# Oracle ZFS Storage Appliances

Daniel A. Morgan | damorgan12c@gmail.com | www.morganslibrary.org

# Disclaimer

- This room is an unsafe harbour
- No one from Oracle has previewed this presentation
- No one from Oracle knows what I'm going to say
- No one from Oracle has supplied any of my materials

- ... because the technology is currently available and
- works extremely well
- You may rely upon this presentation to make decisions for your enterprise

**This disclaimer has <u>not</u> been approved by Oracle Legal**

# Daniel A. Morgan

- Oracle ACE Director
- Consultant to Harvard University
- University of Washington Oracle Instructor, ret.
- The Morgan of Morgan's Library on the web
- Board Member: Western Washington OUG
- Upcoming Presentations
  - Jun 21: VicOUG
  - Sep: OpenWorld 2012: San Francisco
  - Dec 3-5: UKOUG

Official Beta Site ORACLE 11g DATABASE

ORACLE RACSIG

International zSeries Oracle SIG

# Syllabus

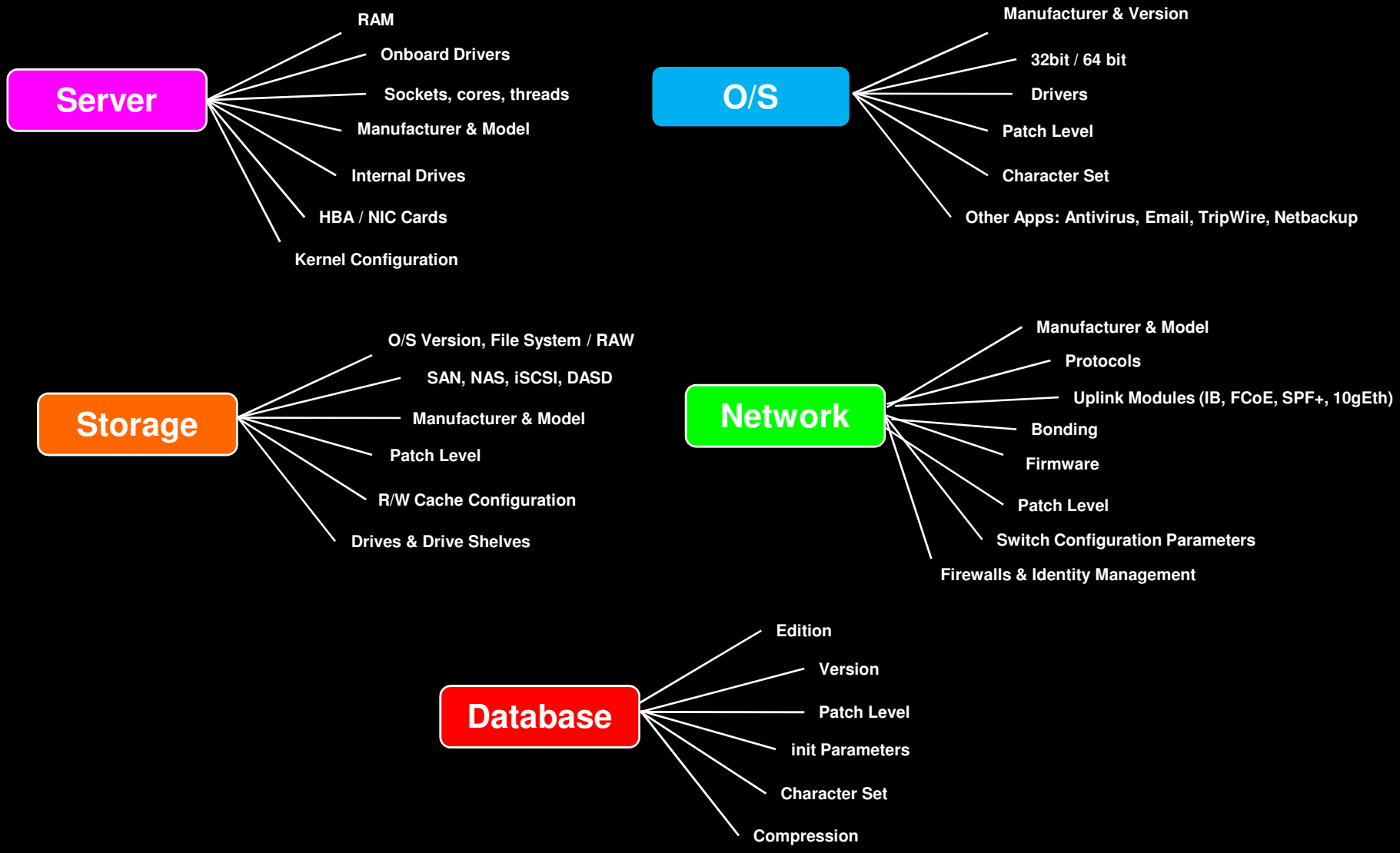
- June Presentation Follow-up
  - At OpenWorld I replaced Britney Spears with a barrel of Squid
- October ZFS at OpenWorld

# Oracle didn't the ODA childproof

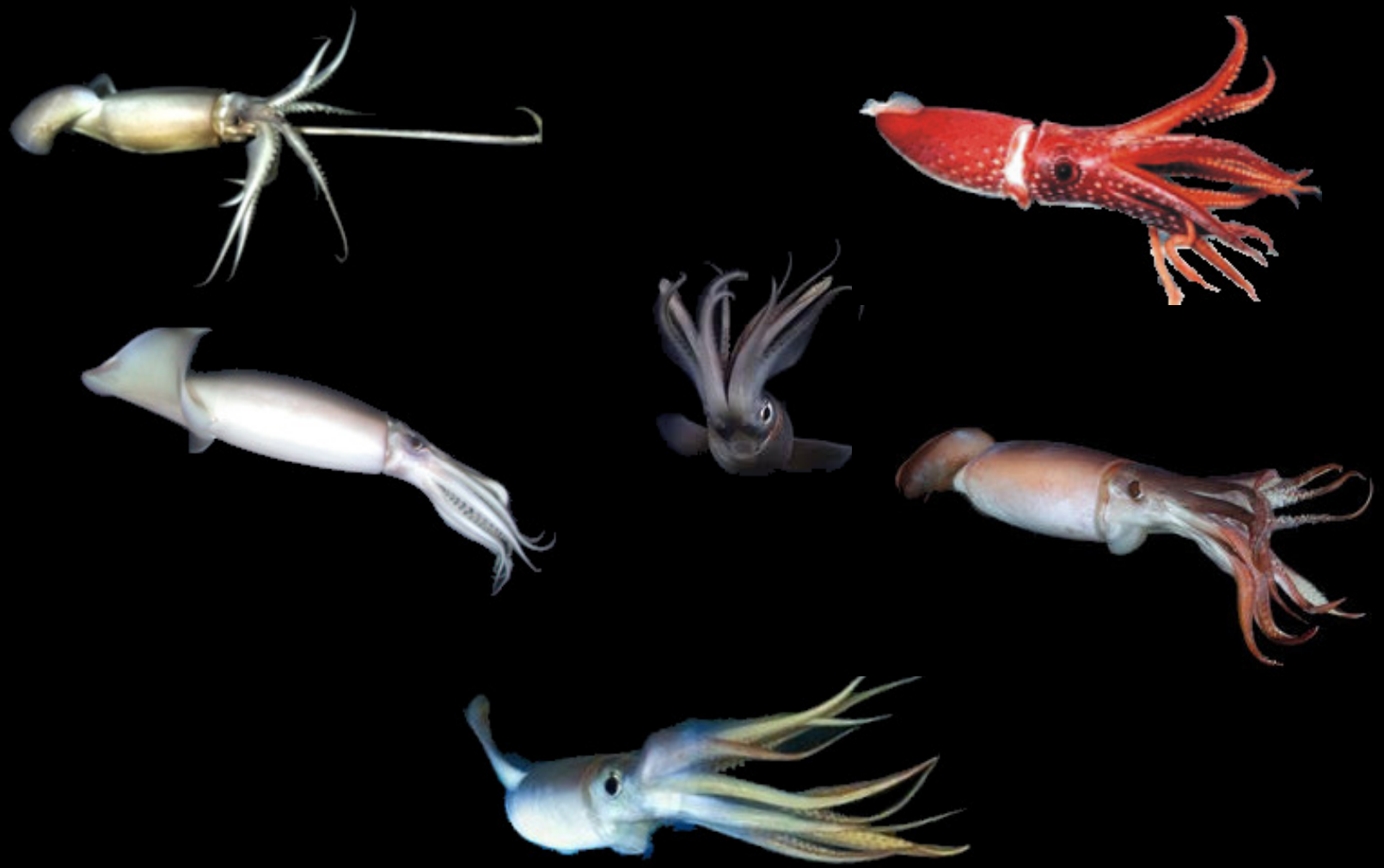
# At OpenWorld I replaced LL with ...



