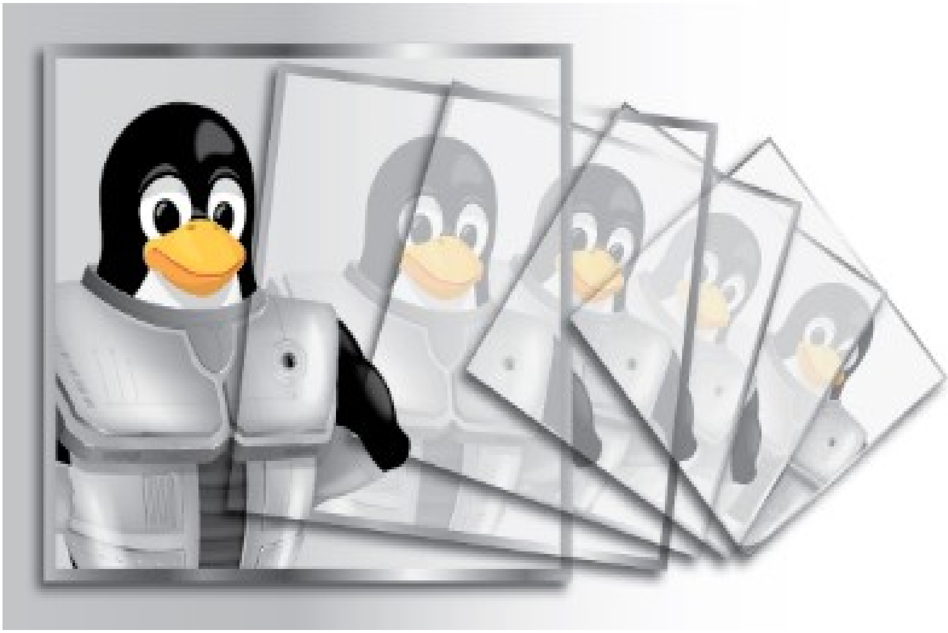
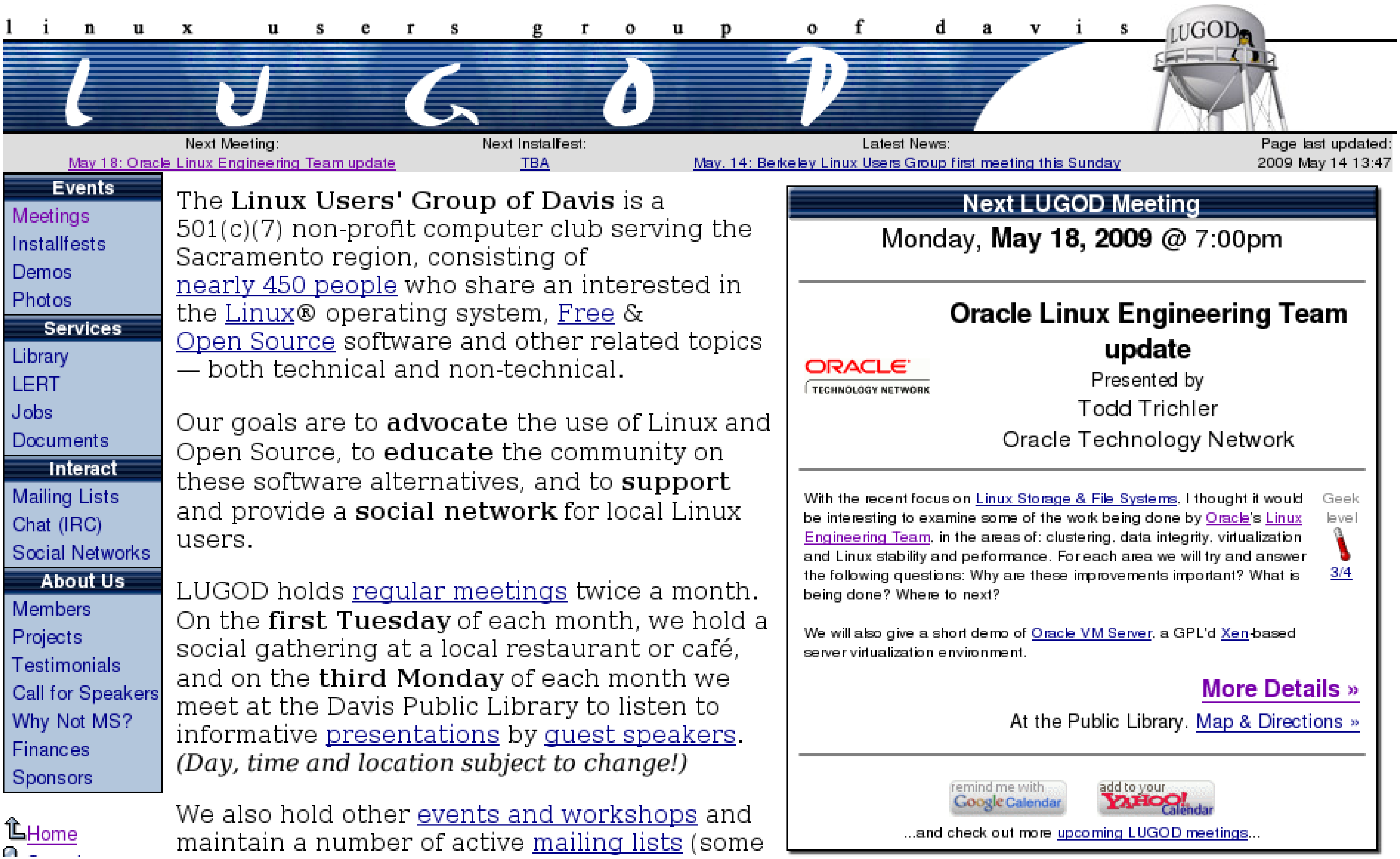
Oracle Linux Engineering Team
Update

What has ORACLE been doing to Linux!?



@ LUGOD
18th May 2009

Todd Trichler
Oracle Technology Network



Next Meeting: May 18: Oracle Linux Engineering Team update

Next Installfest: TBA

Latest News: May. 14: Berkeley Linux Users Group first meeting this Sunday

Page last updated: 2009 May 14 13:47

Events
- Meetings
- Installfests
- Demos
- Photos

Services
- Library
- LERT
- Jobs
- Documents

Interact
- Mailing Lists
- Chat (IRC)
- Social Networks

About Us
- Members
- Projects
- Testimonials
- Call for Speakers
- Why Not MS?
- Finances
- Sponsors

Home

The **Linux Users' Group of Davis** is a 501(c)(7) non-profit computer club serving the Sacramento region, consisting of nearly 450 people who share an interested in the Linux® operating system, Free & Open Source software and other related topics — both technical and non-technical.

Our goals are to **advocate** the use of Linux and Open Source, to **educate** the community on these software alternatives, and to **support** and provide a **social network** for local Linux users.

LUGOD holds regular meetings twice a month. On the **first Tuesday** of each month, we hold a social gathering at a local restaurant or café, and on the **third Monday** of each month we meet at the Davis Public Library to listen to informative presentations by guest speakers. *(Day, time and location subject to change!)*

We also hold other events and workshops and maintain a number of active mailing lists (some

Next LUGOD Meeting

Monday, **May 18, 2009** @ 7:00pm

Oracle Linux Engineering Team update

ORACLE TECHNOLOGY NETWORK

Presented by
Todd Trichler
Oracle Technology Network

With the recent focus on Linux Storage & File Systems, I thought it would be interesting to examine some of the work being done by Oracle's Linux Engineering Team, in the areas of: clustering, data integrity, virtualization and Linux stability and performance. For each area we will try and answer the following questions: Why are these improvements important? What is being done? Where to next?

