PLUG West

Mission-Critical Enterprise Linux

April 17, 2006

UNISYS

Agenda

- Welcome
  - – Who we are & what we do
    - Steve Meyers, Director – Unisys Linux Systems Group (steven.meyers@unisys.com)
- Technical Presentations
  - – Xen Virtualization In The Enterprise
    - Luke Szymanski, Software Engineer – Unisys Linux Systems Group (lukasz.szymanski@unisys.com)
  - – Linux File System Performance In Enterprise Environments
  - – Linux Scalability Challenges
    - Amul Shah – Software Engineer – Unisys Linux Systems Group (amul.shah@unisys.com)

PLUG West

Xen Virtualization In The Enterprise

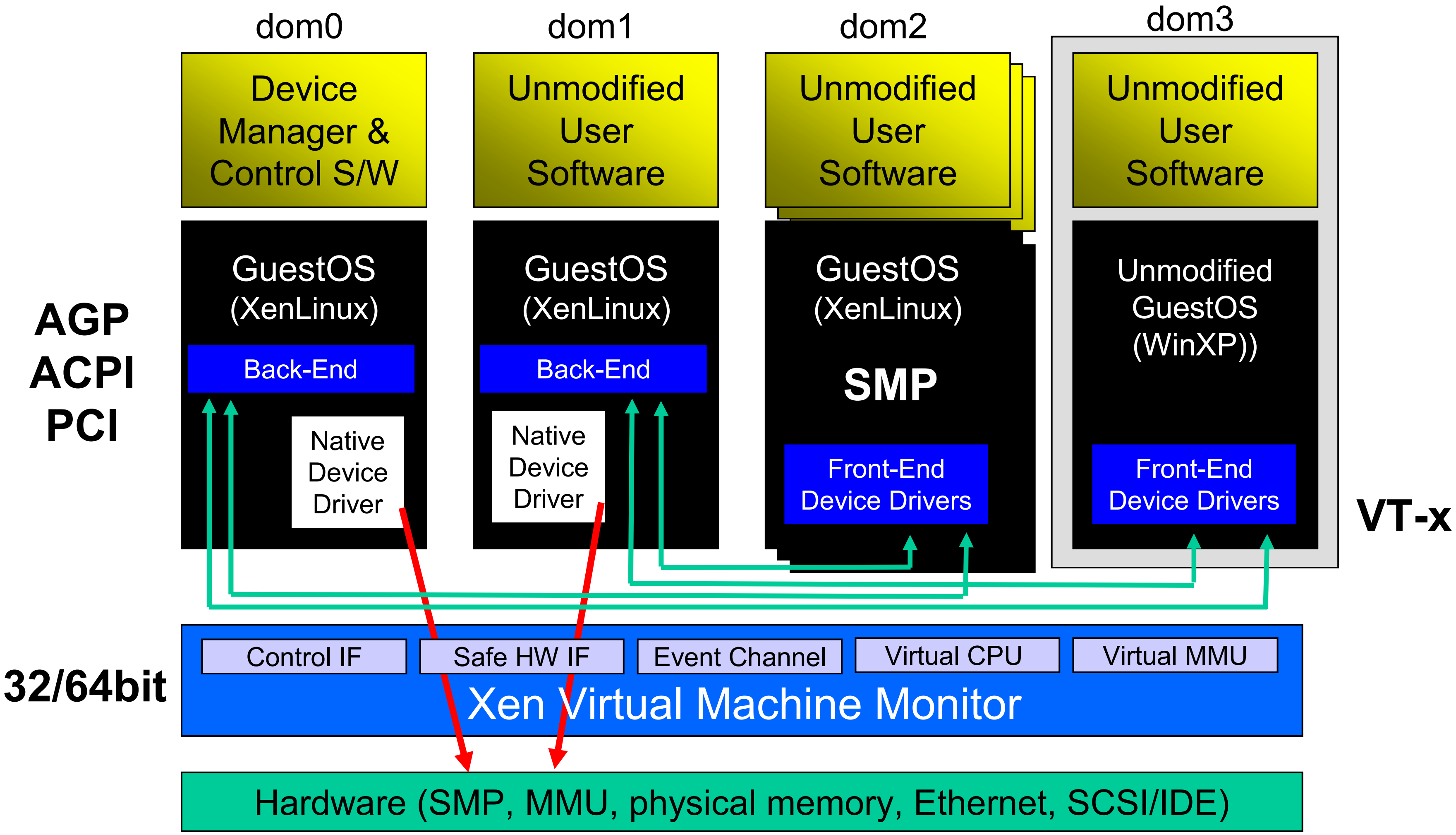
April 17, 2006

UNISYS

Current Solutions

Product	OS-based or Hypervisor-based	Full virtualization or Paravirtualization
Microsoft Virtual Server	OS-based	Full virtualization
VMware GSX	OS-based	Full virtualization
VMware ESX	Hypervisor-based	Full virtualization
Xen	Hypervisor-based	Paravirtualization (and H/W Assisted)

Xen 3.0 Architecture



Xen 3.0 Features

- Intel VT-x support
- Live VM relocation
- Optimized inter-VM networking
- Continued reduction of hypervisor
- Improved management tools

Xen 3.0 Features

- Improved ACPI support
- ia-32, ia-32 with PAE, x86_64, ia-64, PPC
- Host
  - Up to 32 processors
