Subject: BrandZ Support for non-native zones

Submitted by: Nils Niewejaar

File: PSARC/2005/471/opinion.ms

Date: July 28th, 2006

Committee: Ed Gould (Opinion by Alan Hargreaves), James D Carlson, Glenn Skinner, William Sommerfeld, Gary Winiger.

Product Approval Committee:
solaris-pac-opinion@sun.com

1. Summary

BrandZ is an extension of the zones infrastructure that allows the creation of zones that emulate non-native operating system environments, such as Linux. Future projects may extend this project to build other non-native operating environments.

2. Decision & Precedence Information

The project is approved as specified in reference [1].

The project may be delivered in a patch release of Solaris.

The project supercedes PSARC/2003/445: Janus: Linux binary compatibility for Solaris x86.

The project depends on PSARC/2006/440: BrandZ-aware Installer and may not be delivered before it.

3. Interfaces

The project exports the following interfaces.

Interfaces Exported		
Interface	Classification	Comments
Linux Interfaces: System calls (structure, semantics and calling conventions) /dev (names and major/minor #'s) /proc signal numbers error numbers	External	This category includes all of the different Linux interfaces that the lx brand emulates.

Interfaces Exported		
Interface	Classification	Comments
AT_SUN_BRAND_BASE AT_SUN_BRAND_LDDATA AT_SUN_BRAND_LDENTRY AT_SUN_BRAND_BRANDNAME AT_SUN_BRAND_PHDR AT_SUN_BRAND_PHENT AT_SUN_BRAND_PHNUM AT_SUN_BRAND_ENTRY	Project Private	Additional AUX vector flags used to convey brand information to the Solaris linker
config.xml	Project Private	Brand definition
platform.xml	Project Private	Virtual platform definition
struct modlbrand	Consolidation Private	kernel/brand module linkage interface
struct brand	Project Private	Kernel/brand operational interface
struct brand_ops	Project Private	Kernel/brand operational interface
struct brand_mach_ops	Project Private	Arch-specific kernel/brand operational interface
struct brand_attr	Project Private	Userspace/kernel interface
struct lx_brand_registration	Project Private	Userspace/kernel interface
rd_helper_ops_t	Consolidation Private	librtld_db.so helper plugin interface
brand_open() brand_close() brand_is_native() brand_get_boot() brand_get_halt() brand_get_initname() brand_get_install() brand_get_modename() brand_get_postclone() brand_get_verify() brand_platform_iter_gmounts() brand_platform_iter_lmounts() brand_platform_iter_devdir() brand_platform_iter_link()	Project Private	libbrand.so.1 is a new library for parsing the BrandZ .xml files
zonecfg_get_brand() zone_get_brand()	Contracted Project Private	Added to libzonecfg.so.1 Contract in reference [4]
zonecfg(1M)	Evolving	Added -B <brand> option
zoneadm(1M)	Project Private	Added -f (force) option to mount and boot commands Added "brand" column to verbose "list" output
zonecfg(1M)	Evolving	Added -B <brand> option
lockd(1M) statd(1M)	Consolidation Private	Added -P option to indicate portportmapper usage
libnsl(3LIB)	Consolidation Private	Add __use_portmapper() to resurrect old portmapper support

Interfaces Exported		
Interface	Classification	Comments
streamio(7I)	Evolving	Add support for TIOCSCTTY, TIOC-NOTTY, TIOCSETLD and TOICGETLD
uucopy(2)	Evolving	Added to libc.so.1 See design doc: 3.5.2
set_setcontext_enforcement(3C)	Consolidation Private	Added to libc.so.1 See design doc 3.6.2
setsigacthandler(3C)	Consolidation Private	Added to libc.so.1 See design doc 3.6.1
lx-install(1M)	Evolving	Invoked by zoneadm(1M), but options are user-visible
lx-syscall(7D)	Evolving	Linux syscall provider
lx_ptm(7D)	Project Private	Linux pty master driver
ldlinux(7M)	Project Private	STREAMS module that provides Linux termio(7I) semantics
lx_afs(7D)	Project Private	Linux automounter support
lx_audio(7D)	Project Private	Layered driver to convert Linux semantics to Solaris

The project imports the following interfaces.

Interfaces Imported		
Interface	Classification	Comments
Linux syscall Interface	External	
rpm2cpio(1M) CLI rpm CLI	External	Used to install RedHat software
Linux statd(1M) and lockd(1M) uid/gid #'s	External	Used to support NFS locking within lx branded zones
glibc ABI gethostbyname_r gethostbyaddr_r getservbyname_r getservbyport_r openlog syslog closelog __progname	External	Used to provide naming services to Solaris statd(1M) and lockd(1M) daemons. See section 3.8 of the design doc.
RHEL 3.x contents	External	/etc files, rc.d scripts, etc. which we modify at install time.
Linux ELF format	External	Object file format for Linux binaries

4. Opinion

4.1. The *lx* Brand

The word *Linux* is a trademark. To avoid issues, the name *lx* is used to reference linux branded zones.