Geek level 3/4

We will also give a short demo of Oracle VM Server, a GPL'd Xen-based server virtualization environment.

More Details »

At the Public Library. Map & Directions »

remind me with Google Calendar | add to your YAHOO! Calendar

...and check out more upcoming LUGOD meetings...

Cautionary Statement Regarding Forward-Looking Statements

This document contains certain forward-looking statements about Oracle and Sun, including statements that involve risks and uncertainties concerning Oracle's proposed acquisition of Sun, anticipated product information, estimates of future results of operations, anticipated customer and partner advantages and benefits, and general business outlook. When used in this document, the words "anticipates", "plans", "estimates", "may", "can", "will", "believes", "expects" or "expected", "projects", "intends", "likely", similar expressions and any other statements that are not historical facts are intended to identify those assertions as forward-looking statements. Any such statement may be influenced by a variety of factors, many of which are beyond the control of Oracle or Sun, that could cause actual outcomes and results to be materially different from those projected, described, expressed or implied in this document due to a number of risks and uncertainties. Potential risks and uncertainties include, among others, the possibility that the transaction will not close or that the closing may be delayed, the anticipated synergies of the combined companies may not be achieved after closing, the combined operations may not be successfully integrated in a timely manner, if at all, general economic conditions in regions in which either company does business, and the possibility that Oracle or Sun may be adversely affected by other economic, business, and/or competitive factors. Accordingly, no assurances can be given that any of the events anticipated by the forward-looking statements will transpire or occur, or if any of them do so, what impact they will have on the results of operations or financial condition of Oracle or Sun.

In addition, please refer to the documents that Oracle and Sun, respectively, file with the Securities and Exchange Commission (the "SEC") on Forms 10-K, 10-Q and 8-K. These filings identify and address other important factors that could cause Oracle's and Sun's respective financial and operational results to differ materially from those contained in the forward-looking statements set forth in this document. You are cautioned to not place undue reliance on forward-looking statements, which speak only as of the date of this report. Neither Oracle nor Sun is under any duty to update any of the information in this document.

Additional Information about the Merger and Where to Find It

In connection with the proposed merger, Sun will file a proxy statement with the SEC. Additionally, Sun and Oracle will file other relevant materials in connection with the proposed acquisition of Sun by Oracle pursuant to the terms of an Agreement and Plan of Merger by and among Oracle, Soda Acquisition Corporation, a wholly-owned subsidiary of Oracle, and Sun. The materials to be filed by Sun with the SEC may be obtained free of charge at the SEC's web site at www.sec.gov. Investors and security holders of Sun are urged to read the proxy statement and the other relevant materials when they become available before making any voting or investment decision with respect to the proposed merger because they will contain important information about the merger and the parties to the merger.

Oracle, Sun and their respective directors, executive officers and other members of its management and employees, under SEC rules, may be deemed to be participants in the solicitation of proxies of Sun stockholders in connection with the proposed merger. Investors and security holders may obtain more detailed information regarding the names, affiliations and interests of certain of Oracle's executive officers and directors in the solicitation by reading the proxy statement and other relevant materials filed with the SEC when they become available. Information concerning the interests of Sun's participants in the solicitation, which may, in some cases, be different than those of Sun's stockholders generally, is set forth in the materials filed with the SEC on Form 10-K and will be set forth in the proxy statement relating to the merger when it becomes available.

The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decision. The development, release, and timing of any features or functionality described for Oracle's products remains at the sole discretion of Oracle.

This document is for information purposes only and may not be incorporated into a contract.

Support, Comfort & Proof Points

Your perspective – will determine how you value the importance of things.

- Large Scale Customers
  - Who bet a significant portion of their business, demand quality support and someone to hold accountable.
  - New technology projects often only get started, once people are confident of the safety net underneath them.
  - The cost of opportunity lost in business downtime is often many times the initial cost of the infrastructure.

Terminology

- Open source software; not a new Linux distribution
- Terminology
  - Unbreakable Linux = Support Program
  - Enterprise Linux = Software
- Tracks Red Hat product releases
- Freely available source and binaries
- x86, x86-64 (AMD and Intel) and ia64 architectures

