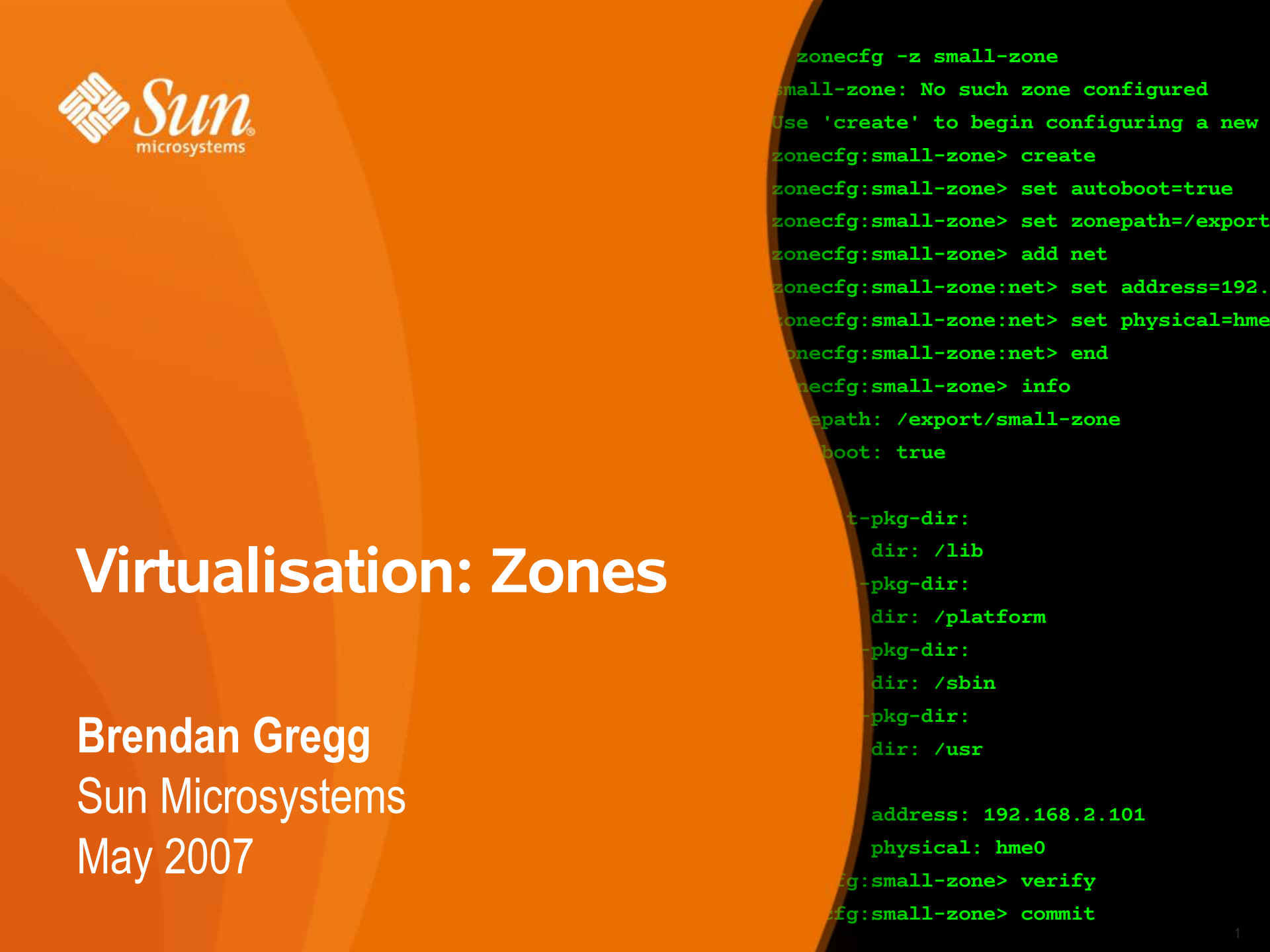


Virtualisation: Zones

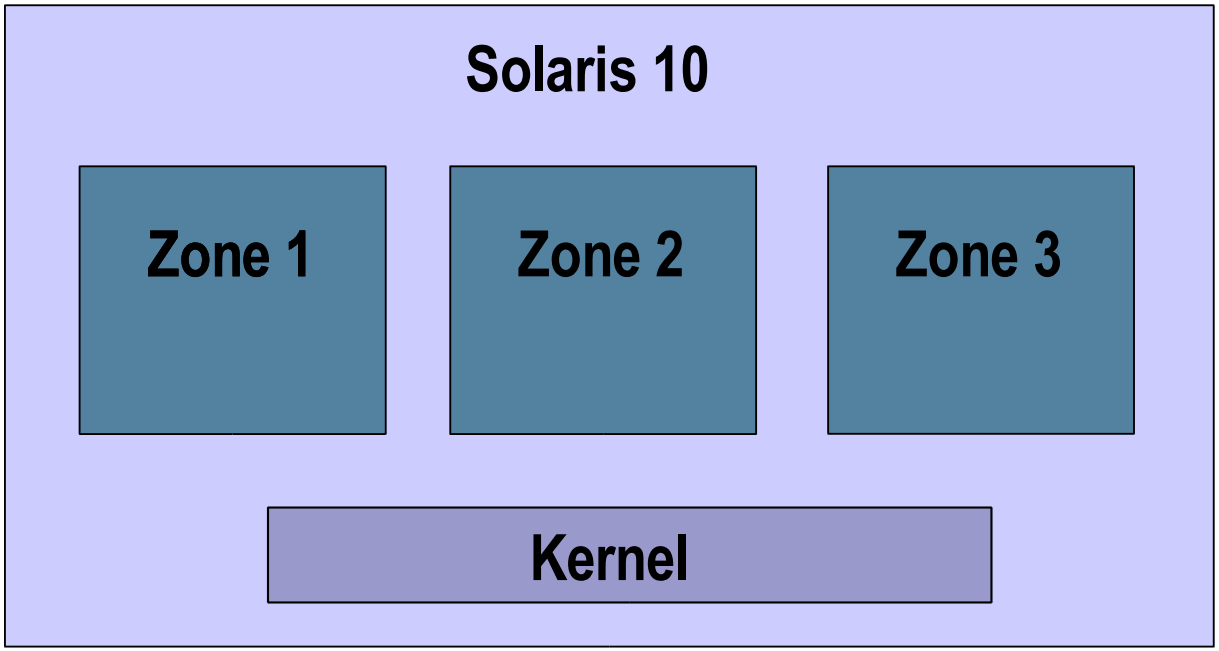
Brendan Gregg
Sun Microsystems
May 2007

Virtualisation: Zones

- This presentation is about Solaris 10 Zones and Containers.
- These slides cover:
  - What are Zones? Containers?
  - Zone Features
  - Zone Types
  - Maintenance
  - Security
  - Resource Management
  - Monitoring

What are Zones?

- Virtual instance of Solaris
- Software Partition of the OS
- A virtualisation solution (along with LDoms, Xen, ...)



Zone Features

- Great Performance
- Easy Administration
- Resource Controls
- Observability
- Security
- Low on-disk footprint
- Supported since Solaris 10 3/05

Not Zone Features

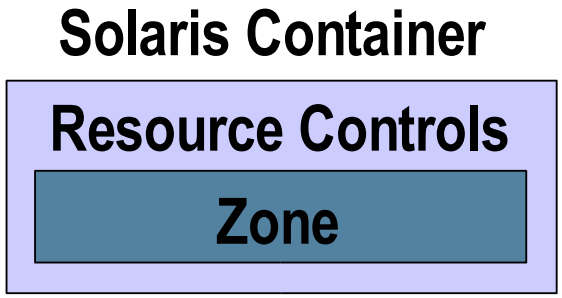
- Since there is only one kernel, the following *cannot* currently be achieved using Zones:
  - Zones for testing kernel patches
    - There are no separate “test kernels” to try patches on
  - Zones for different OSes and Solaris versions
    - BrandZ for creating Linux zones is one exception (so far)

What is best: Zones or VM?

