SUSE® Linux Enterprise Server

IBM System z – Current & Future Features

Session 12364

Marcus Kraft
Product Manager
mkraft@suse.com

Date: 2013-02-04
Distribution: PDF any

SUSE and the Attachmate Group

- SUSE, headquartered in Nürnberg / Germany, is an independently operating business unit of the Attachmate Group, Inc.
- The Attachmate Group is a privately held 1 billion+ $ revenue software company with four brands:

SUSE® at a Glance

SETTING THE BAR

LEADING PROVIDER
OF ENTERPRISE LINUX SOLUTIONS

GLOBAL MARKET

CUSTOMERS WORLDWIDE

15,000+

GLOBAL ORGANIZATION

EMPLOYEES IN 43 COUNTRIES

KNOW HOW

20+

YEARS OF LINUX ENGINEERING EXPERIENCE

PARTNERS

5,000+ MEMBER PARTNER ECOSYSTEM

THE GOLD STANDARD

AWARD WINNING

TECHNICAL SUPPORT AND CUSTOMER SERVICE

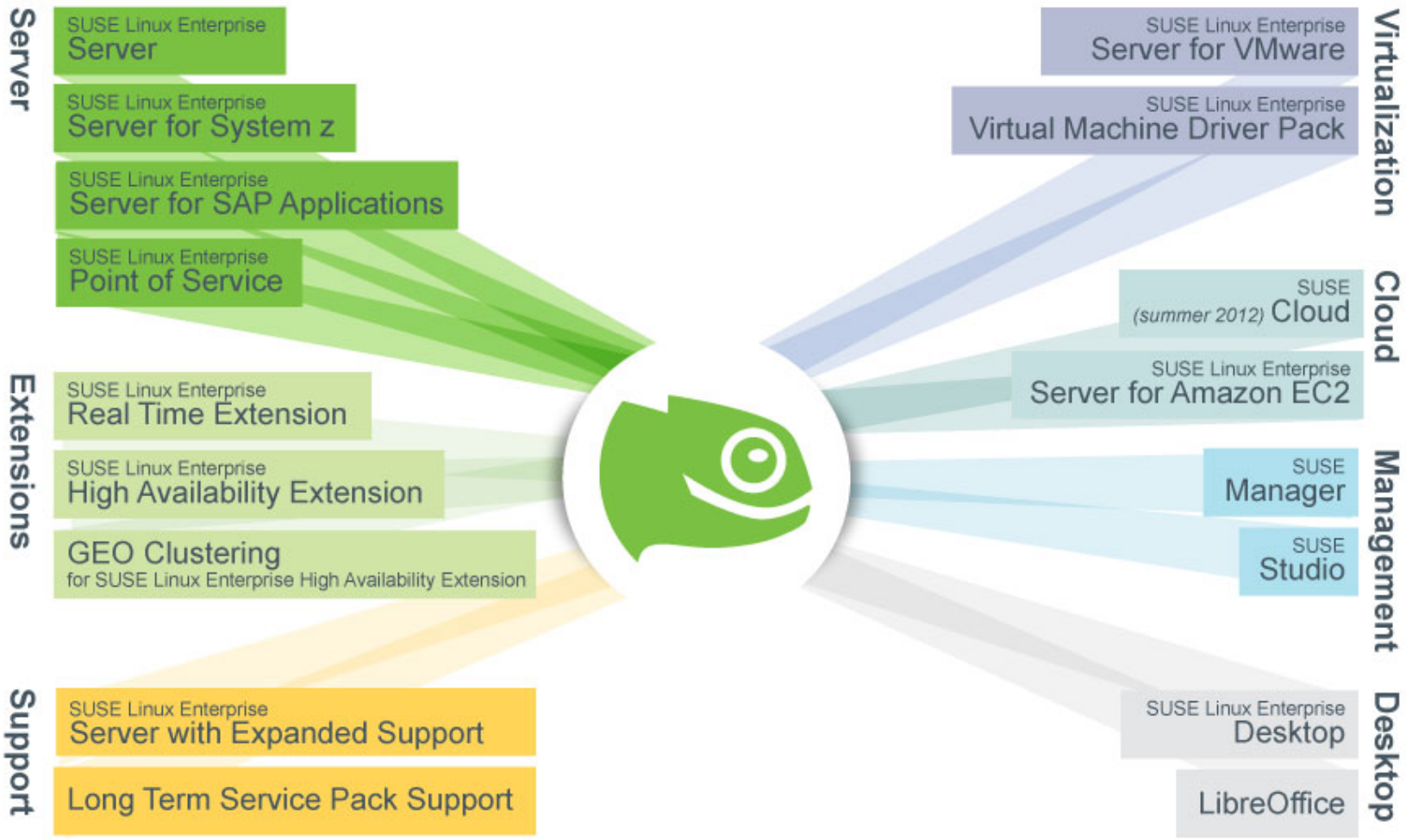
SUSE® Strategy

Cloud Infrastructure

Enterprise Computing

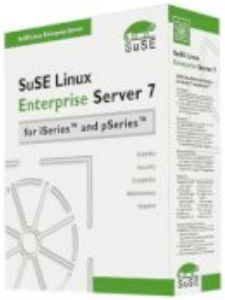
Integrated Systems

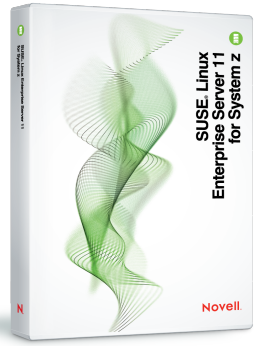
Comprehensive Portfolio



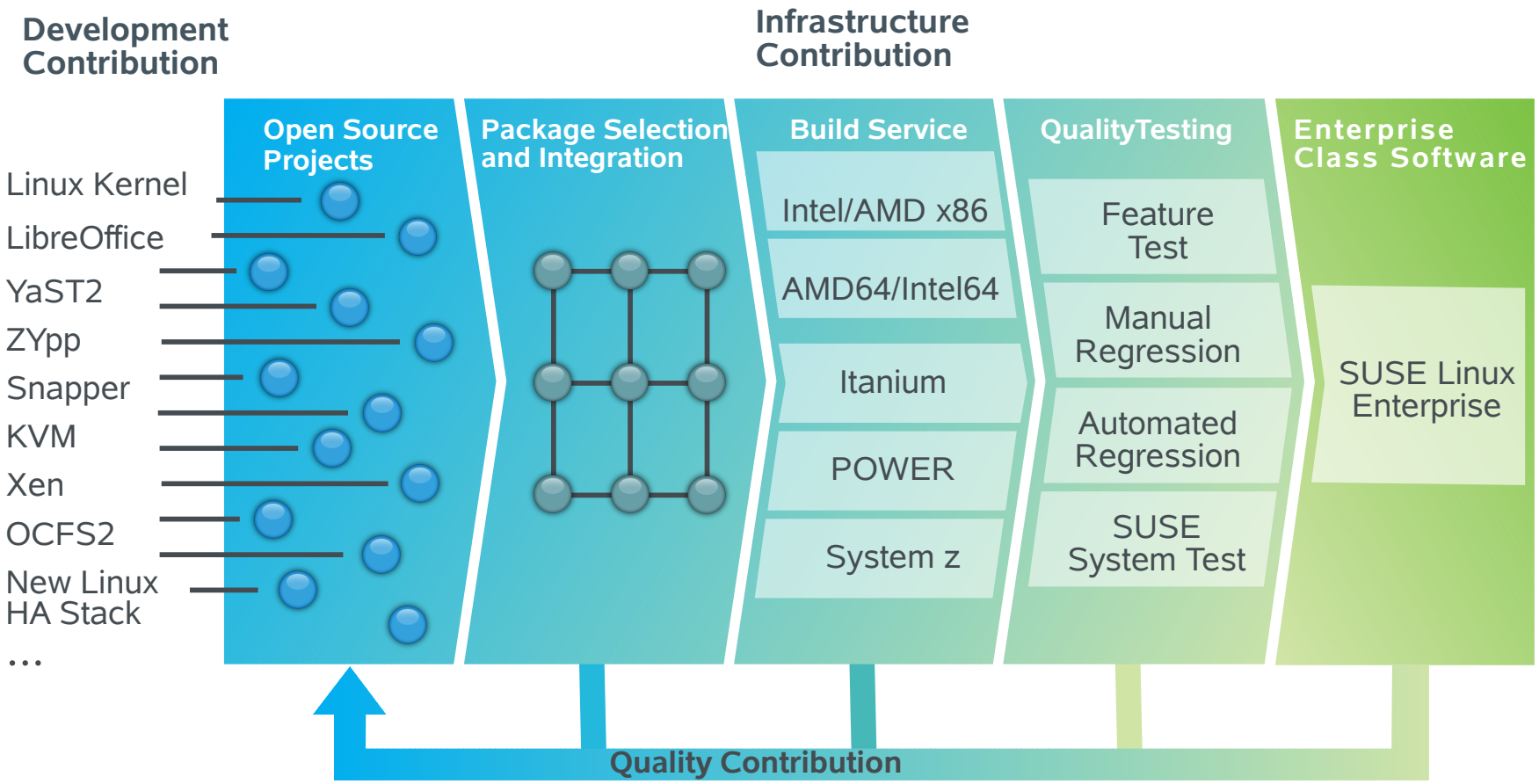
SUSE® Linux Enterprise Server

- SUSE Linux Enterprise Server 10/2000
- SUSE Linux Enterprise Server 7 08/2001
- SUSE Linux Enterprise Server 8 10/2002
- SUSE Linux Enterprise Server 9 08/2004
- SUSE Linux Enterprise Server 10 07/2006
- SUSE Linux Enterprise Server 11 03/2009
- SUSE Linux Enterprise Server 12 ~2014





