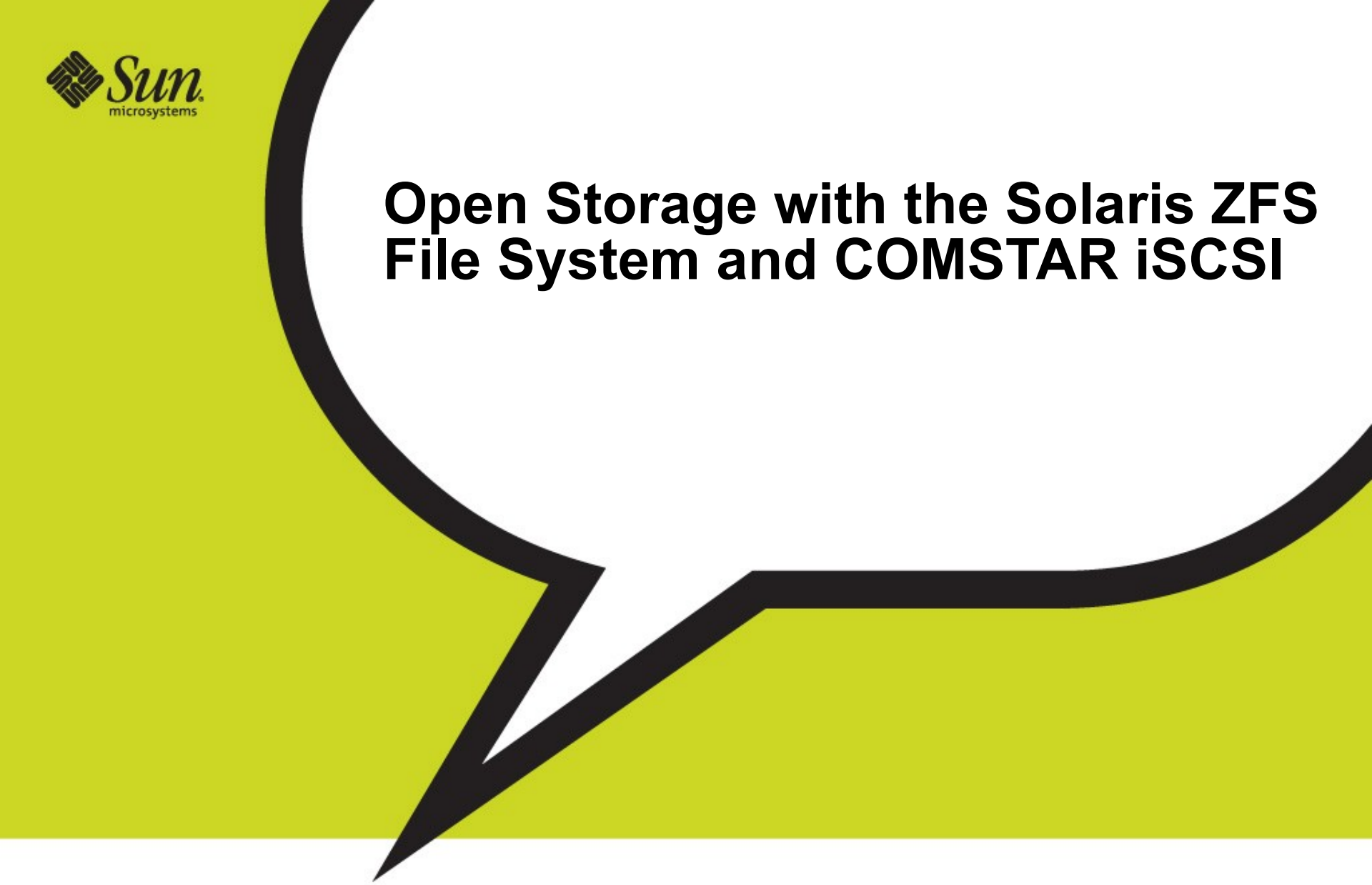


Open Storage with the Solaris ZFS File System and COMSTAR iSCSI

Scott Tracy and Dan Maslowski

Director and Senior Manager for Sun Microsystems

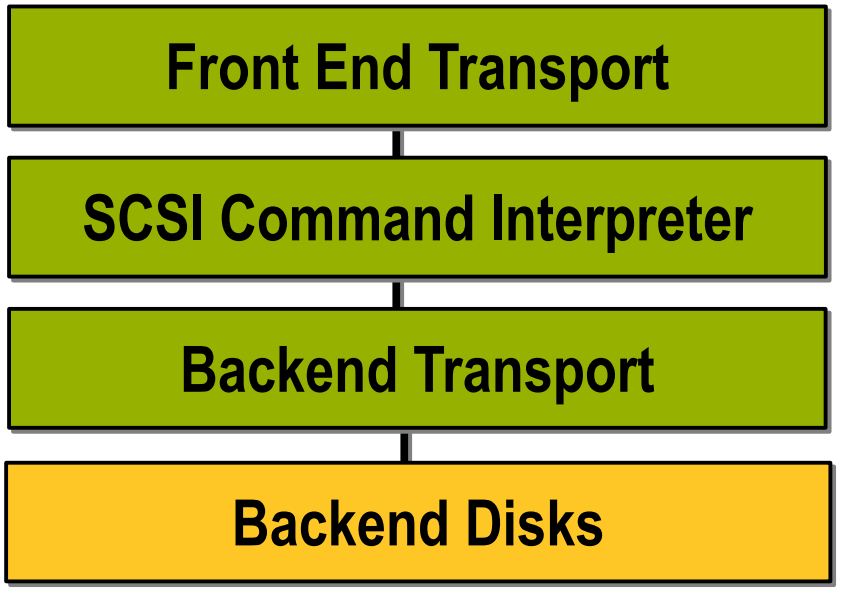
