enough it can be implemented with simple modifications to the Xen hypervisor’s credit scheduler. Our experiments demonstrate that the Xen scheduler with our modifications can support soft real-time guests well, without penalizing non-real-time domains.

Complete Topic Document at [http://www.xen.org/files/xensummit\\_amd10/softrealtime.pdf](http://www.xen.org/files/xensummit_amd10/softrealtime.pdf)

**Topic:** Energy-Efficient Storage in Virtual Machine Environments  
**Speaker:** Lei Ye, Gen Lu, Sushanth Kumar, Chris Gniady, John Hartman  
**Abstract:**

Current trends in increasing storage capacity and virtualization of resources combined with the need for energy efficiency put a challenging task in front of system designers. Previous studies have suggested many approaches to reduce hard disk energy dissipation in native OS environments; however, those mechanisms do not perform well in virtual machine environments because a virtual machine (VM) and the virtual machine monitor (VMM) that runs it have different semantic contexts. This paper explores the disk I/O activities between VMM and VMs using trace driven simulation to understand the I/O behavior of the VM system. Subsequently, this paper proposes three mechanisms to address the isolation between VMM and VMs, and increase the burstiness of hard disk accesses to increase energy efficiency of a hard disk. Compared to standard shutdown mechanisms, with eight VMs the proposed mechanisms reduce disk spin-ups, increase the disk sleep time, and reduce energy consumption by 14.8% with only 0.5% increase in execution time. We implemented the proposed mechanisms in Xen and validated our simulation results.

Complete Topic Document at [http://www.xen.org/files/xensummit\\_amd10/energy.pdf](http://www.xen.org/files/xensummit_amd10/energy.pdf)

**Topic:** VastSky – Cluster Storage System for XCP  
**Speaker:** Hirokazu Takahashi, Takashi Yamamoto, Tomoaki Sato  
**Abstract:**

VastSky is a cluster storage system which is made up with commodity hardware --- PC servers and SATA disks. This is designed to be used in cloud environment, which is scalable and fault-tolerant, and has a feature that virtual machines can directly run on the system. In this presentation, I will talk about the concept, design and roadmap of VastSky. We will make the code open by the summit at

<https://sourceforge.net/projects/vastsky/>.

**Topic:** Extending Xen into Embedded and Communications Applications

**Speaker:** Edwin Verplanke, Don Banks

**Abstract:**

Today's Embedded and Communications applications often have different requirements from server and desktop applications. Advances in semi-conductor development have driven broad adoption of Multi-Core processors. Hardware-assisted virtualization has proven to be an excellent method to effectively utilize the additional core computing capabilities. Common usage models include server consolidation on a single Multi-Core platform and client-server based desktop virtualization. Today's VMM's and silicon architecture generally cater very well for these kinds of general purpose environments and applications, however the embedded and communications environments require enhanced functionality such as real time scheduling, high performance I/O, high availability, and co-located cooperating applications. This presentation will cover some new and exciting Virtualization usage models - does Xen have what it takes to address Embedded and Communications requirements?

**Topic:** Evolving New Configuration Tools for IOV Network Devices

**Speaker:** Mitch Williams

**Abstract:**

I/O Virtualization (IOV) technology for network devices is still in its infancy. While the devices are readily available, and drivers have been pushed into the kernel, configuration tools are few and far between. Kernel maintainers and network administrators are still coming to terms with what types of tools are required to make IOV network devices usable in the real world. This paper describes the current state of these configuration tools, shows some use cases, and provides a overview of future development work. It describes what works today, what's still missing, and what can be done to address these issues.

Complete Topic Document at [http://www.xen.org/files/xensummit\\_amd10/IOVnetwork.pdf](http://www.xen.org/files/xensummit_amd10/IOVnetwork.pdf)

**Topic:** Building an Infrastructure as a Service Cloud on Xen Cloud Platform

**Speaker:** Sheng Liang

**Abstract:**

This talk discusses how an Infrastructure as a Service (IaaS) cloud can be built on the Xen Cloud Platform (XCP.) The XCP already provides robust enterprise-grade hypervisor as well as basic storage and network virtualization support. We will discuss how to: 1. Scale servers by grouping them into availability zones and pods. 2. Organize multiple primary VM disk storage servers and secondary storage servers. 3. Manage isolated guest networks using layer-2 tunneling and hardware VLANs. We will cover service management features such as service definition, usage metering, and OSS/BSS integration. We will share the experience of building public and private cloud for service providers and enterprises.