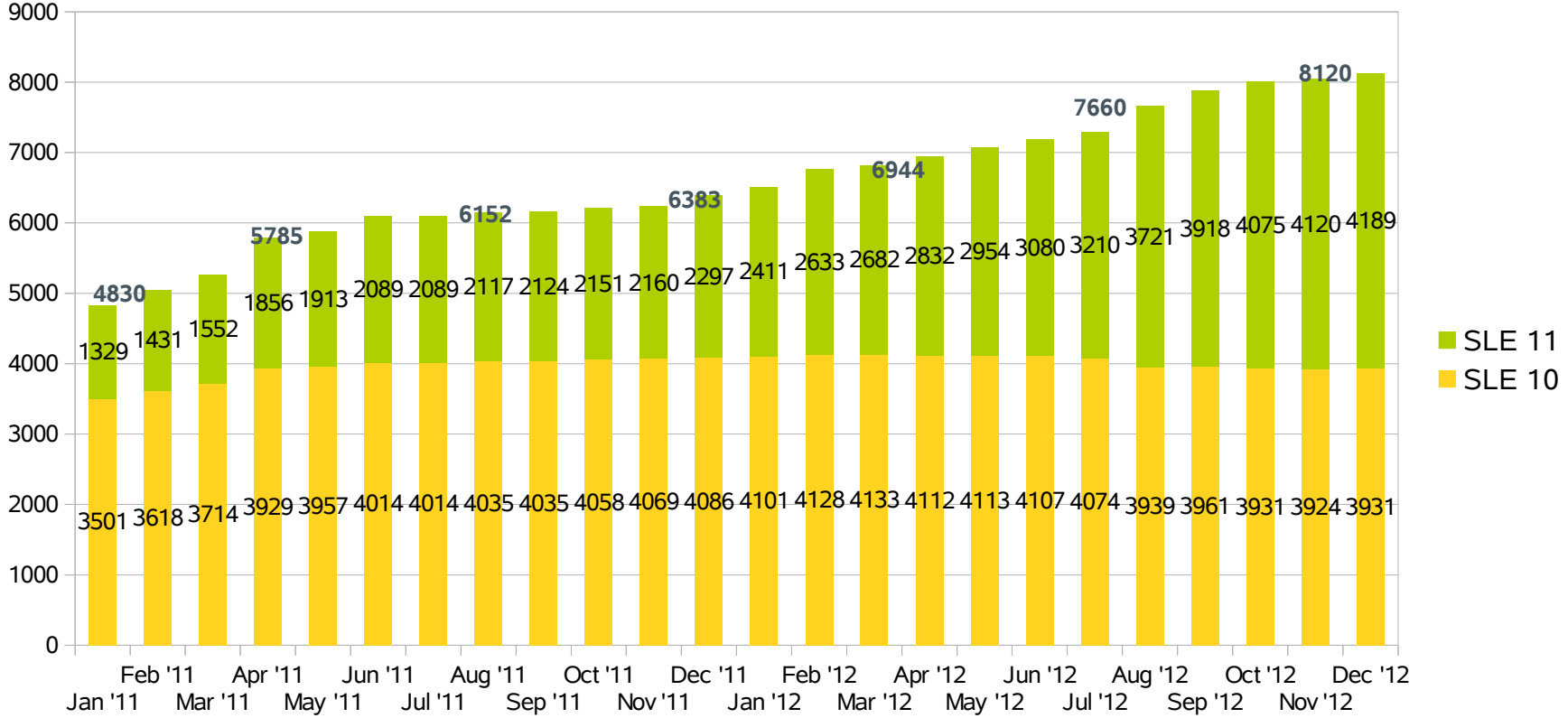
How We Build It



* SUSE Build Service is the internal entity of the Open Build Service

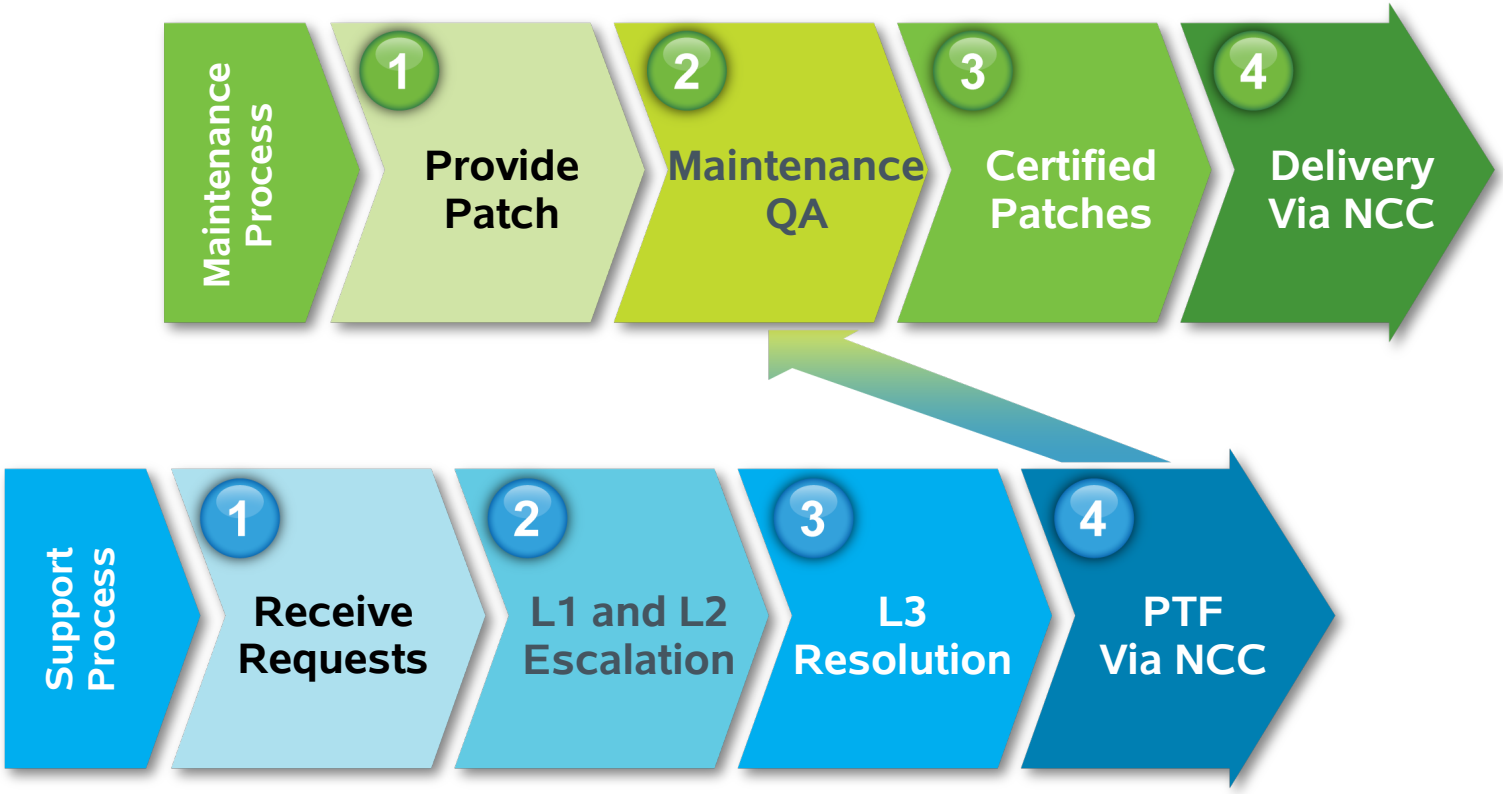
Application Choice

supported for SLE 10/11, Dec 2012

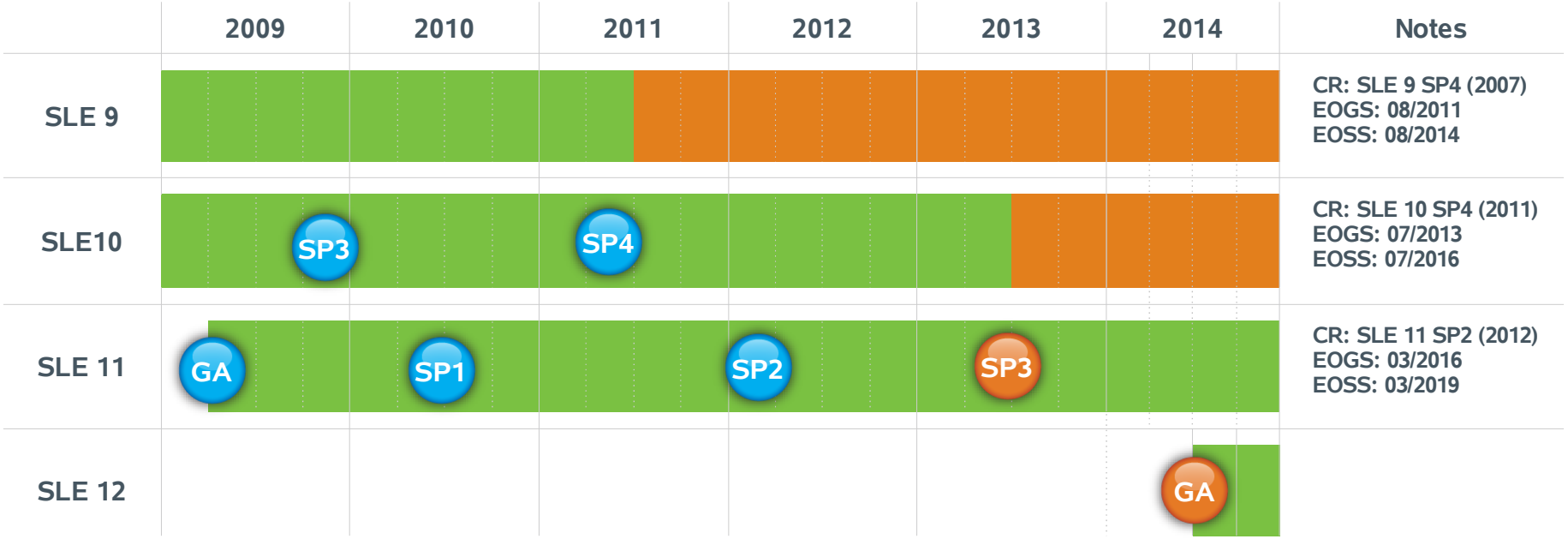


http://www.suse.com/partner/isv/isvcatalog