- Performance: Zones
  - No doubling of syscall and kernel overheads
- Observability: Zones
  - Sysadmins can see inside all zones at once
- Security: Zones
  - Read-only /usr by default, and secure monitoring
- Administration: Zones
  - Zones have easy and fast creation/destruction
- Different OSes: VM
  - There is BrandZ for Zones; but can't do different kernels

What are Containers

- Zones + Resource Controls
- Guide to History,
  - > 1998 - Sun creates Solaris Resource Manager (SRM) as a software package
  - > 2002 - SRM features added to Solaris 9, and additional features added to Solaris 9 updates
  - > 2005 - Resource Control features applied to Solaris 10 Zones, then improved in Solaris 10 updates

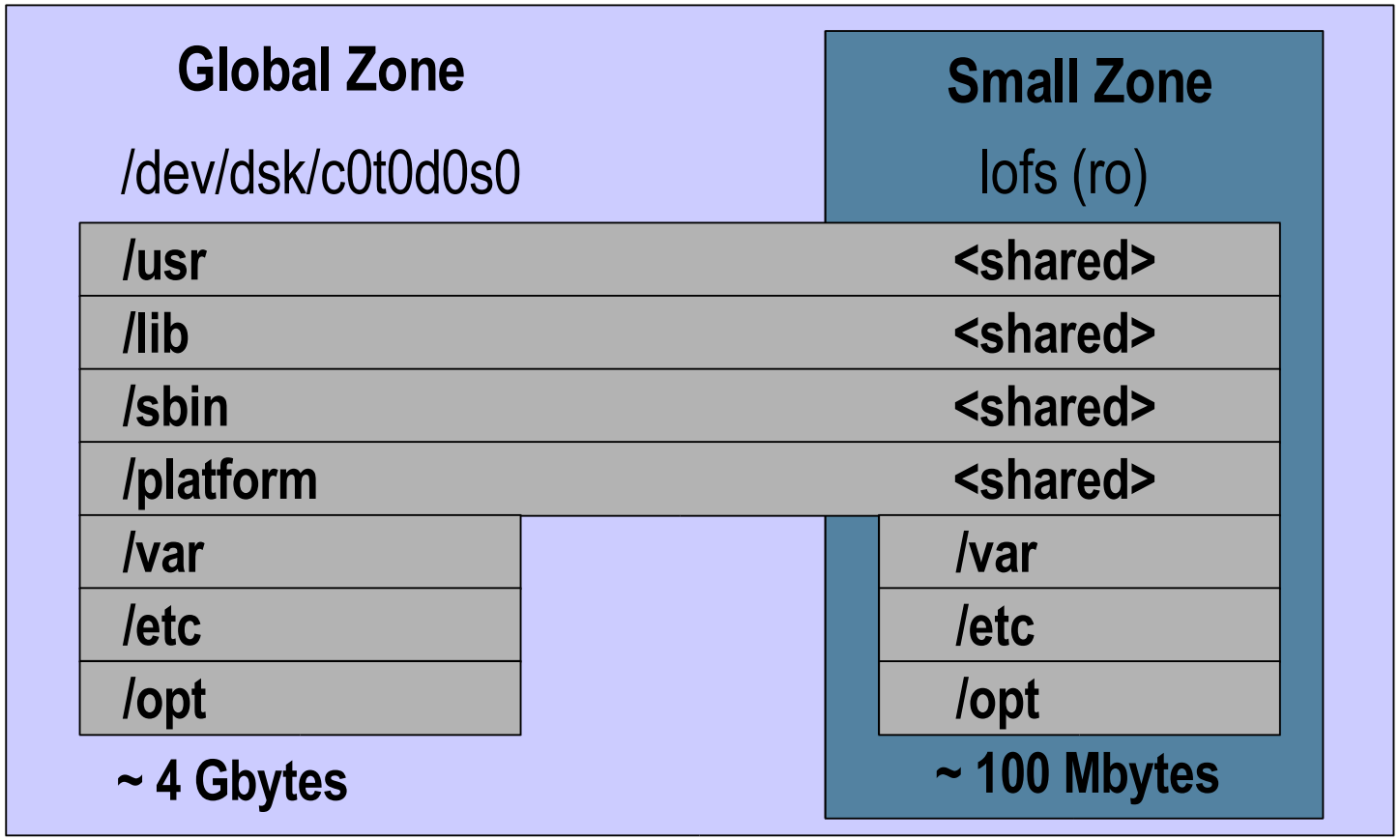


Zone Types

- Global Zone
  - A default Solaris 10 system
  - Can access raw devices
  - Has direct access to the kernel
    - mdb -k
    - patching
  - Exists whether you use zones or not

