N1 Grid Containers: Server Consolidation Made Easy

Sun Microsystems, Inc.

Server Consolidation Goals

- Reduce costs by running multiple workloads on same system
  - Better hardware utilization
  - Reduced infrastructure overhead
  - Lower administration costs (admins/workload)
- Requires support from system
  - Resource controls
  - Security isolation
  - Failure containment
  - Delegated administrative control

N1 Grid Containers

- Basic concept: isolated execution environment *within* a Solaris instance
- Includes resource, security, failure isolation
- Lightweight, flexible, efficient
- One OS to manage
- Components:
  - Resource management (CPU, memory, ...)
  - Security/namespace isolation (*zones*)

Quick Summary

- Containers look like different Solaris instances, but aren't
- Can improve system security
- Isolates applications from each other
- Underlying platform details hidden
- Provides almost arbitrary granularity in isolating and sharing resources
- Application environment is compatible for existing programs

Example Uses

- Data center workload consolidation
- Hostile or untrusted applications
- Hosting environments
- WAN-facing services
  - Break-in containment
- Software development
  - Test vs. Production

Solaris Resource Management

- RM Features:
  - Fair-share scheduler
  - Resource pools
  - Extended accounting
- Now bundled in Solaris
- Based on concept of *project*
  - Basic workload classifier
  - Configuration info can be stored in NIS or LDAP
  - Tools for dynamic project assignment and control

RM: Fair-Share Scheduler

- Controls allocation of CPU cycles
- Each project allocated "shares" of CPU
  - Actual allocation dependent on what else is running
  - Ensures minimum level of service (*entitlement*)
- Migration tools available for older SRM deployments

RM: Resource Pools

- Persistent, named sets of resources
  - CPUs, physical memory*, swap space*
- Partitions resources among consumers
- Automatic assignment of projects to pools
- Dynamic resource assignment in response to events