Maintenance and Support Process

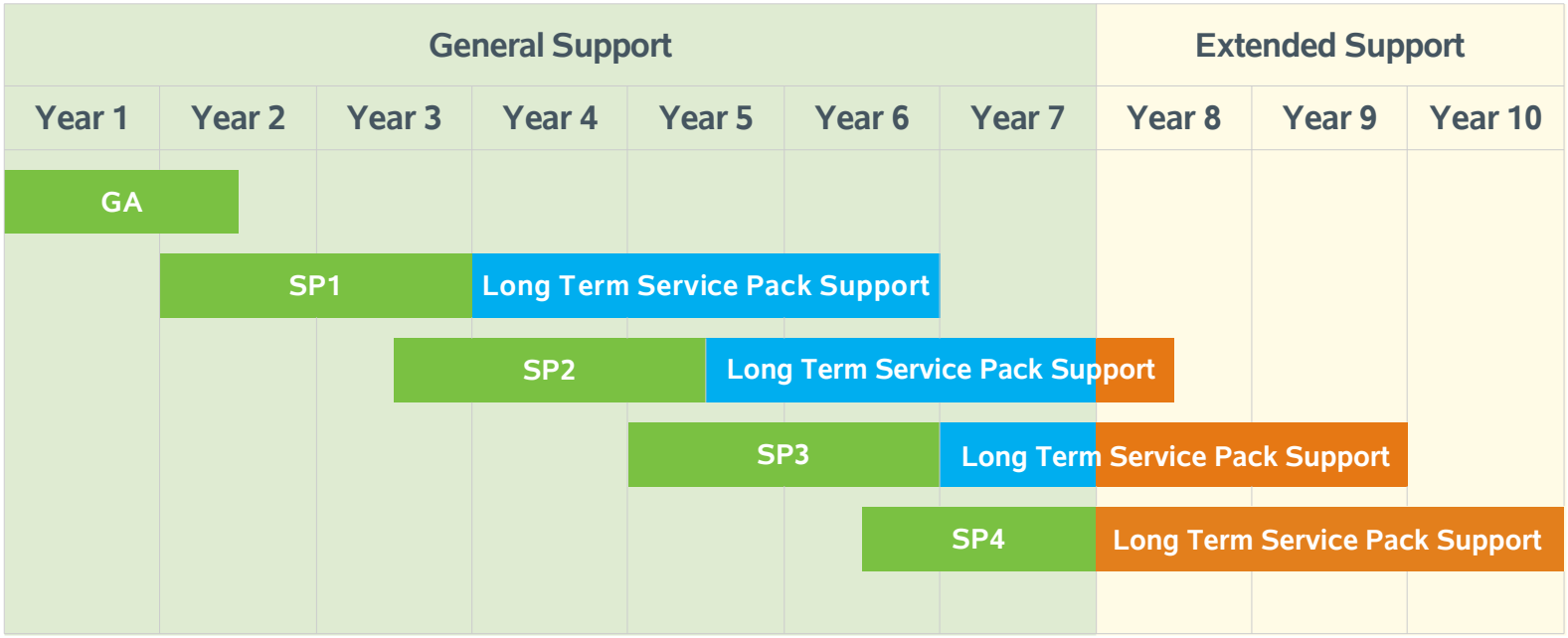


Current SUSE® Linux Enterprise Streams



- Dependable release timing
- Predictability for planning rollouts and migrations
  - Service Pack releases, development and product schedules announced to customers and partners
- Major releases every 4-5 years.