# Static Puzzle Pieces



# Animated Puzzle Pieces



# It's hard to fall in love with a barrel of squid too

# So let's talk about storage

- We MAY want to preserve the 4TB ASM disk for data
- We may want more storage for
  - FRA, Flashback DB files, RMAN files …
  - Clone
  - Data Masking
  - Real Application Testing
  - Staging
  - Logs
  - And so on

# ZFS

# Choices

- ASM
  - Raw devices
- Clustered Storage
  - Which one? OCFS2, VxFS, …
- Non-Clustered Storage
  - Non-blocking visibility on both nodes
  - dNFS, CIFS …

# ASM?

- Excellent decision for database storage
- Perhaps not optimal as a file system
  - ACFS?
- Requires raw disk to be presented to ODA
- Traditional HBA discussion

# Clustered File System?

- Several CFS available for Linux
  - Need expertise
  - Wire it yourself
  - Tech concerns
    - File sizes
    - File counts
- Still traditional HBA discussion

# Non-clustered File System?

- Local File System
  - May be suitable for some applications,
    - But we have two separate hosts in ODA
  - Standard Linux-oriented
    - Still traditional HBA discussion

- [d]NFS
  - Vendor: NetApp, Oracle ZFS Appliance
  - OpenFiler?

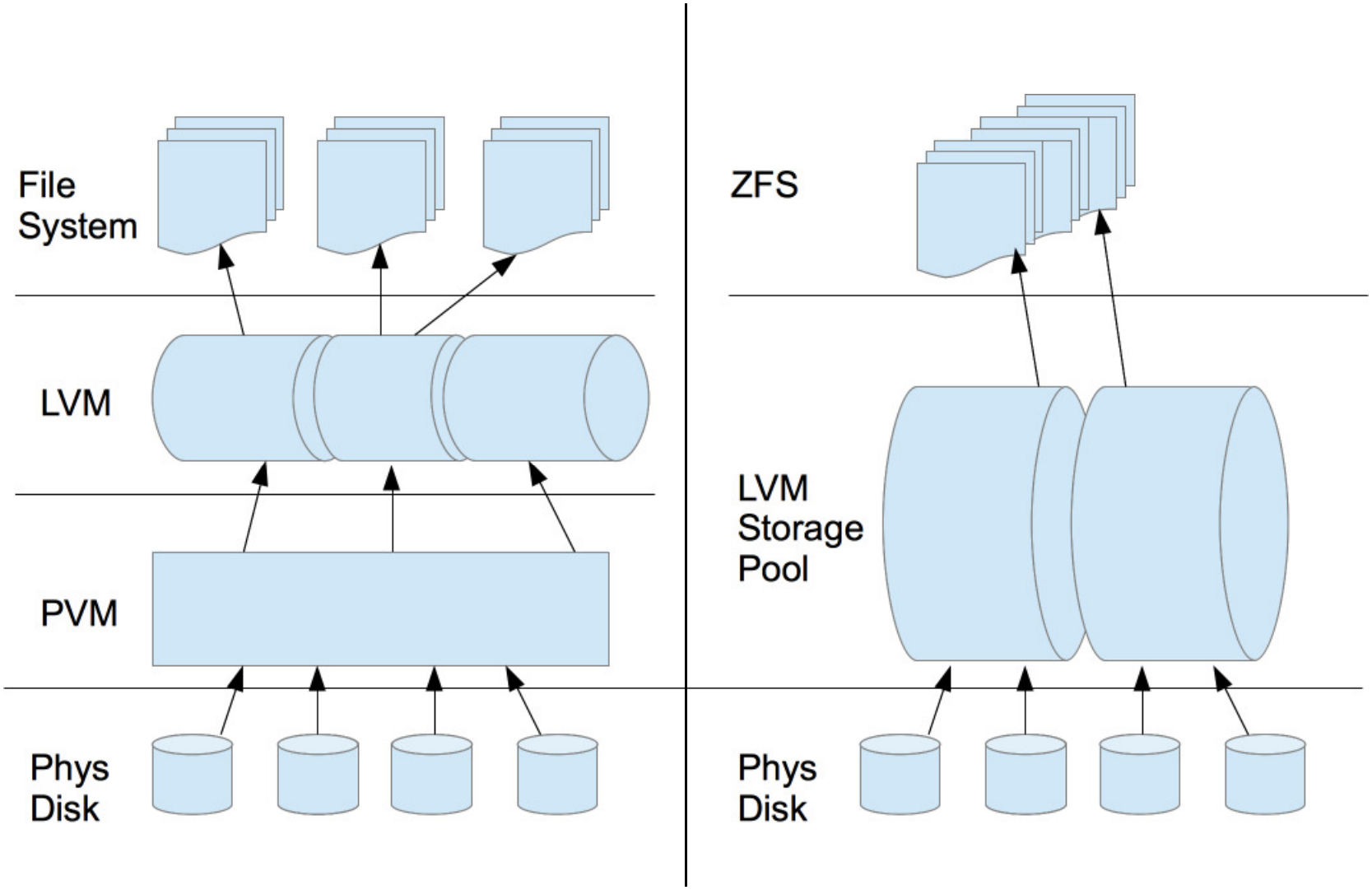
# Additional concern – silent corruption

- An undetected or uncorrectable error can occur on average once every 10-20 TB of data storage OR transfer
  - In modern systems that could mean a corruption in a little as 15 minutes
- ZedFS was designed to combat this challenge
  - Checksum on all blocks
  - Copy on Write (preserve original block, not write in place)
  - Hot spares in pool
  - Auto-healing from ZFS mirror
  - Scrub instead of fsck
    - Monthly (or weekly for consumer disks)

# Additional concern – silent corruption

- An undetected or uncorrectable error can occur on average once every 10-20 TB of data storage OR transfer
  - In modern systems that could mean a corruption in a little as 15 minutes
- ZFS was designed to combat this challenge
  - Checksum on all blocks
  - Copy on Write (preserve original block, not write in place)
  - Hot spares in pool
  - Auto-healing from ZFS mirror
  - Scrub instead of fsck
    - Monthly (or weekly for consumer disks)

# Traditional File System stack vs ZFS



# Quick Notes

- RAID
  - ZFS cannot fully protect the user's data when using a hardware RAID controller, as it is not able to perform the automatic self-healing unless it controls the redundancy of the disks and data.
  - Instead, ZFS provides it's own RAID counterparts within the Storage Pool
- ZFS provides a hot-spare storage pool manager and a 128-bit, Copy on Write File System
- Capacity
  - Single file: 16 exabytes
  - Files in a pool: 264
  - Disks in a pool: 264
  - Pools in a system: 264

# Where do you want to invest your time and treasure?

- Reinventing the wheel?
- Designing physical architecture?
- Applying one-off patches?
- Becoming Linux security experts?
- Writing shell scripts?

or would you rather be ...