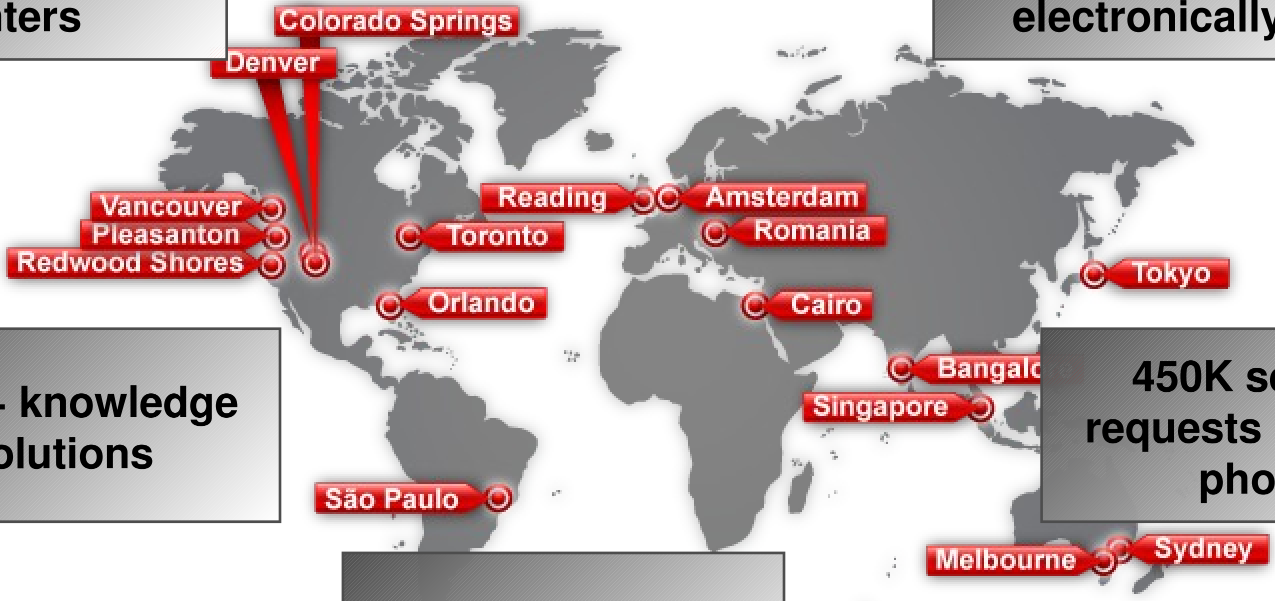
Making Linux Better

One Support Call for the Complete Stack

7,000+ support staff in 17 global support centers

Support in 27 local languages

1.25M service requests filed electronically



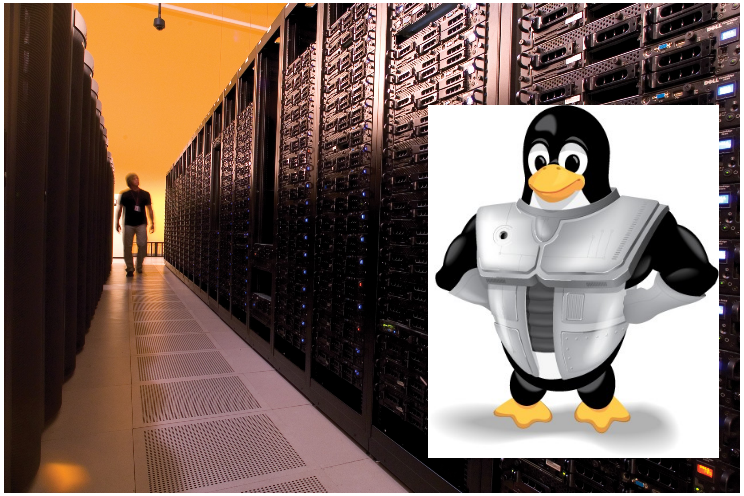
400K+ knowledge solutions

450K service requests over the phone

110M+ web knowledge transfers

Software Channel: Unbreakable Linux Network (ULN)

- linux.oracle.com
- Updates, bug fixes, security fixes
- Access via Yum and Web browser
- Errata mailing list
- Search packages
- Download packages
  - Binary and source
- Manage channel subscriptions
- See how current your registered servers are



What is Premier Backporting?

- **Traditional Backporting** = A specific bug fix produced for the latest version of a package may be retroactively created and introduced as part of an earlier release or update level (e.g. a bug fix released in RHEL5 or OEL5 is also released as part of RHEL4 or OEL4)
- Only Oracle offers **Premier Backporting**, which goes far beyond traditional backporting. For example:
  - Current user runs RHEL4 or OEL4 Update 4 release;
  - When update 5 becomes available, they want to stay on Update 4 and get just one specific bug fix in Update 5 backported to their current environment;
  - With **Premier Backporting**, Oracle makes that specific bug fix available without forcing the customer to upgrade;
  - In sharp contrast, a Red Hat support customer must upgrade to the entire Update release to get just the one bug fix they need.

No pressure to upgrade to the latest Update release

Comprehensive Indemnification

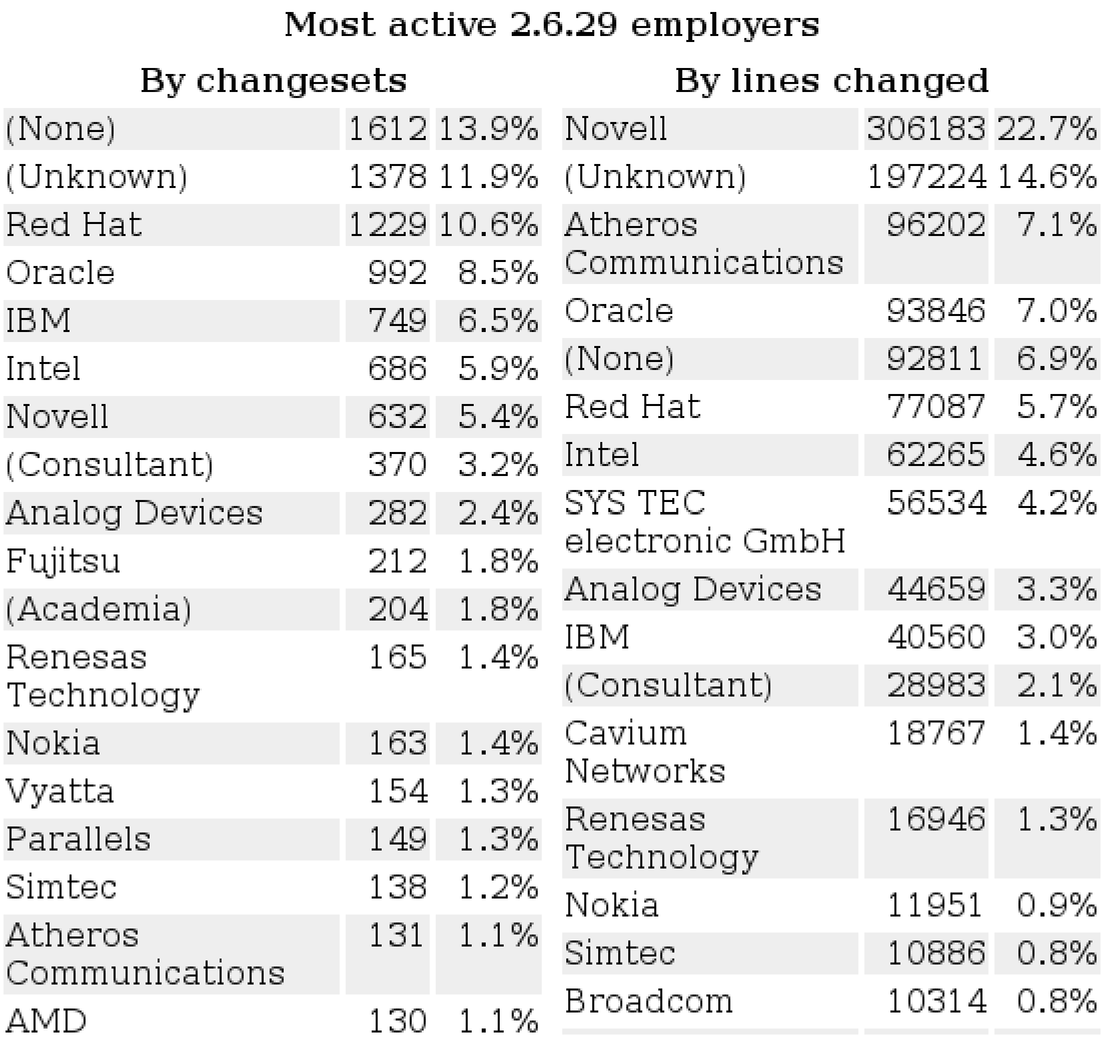
- Indemnification for intellectual property claims raised against our customers
- Available to **all** Oracle-supported customers
  - Network, Basic and Premier
- Includes damages, liabilities, costs and expenses awarded by courts
  - **Not** limited to the amount of money paid by the customer

Deploy Linux with Confidence

Now, back to the original question...

What has ORACLE been doing to Linux!?

Quite a lot ...