Generic Product Lifecycle



- 10-year lifecycle (7 years general support, 3 years extended support)
- Service Packs are released every ~18 months
  - 5 years lifetime with
  - ~2 years general support per Service Pack
  - 6 month upgrade window after release of the next Service Pack
- Long Term Service Pack Support (LTSS) option
  - Extend upgrade window or extend major release lifecycle

Long Term Service Pack Support (LTSS)

Use Cases

I want to run my software stack unchanged for a very long time

- Updating OS does not improve my business process
- Updates can be very expensive to deploy
- Any change may impose additional risk

I need more time to move to the next Service Pack

- Approval process from stake holders
- QA processes
- Very large and/or distributed environment

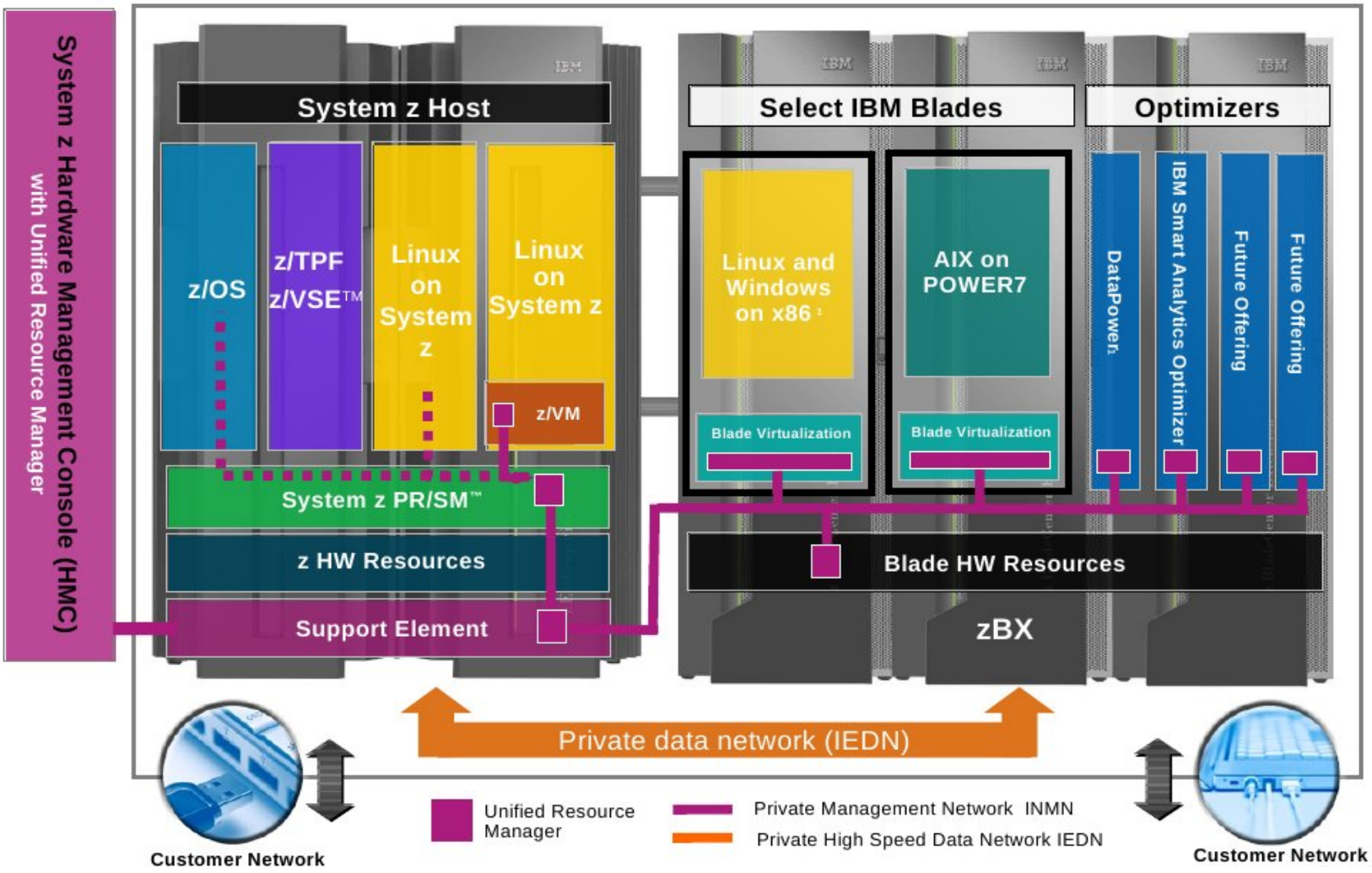
SUSE Linux Enterprise 11 SP2

SUSE Linux Enterprise Server 11 for System z

- Full Dynamic Resource Handling
  - Two levels of virtualizations available: LPAR and z/VM
    - Choose the level of isolation mandated by compliance
    - Flexible resource allocation and reallocation without downtime
  - CPU, memory, I/O hotplug
    - Provide the resource where they are needed in LPAR and z/VM guest
- Abundant memory, IO bandwidth and transaction capability
  - Hipersocket support connects Linux and z/OS applications and data
  - I/O fan out and transaction workload capacity is unmatched
- RAS
  - Detailed I/O device and other performance statistics available
  - Dump generation, handling and inspection tools
  - Centralized and uniform resources support DR recovery setups
  - SUSE Linux Enterprise High Availability Extension included
  - System z specific kernel messages with documentation

SUSE Linux Enterprise Server 11 for System z

IBM zEnterprise System



SUSE Linux Enterprise Server 11 SP2 for System z

- z196 / z114 + zBX = IBM zEnterprise exploitation
  - CPU topology and instruction set exploitation of z196 (SDK)
  - New CHPID support connecting both environments
- Choose the right environment for the right workload
  - ISVs application support might mandate the platform
  - SLES supported for both hardware architectures
- Improved tools and z specific support
  - Disk storage & crypto enhancements
  - Linux RAS support, s390-tools update

SUSE® Linux Enterprise 11 SP2

- Hardware enablement and RAS improvements
- Equivalent or exceeding proprietary Unix capabilities
  - btrfs: file system with “Copy on Write”, checksums, snapshotting, reduce cost of storage management by providing an integration of logical volume management and filesystem, checksums on data and metadata ensure data integrity
  - LXC: container support based on control groups
- Snapshot / rollback for package and configuration updates
  - YaST2 + ZYPP + btrfs
- SUSE Linux Enterprise High Availability Extension: Geo-cluster, automated and pre-configuration
- Unattended upgrade from SUSE Linux Enterprise 10 to SUSE Linux Enterprise 11

Unique Tools Included

- Starter System for System z
  - A pre-built installation server, deployable with z/VM tools
- ***Free*** High Availability Extension
  - Cluster Framework, Cluster FS, DRBD, GEO-cluster*
- AppArmor Security Framework
  - Application confinement
- YaST2 systems management
  - Install, deploy, and configure every aspect of the server
- Subscription Management Tool
  - Subscription and patch management, proxy/mirroring/staging

Kernel 3.0
Selected benefits

- Most recent HW enablement
- Removal of BLK (Big Kernel Lock)
- Control Groups enhancements
  - I/O throttling support for process groups
  - memory cgroup controller
- Integration of AppArmor
- Transparent Huge Pages (THP)
- ...