- Managing your applications, users, and data?
- Optimizing your applications to maximize customer satisfaction?

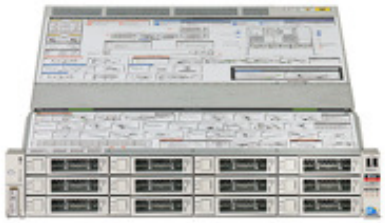
# ZFS Storage Appliance

# ZFS Storage Appliance

- ZFS file system with advanced error detection and self-healing capabilities
- Integrated with Oracle Engineered Systems
- Both ZFS Deduplication and Compression or Hybrid Columnar Compression
- Hybrid Storage Pools
- Simultaneous multiprotocol support across multiple network interconnects, including GbE, 10 GbE, fibre channel and InfiniBand
- Integrated with OEM Grid Control
- Web-based storage management
- Integrated real-time storage analytics

# What is a ZFS Appliance?

- Enterprise class Network Attached Storage (NAS)
- Choose the size that meets your needs
- Hybrid Columnar Compression (w/o an Exadata)
- Hybrid storage pools for DRAM and Flash caches
- DTrace storage analytics
- Use for
  - Backup and Restore
  - Cloning
  - Data Masking



# ZFS Configurations

| Sun ZFS Storage Appliance Configurations | | | | | | |
|---|---|---|---|---|---|---|
| | **Key Requirement** | **Maximum Storage Capacity** | **Space (Rack Units)** | **Write Optimized Flash** | **Read Optimized Flash** | **Cluster Option** |
| Sun ZFS Storage 7120 | Low-priced entry-level system with all software features | 177 TB | 2U/controller, 4U/disk shelf | 73 GB | N | N |
| Sun ZFS Storage 7320 | Entry-level cluster option for high availability | 432 TB | 1U/controller, 4U/disk shelf | Up to 1.2 TB | Up to 2 TB per controller | Y |
| Sun ZFS Storage 7420 | Best price/performance | 1.73 PB | 3U/controller, 4U/disk shelf | Up to 7.0 TB | Up to 2 TB per controller | Y |

# ZFS Specifications

| Sun ZFS Storage Appliance Specifications | | | |
|---|---|---|---|
| | **Sun ZFS Storage 7120** | **Sun ZFS Storage 7320** | **Sun ZFS Storage 7420** |
| **Architecture** | | | |
| Processor | 1x 4-core 2.4 GHz Intel® Xeon® Processor | 2x 4-core 2.4 GHz Intel® Xeon® Processor, per controller | 4x 8-core 2.0 GHz or 10-core 2.4GHz Intel® Xeon® Processors per controller |
| Main memory | 48 GB | Up to 144 GB per controller | Up to 1 TB per controller |
| **Base Configurations** | | | |
| Configuration options | • 3.3 TB to 177 TB using either high-speed (15,000 RPM) or high-capacity (7,200 RPM) SAS-2 disks<br>• Controller contains 11 HDDs and one SSD cache, supports up to two additional disk shelves with 24 disks each (300 GB, 600 GB, 2 TB, or 3 TB) | • 6 TB to 432 TB using either high-speed (15,000 RPM) or high-capacity (7,200 RPM) SAS-2 disks<br>• Supports up to six disk shelves with 20 or 24 disks each (300 GB, 600 GB, 2 TB, or 3 TB) and up to four optional write-optimized SSDs per shelf | • 6 TB to 1.73 PB using either high-speed (15,000 RPM) or high-capacity (7,200 RPM) SAS-2 disks<br>• Supports up to 24 disk shelves with 20 or 24 disks each (300 GB, 600 GB, 2 TB, or 3 TB) and up to four optional write-optimized SSDs per shelf |