Learn how to build a low-cost, full-featured block disk array with data services from OpenSolaris 2009.06 and commodity hardware using ZFS and COMSTAR

goal

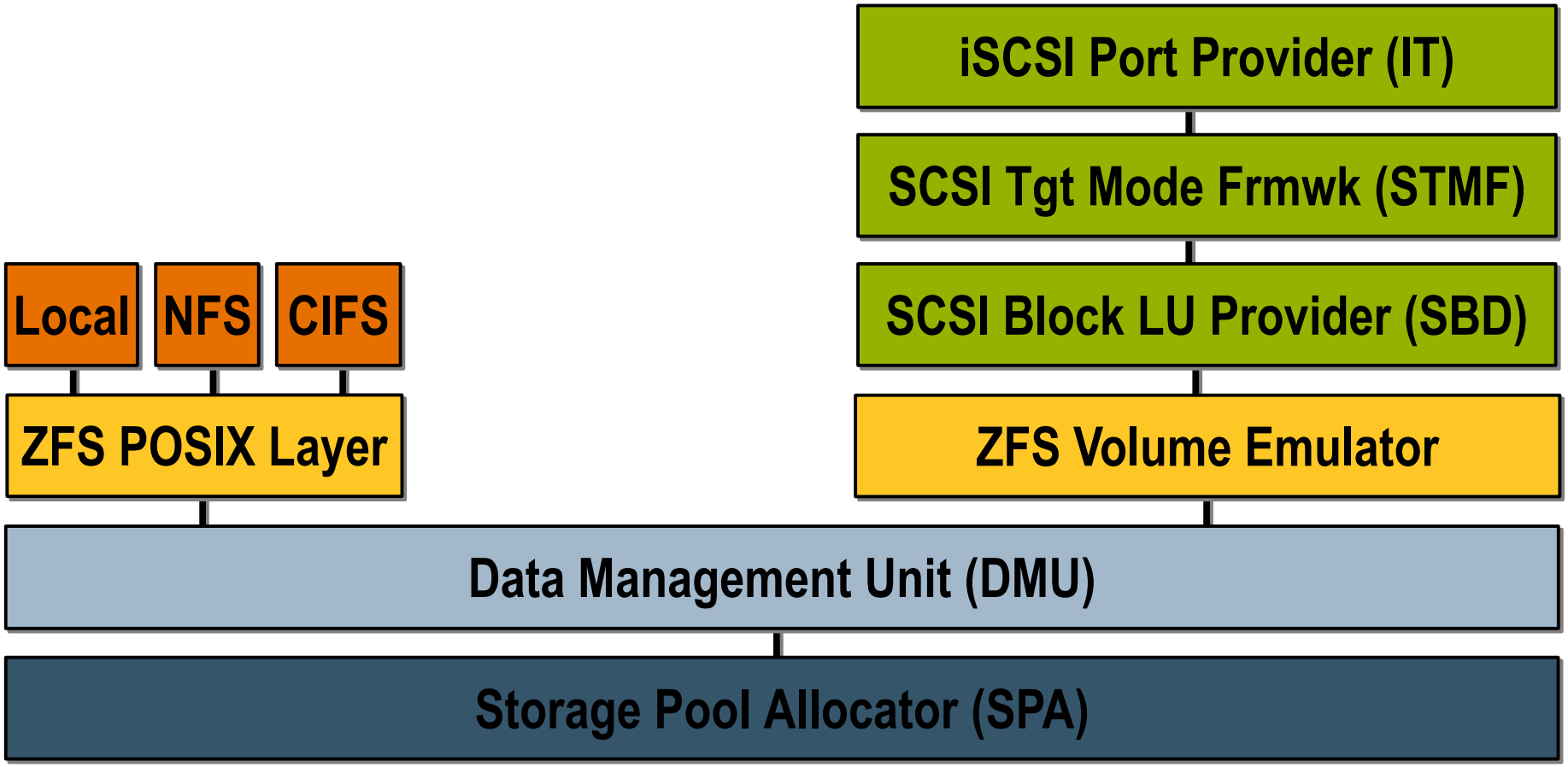
Define Full Featured...