**Topic:** Xen NUMA Guests

**Speaker:** Dulloor Rao

**Abstract:**

The power and performance constraints are pushing platforms towards increasingly NUMA architecture. While such platform architectures provide greater aggregate bandwidth and are more scalable, they also necessitate changes in the system software for optimal performance. For the same

reason, operating systems, such as Linux, are extremely conscious of NUMA behaviour, particularly the latencies in accessing remote nodes and are heavily optimized to ensure locality of memory accesses as much as possible. But, when running on top of a VMM, the domains are completely unaware of the underlying asymmetry, leading to (unpredictable) performance overheads. In this presentation, we discuss the cost of virtualization (with Xen) on such platforms. We also present a global domain memory allocation scheme for Xen, comprised of four different allocation strategies. Then, we present the implementation of the allocation scheme with PV NUMA guests (the same allocation scheme could be used with HVMs too). At the heart of PV NUMA is the virtual NUMA enlightenment, which allows the VM to execute with a NUMA layout correspondingly similar to the underlying platform. While the above mentioned allocation schemes and enlightenment provide the method of setting up and starting domains with the desired domain memory layout, memory over-provisioning features, such as ballooning, also need to ensure that the increase/ decrease/exchange of memory reservations follows the initial memory layout. We add this capability for ballooning in PV NUMA guests. Finally, we open a discussion on the following topics :

- XEN domain memory allocation schemes with NUMA and domain config options. Fewer configs the better !
- NUMA guests with memory over-provisioning, temporal memory allocation (transcendent memory), and other features(?).
- Sweeping across memory in the guest domain (runtime) to identify and exchange the misplaced pages (complementary to the above point).
- More adventurous drills on large enough platforms - like (user-triggered or automatic) inter-node VM migration to consolidate the Vms.

**Topic: Update on Transcendent Memory in Xen**

**Speaker: Dan Magenheimer**

**Abstract:**

At Xen Summit 2008, we described self-ballooning. At Xen Summit 2009, we introduced Transcendent Memory (“tmem”, see <http://oss.oracle.com/projects/tmem>). With Xen 4.0, we have combined the two into a unique enterprise-ready memory utilization optimization solution. For Xen Summit 2010, we will quickly review the goals and basics of the two topics, discuss significant advances in tmem both in Xen and in Linux, and (if the Xen scheduler behaves) present some new performance results.

**Topic: XRM: Event-based Resource Management Framework for XCP**

**Speaker: Pradeep Padala, Ken Igarashi and Ulas C. Kozat, Akshay Kumar Mehta**

**Abstract:**

Xen Cloud Platform (XCP) is emerging as an open source solution to build Infrastructure as a Service (IaaS) platforms. XCP is still in its infancy, and one of the important missing features is “resource management” in virtualized data centers. In this work, we propose XRM, a resource management (RM) framework for XCP, that provides a modular and extensible framework to implement RM strategies for load balancing, optimal VM placement for high utilization, optimal power utilization and high application performance. We have implemented a prototype eventbased RM framework on top of XCP 0.1.1. The RM framework contains a feedback control loop that is driven by monitoring events and pluggable RM algorithms. The event-based framework allows many RM algorithms to be implemented

with minimal programming effort. In the presentation, we will show the design of the framework, issues involved in modularity and extensibility, feedback control loop, our implementation of various algorithms and their effectiveness on a 128 core XCP cluster.

**Topic:**        **Libxenlight: a new low level library for tool stacks**

**Speaker:**     **Stefano Stabellini**

**Abstract:**

Currently many different tool stacks are used to manage a Xen based host, leading to inconsistencies, code duplications and bugs.

Moreover Xend, the default tool stack on xen-unstable, is hard to modify and extend.

Libxenlight was created to fix these issues and this talk will explain the design principles, the architecture and the objectives of this new library. Libxenlight aims to provide a simple and robust API for tool stacks to do Xen operations and to create a common codebase for the lower-level implementation of all the various Xen tool stacks.