* Planned for Solaris 10 update release

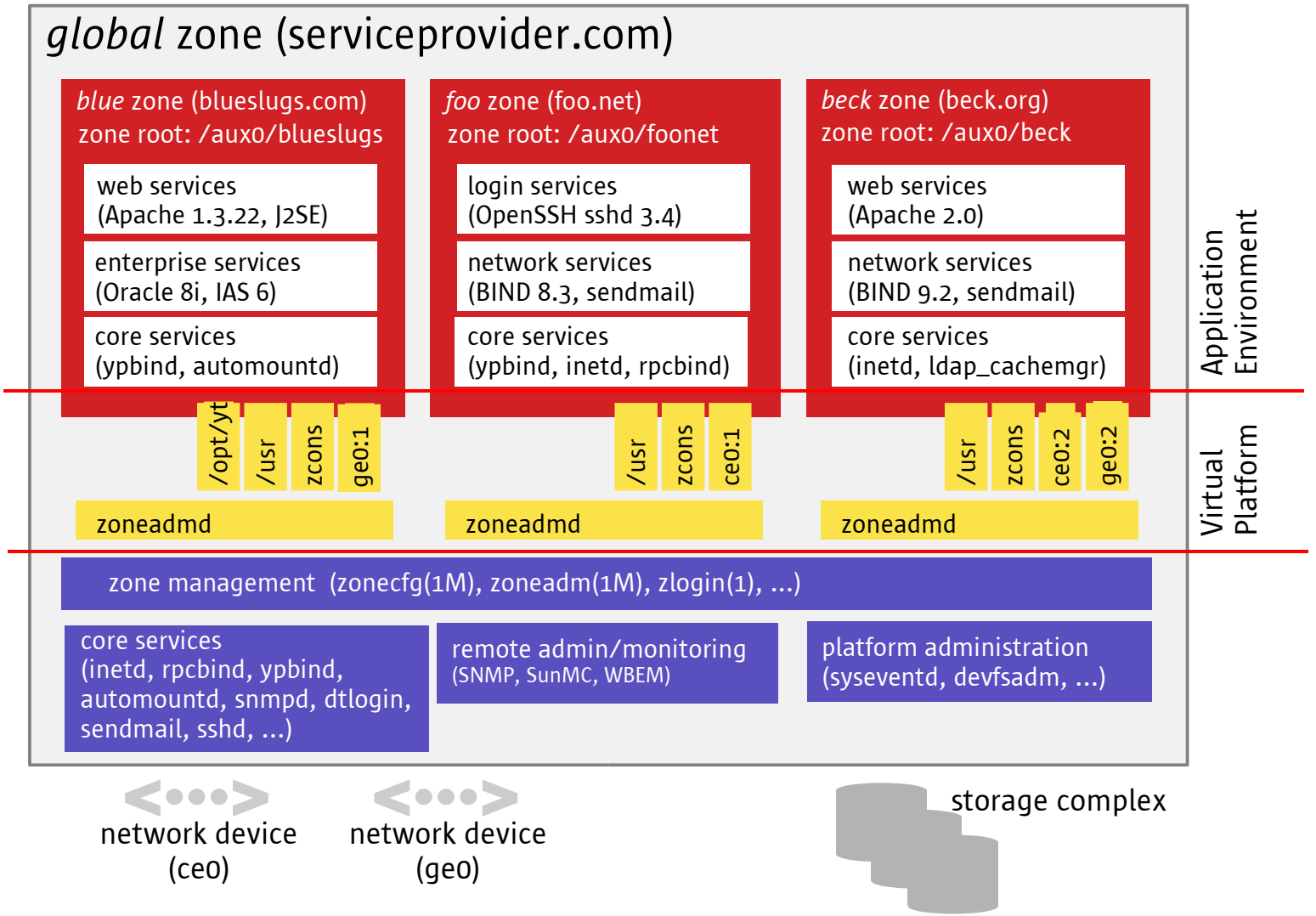
RM: Extended Accounting

- Aggregated accounting records of system activity
- Incorporated into higher level accounting/billing/capacity planning packages
  - Teamquest, Instrumental

Solaris Zones

- Virtualizes OS layer: file system, devices, network, processes
- Secure boundary around instance
- Provides:
  - Privacy: can't see outside zone
  - Security: can't affect activity outside zone
  - Failure isolation: application failure in one zone doesn't affect others
- Lightweight, granular, efficient
- Complements resource management

Zones Block Diagram



Zones: Security

- Root in a zone can't be trusted
  - Many operations requiring root disabled
  - Exceptions: file operations, binding to reserved ports, other "local" operations
  - No way to increase root privileges within zone
- Access limited to resources assigned to zone

Zones: Processes

- Process ID namespace is partitioned
- Processes in the same zone interact as usual
- Processes may not see or interact with processes in other zones
- **`proc(4)`** only provides information about processes in the zone

Zones: File Systems

- Each zone allocated part of file system hierarchy
- One zone can't see another zone's data
- Loopback mounts allow sharing of read-only data (e.g., /usr)
- Can't escape (unlike chroot)
- Zone admin can mount filesystems within zone (NFS, autofs, tmpfs, etc.)

Zones: Networking

- Assign set of IP addresses to each zone
  - Per-zone virtual interfaces multiplexed over physical interfaces
- Processes can't bind to addresses not assigned to their zone
  - INADDR_ANY mapped to local set
- Allows multiple services to bind to same port in different zones

Zones: Devices

- Logical (pseudo) devices within zone
  - Access storage through file system
  - /dev/null, /dev/zero, /dev/random, etc. safe
  - /dev/tcp, /dev/log are "virtualized"
- Some pseudo devices disallowed
  - /dev/kmem, ...
- Can also allow access to physical devices (e.g., tape drives)
  - But be careful of shared HW (adapters, buses, etc.)

Zones: Identity

- Each zone has own hostname, domain, etc.
- Name service can be separately administered
  - Needed to support different administrative domains, ensure data is kept private

    "Give customers their own root password"
  - User ids may have different meanings in different zones

Zones: Interprocess Communication

- Usual IPC mechanisms (System V, pipes, STREAMS, sockets, doors, loopback transport) work within zone
  - Key namespaces are per-zone
- Cross-zone communication only via network interfaces
  - Except with global zone participation
  - Network traffic looped back through IP