- RAID 1, 5 or 6
- Data Integrity Checking
- Infinite snapshots
- Automated Data Validation
- Compression
- Automated Backup
  - Tape
  - Local disks
  - Remote system (DR)
- LUN Mapping/Masking
- Thin Provisioning
- Dynamic LUN expansion
- Multiprotocol Access
  - iSCSI
  - FC
  - iSER
  - FCoE
  - SRP

Block Target Software



COMSTAR/ZFS Block Target Software



COMSTAR – COmmon Multiprotocol Scsi TARget

- Modular Components that reflect the moving parts of an array
  - Port Provider – Thin layer(s) to describe individual transport protocol and port behavior
  - SCSI Target Mode Framework (STMF) – Framework written to handle:
    - Keeps track of all logical units (LU) and port providers
    - Context and resources for SCSI command execution
    - LUN mapping
    - Abnormal termination of commands and clean-up
  - LU Provider – Thin layer to describe any device type logical unit
- Can utilize any filesystem or raw disk as backing store, but most powerful with ZFS!

ZFS Volume Emulator – The ultimate backing store

- Pooled storage
  - Completely eliminates the antique notion of volumes
  - Does for storage what VM did for memory
- Transactional object system
  - Always consistent on disk – no fsck, ever
  - Universal – file, block, COMSTAR, swap ...
- Provable end-to-end data integrity
  - Detects and corrects silent data corruption
  - Historically considered “too expensive” – no longer true
- Simple administration
  - Concisely express your intent

The Array Hardware (http://www.sun.com/bigadmin/hcl/)

- Parts List:
  - Intel DP35DPM Mobo
  - Intel Core 2 Duo E6570 – 2.66 Ghz, 4MB, 1333FSB CPU
  - 4 GB RAM – DDR2 800 Mhz
  - GeForce 8600 GT 512 MB PCIe
  - 4 Seagate 500 GB SATA HD 7200/16MB/3G/s
  - 500 Watt ATX power supply
  - Mid-Tower case
- Total Cost - 658.47
- Borrowed
  - 2 case fans
  - DVD Drive
  - 2 8G Memory Sticks

Taskmap

Create Storage Pool with ZFS

Configure COMSTAR

Configure Data Services

Assumptions
Installed 2009.06
Packages
storage-server
SUNWiscsit

Taskmap

Create Storage Pool with ZFS

- Pool/Vol Creation
  - RAID Level
  - Compression

Configure COMSTAR

Configure Data Services

Taskmap

Create Storage Pool with ZFS

Configure COMSTAR

- LUN Creation
- LUN Provisioning

Configure Data Services

Taskmap

Create Storage Pool with ZFS

Configure COMSTAR

Configure Data Services

- Automate
  - Snapshots
  - Backup/DR
  - Data Validation

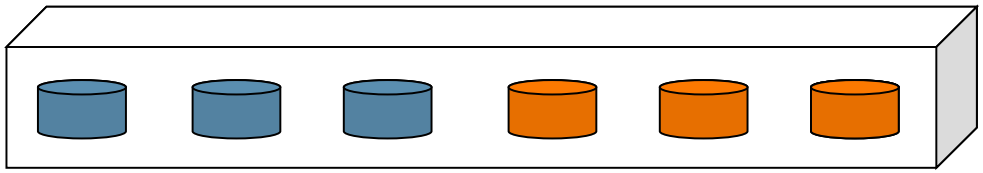
Decisions

Before setting up your full-featured array, it's worth taking some time to walk through choices now to properly match your speed, resiliency, capacity, backup and security requirements.