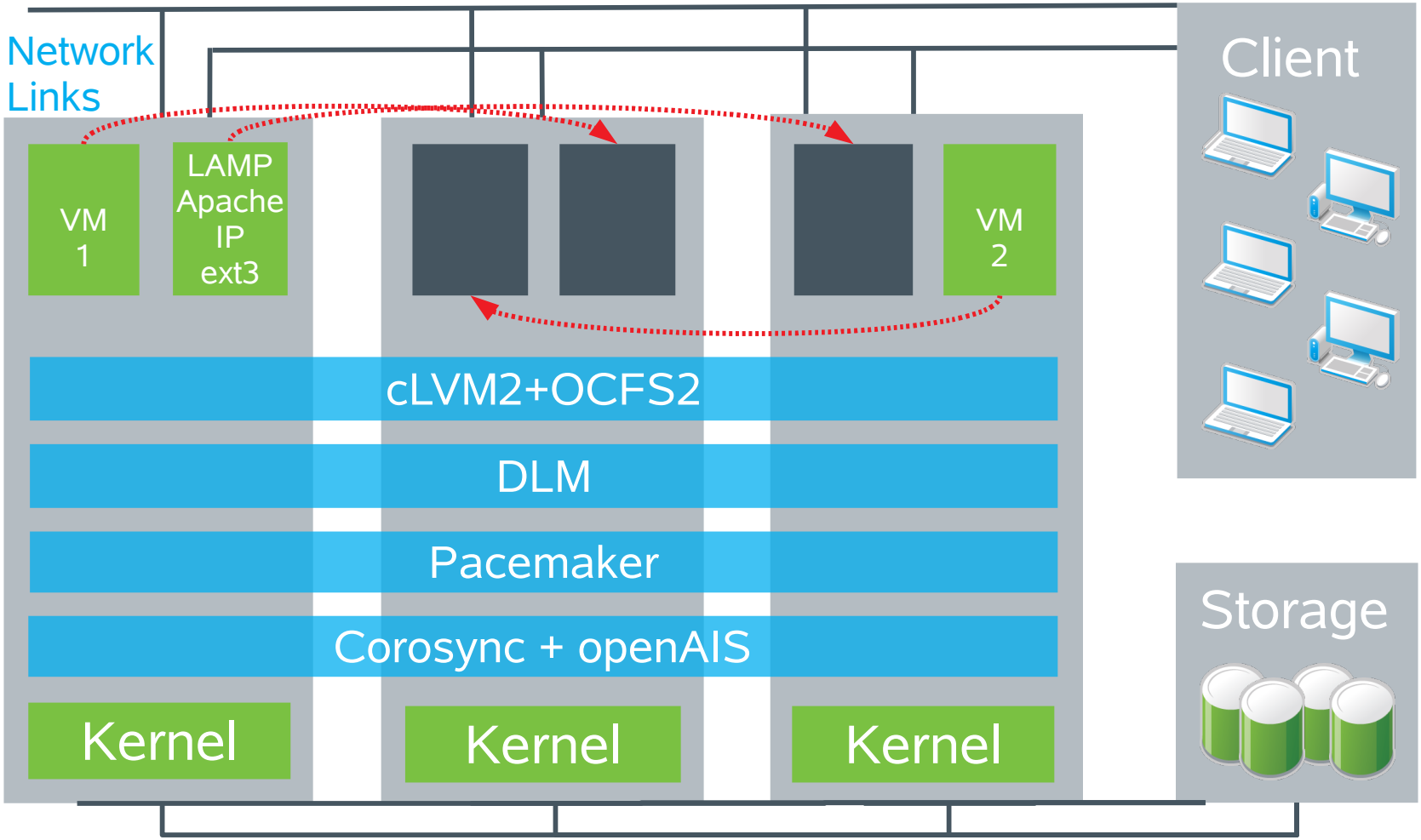
Enhance Your Applications

Examples

- SLE HA: make your workloads High Availability ready
  - Resource agents examples
    - /usr/lib/ocf/resource.d/heartbeat/* → example: Dummy resource agent
  - http://www.opencf.org Standard APIs for clustering functions
- AppArmor: secure your applications
  - Application confinement
  - Easy to use GUI tools with statics analysis and learing-based profile development
  - Create custom policy in hours, not days

Cluster Example

SUSE® Linux Enterprise High Availability Extension



AppArmor: usr.sbin.vsftpd

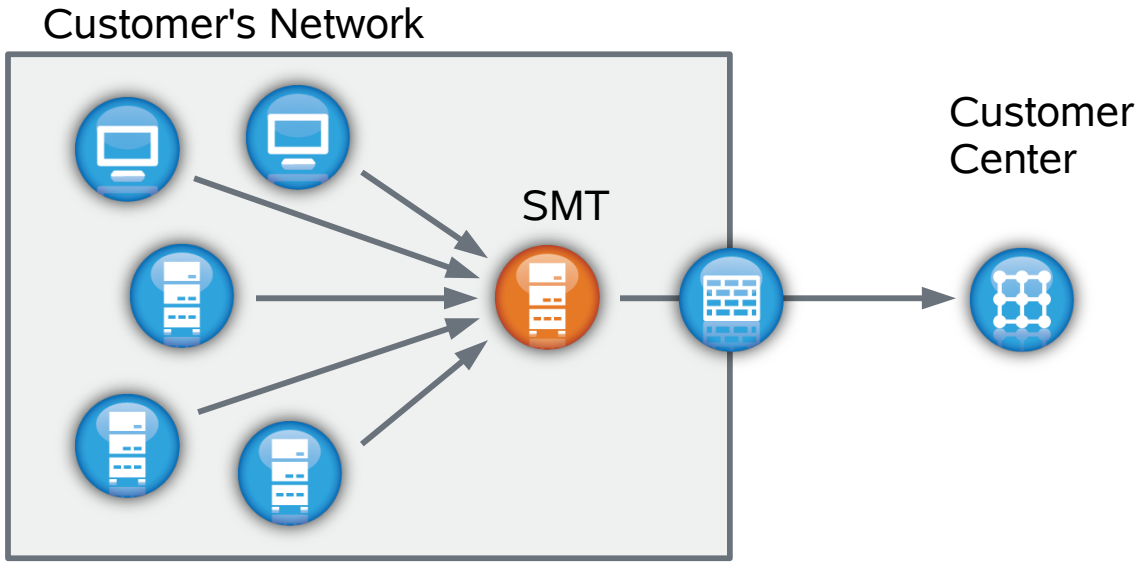
/etc/apparmor/profiles/extras/

```
#include <tunables/global>

/usr/sbin/vsftpd {
  #include <abstractions/base>
  #include <abstractions/nameservice>
  #include <abstractions/authentication>

  /dev/urandom              r,
  /etc/fstab                r,
  /etc/hosts.allow          r,
  /etc/hosts.deny           r,
  /etc/mtab                 r,
  /etc/shells               r,
  /etc/vsftpd.*             r,
  /etc/vsftpd/*             r,
  /usr/sbin/vsftpd          rmix,
  /var/log/vsftpd.log       w,
  /var/log/xferlog          w,
  # anon chroots
  /                         r,
  /pub                      r,
  /pub/**                   r,
  @{HOMEDIRS}               r,
  @{HOME}/**                rwl,
}
```

Subscription Management Tool
Overview

SMT is a proxy and auditing tool that mirrors the Customer Center and tightly integrates with it.

It allows you to accurately register and manage an entire SUSE® Linux Enterprise deployment, guaranteeing the subscription compliance and secure IT process flow organizations require.



Why btrfs?

btrfs (better fs) – Features

- Integrated Volume Management
- Support for Copy on Write
- Powerful Snapshot capabilities
- Scalability (16 EiB) including effective shrink
- Supports offline in-place migration from ext2, ext3
- Other Capabilities:
  - Compression
  - Data integrity (checksums)
  - SSD optimization (TRIM support)

Technology Overview

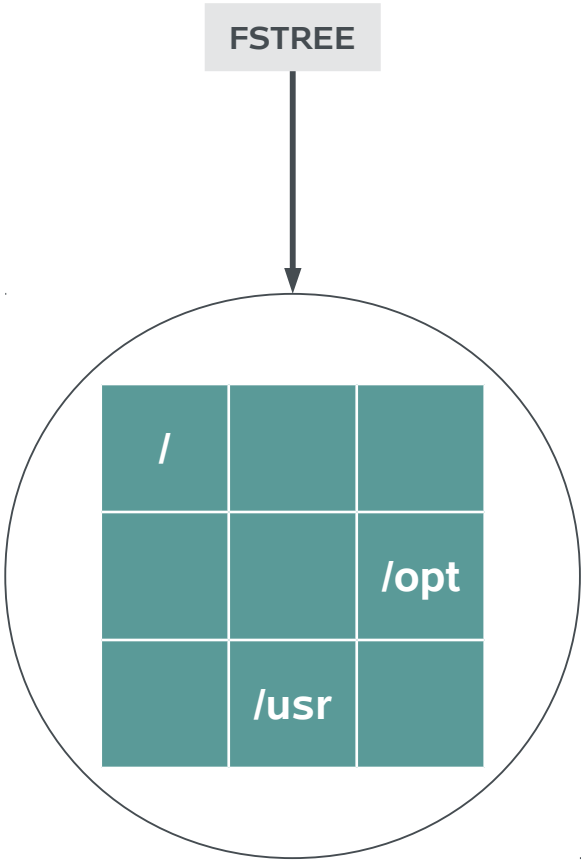
Subvolume (1)

- A complete filesystem tree
- Usually appears as a sub-directory in the “parent” fs
- Can be mounted separately, but not “just a subdirectory”
- Simliar to
  - two “foreign” filesystems, which are
  - using the same pool of data blocks (and other infrastructure)
- Benefits
  - different parts (subvolumes) of a filesystem can have different attributes, such as quotas or snapshotting rules
  - Copy on Write is possible across volumes
- Basic commandline management
  - “btrfs subvolume …”