  - Up to 16 GB memory on ia-32 with PAE
  - Up to 8 TB memory on x86_64
- Guests
  - SMP guests
  - Up to 16 GB memory on ia-32 with PAE
  - Up to 8 TB memory on x86_64

Why do I Care?

- Increased resource utilization
- Greater usage flexibility
- Better availability
- Legacy compatibility
- Improved manageability

Unisys’ Involvement

- Active participant in Xen community since 2004
- Scalability & performance
  - First member to run 32 processors
  - First member to consistently run with >4 GB of memory
  - First member to push Xen to maximum # of VMs
  - Currently raising limit on # of processors
  - Supporting “mini OS” as building block for VT-x I/O performance improvements
- Systems management
  - Contributor to CIM development subgroup

Unisys’ Involvement

- Support of ES7000/one
  - 32 sockets / 64 cores / 128 threads
  - 256 GB memory
  - x86_64 (and ia-32 with PAE)

References

- Xen project at University of Cambridge
  - http://www.cl.cam.ac.uk/Research/SRG/netos/xen/
- XenSource
  - http://www.xensource.com/
    - You can download Xen and a live-cd version of Xen from this site.
- Xen mailing lists
  - http://lists.xensource.com/
- Proceedings from the 2005 Ottawa Linux Symposium
  - http://www.linuxsymposium.org/2005/
  - Two papers in volume 1
  - One paper in volume 2