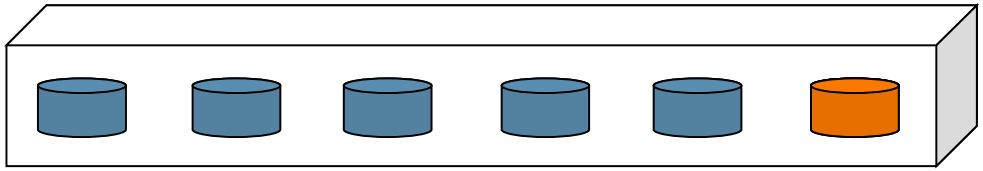
Not to mention your budget...

Some of these choices can change later as your needs change but good to understand those trade-offs now.

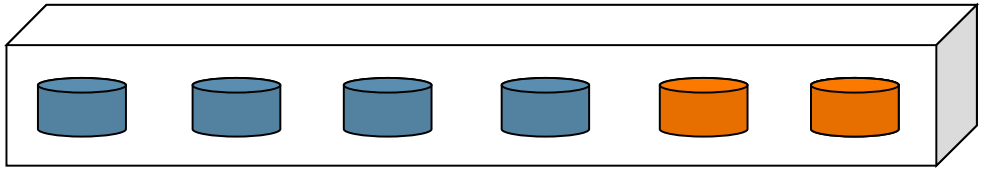
RAID



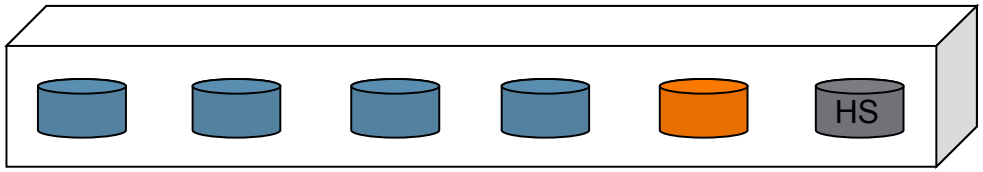
Mirror Volume (RAID 1)



RAIDZ Volume (RAID 5)



RAIDZ2 Volume (RAID 6)



RAIDZ with Hot Spare

Speed	Capacity	Recovery
Fastest	1/N (N-Way)	N-1 Disks (N-Way)
< Mirror	N-1	1 Disk
< RAIDZ (Slightly)	N-2	2 Disks
RAIDZ	N-1	1 Disk Auto Replace

Compression

- Can be set on zPool or zVol
- Data Dependent
  - Suggest setting on appropriate zVol/LUN combination
  - For items that compress well
    - I/O operations will go *faster*
    - Storage will be higher (obvious)
  - For items that don't compress well just the opposite
    - Pictures
    - Music
- Changeable

Backup – It's a snap!