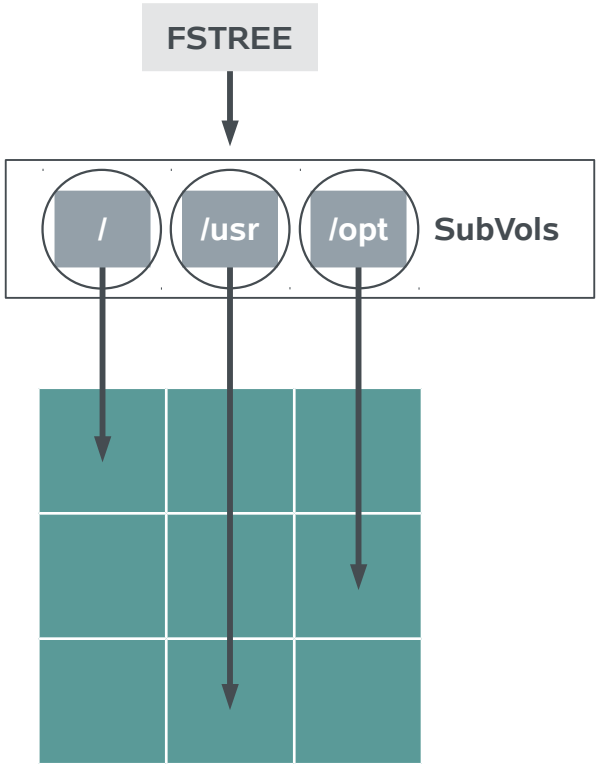
Technology Overview

Subvolume (2)

Normal Filesystem



With Subvolumes



Rollback – per Subvolume

How it works

- Instead of the original subvolume, the snapshot is mounted with the options “subvol=<name>”
  - Remember: snapshots are subvolumes
- Talking about the “/” filesystem, the “subvol” can also be hardcoded using “btrfs subvolume set-default ...”

Benefits

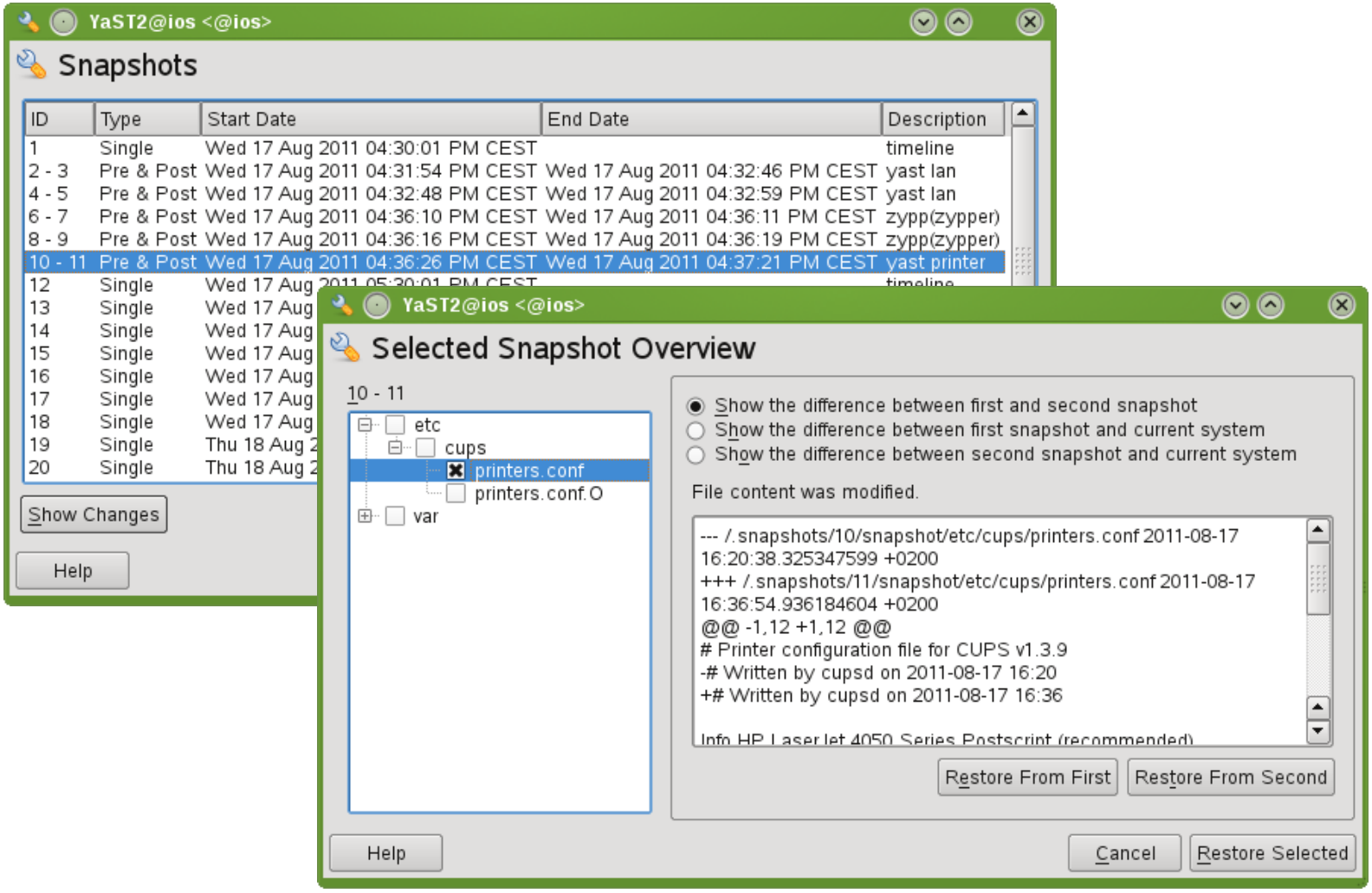
- “atomic” operation
- Very fast

Disadvantages

- Additional complexity
  - May require explicit mounting of subvolumes
- No “rollback” per single file

Snapshots in SUSE® Linux Enterprise 11 SP2

YaST2 Management



cgroups - Resource Control

Consider a large university server with various users - students, professors, system tasks etc. The resource planning for this server could be along the following lines:

CPUs

Top cpuset (20%)

/ \

CPUSet1 CPUSet2

| |

(Profs) (Students)

60% 20%

Memory

Professors = 50%

Students = 30%

System = 20%

Disk I/O

Professors = 50%

Students = 30%

System = 20%

Network I/O

WWW browsing = 20%

/ \

Prof (15%) Students (5%)

Network File System (60%)

Others (20%)

Source: /usr/src/linux/Documentation/cgroups/cgroups.txt

Device Subsystem

Isolation

A system administrator can provide a list of devices that can be accessed by processes under cgroup

- Allow/Deny Rule
- Allow/Deny : READ/WRITE/MKNOD

Limits access to device or file system on a device to only tasks in specified cgroup

Source: http://jp.linuxfoundation.org/jp_uploads/seminar20081119/CgroupMemcgMaster.pdf

cgroups - Memory Subsystem

- For limiting memory usage of user space processes.
- Limit LRU (Least Recently Used) pages
  - Anonymous and file cache
- No limits for kernel memory
  - Maybe in another subsystem if needed
  - Note: cgroups need ~2% of (resident) memory
    - can be disable at boot time with kernel paramenter "cgroup_disable=memory"

Source: http://jp.linuxfoundation.org/jp_uploads/seminar20081119/CgroupMemcgMaster.pdf

Tools / SDK

Developer Tools

Dynamic analysis tools

- valgrind
  - Cachegrind
  - Memcheck
  - Massif
  - Helgrind
  - DRD
  - None
  - Exp-ptrcheck
  - Callgrind
- http://valgrind.org