Most active 2.6.29 employers

By changesets

(None)	1612	13.9%
(Unknown)	1378	11.9%
Red Hat	1229	10.6%
Oracle	992	8.5%
IBM	749	6.5%
Intel	686	5.9%
Novell	632	5.4%
(Consultant)	370	3.2%
Analog Devices	282	2.4%
Fujitsu	212	1.8%
(Academia)	204	1.8%
Renesas Technology	165	1.4%
Nokia	163	1.4%
Vyatta	154	1.3%
Parallels	149	1.3%
Simtec	138	1.2%
Atheros Communications	131	1.1%
AMD	130	1.1%

By lines changed

Novell	306183	22.7%
(Unknown)	197224	14.6%
Atheros Communications	96202	7.1%
Oracle	93846	7.0%
(None)	92811	6.9%
Red Hat	77087	5.7%
Intel	62265	4.6%
SYS TEC electronic GmbH	56534	4.2%
Analog Devices	44659	3.3%
IBM	40560	3.0%
(Consultant)	28983	2.1%
Cavium Networks	18767	1.4%
Renesas Technology	16946	1.3%
Nokia	11951	0.9%
Simtec	10886	0.8%
Broadcom	10314	0.8%

Source: Jonathan Corbet's article, March 18th 2009
http://lwn.net/Articles/324046/

Code, Commitment & Contributions

- Clustering
  - OCFS2
- Data Integrity
  - T10Dif
- Stability & Performance
  - NUMA aware semtimedop
  - Automated kernel testing
  - Kthread backed AIO
  - IO queuing and completion affinity
  - CFQ IO context sharing
  - Real time scheduler enhancements

Code, Commitment & Contributions

- Stability & Performace
  - SSD simulation and tuning
  - RDS – Reliable Data Streams
- File System Improvements
  - NFS Ipv6
  - Btrfs
  - CRFS – Coherent Remote File System
- Virtualization
  - PV Networking driver for windows
  - PV Hugepages
  - Loop improvements for Xen images
  - JeOS – Just Enough OS

Code, Commitment & Contributions

- Oracle VM Server a Xen based virtualization environment.
  - All fixes contributed back to the Xen community
  - Freely available source and ISO's
- Open Source Projects
  - Assist internal teams with kernel debugging and tuning
  - Run visible mainline projects that improve Linux for all workloads
  - Develop specific features for Oracle products
  - Participate heavily in the kernel community.

OCFS2 - Clustering

- Oracle Cluster File System 2 – was the first general purpose cluster file system to be added to the mainline linux kernel (2.6.16).
- Support for 3rd party clusterware
- Applications like Oracle RAC and Oracle VM, can leverage the advanced features in OCFS2
- Next steps
  - Add additional features, that will make snapshoting in virtualization environments easier.
  - Check out Wim's blog entry on OCFS2_reflink

Source: Wim Coekaert's blog, May 4th 2009
http://blogs.oracle.com/wim/2009/05/ocfs2_reflink.html

Now, about reflink. The reason we implemented reflink is for Oracle VM. As you know, a virtual machine/guest owns one or more virtual disks. These virtual disks are represented as files on a filesystem hosted by the hypervisor. In the case of Oracle VM, if you have SAN or iSCSI storage, we put an OCFS2 filesystem on top of this, managed by the management domain (dom0). The virtual disks live on top of this OCFS2 volume.

These virtual disks can become very large, they usually are many GB's in size. So when a user wants to create a clone of a virtual machine or create a virtual machine based on an existing template, we copy the content of the original virtual disks to a new set of virtual disks. By default this duplicates the amount of storage used.

ie. you have VM1 with a 40gb virtual disk (vm1/system.img) and you want to copy that to create VM2 based on the same virtual disk image (vm2/system.img).

The reflink feature in OCFS2 which was published to fs-devel and ocfs2-devel a while back, supports this operation through effectively creating hard links but with copy-on-write (or basically a point-in-time data hard link).

Today, we copy the file vm1/system.img to vm2/system.img. Tomorrow, we do reflink vm1/system.img vm2/system.img. At initial create time no additional space is used, no actual copying is done, it just creates a totally new inode/file and shares the data extents. As soon as a write is done to one or the other side, 1mb chunks are copied over where the writes occur.

This allows us to create instant copies of files (or in the case of Oracle VM, virtual disk images).

Some of the advantages of reflink are :

- Each “hard link" or point-in-time copy, is a regular file for the OS, for an application etc, so there are no changes needed to applications or backup software. This is totally transparent, there is no container around these files etc. Unlike vmdk and vhd where the snapshots live inside the containers.

- It is fully cluster safe so this works in an OCFS2 filesystem cluster so the link and the COW works on any node even if the file is used and opened on another node. This allows us in the Oracle VM case to create snapshots and run these new VMs on a different node than the original VM is running.

- This is a generic feature just like symlink. It is available to any user or application.

- It is open source (part of OCFS2 code) free to use for anyone.

T10Dif – Data Integerity

- First Linux implementation of T10 standard for end to end checksumming done by Oracle.
- Solves the problem of only finding out if your backup is good when you go to restore. Very valuable for super critical systems.
- http://oss.oracle.com/projects/data-integrity/
- Next steps
  - Test with early implementations from LSI and others
  - Get code into mainline

NUMA Aware Semtimedop

- NUMA aware implementation of semtimedop greatly reduces lock contention on large machines. Memory allocation optimizations, very important for large systems.
- Patch improves performance by 18%, but we needed 50% to match nanosleep hotfix
- Next steps
  - Tune patch based on Intel performance data
  - Port to mainline

Solid State Disk Simulation and Tuning

- SSD devices will soon expose bottlenecks in IO subsystems as the iops/sec increase
- Simulate an SSD device by connecting two Marvel sata cards in two machines. Host memory on one machine is used to simulate a drive
- Detect whether storage is flash or spindle
  - allows for intelligent use of storage by app.
- Next Steps
  - Finish driver updates to enable the sata host mode
  - Benchmark & Send fixes to mainline

Infiniband & Reliable Data Streams

- Through the Reliable Data Sreams work we have done a lot of bugfixing on the generic linux infiniband stack(ofed) but also fully implemented the RDS driver, and this driver is both in core linux and in the ofed (bsd) OS vendor neutral stack.
- We are maintainers of RDS.
  - Even though RDS is used by Oracle in Exadata, it is not oracle “specific”