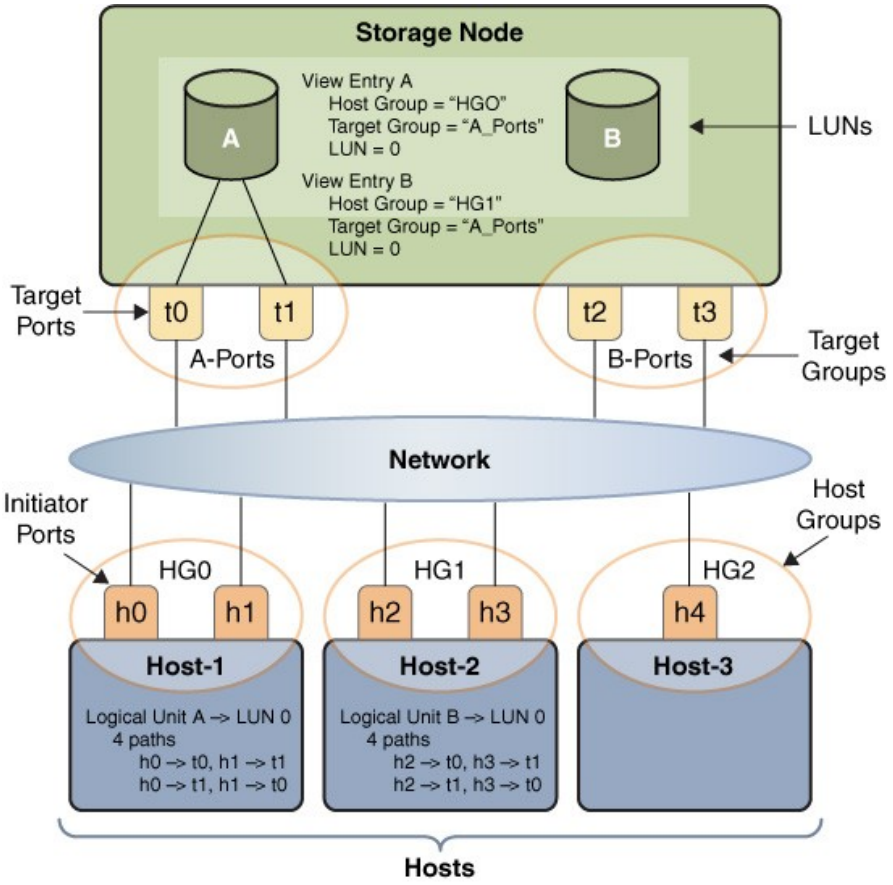
- Snapshots – read-only copy of the volume
  - Automate taking these at regular intervals
  - Suggest using a rolling window of these
  - Use Timeslider or your own script...
- Send/Receive – send a snapshot and receive it to another volume
  - Automate sending these daily
    - Another system (near or remote)
    - Tape (/dev/rmt/0)
  - Use -i option to send these incrementally
- Recovery
  - Rollback – Previous state of file(s)
  - Clone – Create a read/write volume of snapshot

Data Validation – Scrub it!

- Utilize the 'zpool scrub' command
  - Performs a data validation based on a compare between the data and checksum at rest on disk
  - Should factor into your backup plan
    - Think of it this way – if you do find an uncorrectable error, what restore operation would bring it back?
    - Must be more often than a full tape cycle
    - Best Practice - ½ tape cycle
- Why you don't need to do this
  - All IO operations do the comparison
    - If you cycle through your data often, you get this for free
  - RAID plays a role – In a RAIDZ2 configuration, bit rot has to occur on 2 disks on the same data to prevent reconstruction.

LUN Provisioning with COMSTAR

- What are we doing?
  - Masking LUNs to certain Hosts through specific ports
  - Setting LUN numbers for Hosts to view
- Steps
  - Create Host Group
  - Add Host Group members
  - Create Target Group
  - Add Target Group Members
  - Use HG and TG and LUN# in view entry



iSCSI Authentication

- Security – Determine whether a connection command is truly coming from a trusted host
- iSCSI Port Provider uses CHAP (Challenge-Handshake Authentication Protocol)
  - Username and Password – UserName is node name (iqn)
  - Unidirectional (Default) – Target verifies Initiator from stored secret
  - Bidirectional – Mutual authentication between target and initiator
- If connecting many nodes – use a RADIUS server to simplify

Simple iSCSI Array Setup

How hard is this stuff?

Simple iSCSI Array Setup

- Cheap and small
  - 3 Disk Machine – 1 for the OS, 2 for mirror
- Create Storage Pool
  - Create mirrored cpool for COMSTAR data
  - Create single ZFS Volume of mirror size
- Configure COMSTAR
  - Enable the services
  - Create 1 LUN with ZFS Volume
  - Make LUN available to all hosts
- No Data Services

Advanced Array Setup

Let's put it all together!

demo

Advanced Array Setup

- Still Cheap, a little bigger
  - 2 Flash drives for mirrored OS
  - 4 Disk Machine
- Create Storage Pool
  - Create raidz cpool for COMSTAR data
  - Create multiple ZFS Volumes with different characteristics
- Configure COMSTAR
  - Enable the services
  - Create LUNs with ZFS Volume backing store
  - LUN Provisioning
- Data Services
  - Automate
    - Backup
    - Data Validation
  - Authentication

Summary

- COMSTAR + ZFS makes a powerful combination to put together
  - Low-cost, commodity-based block disk array with
  - Advanced Data Services that typically cost extra $$
- More Info
  - http://www.opensolaris.org/os/project/comstar/

Open Storage with the Solaris ZFS File System and COMSTAR iSCSI

Scott Tracy and Dan Maslowski
scott.tracy@sun.com, dan.maslowski@sun.com

Simple iSCSI Array Setup

How hard is this stuff?