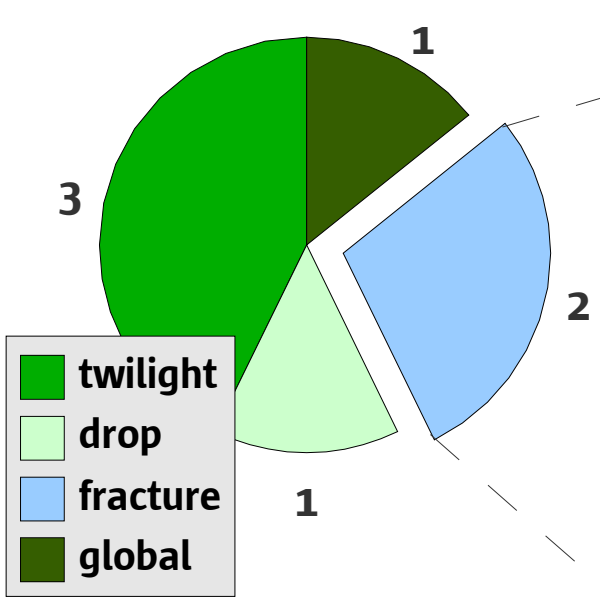
Zones: Installation

- **zoneadm(1M)** utility constructs “clean” zone image from global zone
  - Resets config files to out-of-the-box state
  - Skips files that only make sense in global zone
  - Default is to share **/usr**
- Single operation to roll out patches and upgrades across all zones
  - Need to keep in sync due to kernel dependencies
- Packages can be installed in all zones, or just one

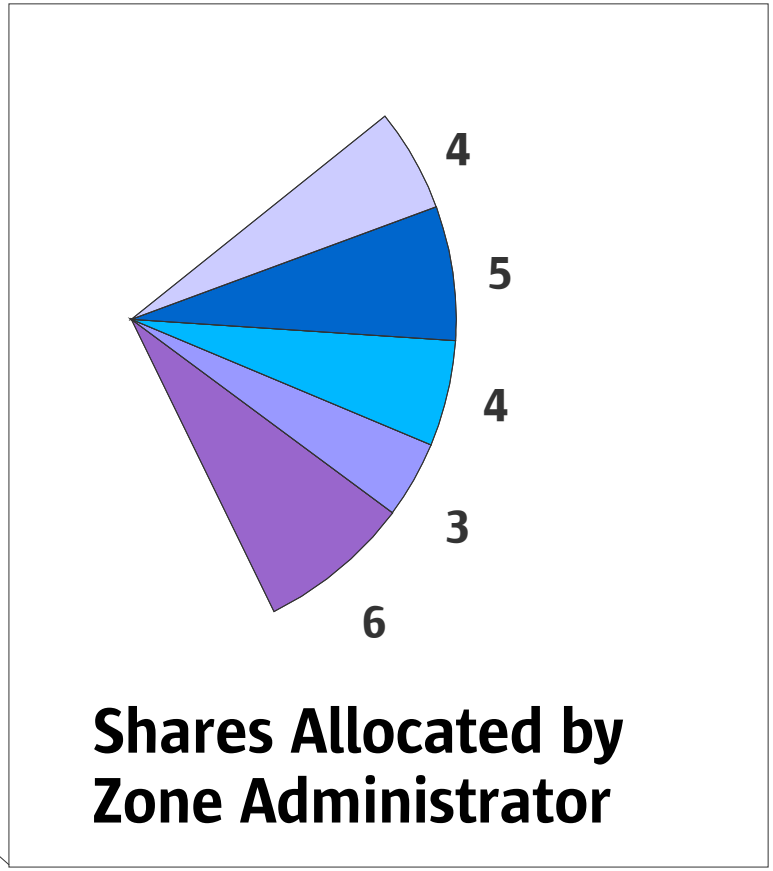
Zones and Resource Management

- Complementary technologies
- 2-level fair-share CPU scheduler
  - Per-*zone* shares configured in global zone
  - Per-*project* shares configured within zone
- Binding from zone to resource pool
  - 1 zone ⇒ 1 pool
  - n zones ⇒ 1 pool
- Per-zone resource limits