PLUG West

Linux File System Performance In Enterprise Environments

April 17, 2006

UNISYS

Overview

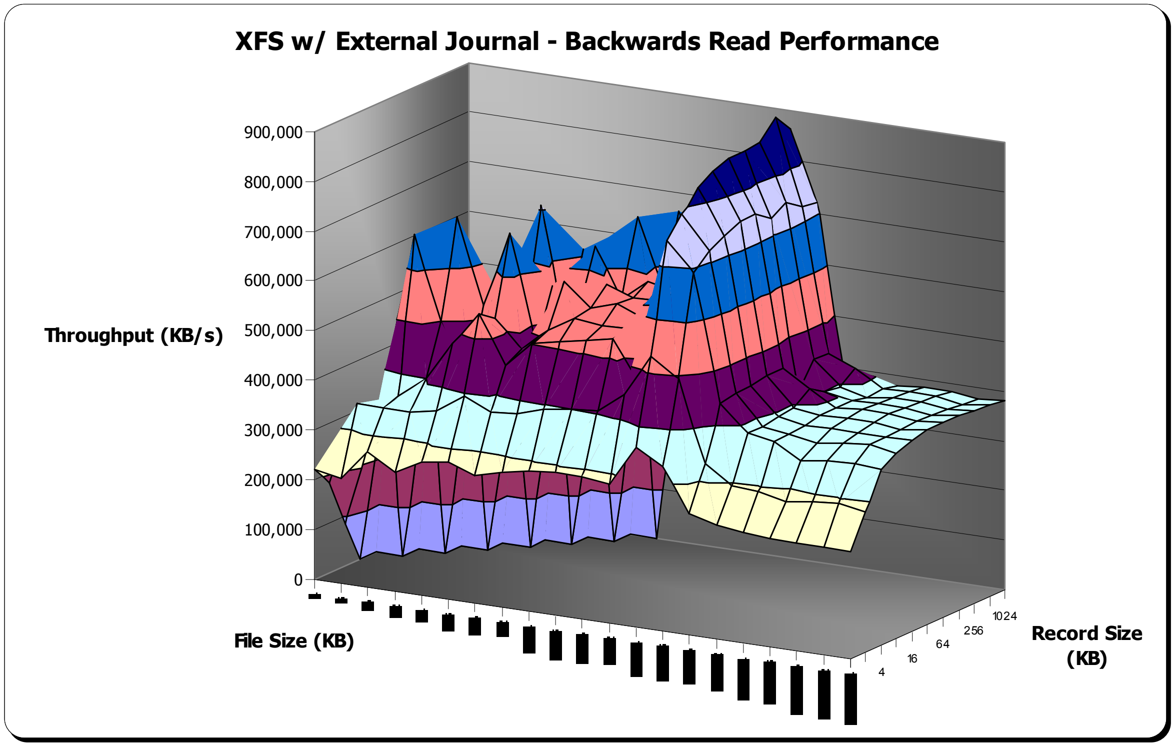
- What is a File System?
  - Allows multiple file storage
    - Raw partitions store single files only
  - Method for file management
    - Physical location can be optimized
    - ACLs
    - Quotas
  - Advanced functionality
    - Undelete
    - Security
    - FS specific features
      - GRIO
      - Atomic Operations

File System Choices

- Many to choose from
  - Linux supports dozens of file systems
  - Our focus is on enterprise file systems
- Enterprise Class File Systems
  - ext2/ext3
    - Included as the performance high watermark
    - Can journaling file systems compete with ext2?
  - Reiser3/4
    - First journaling fs to be included in the kernel
  - JFS
    - Open sourced IBM AIX file system
  - XFS
    - Open sourced SGI IRIX file system

File System Testing

- 5 File Systems tested
- 1000+ Hours of SMP-based testing on 2.6 Linux
- Approx. 200,000 data points