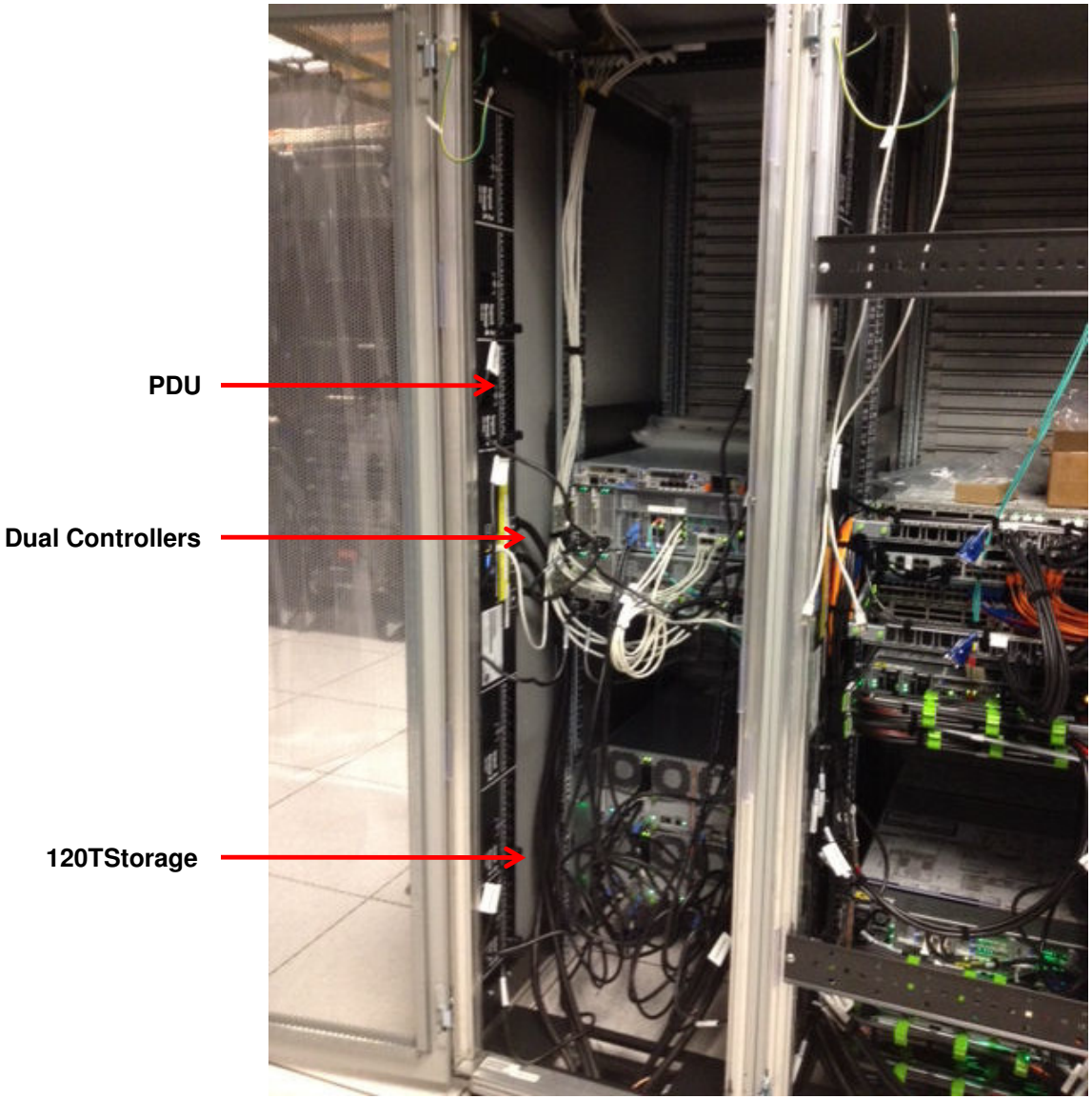
# ZFS In The Data Center

# ODA Front

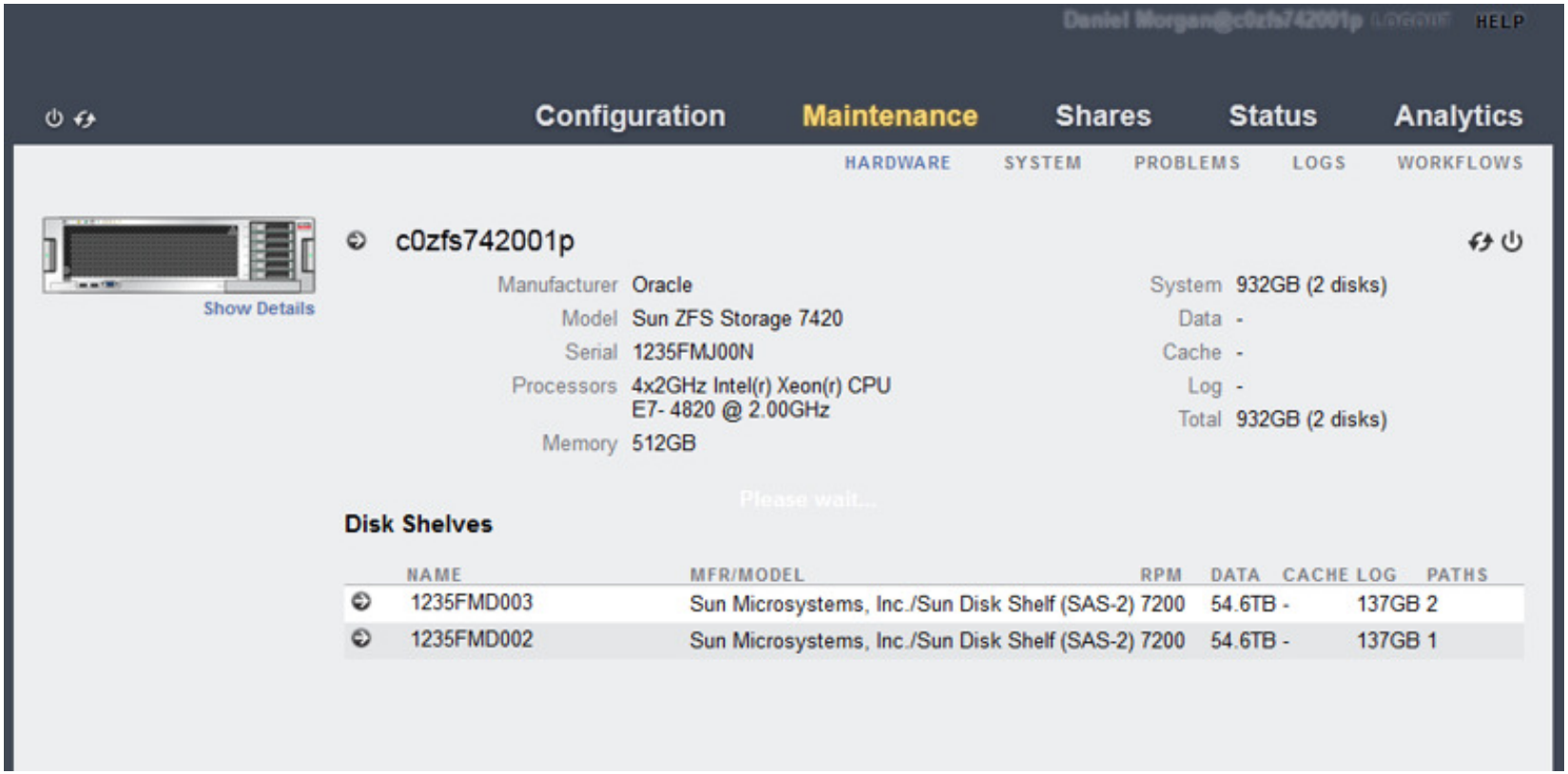
# ZFS 7420


Sun
ORACLE
SUN ZFS STORAGE
7420

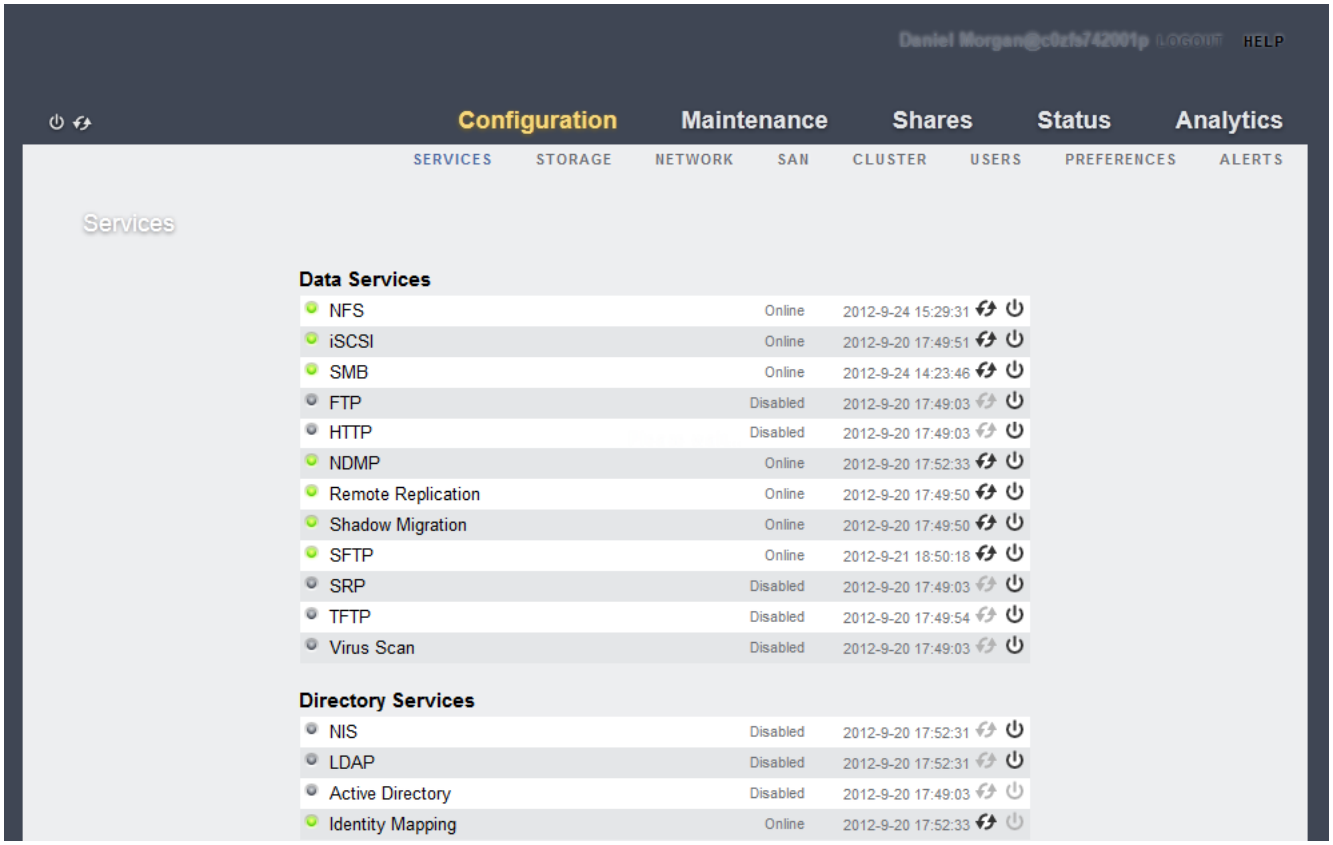
# ZFS Internals



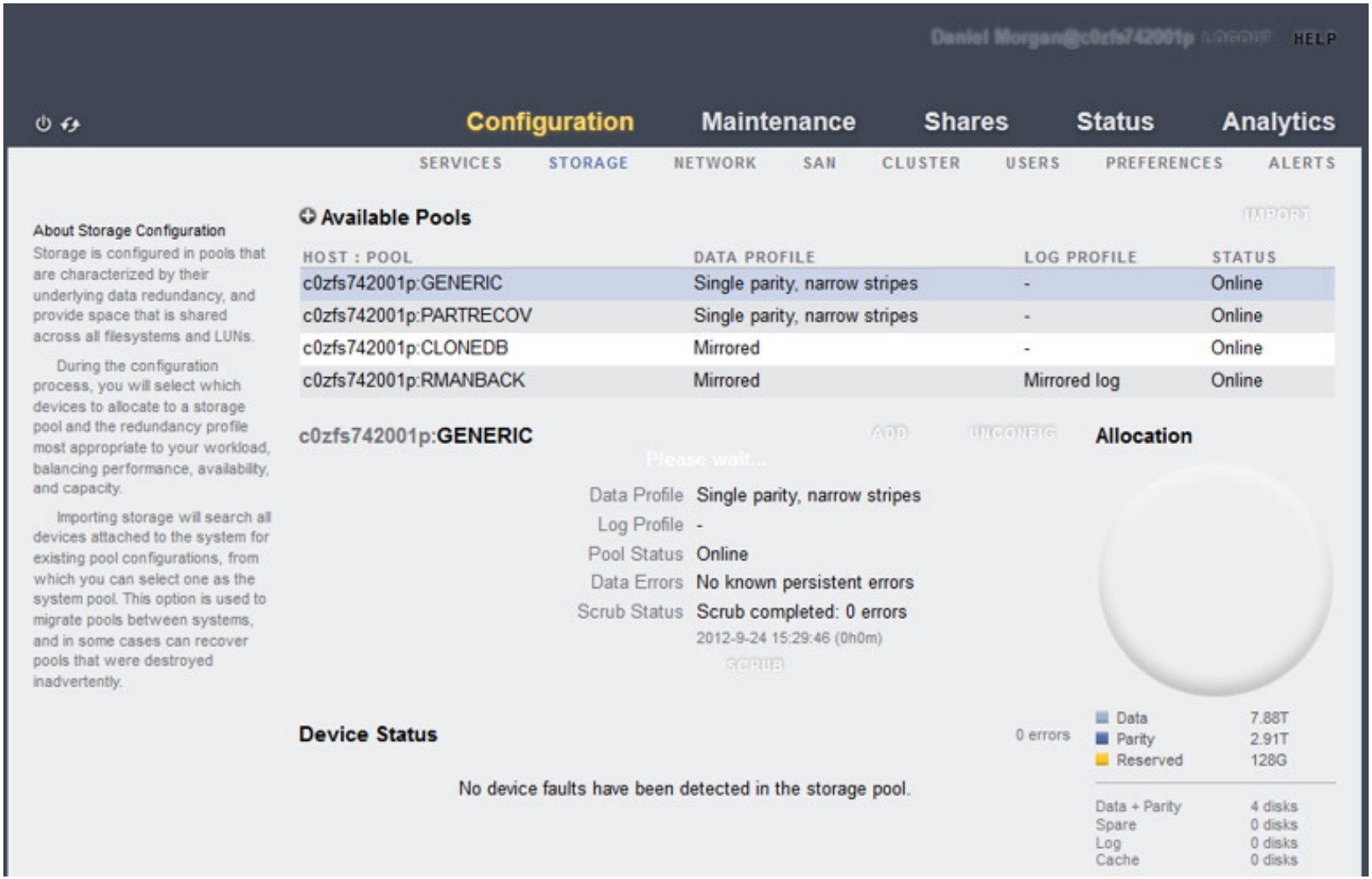
# ZFS BUI


HELP
Configuration
Maintenance
Shares
Status
Analytics
HARDWARE
SYSTEM
PROBLEMS
LOGS
WORKFLOWS
Show Details
c0zfs742001p
Manufacturer Oracle
Model Sun ZFS Storage 7420
Serial 1235FMJ00N
Processors 4x2GHz Intel(r) Xeon(r) CPU E7- 4820 @ 2.00GHz
Memory 512GB
System 932GB (2 disks)
Data -
Cache -
Log -
Total 932GB (2 disks)
Please wait...
Disk Shelves
NAME MFR/MODEL RPM DATA CACHE LOG PATHS
1235FMD003 Sun Microsystems, Inc./Sun Disk Shelf (SAS-2) 7200 54.6TB - 137GB 2
1235FMD002 Sun Microsystems, Inc./Sun Disk Shelf (SAS-2) 7200 54.6TB - 137GB 1