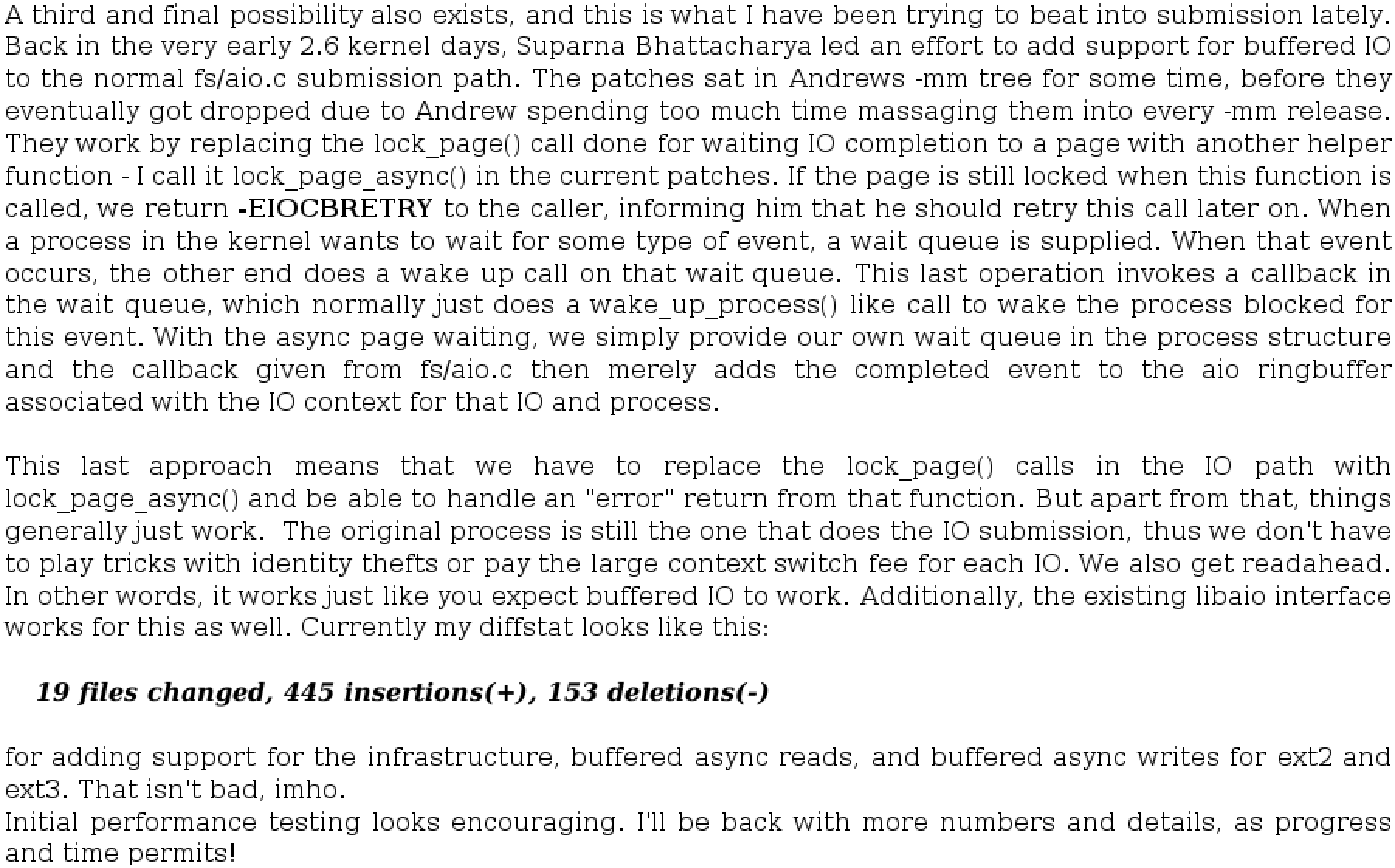
IO Layer

- Maintainership for Linus of the Linux block IO layer.
- Lots of work on performance enhancements.
- Totally new linux block io code written by Jens. (2.6)
- Async IO changes for intensive operations, standard AIO used by pretty much every database (oracle,db2,msql etc.)
- Linux implementation for O_DIRECT was doen by Oracle.

IO Affinity

- Framework to force IO completion onto the same CPU that started the IO
- Can significantly reduce lock contention and cache line bouncing, improving NUMA performance, just for large cpu systems.
- Status: early implementation
- Next steps
  - Gather performance results and tune the patches (HP is benchmarking on large hardware)

Source: Jens Axboe's blog, Jan 19th 2009
http://axboe.livejournal.com/1718.html



A third and final possibility also exists, and this is what I have been trying to beat into submission lately. Back in the very early 2.6 kernel days, Suparna Bhattacharya led an effort to add support for buffered IO to the normal fs/aio.c submission path. The patches sat in Andrews -mm tree for some time, before they eventually got dropped due to Andrew spending too much time massaging them into every -mm release. They work by replacing the lock_page() call done for waiting IO completion to a page with another helper function - I call it lock_page_async() in the current patches. If the page is still locked when this function is called, we return **-EIOCBRETRY** to the caller, informing him that he should retry this call later on. When a process in the kernel wants to wait for some type of event, a wait queue is supplied. When that event occurs, the other end does a wake up call on that wait queue. This last operation invokes a callback in the wait queue, which normally just does a wake_up_process() like call to wake the process blocked for this event. With the async page waiting, we simply provide our own wait queue in the process structure and the callback given from fs/aio.c then merely adds the completed event to the aio ringbuffer associated with the IO context for that IO and process.

This last approach means that we have to replace the lock_page() calls in the IO path with lock_page_async() and be able to handle an "error" return from that function. But apart from that, things generally just work. The original process is still the one that does the IO submission, thus we don't have to play tricks with identity thefts or pay the large context switch fee for each IO. We also get readahead. In other words, it works just like you expect buffered IO to work. Additionally, the existing libaio interface works for this as well. Currently my diffstat looks like this:

19 files changed, 445 insertions(+), 153 deletions(-)

for adding support for the infrastructure, buffered async reads, and buffered async writes for ext2 and ext3. That isn't bad, imho.
Initial performance testing looks encouraging. I'll be back with more numbers and details, as progress and time permits!

GCC and libstdc++

- Completed implementation of UTF-16 and UTF-32 support in gcc
- Work with the new GNU linker (Gold)
- Use Oracle products to do compiler testing
- Restructuring debuginfo project
- libstdc++ upstream maintainership
- performance analysis of the low level IO libraries

NFS Layer – statd and NFS IPv6

- Co-maintainer together with Trond (Netapp)
- A large collection of changes required to enable NFS and RPC on IPv6
- IPv6 export access control, and user-level command updates.
- Ipv6 support is important for goverments
- Status
  - In final stages, most patches already included in mainline
  - NFS testing and interoperability requirements (Bull and Netapp are doing the testing)