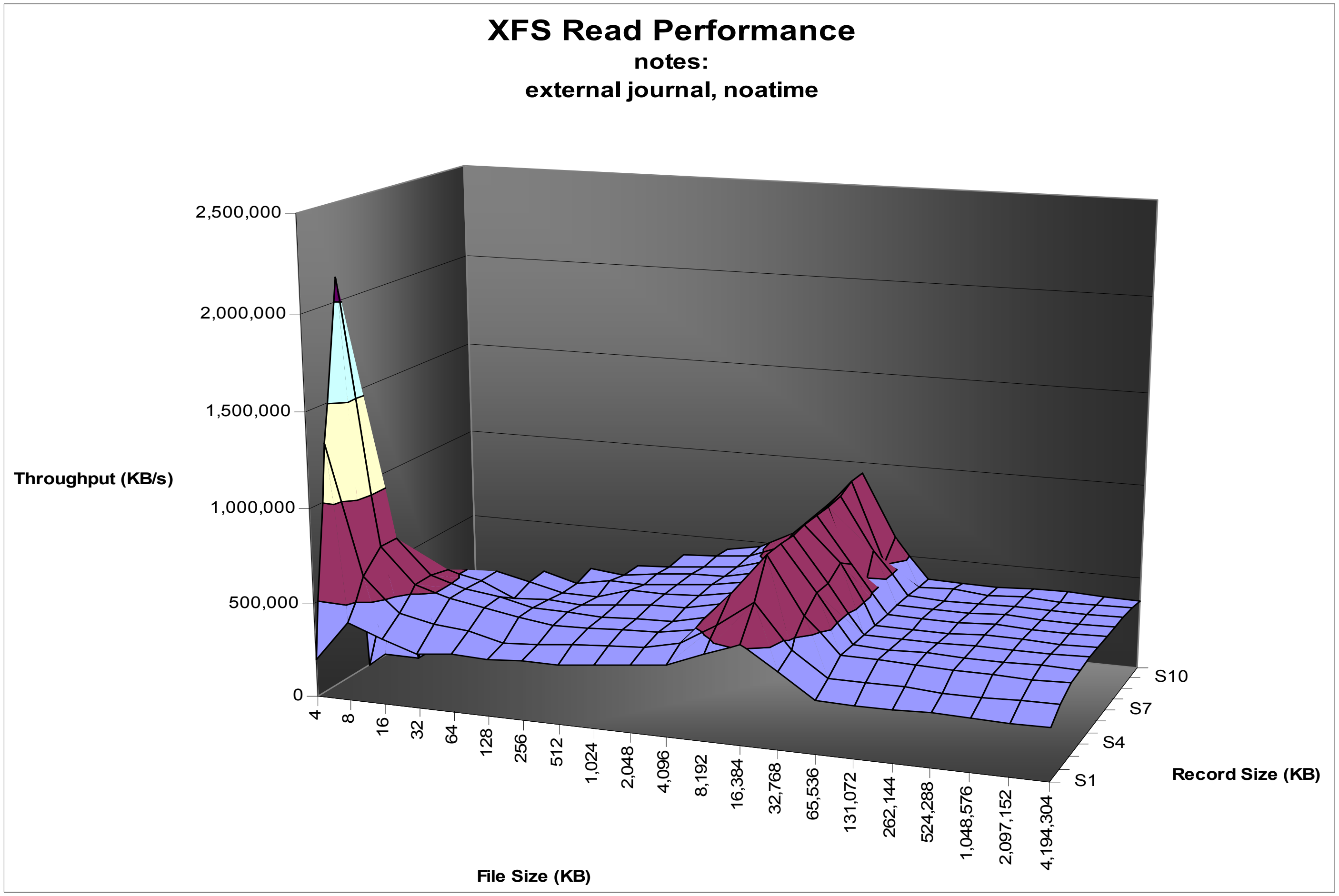
Testing on Datacenter Hardware

- Server
  - Single 8-socket cell of an ES7000/540 “Orion” IA32 server
    - 8 GB RAM
  - QLogic 2310F Fibre Card
    - Single path to eliminate MPIO variables
- Disk Subsystem
  - EMC CLARiiON CX600
  - 16-Disk RAID-0 Array (target)
- Software
  - SUSE Linux Enterprise Server 9 (ia32)
  - IOZone
    - Open-source file system benchmark

Results

- Ext2, with its lack of journaling, was very fast--but not always the fastest.
- XFS was the fastest in overall **write performance.**
- JFS was the fastest in overall **random write performance.**
- Ext2 was the fastest for **Oracle performance**, but XFS was a close second.
- Ext2 was the fastest for Desktop workloads, but JFS was a close second
  - Recommended: JFS offers ext2-like performance for the desktop, but with the added integrity of file system journaling

JFS ***minimum*** read performance – 198 MB/s



Whitepaper

- The whitepaper and the performance metrics on which it is based are open to the public.
- See the Unisys eCommunity website for the paper or send a request to troy.stepan@unisys.com for more details.

http://ecommunity.unisys.com

PLUG West

Linux Scalability Challenges

April 17, 2006

UNISYS

Why are there problems?