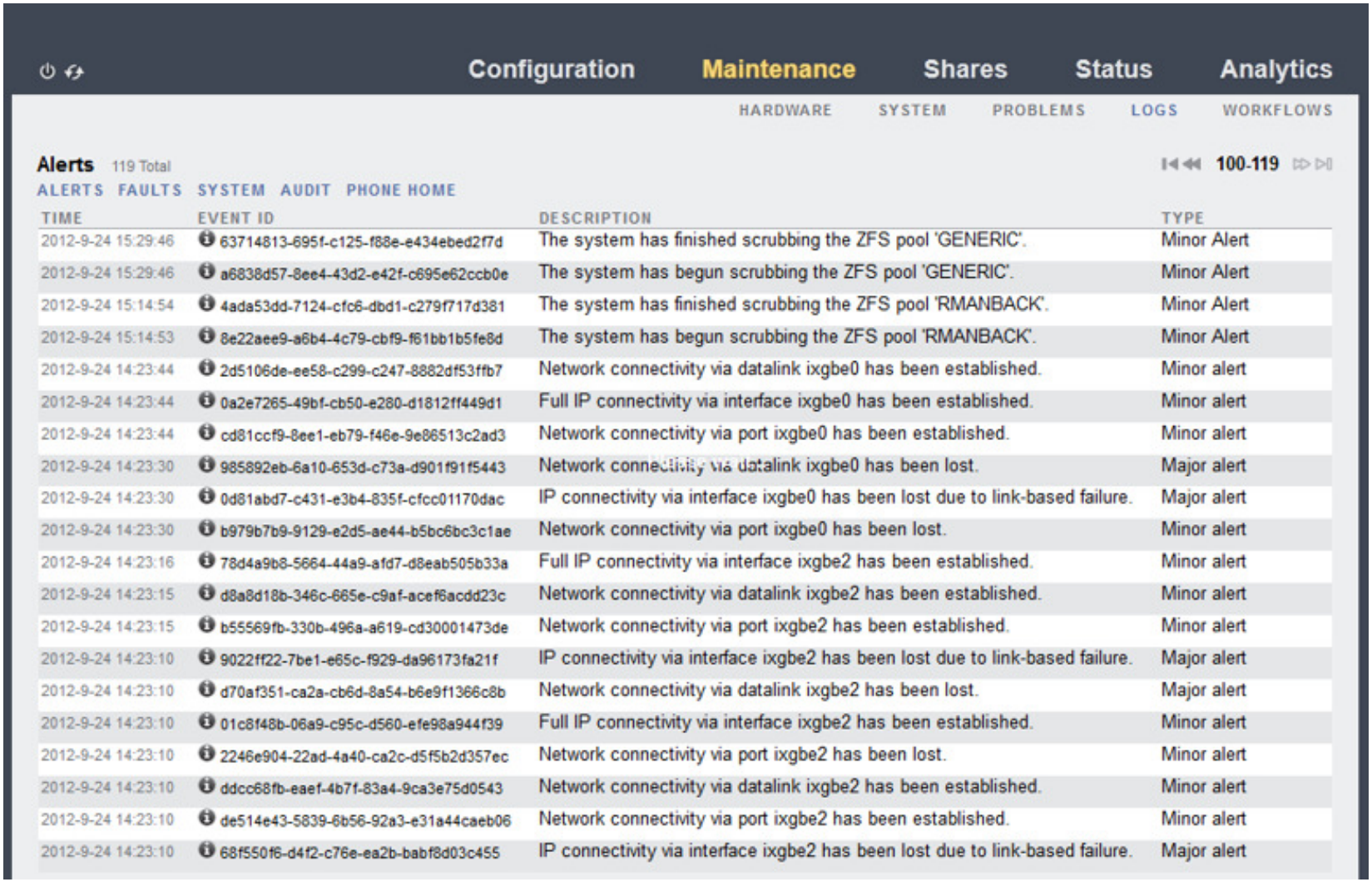
# ZFS Config Services



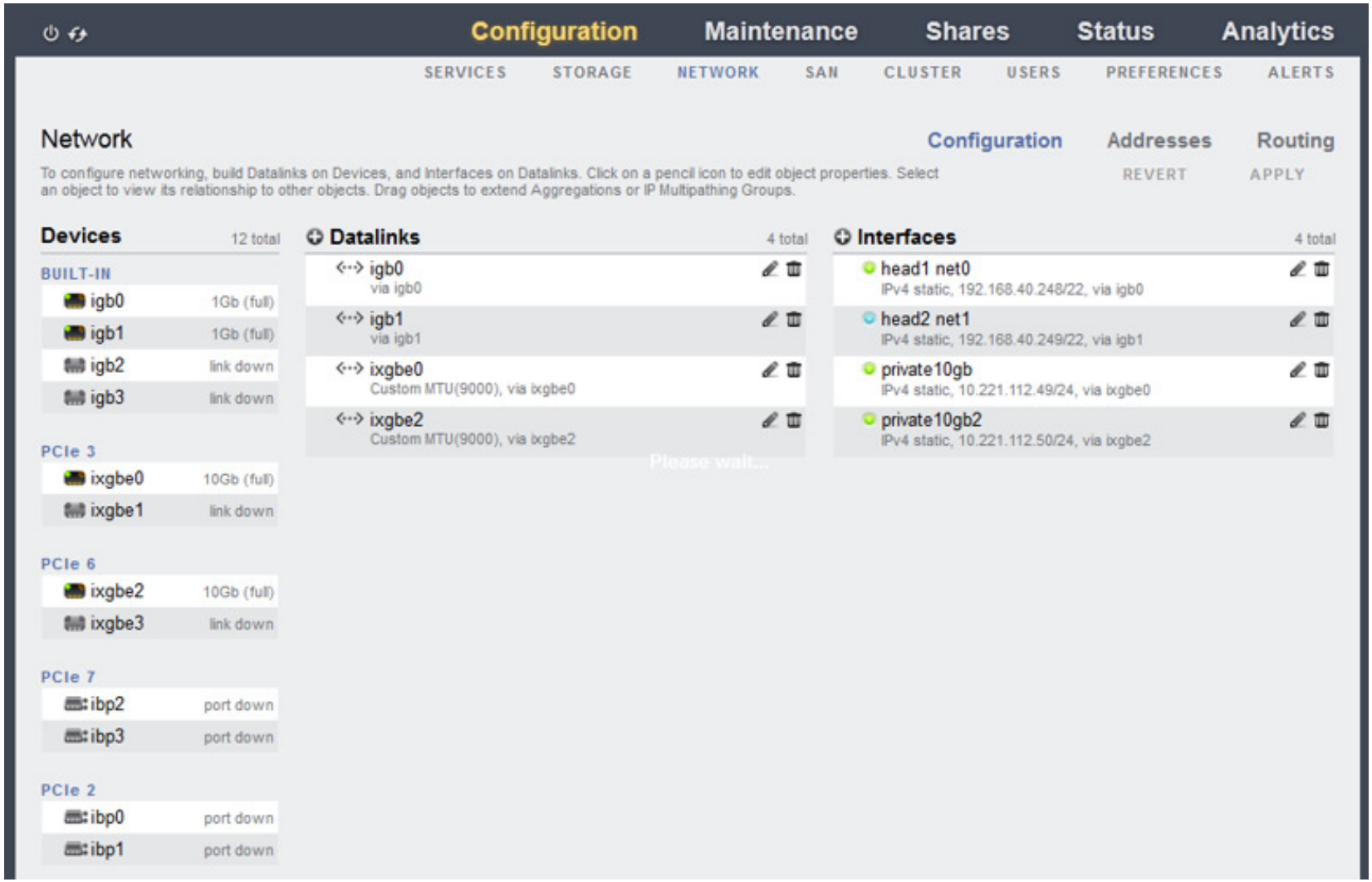
# ZFS BUI


HELP
Configuration
Maintenance
Shares
Status
Analytics
SERVICES
STORAGE
NETWORK
SAN
CLUSTER
USERS
PREFERENCES
ALERTS
About Storage Configuration
Storage is configured in pools that are characterized by their underlying data redundancy, and provide space that is shared across all filesystems and LUNs.
During the configuration process, you will select which devices to allocate to a storage pool and the redundancy profile most appropriate to your workload, balancing performance, availability, and capacity.
Importing storage will search all devices attached to the system for existing pool configurations, from which you can select one as the system pool. This option is used to migrate pools between systems, and in some cases can recover pools that were destroyed inadvertently.
Available Pools
HOST : POOL
DATA PROFILE
LOG PROFILE
STATUS
c0zfs742001p:GENERIC
Single parity, narrow stripes
-
Online
c0zfs742001p:PARTRECOV
Single parity, narrow stripes
-
Online
c0zfs742001p:CLONEDB
Mirrored
-
Online
c0zfs742001p:RMANBACK
Mirrored
Mirrored log
Online
c0zfs742001p:GENERIC
Data Profile Single parity, narrow stripes
Log Profile -
Pool Status Online
Data Errors No known persistent errors
Scrub Status Scrub completed: 0 errors
2012-9-24 15:29:46 (0h0m)
Allocation
Data 7.88T
Parity 2.91T
Reserved 128G
Data + Parity 4 disks
Spare 0 disks
Log 0 disks
Cache 0 disks
Device Status
0 errors
No device faults have been detected in the storage pool.