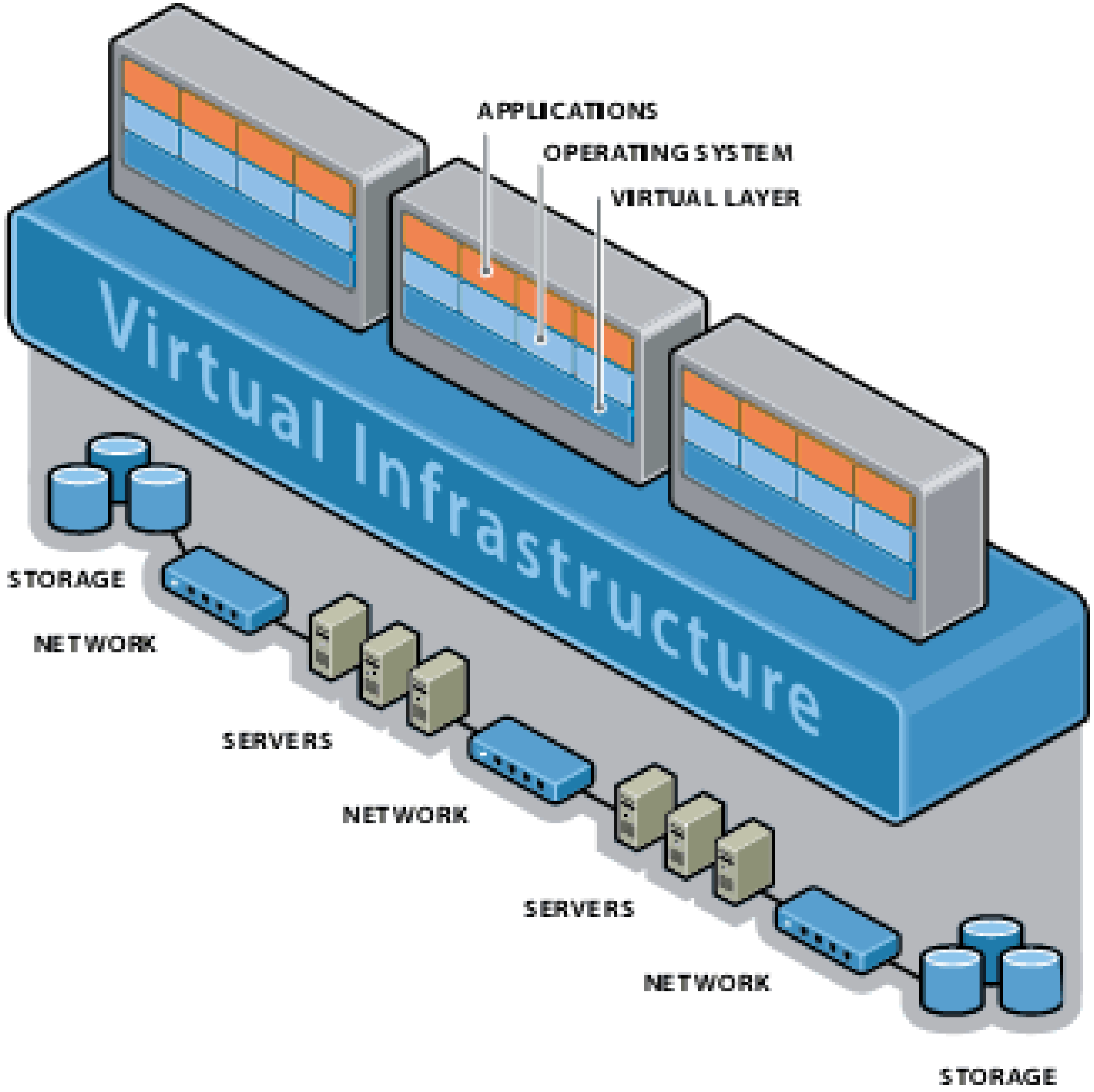
Btrfs

- Hailed by Linux kernel developers as
  “The Next Generation File System”
  included in 2.6.29 kernel
- Copy on write filesystem with snapshotting and checksumming
- Designed for repair and reliability
- Many features that are difficult or impossible with today's filesystems
- Includes flexible storage management across multiple devices
- http://btrfs.wiki.kernel.org/

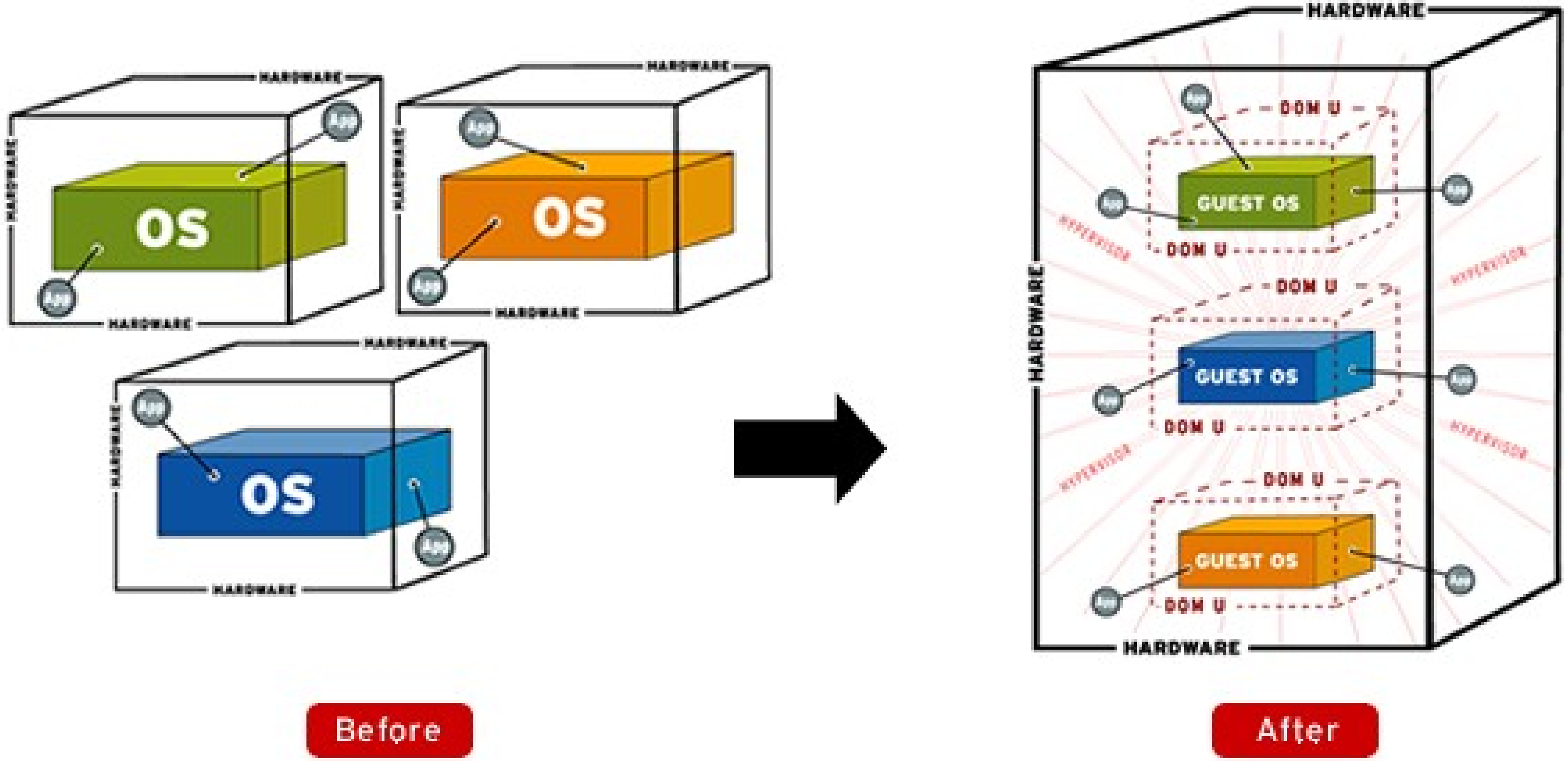
CRFS

- Coherent Remote Filesystem
- Network filesystem using the Btrfs disk format
- Local caching with POSIX semantics
- Aimed at the most common NFS use cases
- Protocol sends Btrfs key/value pairs over the wire, making it ideal for synchronizing two Btrfs filesystems as well
- May cluster enable the Btrfs volume manager