- ES7000/one has
  - – 32 dual-core hyper-threaded Xeon processors (128 CPUs)
  - – 256GB of RAM
  - – 48 PCI-X slots
    - Total system IO space is 64KB
  - – 16 built-in Gigabit Ethernet ports
  - – 65 I/O Advanced Programmable Interrupt Controllers
    - 1560 Interrupts
  - – Non-Uniform Memory Architecture

Why are there problems? Cont.

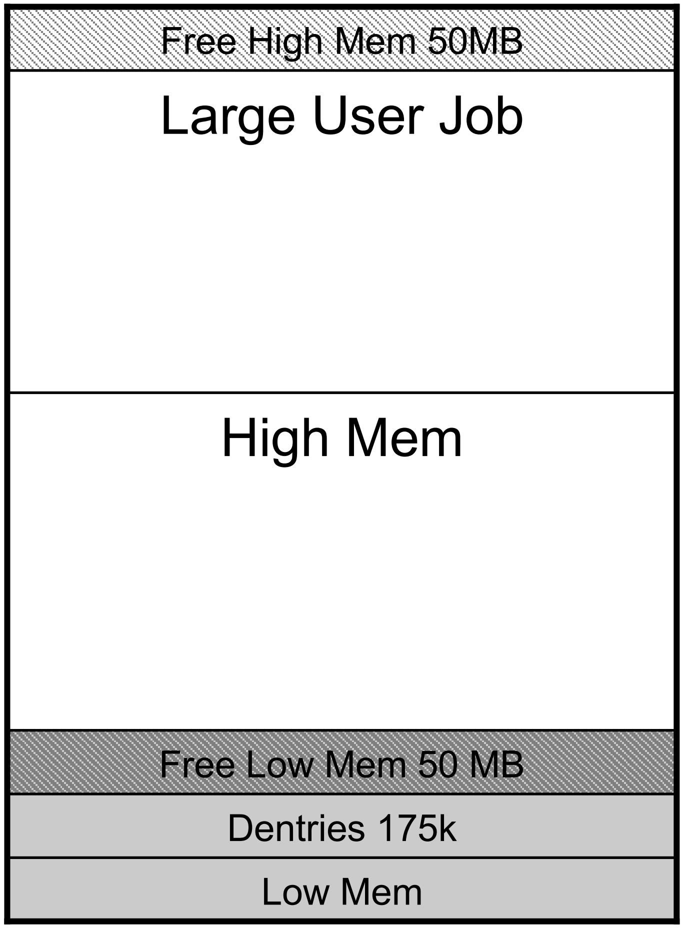
- Red Hat RHEL 4 and SuSE SLES 9 didn’t support all of our IOAPICs
- Red Hat RHEL 4 x86_64 only supported 8 CPUs
- Red Hat claimed RHEL 4 Update 3 supports 256GB
- Distribution install kernels do not support APIC mode
- Patch the kernel to handle resource conflicts in I/O space
- Issues with BIOS table memory locations and format
  - Hot-Add Memory
  - MP Tables

Sample Customer Issue

- System has 8 x86_64 processors, 28GB of RAM.
- Running SLES 9 SP2 for ia32 processors.
- System hangs every evening.

Diagnosis

- System runs all day with roughly 50MB of free memory high memory, and 50MB of free low memory.
- During the day, there are roughly 175k directory entries in the cache.