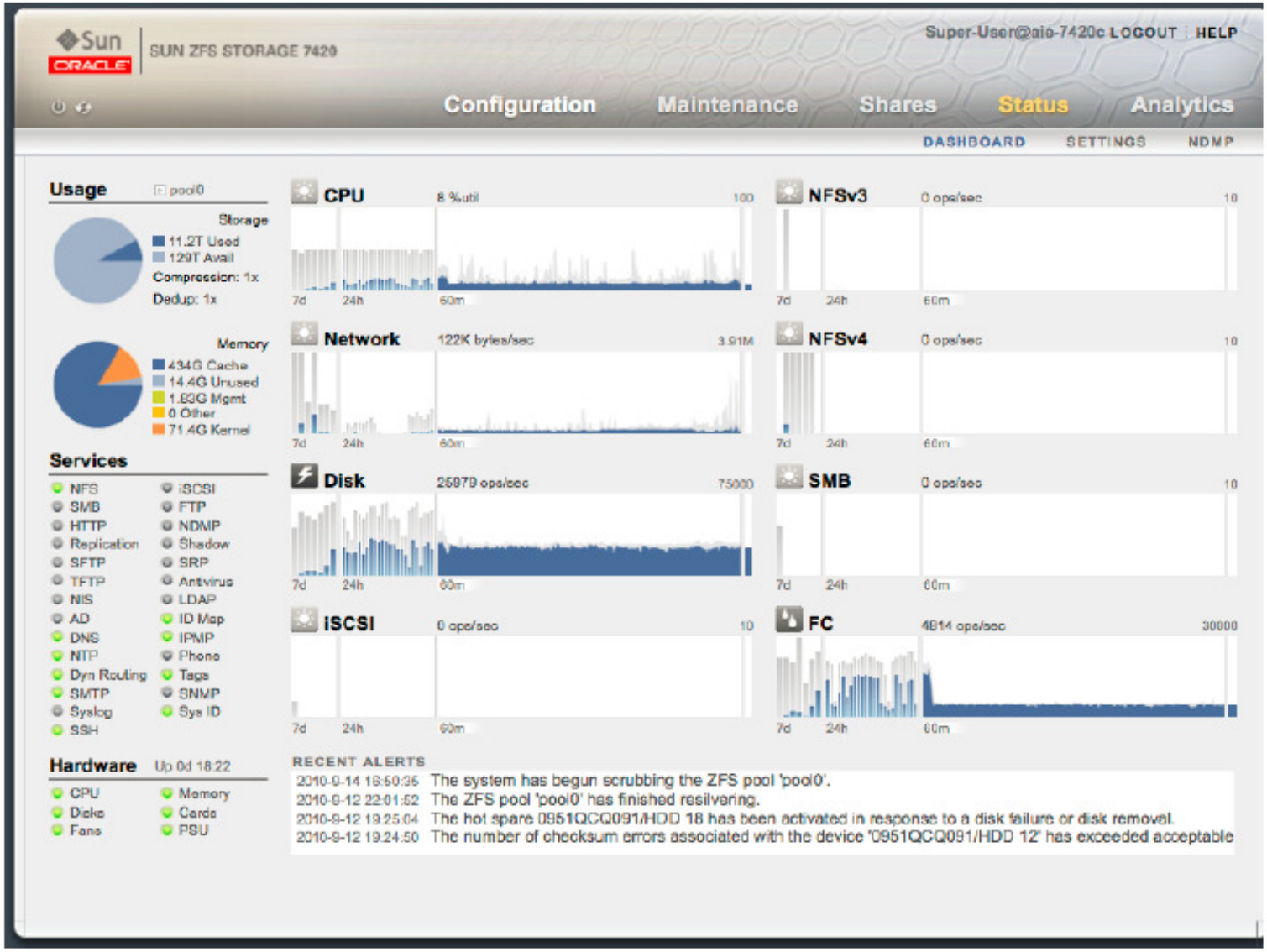
# ZFS BUI


Configuration
Maintenance
Shares
Status
Analytics
HARDWARE
SYSTEM
PROBLEMS
LOGS
WORKFLOWS
Alerts 119 Total
100-119
ALERTS FAULTS SYSTEM AUDIT PHONE HOME
TIME EVENT ID DESCRIPTION TYPE
2012-9-24 15:29:46 63714813-695f-c125-f88e-e434ebed2f7d The system has finished scrubbing the ZFS pool 'GENERIC'. Minor Alert
2012-9-24 15:29:46 a6838d57-8ee4-43d2-e42f-c695e62ccb0e The system has begun scrubbing the ZFS pool 'GENERIC'. Minor Alert
2012-9-24 15:14:54 4ada53dd-7124-cfc6-dbd1-c279f717d381 The system has finished scrubbing the ZFS pool 'RMANBACK'. Minor Alert
2012-9-24 15:14:53 8e22aee9-a6b4-4c79-cbf9-f61bb1b5fe8d The system has begun scrubbing the ZFS pool 'RMANBACK'. Minor Alert
2012-9-24 14:23:44 2d5106de-ee58-c299-c247-8882df53ffb7 Network connectivity via datalink ixgbe0 has been established. Minor alert
2012-9-24 14:23:44 0a2e7265-49bf-cb50-e280-d1812ff449d1 Full IP connectivity via interface ixgbe0 has been established. Minor alert
2012-9-24 14:23:44 cd81ccf9-8ee1-eb79-f46e-9e86513c2ad3 Network connectivity via port ixgbe0 has been established. Minor alert
2012-9-24 14:23:30 985892eb-6a10-653d-c73a-d901f91f5443 Network connectivity via datalink ixgbe0 has been lost. Major alert
2012-9-24 14:23:30 0d81abd7-c431-e3b4-835f-cfcc01170dac IP connectivity via interface ixgbe0 has been lost due to link-based failure. Major alert
2012-9-24 14:23:30 b979b7b9-9129-e2d5-ae44-b5bc6bc3c1ae Network connectivity via port ixgbe0 has been lost. Minor alert
2012-9-24 14:23:16 78d4a9b8-5664-44a9-afd7-d8eab505b33a Full IP connectivity via interface ixgbe2 has been established. Minor alert
2012-9-24 14:23:15 d8a8d18b-346c-665e-c9af-acef6acdd23c Network connectivity via datalink ixgbe2 has been established. Minor alert
2012-9-24 14:23:15 b55569fb-330b-496a-a619-cd30001473de Network connectivity via port ixgbe2 has been established. Minor alert
2012-9-24 14:23:10 9022ff22-7be1-e65c-f929-da96173fa21f IP connectivity via interface ixgbe2 has been lost due to link-based failure. Major alert
2012-9-24 14:23:10 d70af351-ca2a-cb6d-8a54-b6e9f1366c8b Network connectivity via datalink ixgbe2 has been lost. Major alert
2012-9-24 14:23:10 01c8f48b-06a9-c95c-d560-efe98a944f39 Full IP connectivity via interface ixgbe2 has been established. Minor alert
2012-9-24 14:23:10 2246e904-22ad-4a40-ca2c-d5f5b2d357ec Network connectivity via port ixgbe2 has been lost. Minor alert
2012-9-24 14:23:10 ddcc68fb-eaef-4b7f-83a4-9ca3e75d0543 Network connectivity via datalink ixgbe2 has been established. Minor alert
2012-9-24 14:23:10 de514e43-5839-6b56-92a3-e31a44caeb06 Network connectivity via port ixgbe2 has been established. Minor alert
2012-9-24 14:23:10 68f550f6-d4f2-c76e-ea2b-babf8d03c455 IP connectivity via interface ixgbe2 has been lost due to link-based failure. Major alert