Tools
cachegrind

- Analysis of cache behaviour of applications
  - z10 cache sizes used as default, changeable (eg. z9, z196)
  - Two cache levels (1st and last level) for instructions & data
  - Writes cachegrind.out.<pid> files

```
r1745045:~ # valgrind --tool=cachegrind ls
==21487== Cachegrind, a cache and branch-prediction profiler
==21487== Copyright (C) 2002-2010, and GNU GPL'd, by Nicholas Nethercote et al.
==21487== Using Valgrind-3.6.1 and LibVEX; rerun with -h for copyright info
==21487== Command: ls
==21487==
--21487-- Warning: Cannot auto-detect cache config on s390x, using one or more defaults
bin  inst-sys  repos  testtools
==21487==
==21487== I   refs:      656,270
==21487== I1  misses:        792
==21487== LLi misses:        656
==21487== I1  miss rate:    0.12%
==21487== LLi miss rate:    0.09%
==21487==
==21487== D   refs:      453,124  (361,066 rd   + 92,058 wr)
==21487== D1  misses:      1,869  (  1,589 rd   +    280 wr)
==21487== LLd misses:      1,313  (  1,061 rd   +    252 wr)
==21487== D1  miss rate:     0.4% (    0.4%     +    0.3%  )
==21487== LLd miss rate:     0.2% (    0.2%     +    0.2%  )
==21487==
==21487== LL refs:         2,661  (  2,381 rd   +    280 wr)
==21487== LL misses:       1,969  (  1,717 rd   +    252 wr)
==21487== LL miss rate:      0.1% (    0.1%     +    0.2%  )
```

zPDT

IBM System z Personal Development Tool

https://www.ibm.com/partnerworld/page/pw_com_zpdt

- zPDT is a software-based application tool
  - Low cost IBM System z platform for ISV application development, testing, demo
  - A virtual System z architecture environment that allows select mainframe operating systems, middleware and software to run unaltered on x86 processor-compatible platforms.
  - Portable System z platform for training & education of applications and operating system environments
  - Supports openSUSE 11+, SLES 11 SP2 x86_64, and others
  - SUSE's evaluation versions for x86_64 and s390x available at http://www.suse.com/products/server/eval.html

Tentative Features for SLES 11 SP3

SUSE® Linux Enterprise Server 11 SP3 for System z

- EC12 + zBX = IBM zEnterprise exploitation continued
  - Update to Java 7 and supportive kernel enhancements
  - GCC 4.7 for applications targeting EC12 processor
  - Cross memory attach APIs for middleware
  - zBX HX5 support
- Improved tools and z specific support
  - 2 stage & network dump storage sharing, plus compression
  - Disk mirroring with real-time enhancement for z
  - Enhanced DASD statistics for PAV & HPF
  - Thin provisioning support (LVM and reference links)
  - s390-tools update, terminal server appliance for z/VM

EC12 Exploitation

- Kernel support to improve Java performance (Transactional Execution)
  - Middleware & applications using Java will benefit
- Backport GCC 4.7.x patches (SDK)
  - Add new instructions to the compiler
  - Added new pipeline description to generate optimal code
- Storage class memory – Flash Express
  - Support new storage device: /dev/scm
  - IPL save 'disk storage' with low latency and high throughput
- Support for Crypto Express 4S cards
- Leverage Cross Memory Attach Functionality
  - Middleware connections

Enhanced Dump Capabilities

- Two Stage Dumper framework
  - More flexible and efficient handling of dumps
- Fuzzy live dump
  - Extract current memory state of th kernel
- Compression of kernel dumps
  - Storage requirement reduction
- Allow to compare dump system with boot system
  - Did the dump occurred on the system it started ?
- Add option to mkdumprd to clean up older initrd's
  - Dump and initrd handling in /boot
- [FICON] DASD: add sanity check to detect path connection error

Misc

- Optimized compression library zlib
  - Enhanced to speed up Java, report generation, backup and installation
- ZYpp transaction auditing
  - Track transaction id also for client side
- libhugetlbfs support
  - Allow applications to benefit from hugetbls w/o recompile
- Enable larger shm segments than 256GB
  - Allows data bases to share larger areas
- Enhanced DASD statistics for PAV and HPF
  - Improved diagnosis and analysis
  - Supports recommendations on the use of eg aliases

Misc

- Data deduplication support (UserLand)
- Tool to display disk using/savings with reflinks
- Add support for thin provisioning to device mapper kernel modules
- Upgrade lvm2 and device-mapper to version that can handle thin provisioning

Tech Preview: KVM for s390x

- Technical Preview Policies
  - Tech Previews are intended to allow early access to evolving technologies, which will potentially be integrated into future product releases or products.
  - They help to understand and get familiar with new technologies and functions.
  - They are sufficiently mature to be used and explored, but might receive significant changes and updates.
  - Problem reports are accepted and handled.
  - Tech Previews are not intended for production (no SLA).

Tech Preview: KVM for s390x

- Kernel Based Virtual Machine
  - KVM (for Kernel-based Virtual Machine) is a virtualization solution for Linux on x86, POWER and z/Architecture hardware containing virtualization extensions
  - It consists of a loadable kernel module, kvm.ko, that provides the core virtualization infrastructure and a processor specific module (eg. kvm-intel.ko or kvm-amd.ko)
  - KVM also requires a modified QEMU to connect to the I/O world of the hosting system.

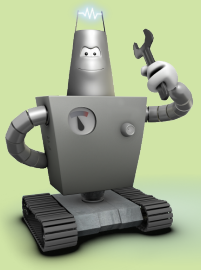
  - **Session 13130: KVM on IBM System z**

Summary


zEnterprise 196
z BladeCenter Extension

How to build a SUSE environment

BUILD
your workloads



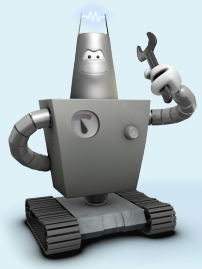
SUSE Studio
Build workloads for any platform and the cloud

MANAGE
your environment

SUSE Manager
Manage Linux workloads across platforms

SUSE® Building Blocks

for the Linux OS Lifecycle



SUSE Studio
Building workloads
for physical and
cloud environments

SUSE Manager
Provisioning
Management
Monitoring

SUSE Linux Enterprise
The foundation for your datacenter workloads
and virtualization, from x86 to the mainframe

SUSE® Linux Enterprise Server for System z

Summary

Available today

- IBM zEnterprise System exploitation
- Enhanced tools and z support
- Choose the right environment for the right workload

Meet us at the booth.

Additional SUSE Sessions at SHARE San Francisco:

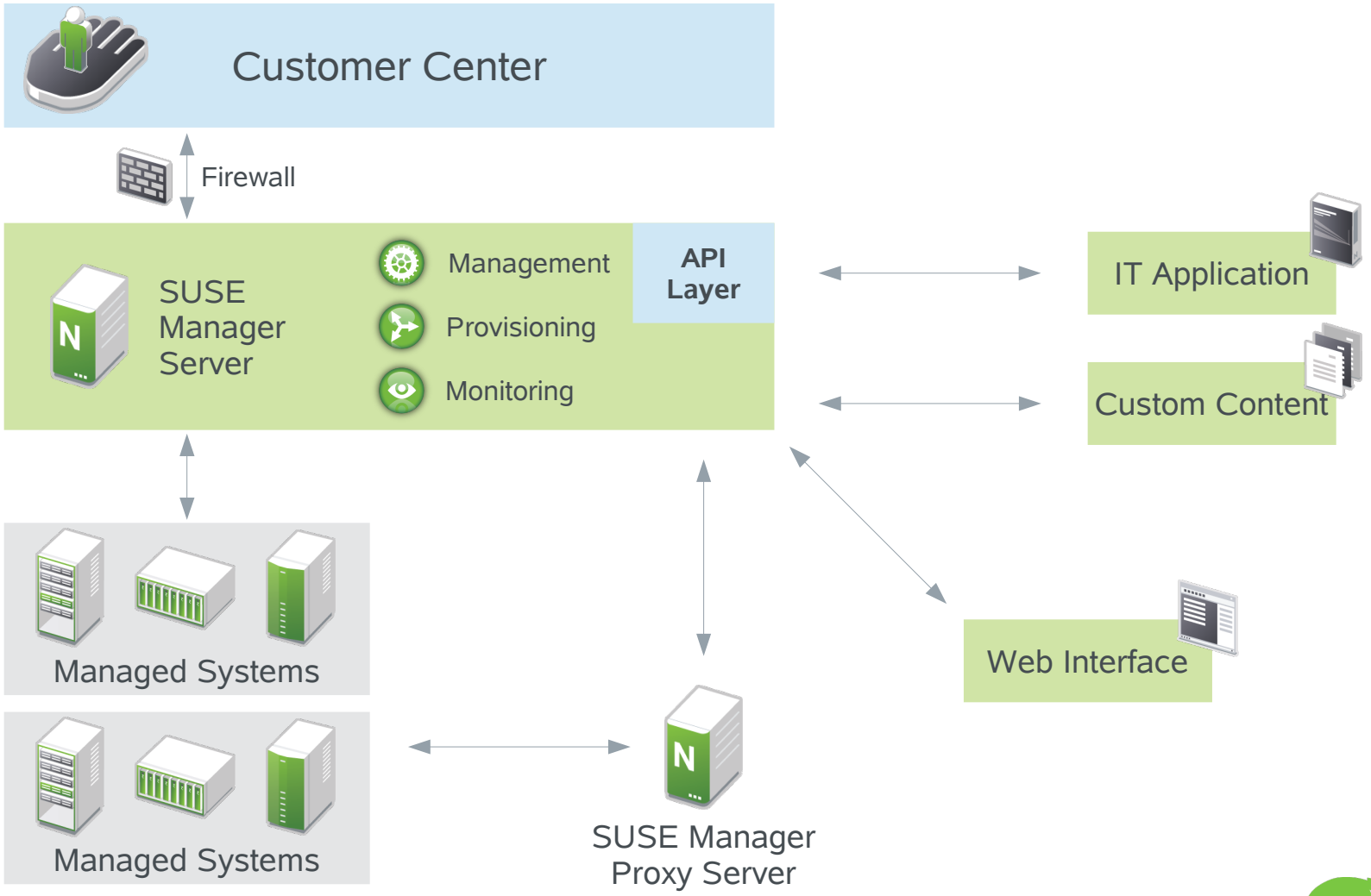
12367: High Availability for Highly Reliable Systems
12876: Using LXC and Btrfs with SLES