Two Level FSS

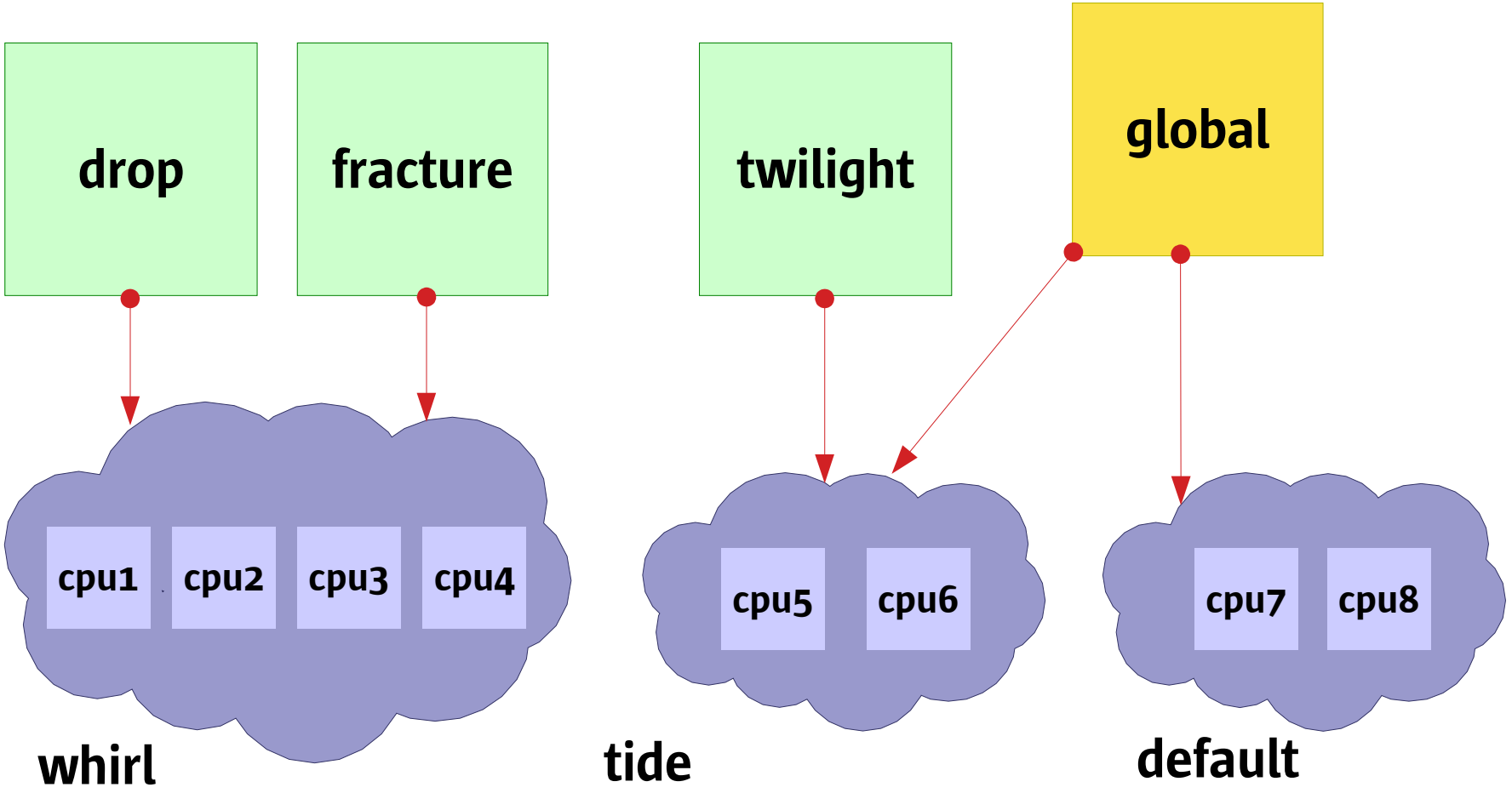


Shares Allocated to Zones



Shares Allocated by Zone Administrator

Zones and Pools



Zones and Fault Isolation

- Zone represents failure boundary for applications
  - Can't affect other apps
  - Per-zone core file configuration
  - Zone “reboot” cleans up application environment (System V IPC, file systems, etc.)
- Can also limit effect of hardware faults
  - If fault affects only application within zone, reboot zone rather than entire system

Zones: Examples

```
d-mpk17-86-237 # zoneadm info -v
     ZID ZONENAME          NODENAME          ROOT
       0 global            d-mpk17-86-237    /
       2 zooropa           zooropa           /export/home/zooropa
       1 kokakola          kokakola          /export/home/kokakola
d-mpk17-86-237 # zlogin zooropa w
  1:12pm  up  1:11,  1 user,  load average: 3.30, 3.67, 2.73
User     tty           login@  idle   JCPU   PCPU  what
comay    pts/5        12:07pm    17     41     41  java_vm
d-mpk17-86-237 # zlogin kokakola df -hl
Filesystem             size   used  avail capacity  Mounted on
/                      9.1G    94M   8.9G     2%    /
/export/home           7.0G   342M   6.6G     5%    /export/home
fd                       0K     0K     0K     0%    /dev/fd
/opt                   3.9G   2.4G   1.5G    62%    /opt
/sbin                  3.9G   2.4G   1.5G    62%    /sbin
swap                   1.1G   8.4M   1.1G     1%    /tmp
swap                   1.1G    56K   1.1G     1%    /var/run
/usr                   3.9G   2.4G   1.5G    62%    /usr
mnttab                   0K     0K     0K     0%    /etc/mnttab
proc                     0K     0K     0K     0%    /proc
d-mpk17-86-237 #
```