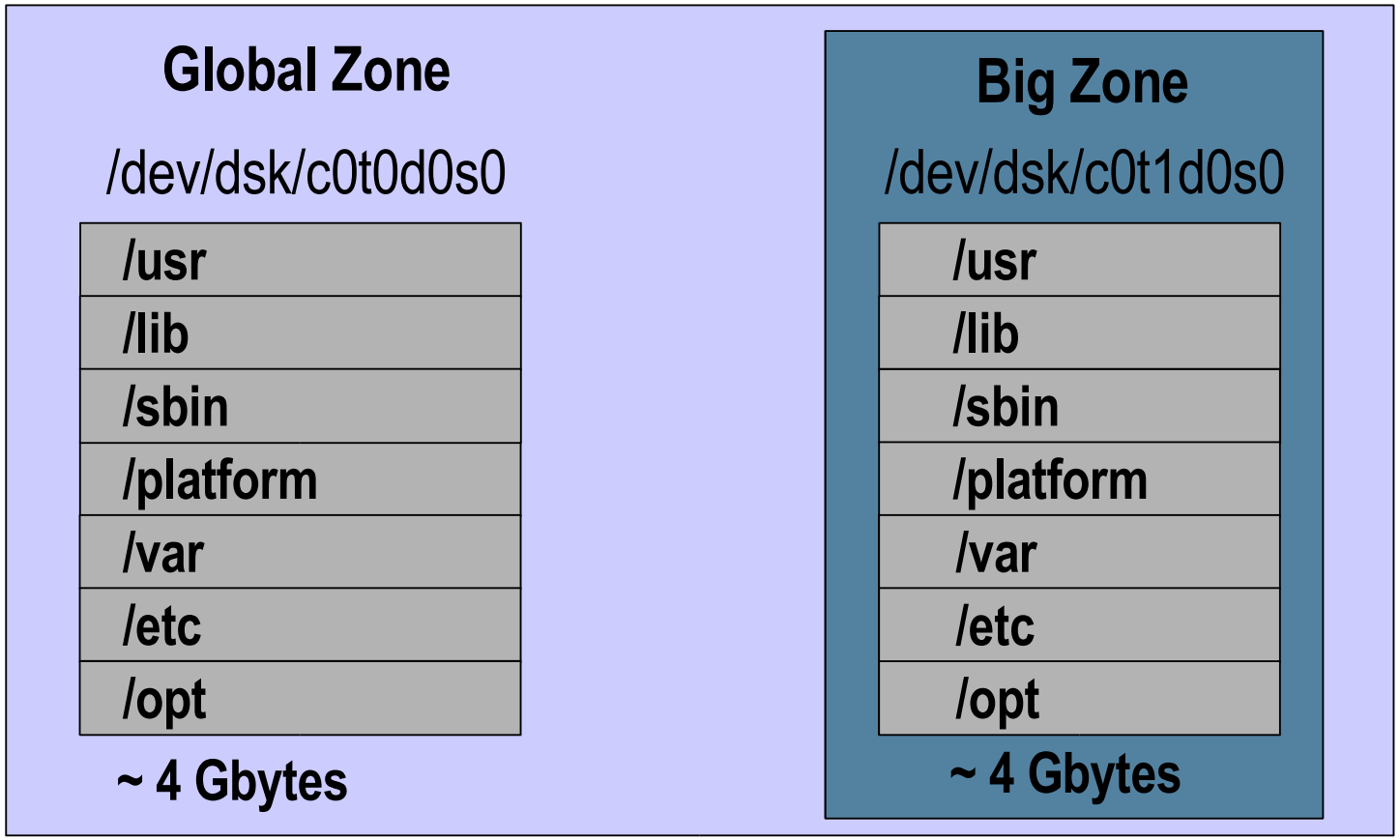
Zone Types

- Sparse Root Zone
  - > This “small zone” shares binaries with the global zone



Zone Types

- Whole Root Zone
  - > This “big zone” has its own OS files



Zone Types

- BrandZ
  - A “Branded Zone”, allows a zone to run non-native operating environments
  - `lx` brand - for Linux zone, provides syscall translation
  - Can run CentOS 3.x, Red Hat Enterprise Linux 3.x
    - Versions 3.5 to 3.8 for both
  - DTrace can trace Linux applications using the `lxsyscall` provider from the global zone

Zone Types

- Sparse Root Zone is default
- Sparse Root Zone advantages
  - Low disk overhead
  - Faster creation, destruction, boot
  - Better performance (higher OS file cache hit rate)
  - Secure - read-only binary files
- When to use the Whole Root Zone
  - When OS binaries need to be modified, customised.