Free High Mem 50MB

Large User Job

High Mem

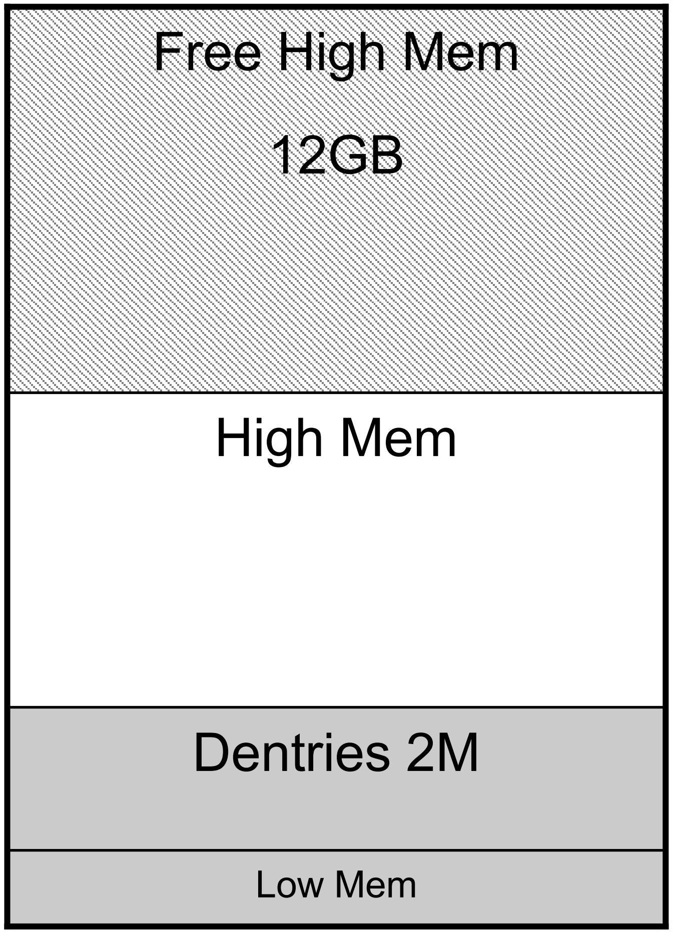
Free Low Mem 50 MB

Dentries 175k

Low Mem

Diagnosis cont.

- At the end of the work day, a large job ends, and the high free memory jumps to 12GB.
- Over the next couple of hours, the number of directory entries in the cache climbs to 2M.
- Free high memory remains around 12GB, but free low memory drops below 10MB.
- The system becomes sluggish, and finally hangs.



Conclusion

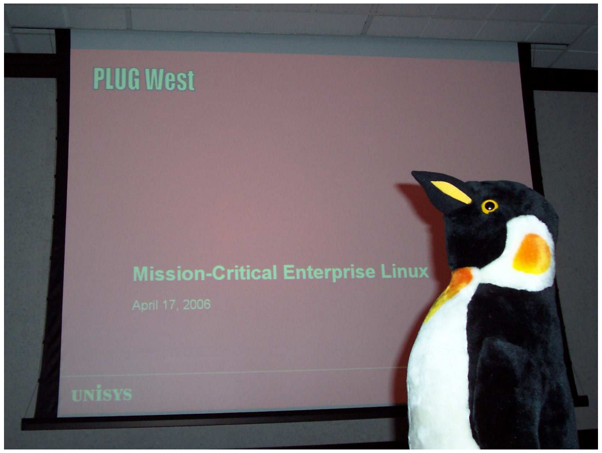
- In the ia32 kernel, all kernel data structures must reside in low memory.
- The algorithms to purge the caches are based on percentage of memory free, not percentage of low memory free.
- With 12GB of free memory, the directory entry cache is never purged.
- The directory entry cache consumed all free low memory, until the kernel could no longer function.

Solution

- We created a kernel module that constantly monitors the directory entry cache.
- When the directory entry cache exceeds 200k entries, the module calls the cache’s shrink routine to free old entries.
- We are investigating the ia32 virtual memory code in the kernel to propose a permanent solution to the kernel community.

Q & A

Event Pictures – Executive Conference Room



Event Pictures – Engineering Lab Tour