Virtualization Infrastructure

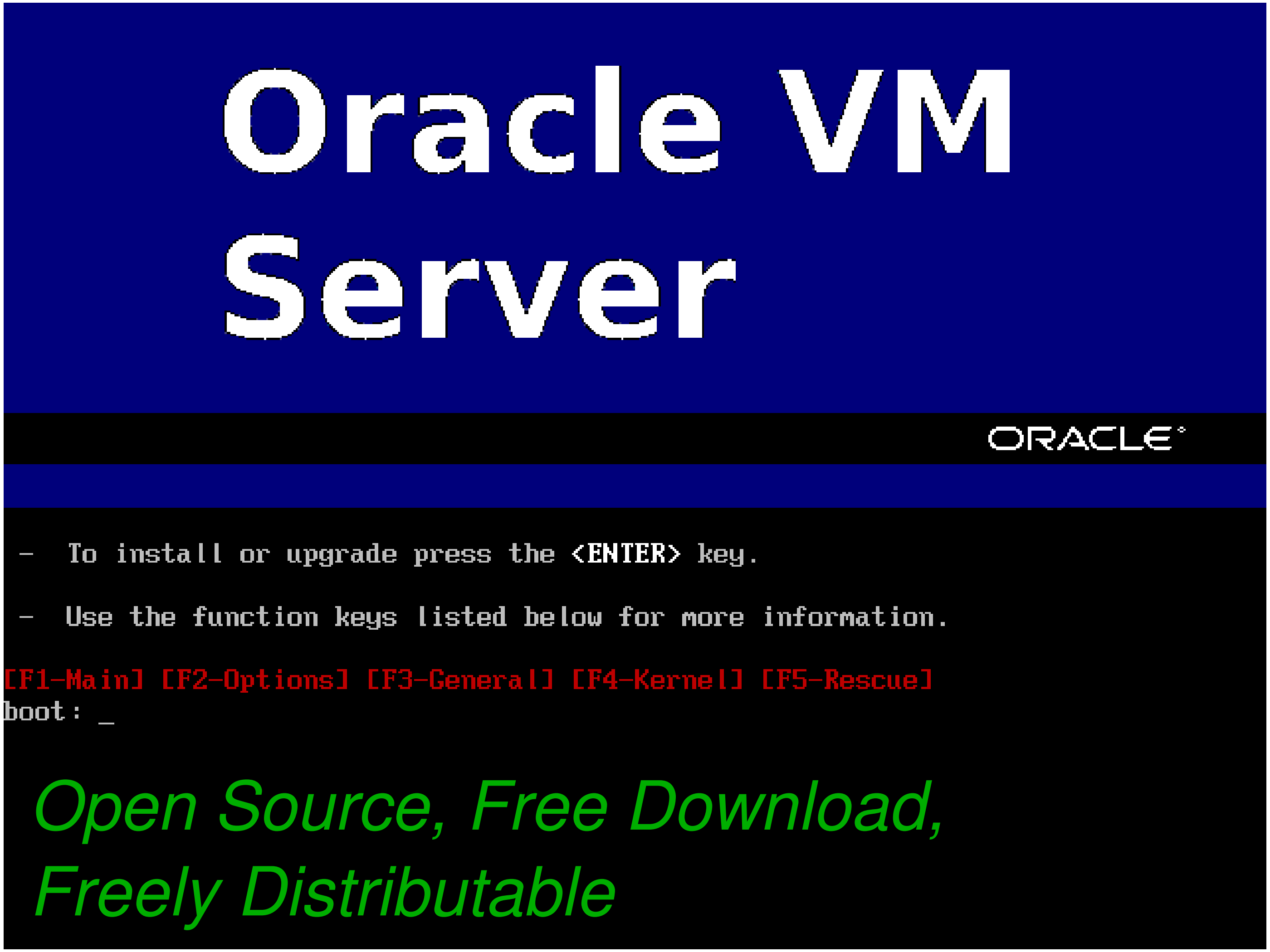


Server Consolidation



Server Consolidation is very Green $$$$

Oracle VM Server

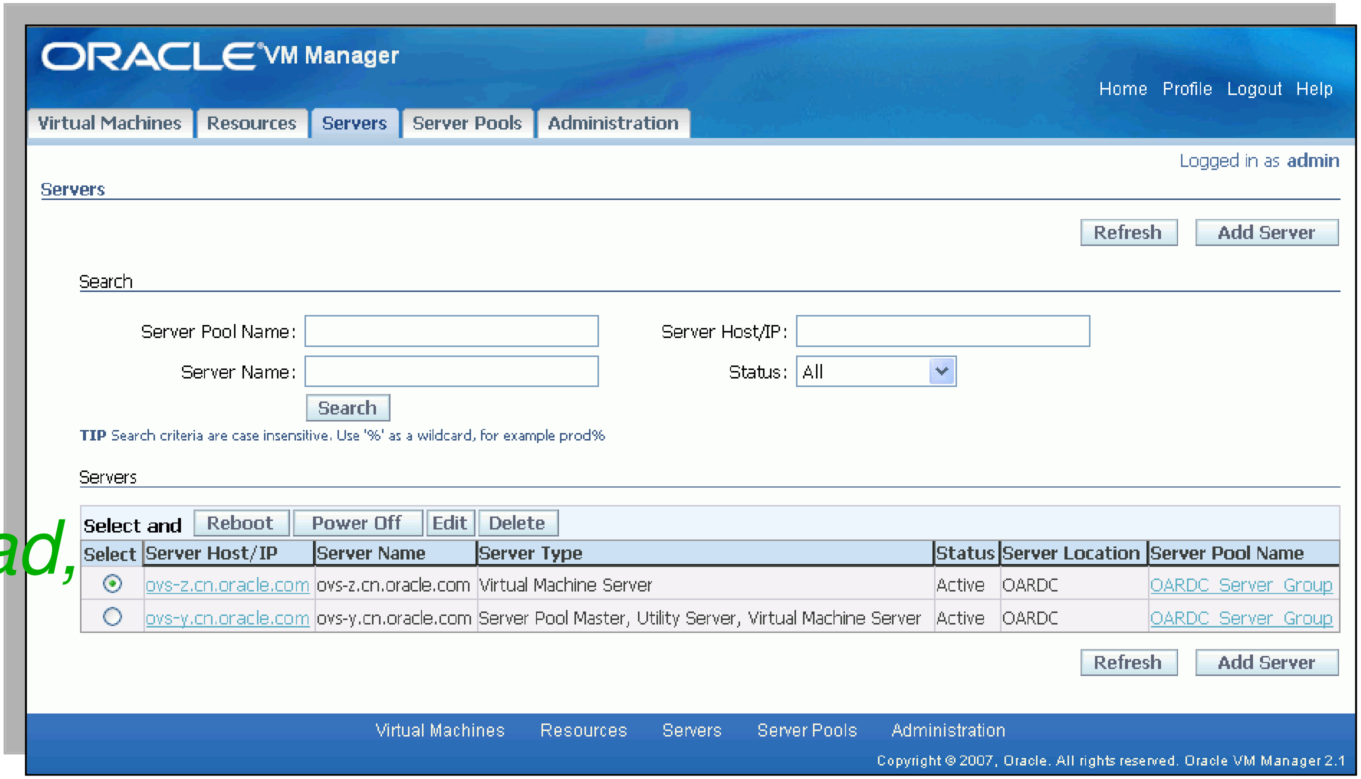


```
 -   To install or upgrade press the <ENTER> key.

 -   Use the function keys listed below for more information.

[F1-Main] [F2-Options] [F3-General] [F4-Kernel] [F5-Rescue]
boot: _
```

Open Source, Free Download,
Freely Distributable

Oracle VM Manager

- Browser-based management solution
- Included with Oracle VM
- Full VM lifecycle management:
  - Create
  - Configure
  - Clone
  - Share
  - Boot
  - Migrate



Logged in as admin

Servers

Search

Server Pool Name:

Server Host/IP:

Server Name:

Status: All

TIP Search criteria are case insensitive. Use '%' as a wildcard, for example prod%