Simple iSCSI Array Demo

- Create mirrored zpool and create LUN backing store

```
# zpool create cpool mirror c8d0s0 c9d0s0
# zfs create -V 500G cpool/ScottLUN
```

- Configure COMSTAR iSCSI Target and LUN

```
# sbdadm create-lu /dev/zvol/rdsk/cpool/ScottLUN
# stmfadm add-view <GUID>
# itadm create-target
```

Advanced Array Setup

Let's put it all together!

demo

Advanced iSCSI Array Demo

- Mirror your OS

```
# zpool attach rpool c8d0s0 c9d0s0
```

- Configure COMSTAR Pool and create LUN backing store

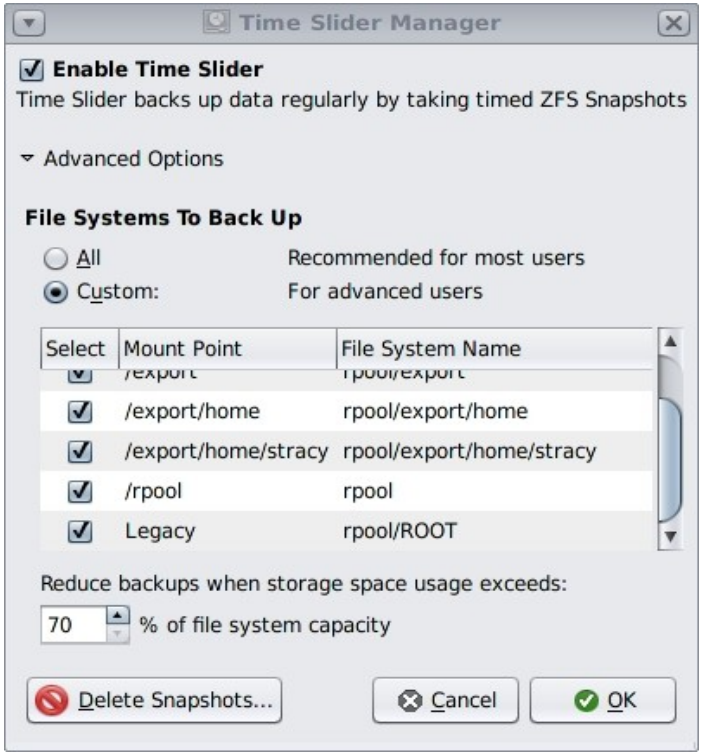
```
# zpool create cpool raidz c5d0s0 c6d0s0 c7d0s0
# zfs create -V 100G cpool/Music
# zfs create -V 100G cpool/Pic
# zfs create -V 100G cpool/Work
# sbdadm create-lu /dev/zvol/rdsk/cpool/Music
# sbdadm create-lu /dev/zvol/rdsk/cpool/Pic
# sbdadm create-lu /dev/zvol/rdsk/cpool/Work
```

Advanced iSCSI Array Demo

- Turn on Compression for the Work LUN

```
# zfs set compression=yes cpool/Work
```

- Configure Timeslider for backup



Advanced iSCSI Array Demo

- Cronjob your scrub – 2x per month on 1st and 15th at 1:23 AM

```
# cat scrub (in /var/spool/cron/crontabs)
# 23 01 1,15 * * for i in 'zpool list -H -o
  name';do zpool scrub $i; done
```

- LUN Mapping/Masking

```
# stmfadm create-hg HG0
# stmfadm add-hg-member -g HG0 iqn.1986-
  03.com.sun:01:e00000000000.4a098627
# stmfadm create-tg A_Ports
# stmfadm add-tg-member -g A_Ports iqn.1986-
  03.com.sun:02:6d76e02f-3c2f-4543-f917-
  97fad79603b8
# stmfadm add-view -h HG0 -t A_Ports -n 0
  600144F0D8D78D0000004A0B60E70001
```

Advanced iSCSI Array Demo

- Setup Authentication for iSCSI

```
# iscsiadm modify initiator-node --CHAP-secret
Enter CHAP secret:************
Re-enter secret:************
# iscsiadm modify initiator-node -authentication
  CHAP
# itadm modify-target -a chap iqn.1986-
  03.com.sun:02:6d76e02f-3c2f-4543-f917-
  97fad79603b8
# itadm create-initiator -s iqn.1986-
  03.com.sun:01:e00000000000.4a098627
Enter CHAP secret:************
Re-enter secret:************
```