No, its not the Magnet: 

*Surviving Data Disasters on AFS*

Jed Dobson

James.E.Dobson@Dartmouth.EDU

Dartmouth College
Hanover, NH
October 26th, 2003
AFS & Dartmouth

- Ivy League College in New Hampshire
- 3800 Students, 1200 Grad Students
- Currently 5 Cells:
  - Thayer, Northstar, SBS, fMRIDC, DBIC
- AIX, Solaris and Linux Servers

Dartmouth Brain Imaging Center
- About 150 Workstations
- 3 Servers (both DB & FS, soon 6)
- Large partitions (.5TB)
- 281 Volumes
- Avg Vol: 7.5GB
- 74% full (2.8TB)
fMRI Data Needs

- Thousands of files
- Very small files (33% are 348 Bytes)
- Very Sequential Access
- Access from multiple machines
- Share data with multiple users
- Long(ish) term archival
Typical User Volume

At “screw” level:
- 41GB bytes
- 395 dirs
- 465,620 files

Entire volume
- 73GB bytes
- 1,047 dirs
- 654,409 files

avg file: ~100k
<table>
<thead>
<tr>
<th>System Volumes</th>
<th>User Volumes</th>
<th>Protocol Volumes</th>
<th>Inboxes</th>
</tr>
</thead>
<tbody>
<tr>
<td>usr.local, afsws</td>
<td>home directories</td>
<td>project volumes</td>
<td>raw data</td>
</tr>
<tr>
<td>Several GB</td>
<td>40G Avg</td>
<td>20G Avg</td>
<td>10G Max</td>
</tr>
<tr>
<td>Few (10-15)</td>
<td>Around 80</td>
<td>Will be around 800</td>
<td>80</td>
</tr>
<tr>
<td>Critical</td>
<td>Not</td>
<td>Not (well mine!)</td>
<td>Critical</td>
</tr>
<tr>
<td>To local disks</td>
<td>We don’t</td>
<td>We don’t</td>
<td>To HSM</td>
</tr>
</tbody>
</table>
What Problems have we had?

- Server crashes
- Power Outages
- Massive RAID failures (loss of parity)
- Users deleting files
- Salvager weirdness
- vfsck weirdness
Cause for Data Loss?

- Salvager removed data (200G)
- Only on large volumes
- Lots of files
- Inode fileserver (1.2.x->.9)
- Inbox (raw data) destroyed
- User(s) didn’t have backups
What have we done?

- Namei Fileserver!
- RAID-5 with hotspares
- Redundant FC/SCSI controllers
- Mirrored boot devices
- Enabled UFS Logging
- Replaced Hardware
- Replicated all critical volumes
- Backup volumes for all users (snapshots)
- UPS on all servers & disks
Attempts to Recover...

volinfo -saveinode
- Creates too many files in one dir
- Caused awk/sh to core dump
- Most of my files look alike

voldump
- Cool tool written by Nathan Neulinger
- Dump volume from partition
- Wasn’t able to recover data
- Keep in my toolbox
More Attempts to Recover

- Scan salvager log (1GB!)
- Try to convert to inode
- icat/unrm (tct)
- No luck (files hard to identify)
- Virtualization/Abstract hard to understand at 2AM
- Feel more comfortable with namei
Replication...

- Follow the paths
- Prevent single server from hanging all clients
- RO replicas great for apps
- RW "replica" created with vos dump |vos restore
- Read-only install issues
- Forgot to modify some scripts...

/afs/dbic.dartmouth.edu/i386_linux24/usr/afs(ws
/afs/dbic.dartmouth.edu/images/Neuroimaging434/image/etc
Snapshot Volumes

- “Backup” gives wrong idea
- Mounted as .snapshot in all user volumes
- Was concerned about disk space, few crashes later – plenty of space
- Life-saver for careless users
- Multiple versions desired by users
Backups

- Limited $ for centralized solution (or interest)
- We provide system volume backups
- Inbox data (raw) archived to HSM *
- Critical vols dumped to local FS (root.*, system vols)
- We archive from *.backup volumes
- Volumes are too large for single tape (many)
- Previous policy used tar (often bad)
- Some uses archive nightly to local disk
- Some users use tape or cds
- Some users don’t care

* Where is HSM support at for AFS?
Our Servers

Sun Fire V240
T3 Array
18x73G, 2 HS

Darmouth Public Network

OoB Management Network

/vicepb
/vicepa

1.02GHz
Our Servers, cont.

- Sun Fire V240s (File Servers)
- Sun Fire V120 (Future DB, old FS/DB)
- 2x36G mirrored boot disks
- Redundant FC & Network
- Solaris 9 04/03
- OpenAFS 1.2.9 (1.2.10 shortly)
- Few kernel tweaks
- No remote access (console ts only)
The Magnet
Contact Information

James.E.Dobson@Dartmouth.EDU
HB 6162 Moore Hall
Hanover, NH 03755
(603) 646-9324