Servers

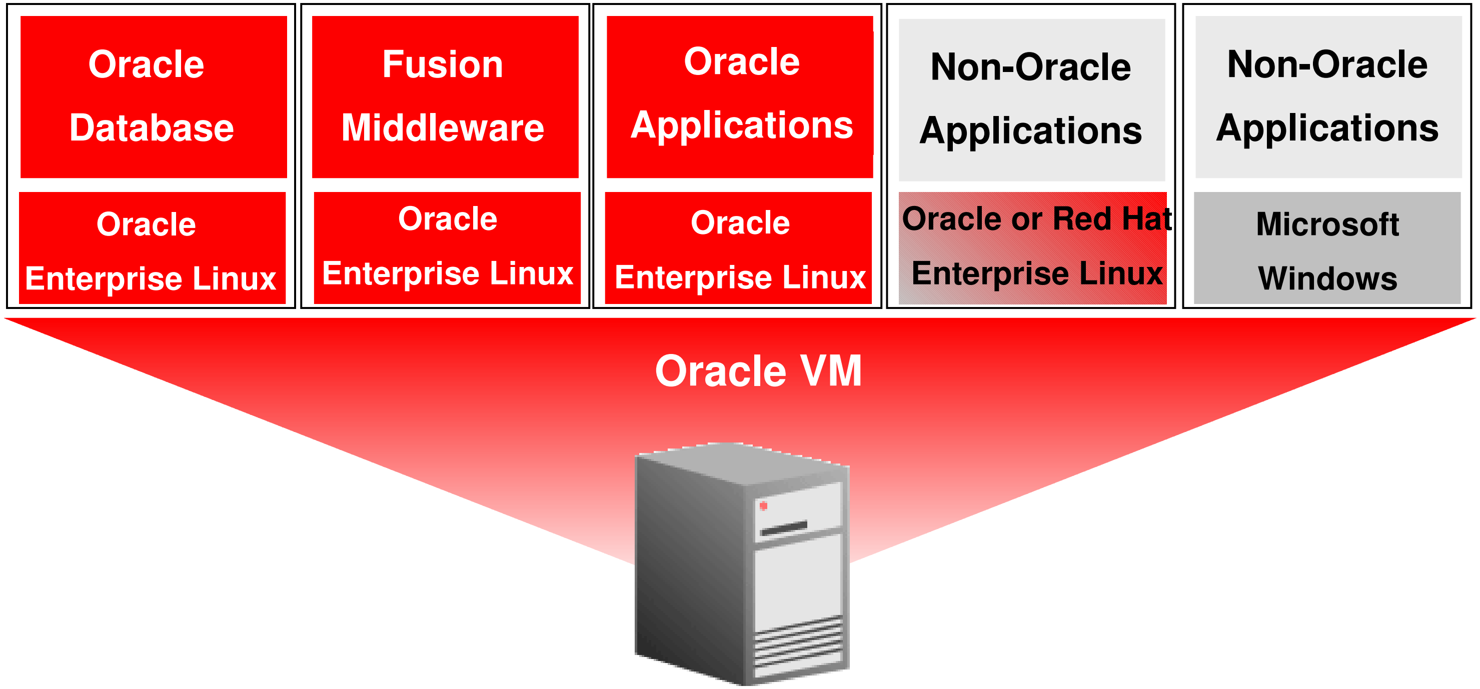
Select and

Select	Server Host/IP	Server Name	Server Type	Status	Server Location	Server Pool Name
●	ovs-z.cn.oracle.com	ovs-z.cn.oracle.com	Virtual Machine Server	Active	OARDC	OARDC_Server_Group
○	ovs-y.cn.oracle.com	ovs-y.cn.oracle.com	Server Pool Master, Utility Server, Virtual Machine Server	Active	OARDC	OARDC_Server_Group

Copyright © 2007, Oracle. All rights reserved. Oracle VM Manager 2.1

Free Download, Freely Distributable

Oracle VM



Oracle Database

Oracle Enterprise Linux

Fusion Middleware

Oracle Enterprise Linux

Oracle Applications

Oracle Enterprise Linux

Non-Oracle Applications

Oracle or Red Hat Enterprise Linux

Non-Oracle Applications

Microsoft Windows

Oracle VM

- Oracle tested and supported server virtualization technology
- Maximizes consolidation of Linux and Windows servers
- Saves on power, cooling and space

JeOS

- Just enough OS - is a secure, minimized OS that is freely re-distributable and backed by enterprise-class support - developers and ISVs can now easily build Oracle VM Templates.
- With Oracle Enterprise Linux JeOS, anyone can put their applications on top of a small footprint, enterprise operating system and build a full stack virtual machine or Oracle VM Template. The resulting Oracle VM Template is freely re-distributable, without trial license requirements, and is backed by enterprise-class support.

Oracle’s Contributions

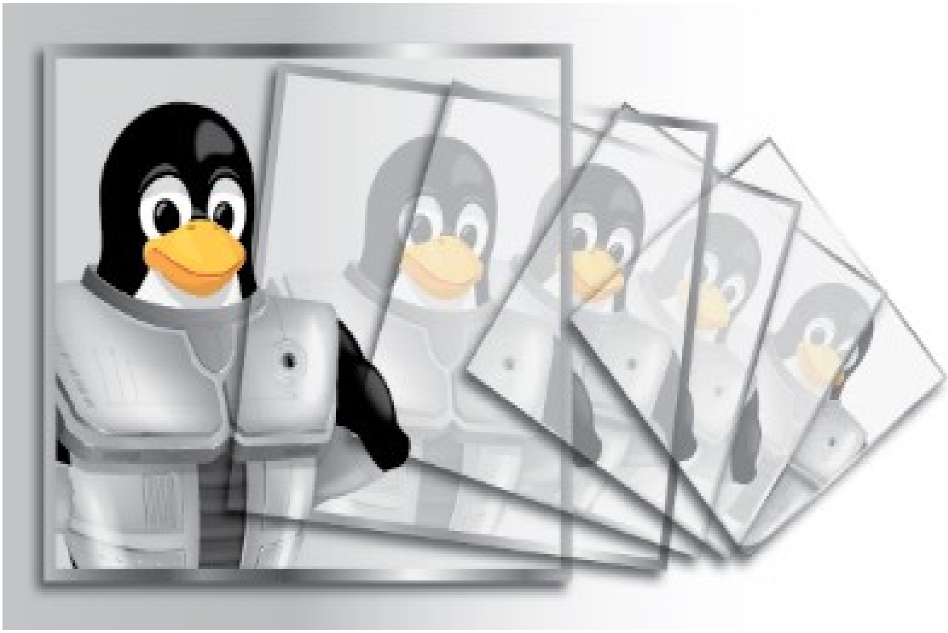
- Enhanced and optimized Xen technology
  - I/O overhead
  - Memory overhead (also ... tmem coming soon)
  - Process Scheduling
- Community involvement
  - Dedicated Xen development team
  - Code and bug fix contributions to Xen community
  - Members of Xen community at Oracle
  - Member of XenSummit committee
- Significant testing
  - Real-world testing with Oracle On Demand workloads
  - Testing with Oracle Validated configuration workloads

Oracle's Internal Usage
Oracle VM

- Oracle VM widely deployed across Oracle
  - Product Development, Test, QA
  - Oracle University's public training infrastructure
  - Oracle On Demand:
    - Oracle's subscription and managed applications business
- As of Dec 2008
  - 1,000+ Oracle VM Servers
  - 6,500+ guest VMs
    - Enterprise Linux and Windows
- Still aggressively rolling across the company
  - Expect to increase by 2X+ by mid 2009

Thanks & Resources

- **http://www.oracle.com/technology/products/vm/index.html**
- **http://linuxfestnorthwest.org/slides-2009/oracle.pdf**
- **http://otn.oracle.com/linux**
- **http://oss.oracle.com**



@ LUGOD
18th May 2009

“So long, and thanks for the fish”

Credits: Adam Hawley, Dan Magenheimer and Wim Coekaerts for allowing me to blatantly plagiarize parts of their respective presentations.

Remember: re-use is at the core of virtualization