PSARC had concerns about the management of this namespace for future brands and releases of linux. As a result, references to specific releases of the linux kernel were removed from the documentation and the lx brand will not be associated with specific releases of a linux kernel.

4.2. Executable Stacks

For compatibility BrandZ has to allow Linux applications to run with executable stacks, so those applications are vulnerable to any security holes that are opened by those stacks. However, since it is running inside a zone, any damage would be confined to that BrandZ instance. A compromised zone will not be able to bring down the system, and will neither have access to, nor be able to damage applications or data in other zones.

Considered more generally, a BrandZ-hosted linux environment will be subject to any security holes in the Linux user-space. However, it will not be vulnerable to any security holes that depend on kernel support or kernel bugs. This would arguably make a BrandZ-hosted RHEL 3 environment more secure than a native RHEL 3 environment.

4.3. Truss, Apptrace and Dbx

Truss has been updated to recognize the new Solaris system calls; it has not been, and will not be, updated to understand and display the Linux system calls issued by the application. An lx-syscall DTrace provider makes that information available.

dbx does not currently work, but this appears to be a bug rather than a fundamental limitation of the design. This is still under investigation and is being tracked as:

> 6445248 dbx cannot grok Linux processes

4.4. Live Upgrade and Packaging Tools

Live upgrade doesn't run with zones. The packaging tools will go into the install gate

> 63242179 packaging tools need to be brand aware

has been filed and links to this case.

PSARC/2006/440 has been submitted and approved for working with live upgrade.

4.5. Audio

There is no notion of a device-specific attribute, which is needed to support systems with multiple audio devices, in the zone's infrastructure now. Adding such a capability would have required an extensive overhaul of how devices are configured and managed.

Rather than redesign the core of the zones configuration tools simply to solve one Linux corner case, the project team chose to use the generic attributes mechanism to support audio devices.

4.6. Solaris Trusted Extensions

After discussions between the project teams for this case and for the Trusted Extensions, it was determined that lx branded zones will not be supported on trusted systems where labels are active.

4.7. lxrun

PSARC/2006/441 has been submitted and approved to EOL *lxrun*.

4.8. Process Auditing

Processes running in an lx-branded zone do not have their Linux system calls audited. Otherwise, they are subject to all the standard auditing. For example, Linux process creation/exit events are captured as for any other process. The Solaris system calls that the brand library uses to emulate the Linux system calls are subject to auditing.

The only restriction is that the Solaris audit processing tools cannot run inside the Linux zone, so the audit records must be consumed by tools running in the global zone.

4.9. Signals to init

During inception, PSARC expressed a concern about how the lx init would deal with system generated signals that it was not expecting. The project team has addressed these concerns as follows.

With standard Solaris zones, the kernel and init are in agreement on how to handle the death of init: the kernel restarts the process, and the resurrected init process uses a state file to pick up where its predecessor left off.

The Linux init is not prepared to handle this kind of restart. When it is restarted, it works its way through the entire boot process again. This means that all the rc.d scripts are rerun, and we end up with multiple instances of services like crond, syslogd, and so on.

Since it cannot simply ignore SIGSEGV, and since the Linux init is not prepared to handle a warm restart, the only action that will deliver a sensible result is to reboot the zone. Regardless of whether this is the expected behavior on a native Linux system, it’s the behavior that will be implemented inside a Linux zone.

4.10. Delegated Administration of Solaris-specific capabilities

Linux-branded zones will always be second-class citizens in many ways. As our real goal is to increase Solaris adoption, using BrandZ as one part of a migration strategy, we view this as a feature rather than a bug.

To address these specific issues: ZFS delegation will not work within a Linux zone. Given sufficient customer interest, we could possibly support the ZFS utilities, but it would take a significant amount of engineering work, and would violate our "one binary type per zone" model. It should be noted that this in no way affects being able to install and run a Linux zone on a ZFS filesystem.

Supporting network delegation is significantly more feasible. By emulating the ioctl()s needed to perform network configuration tasks, we should be able to support network delegation using Linux configuration tools. This would not be a trivial engineering effort, but it would certainly fit within the overall BrandZ model.

4.11. Impact on Zones Upgrade

The zones test suite, which is run as a regular part of the PIT suite, will be extended to include testing of lx-branded zones.

5. Minority Opinion(s)

None.

6. Advisory Information

None.

7. Appendices

7.1. Appendix A: Technical Changes Required

None.

7.2. Appendix B: Technical Changes Advised

None.

7.3. Appendix C: Reference Material

Unless stated otherwise, path names are relative to the case directory PSARC/2005/471.

1. **Specification**
   File: final.materials/design.pdf
   File: committment.materials/onepager
   File: committment.materials/what_works
2. **20 Questions**
   File: final.materials/20_questions
3. **Man Pages**
   File: committment.materials/brand.dtd.1
   File: committment.materials/brands.5
   File: committment.materials/design.pdf
   File: committment.materials/lx.5
   File: committment.materials/zone_platform.dtd.1
   File: committment.materials/zoneadm.1m
   File: committment.materials/zonecfg.1m
   File: committment.materials/zones.5
4. **Contract between Solaris Core Technologies and Solaris Install**
   File: contract-01