Thank you.

SUSE
We adapt. You succeed.

Appendix

How Does SUSE Manager Work?



Specs

SUSE® Linux Enterprise 11 SP2

Kernel Capabilities

SLE 11 SP 2 (3.x)	x86	ia64	x86_64	s390x	ppc64
CPU bits	32	64	64	64	64
max. # logical CPUs	32	up to 4096	up to 4096	64	up to 1024
max. RAM (theoretical/practical)	64/ 16 GiB	1 PiB/ 8+ TiB	64 TiB/ 16TiB	4 TiB/ 256 GiB	1 PiB/ 512 GiB
max. user-/ kernelspace	3/1 GiB	2 EiB/ϕ	128 TiB/ 128 TiB	ϕ/ϕ	2 TiB/ 2 EiB
max. swap space	up to 31 * 64 GB				
max. #processes	1048576				
max. #threads per process	tested with more than 120000; maximum limit depends on memory and other parameters				
max. size per block device	up to 16 TiB	and up to 8 EiB on all 64-bit architectures			

Supported on certified hardware only

SUSE® Linux Enterprise 11 SP2

Filesystems

Feature	Ext 3	reiserfs	XFS	OCFS 2	btrfs
Data/Metadata Journaling	•/•	○/•	○/•	○/•	N/A [3]
Journal internal/external	•/•	•/•	•/•	•/○	N/A
Offline extend/shrink	•/•	•/•	○/○	•/○	•/•
Online extend/shrink	•/○	•/○	•/○	•/○	•/•
Inode-Allocation-Map	table	u. B*-tree	B+-tree	table	B-tree
Sparse Files	•	•	•	•	•
Tail Packing	○	•	○	○	•
Defrag	○	○	•	○	•
ExtAttr / ACLs	•/•	•/•	•/•	•/•	•/•
Quotas	•	•	•	•	○
Dump/Restore	•	○	•	○	○
Blocksize default	4KiB				
max. Filesystemsize [1]	16 TiB	16 TiB	8 EiB	4 PiB	16 EiB
max. Filesize [1]	2 TiB	1 EiB	8 EiB	4 PiB	16 EiB
Support Status	SLES	SLES	SLES	SLE HA	SLES

SUSE® Linux Enterprise was the first enterprise Linux distribution to support journaling filesystems and logical volume managers back in 2000. Today, we have customers running XFS and ReiserFS with more than 8TiB in one filesystem, and the SUSE Linux Enterprise engineering team is using our 3 major Linux journaling filesystems for all their servers. We are excited to add the OCFS2 cluster filesystem to the range of supported filesystems in SUSE Linux Enterprise. For large-scale filesystems, for example for file serving (e.g., with with Samba, NFS, etc.), we recommend using XFS. (In this table "+" means "available/supported"; "-" is "unsupported")

[1] The maximum file size above can be larger than the filesystem's actual size due to usage of sparse blocks. It should also be noted that unless a filesystem comes with large file support (LFS), the maximum file size on a 32-bit system is 2 GB (2^{31} bytes). Currently all of our standard filesystems (including ext3 and ReiserFS) have LFS, which gives a maximum file size of 2^{63} bytes in theory. The numbers given in the above tables assume that the filesystems are using 4 KiB block size. When using different block sizes, the results are different, but 4 KiB reflects the most common standard.

[2] 1024 Bytes = 1 KiB; 1024 KiB = 1 MiB; 1024 MiB = 1 GiB; 1024 GiB = 1 TiB; 1024 TiB = 1 PiB; 1024 PiB = 1 EiB (see also http://physics.nist.gov/cuu/Units/binary.html)

[3] Btrfs is a copy-on-write logging-style file system, so rather than needing to journal changes before writing them in-place, it writes them in a new location, and then links it in. Until the last write, the new changes are not "committed."

[4] Btrfs quotas will operate differently than traditional quotas. The quotas will be per-subvolume rather than operating on the entire filesystem at the user/group level. They can be made functionally equivalent by creating a subvolume per- user or group.

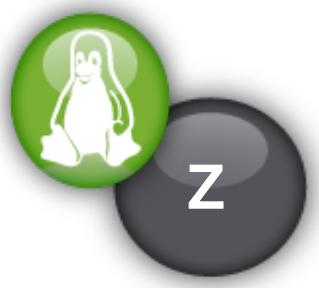
Resources

SUSE® Linux Enterprise

Documentation and Release Notes

- Product Pages
  - http://www.suse.com/products/server/
  - http://www.suse.com/products/sles-for-sap/
  - http://www.suse.com/products/highavailability/
  - http://www.suse.com/products/realtime/
- Unix to Linux Migration
  - http://www.suse.com/solutions/enterprise-linux-servers/unixtolinux.html
- Documentation
  - http://www.suse.com/documentation/
- Release Notes
  - http://www.suse.com/releasenotes/

Resources



- Product website
  www.suse.com/products/systemz
- Customer References
  www.suse.com/success → extended search for SUSE Linux Enterprise Server for System z
- Download SUSE Linux Enterprise Server for System z
  www.suse.com/products/server/eval.html
- Promotion Website
  www.novell.com/products/systemz/els.html
- Partner Website
  www.suse.com/mainframe
- Starter System for System z
  www.suse.com/partner/ibm/mainframe/startersystem.html

Resources

- SUSE Linux Enterprise Server and IBM zEnterprise
  http://www.novell.com/docrep/2010/11/suse_linux_enterprise_server_and_ibm_zenterprise_system.pdf
- zBX entitlement for SUSE Linux Enterprise Server offering
  http://www.suse.com/promo/zbx.html
- SUSE Linux Enterprise Server for System z
  http://www.suse.com/products/systemz/
- IBM zEnterprise Success Story: Sparda-Datenverarbeitung eG
  http://www.novell.com/success/sparda.html
- Chalk Talk: Server consolidation on IBM System z
  http://www.novell.com/media/content/chalktalk-server-consolidation-on-system-z.html
- SUSE Manager
  http://www.suse.com/products/suse-manager
- SUSE Studio
  http://www.susestudio.com

Corporate Headquarters
Maxfeldstrasse 5
90409 Nuremberg
Germany

+49 911 740 53 0 (Worldwide)
www.suse.com

Join us on:
www.opensuse.org