**Topic:**        **Status update on pv-ops Linux kernel**

**Speaker:**     **Konrad Rzeszutek Wilk**

**Abstract:**

Short primer on the para-virtualized operations (pv-ops) Linux kernel, and the pros and cons compared to XenLinux. Also covers the current upstream integration strategy, status and outstanding issues.

**Topic:**        **Neon: System Support for Derived Data Management**

**Speaker:**     **Qing Zhang**

**Abstract:**

Modern organizations face increasingly complex information management requirements. A combination of commercial needs, legal liability and regulatory imperatives has created a patchwork of mandated policies. Among these, personally identifying customer records must be carefully access-controlled, sensitive files must be encrypted on mobile computers to guard against physical theft, and intellectual property must be protected from both exposure and “poisoning.” However, enforcing such policies can be quite difficult in practice since users routinely share data over networks and derive new files from these inputs—incidentally laundering any policy restrictions. In this paper, we describe a virtual machine monitor system called Neon that transparently labels derived data using bytelevel “tints” and tracks these labels end to end across commodity applications, operating systems and networks. Our goal with Neon is to explore the viability and utility of transparent information flow tracking within conventional networked systems when used in the manner in which they were intended. We demonstrate that this mechanism allows the enforcement of a variety of data management policies, including data-dependent confinement, mandatory I/O encryption, and intellectual property management.

Complete Topic Document a [http://www.xen.org/files/xensummit\\_amd10/neon.pdf](http://www.xen.org/files/xensummit_amd10/neon.pdf)

**Topic:** Parallax: Past, Present, & Future

**Speaker:** Mohammad Shamma

**Abstract:**

Parallax [2, 3] is a distributed storage system that serves virtual disk abstractions for hosts in virtual environments. A proof of concept implementation have been in place that does high-speed disk checkpointing and provisioning.

Complete Topic Document at [http://www.xen.org/files/xensummit\\_amd10/parallax.pdf](http://www.xen.org/files/xensummit_amd10/parallax.pdf)

**Topic:** Improving Paravirtualized I/O performance with EPT based page flip

**Speaker:** Xiaowei Yang, Eddie Dong, Jun Nakajima

**Abstract:**

Paravirtualized I/O virtualization imposes challenge on virtualization overhead. The historical page flip solution swapped the pages between frontend guest and backend guest to multiplex the host side received packets. However page flip was proven to be suboptimal due to the excessive operation of page table base on the hardware of that time. Both frontend guest and backend guest need to zap the original mapping, walking through all the reverse maps of the original pages, and insert new maps. The solution has been replaced with page copy since 4 years ago. However page copy also involves excessive virtualization overhead due to packet copy, and thus limits the overall throughput.

We improve the paravirtualized I/O performance to do page flip base on latest hardware virtualization technologies (that is Extended Page Table mechanism or EPT), which is widely adopted in commercial product now. With EPT, the previous virtualization overhead, spending in excessive page table operation, can be greatly reduced, by simply modifying the EPT table both for frontend side and backend side. This solution requires the backend driver to be run in HVM guest, which has proven to have equal or even better performance, with latest hardware acceleration, than conventional paravirtualized virtual machine, implemented based on 7+ years old hardware.

**Topic:** SleepServer: A Software-Only Approach for Reducing the Energy Consumption of PCs within Enterprise Environments

**Speaker:** Yuvraj Agarwal

**Abstract:**

Desktop computers are an attractive focus for energy savings as they are both a substantial component of enterprise energy consumption and are frequently unused or otherwise idle. Indeed, past studies have shown large power savings if such machines could simply be powered down when not in use. Unfortunately, while contemporary hardware supports low power “sleep” modes of operation, their use in desktop PCs has been curtailed by application expectations of “always on” network connectivity. In this paper, we describe the architecture and implementation of SleepServer, a system that enables hosts to transition to such low-power sleep states while still maintaining their application’s expected network presence using an on demand proxy server. Our approach is particularly informed by our focus on practical deployment and thus SleepServer is designed to be compatible with existing networking infrastructure, host hardware, operating system and application software and introduces only a trivial software agent on each system under management. We detail results from our experience in deploying SleepServer in a medium scale enterprise with a sample set of thirty machines instrumented to provide accurate real-time measurements of energy consumption. Our measurements show significant energy savings for PCs ranging from 60%-80%, depending on their use model.