Zone Example

- Creating a sparse root zone,

```
# zonecfg -z small-zone
small-zone: No such zone configured
Use 'create' to begin configuring a new zone.
zonecfg:small-zone> create
zonecfg:small-zone> set autoboot=true
zonecfg:small-zone> set zonepath=/export/small-zone
zonecfg:small-zone> add net
zonecfg:small-zone:net> set address=192.168.2.101
zonecfg:small-zone:net> set physical=hme0
zonecfg:small-zone:net> end
zonecfg:small-zone> verify
zonecfg:small-zone> commit
zonecfg:small-zone> exit
# zoneadm list -cv
  ID NAME             STATUS         PATH
   0 global           running        /
   - small-zone       configured     /export/small-zone
```

Zone Example

- Installing a sparse root zone,

```
# zoneadm -z small-zone verify
# zoneadm -z small-zone install
Preparing to install zone <small-zone>.
Creating list of files to copy from the global zone.
Copying <2574> files to the zone.
Initializing zone product registry.
Determining zone package initialization order.
Preparing to initialize <987> packages on the zone.
Initialized <987> packages on zone.
Zone <small-zone> is initialized.
Installation of these packages generated warnings: <SUNWcsr SUNWdtdte>
The file </export/small-zone/root/var/sadm/system/logs/install_log>
contains a log of the zone installation.
# zoneadm -z small-zone boot
# zoneadm list -cv
  ID NAME             STATUS         PATH
   0 global           running        /
   1 small-zone       running        /export/small-zone
```

Maintenance

- Packages
  - pkgadd is zone aware
    - from global will attempt installing to all zones, unless -G
- Patching
  - patchadd is zone aware
    - from global will attempt installing to all zones if needed
- Upgrading
  - Upgrades on the global zone will upgrade all zones (Solaris 10 1/06); live upgrade, check for support (soon)
- Cloning
  - fast zone creation, especially on ZFS

Security

- Zones are ideal as security containers
- Some applications have a high risk of attack, such as public facing web servers hosting cgi scripts
- What happens if you think your server may be compromised?
  - Your Intrusion Response Plan may involve booting from “known to be good” CDROMs for analysis. Imagine the down time. Picture making that call if you suspect an attack but have no hard proof (it is tough!)
  - Zones can be examined live from a “known to be good” global Zone, which runs no risky software but ssh.

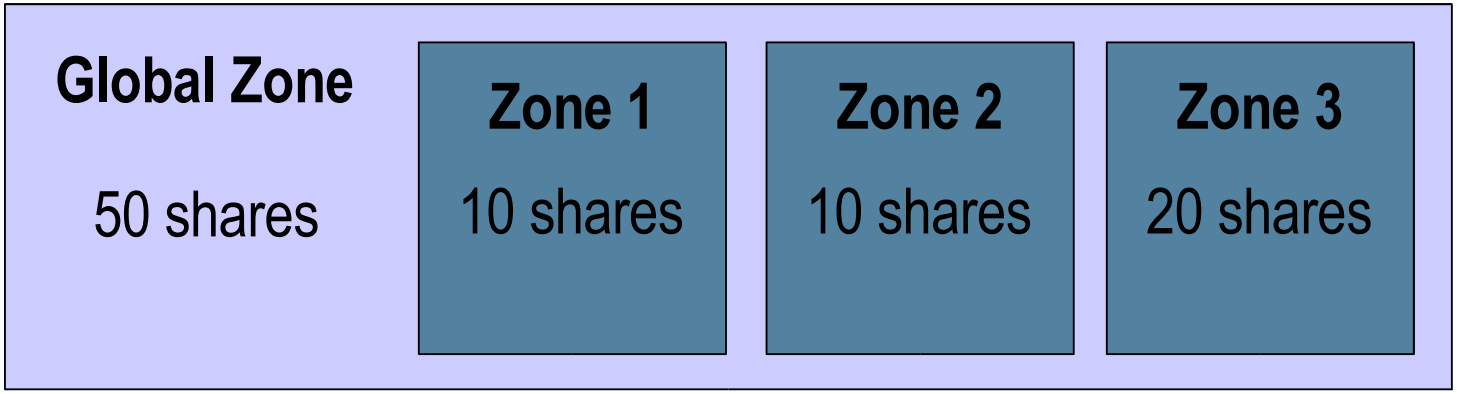
Resource Management

- Many resource management features are available, depending on the version of Solaris 10
  - features in *italic* are in development