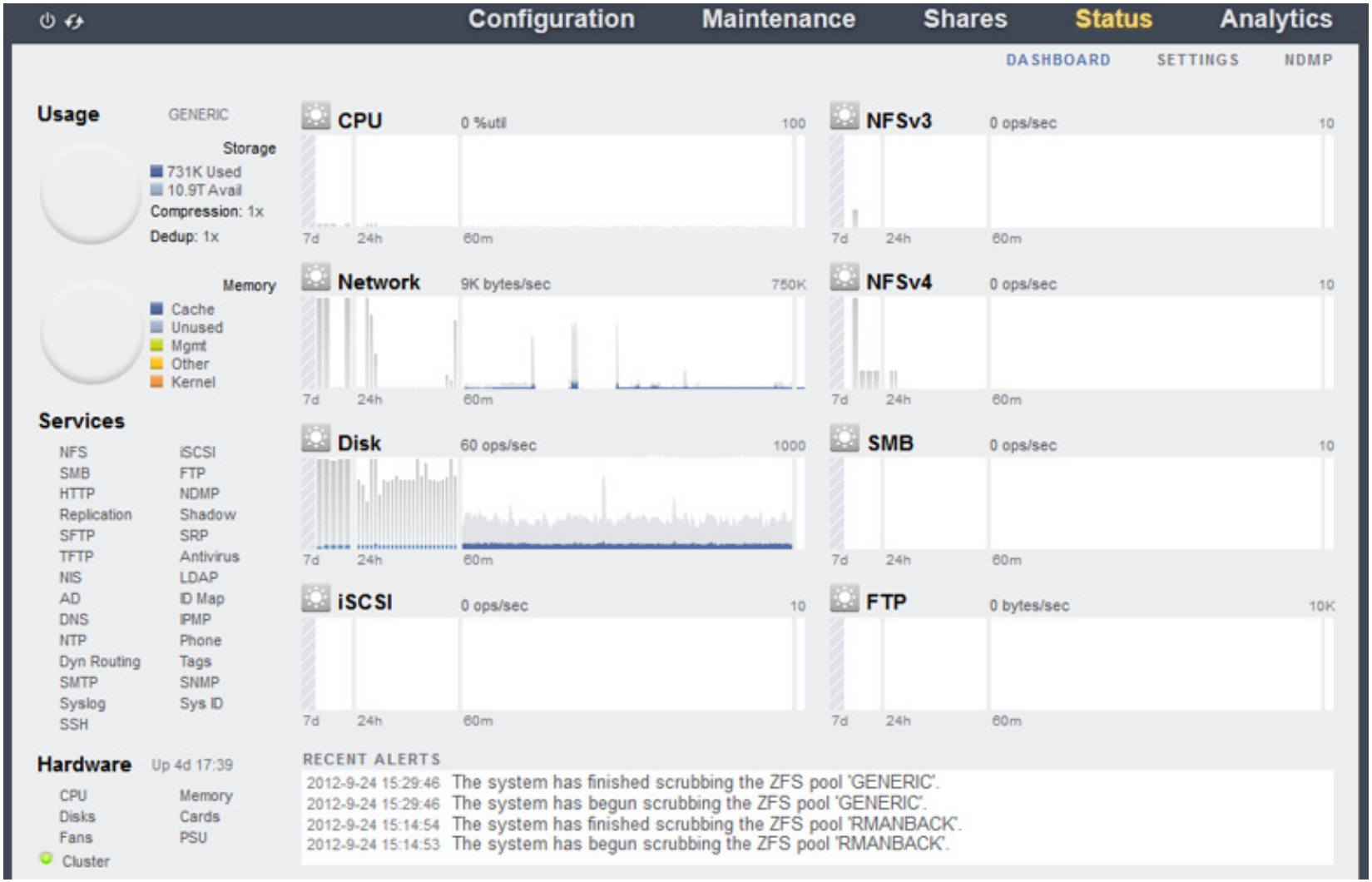
# ZFS BUI


Configuration
Maintenance
Shares
Status
Analytics
SERVICES
STORAGE
NETWORK
SAN
CLUSTER
USERS
PREFERENCES
ALERTS
Network
Configuration
Addresses
Routing
To configure networking, build Datalinks on Devices, and Interfaces on Datalinks. Click on a pencil icon to edit object properties. Select an object to view its relationship to other objects. Drag objects to extend Aggregations or IP Multipathing Groups.
REVERT
APPLY
Devices
12 total
BUILT-IN
igb0 1Gb (full)
igb1 1Gb (full)
igb2 link down
igb3 link down
PCIe 3
ixgbe0 10Gb (full)
ixgbe1 link down
PCIe 6
ixgbe2 10Gb (full)
ixgbe3 link down
PCIe 7
ibp2 port down
ibp3 port down
PCIe 2
ibp0 port down
ibp1 port down
Datalinks
4 total
igb0 via igb0
igb1 via igb1
ixgbe0 Custom MTU(9000), via ixgbe0
ixgbe2 Custom MTU(9000), via ixgbe2
Please wait...
Interfaces
4 total
head1 net0 IPv4 static, 192.168.40.248/22, via igb0
head2 net1 IPv4 static, 192.168.40.249/22, via igb1
private10gb IPv4 static, 10.221.112.49/24, via ixgbe0
private10gb2 IPv4 static, 10.221.112.50/24, via ixgbe2

# ZFS Storage Appliances



# ZFS BUI


Configuration
Maintenance
Shares
Status
Analytics
DASHBOARD
SETTINGS
NDMP
Usage
GENERIC
Storage
731K Used
10.9T Avail
Compression: 1x
Dedup: 1x
Memory
Cache
Unused
Mgmt
Other
Kernel
Services
NFS
SMB
HTTP
Replication
SFTP
TFTP
NIS
AD
DNS
NTP
Dyn Routing
SMTP
Syslog
SSH
iSCSI
FTP
NDMP
Shadow
SRP
Antivirus
LDAP
ID Map
IPMP
Phone
Tags
SNMP
Sys ID
Hardware
Up 4d 17:39
CPU
Disks
Fans
Cluster
Memory
Cards
PSU
CPU
0 %util
100
NFSv3
0 ops/sec
10
Network
9K bytes/sec
750K
NFSv4
0 ops/sec
10
Disk
60 ops/sec
1000
SMB
0 ops/sec
10
iSCSI
0 ops/sec
10
FTP
0 bytes/sec
10K
7d
24h
60m
RECENT ALERTS
2012-9-24 15:29:46 The system has finished scrubbing the ZFS pool 'GENERIC'.
2012-9-24 15:29:46 The system has begun scrubbing the ZFS pool 'GENERIC'.
2012-9-24 15:14:54 The system has finished scrubbing the ZFS pool 'RMANBACK'.
2012-9-24 15:14:53 The system has begun scrubbing the ZFS pool 'RMANBACK'.

# How Does This Change Our Jobs?

| Job Title | Loses | Gains |
|---|---|---|
| **Storage Admins** | Time wasted monitoring competing loads on the storage appliance balancing competing need to read/write cache, and allocation of disk. | More efficient storage environment as it is all file system. |
| **Network Admins** | Pain and suffering | Time to devote to troubleshooting, security monitoring, and other value-added tasks. |
| **System Admins** | ▪Gives up appliance root password<br>▪Gives up 2:00am support calls | |
| **Database Admins** | | Patching operating system, firmware, and database as a single unit with patches previously tested for compatibility |

**Your ODA is not a general purpose computer, will not be hosting files, applications, middleware, etc.**

# How Does This Change Our Jobs?

- Storage Admin
  - No longer required
- Network Admin
  - Only required for public network interface
- System Admin
  - Advise on configuration
  - Install backup agent (ie Networker)
  - Install security software (ie TripWire)
- DBA
  - Just like with ASM ... assumes broader responsibility for deployment and patching
  - Gives up large amounts of unproductive time debugging configurations

# Questions

**ERROR at line 1:**
**ORA-00028: your session has been killed**

Thank you