Zones: Examples

```
zooropa $ uname -a
SunOS zooropa 5.10 kevlar-myclone sun4u sparc SUNW,Sun-Blade-100
zooropa $ ls /dev
arp         icmp6       lo1         null        sad         systty      tty
conslog     ip          lo2         poll        stderr      tcp         udp
console     ip6         lo3         ptmx        stdin       tcp6        udp6
dsk         kstat       log         pts         stdout      ticlts      urandom
fd          ksyms       logindmux   random      syscon      ticots      zero
icmp        lo0         msglog      rdsk        sysmsg      ticotsord
zooropa $ ifconfig -a
lo0:2: flags=1000849<UP,LOOPBACK,RUNNING,MULTICAST,IPv4> mtu 8232 index 1
        inet 127.0.0.1 netmask ff000000
eri0:2: flags=1000843<UP,BROADCAST,RUNNING,MULTICAST,IPv4> mtu 1500 index 2
        inet 129.146.86.231 netmask ffffff00 broadcast 129.146.86.255
lo0:2: flags=2000849<UP,LOOPBACK,RUNNING,MULTICAST,IPv6> mtu 8252 index 1
        inet6 ::1/128
eri0:4: flags=2000841<UP,RUNNING,MULTICAST,IPv6> mtu 1500 index 2
        inet6 fe80::8192:56d3:2/10
eri0:5: flags=2000841<UP,RUNNING,MULTICAST,IPv6> mtu 1500 index 2
        inet6 2002:8192:56bb:9256:0:8192:56d3:2/64
zooropa $
```

Zones: Examples

```
d-mpk17-86-237 # zlogin -C kokakola

INIT: New run level: 6
System services are now being stopped.
umount: /home busy
nfs umount: /home/comay: is busy


SunOS Release 5.10 Version kevlar-myclone 64-bit
Copyright 1983-2002 Sun Microsystems, Inc.  All rights reserved.
Use is subject to license terms.
NIS domain name is it.sfbay.sun.com
starting rpc services: rpcbind keyserv ypbind done.
syslog service starting.
The system is ready.

kokakola console login:
```

Zones: Examples

```
   PID USERNAME  SIZE   RSS STATE   PRI NICE      TIME  CPU PROCESS/NLWP
108987 comay      58M   32M run      59    0   0:00:12  10% java_vm/25
108694 comay    7096K 6624K sleep    59    0   0:00:04 3.1% make/1
108789 comay      38M   29M sleep    59    0   0:00:16 2.5% mozilla-bin/7
109528 comay      10M 7168K run      31    0   0:00:00 2.3% cg/1
109534 comay      10M 6728K run      31    0   0:00:00 1.6% cg/1
109535 comay    2784K 2384K run      31    0   0:00:00 1.5% ctfconvert/1
109536 comay    7360K 3696K run      39    0   0:00:00 0.4% iropt/1
100722 comay    4680K 3288K cpu0     59    0   0:00:12 0.2% prstat/1
109524 comay    1184K 1040K sleep    59    0   0:00:00 0.1% cc/1
100414 root     2072K  568K sleep   100    -   0:00:02 0.1% xntpd/1
109531 comay    1184K 1040K sleep    49    0   0:00:00 0.1% cc/1
109518 comay    1184K 1040K sleep    59    0   0:00:00 0.1% cc/1
109530 comay    1064K  816K sleep    59    0   0:00:00 0.1% sh/1
109529 comay    7008K 1040K sleep    59    0   0:00:00 0.1% make/1
109523 comay    1064K  816K sleep    59    0   0:00:00 0.1% sh/1
ZONEID    NPROC  SIZE   RSS MEMORY      TIME  CPU ZONE
     2       27  160M   83M     17%   0:00:31  13% zooropa
     1       48  148M   65M     13%   0:00:08 9.9% kokakola
     0       45  113M   18M    3.7%   0:00:18 0.5% global

Total: 120 processes, 311 lwps, load averages: 4.45, 2.66, 1.26
```

For More Information

- RM features available in Solaris 9
- Zones available *today* through Solaris Express
- Documentation on docs.sun.com
- Active discussion forum on BigAdmin

  http://www.sun.com/bigadmin/content/zones
- More coming...

N1 Grid Containers: Server Consolidation Made Easy

andrew.tucker@sun.com
ozgur.leonard@sun.com