Resource	Fine Control	Course Control
CPU	FSS	Processor Sets
Memory	rcapd	*Memory Sets*
Disk Size	ZFS, SVM soft partitions	volumes, disks
Disk Throughput		disks, controllers
Network	IPQoS	Seperate NICs
Swap	swap-max	*Swap Sets*

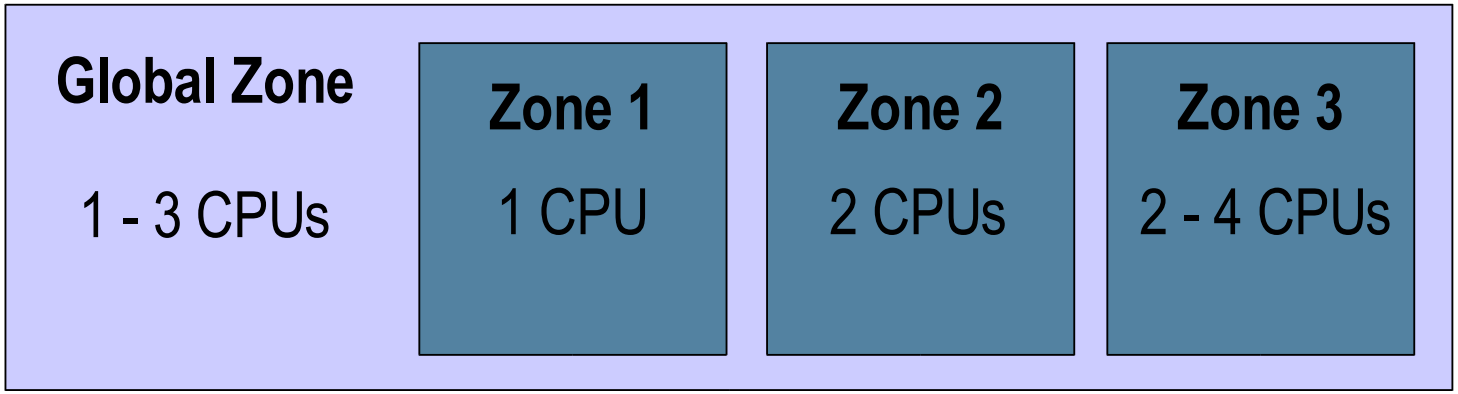
FSS

- Fair Share Scheduler
  - > Fine grained CPU resource control
  - > Allocate each zone a share value
  - > Each zone gets a CPU ration of its shares divided by total busy shares
  - > If only one zone is busy, it gets 100% CPU
  - > Good for CPU utilisation - ROI



Pools

- CPU Resource Pools
  - > Allows fixed CPU allocation
    - – useful for by-CPU licensing
  - > Allows min/max CPU configs
    - – CPU allocation can be tweaked manually
    - – CPU allocation can change during dynamic reconfiguration (add/remove system boards)
    - – CPU allocation can move based on configured objectives



Monitoring

- Many Solaris observability tools are zone aware
  - some are only zone aware with psets (this will get better)
- `ps -Z, df -hZ`
- `prstat -Z` - by-zone status

```
# prstat -Z
   PID USERNAME  SIZE   RSS STATE  PRI NICE      TIME  CPU PROCESS/NLWP
  2008 root     4000K 1168K cpu513  28    0   0:02:11 3.7% cpuhog.pl/1
[...]
ZONEID    NPROC  SIZE   RSS MEMORY      TIME  CPU ZONE
     2       51  182M   93M   0.5%   0:37:27  59% workzone1
     4       51  182M   92M   0.5%   0:16:25  30% workzone2
     3       51  183M   93M   0.5%   0:16:30  10% workzone3
     0       61  359M  194M   1.1%   0:00:11 0.1% global
     1       34  116M   72M   0.4%   0:00:12 0.0% workzone4
Total: 248 processes, 659 lwps, load averages: 51.19, 40.28, 20.52
```

References

- http://www.opensolaris.org/os/community/zones
- http://docs.sun.com
  - Zones and Containers System Administration Guide
- http://www.solarisinternals.com/wiki/index.php/Zones
  - Community wiki

Ctrl-D

Brendan Gregg
brendan@sun.com