ZFS
(Zettabyte File System)

Shawn Wallbridge
Senior Systems Administrator - Frantic Films

What is ZFS?

584 files

92,000 changed lines of code

56 patents

5 years

http://blogs.sun.com/bonwick/entry/zfs_the_last_word_in

What is ZFS?

- New filesystem in Solaris 10 (6/06 update)
- Pooled storage
- Transaction based
- ‘Self-Healing’ with Checksums
- Snapshots
- Scalable...

Scalable

Some theoretical limits in ZFS are:

- 2^{48} — Number of snapshots in any file system (2×10^{14})
- 2^{48} — Number of files in any individual file system (2×10^{14})
- 16 exabytes (2^{64} byte) — Maximum size of a file system
- 16 exabytes (2^{64} byte) — Maximum size of a single file
- 16 exabytes (2^{64} byte) — Maximum size of any attribute
- 3×10^{23} petabytes — Maximum size of any zpool
- 2^{56} — Number of attributes of a file (actually constrained to 2^{48} for the number of files in a ZFS file system)
- 2^{56} — Number of files in a directory (actually constrained to 2^{48} for the number of files in a ZFS file system)
- 2^{64} — Number of devices in any zpool
- 2^{64} — Number of zpools in a system
- 2^{64} — Number of file systems in a zpool

http://en.wikipedia.org/wiki/Zetabyte_File_System

Scalable

256 Quadrillion ZetaBytes

256,000,000,000,000,000,000,000,000 ZetaBytes
1ZB = 1,000, 000, 000, 000, 000, 000,000 Bytes
1ZB = 1,000, 000, 000, 000, 000 MB
1ZB = 1,000, 000, 000, 000 GB
1ZB = 1,000, 000, 000 TB

Scalable

Although we'd all like Moore's Law to continue forever, quantum mechanics imposes some fundamental limits on the computation rate and information capacity of any physical device. In particular, it has been shown that 1 kilogramme of matter confined to 1 litre of space can perform at most 10^{51} operations per second on at most 10^{31} bits of information [see Seth Lloyd, "Ultimate physical limits to computation." Nature 406, 1047-1054 (2000)]. A fully populated 128-bit storage pool would contain 2^{128} blocks = 2^{137} bytes = 2^{140} bits; therefore the minimum mass required to hold the bits would be (2^{140} bits) / (10^{31} bits/kg) = 136 billion kg.

To operate at the 10^{31} bits/kg limit, however, the entire mass of the computer must be in the form of pure energy. By $E=mc^2$, the rest energy of 136 billion kg is 1.2×10^{28} J. The mass of the oceans is about 1.4×10^{21} kg. It takes about 4,000 J to raise the temperature of 1 kg of water by 1 degree Celsius, and thus about 400,000 J to heat 1 kg of water from freezing to boiling. The latent heat of vaporization adds another 2 million J/kg. Thus the energy required to boil the oceans is about 2.4×10^{6} J/kg * 1.4×10^{21} kg = 3.4×10^{27} J. Thus, fully populating a 128-bit storage pool would, literally, require more energy than boiling the oceans.

http://en.wikipedia.org/wiki/Zetabyte_File_System

Much ado about ZFS

Sun

microsystems

Much ado about ZFS

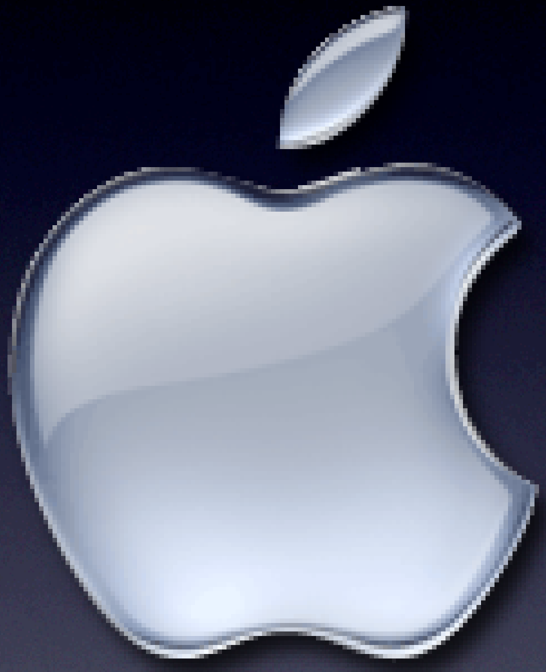
solaris™

Much ado about ZFS

opensolaris™

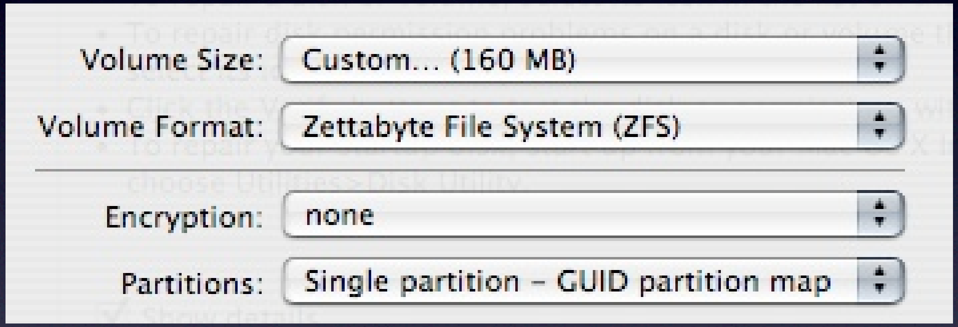
Much ado about ZFS

Much ado about ZFS



?

Much ado about ZFS



Volume Size: Custom... (160 MB)

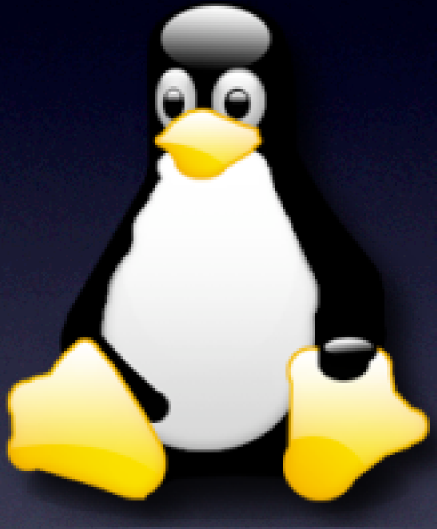
Volume Format: Zettabyte File System (ZFS)

Encryption: none

Partitions: Single partition – GUID partition map

Posted to the etherweb, which makes it true.

Much ado about ZFS



FUSE

Much ado about ZFS

February 2006

March 2006

The Best File System in the World?

http://www.samag.com/documents/s=9950/sam0602j/0602j.htm

Why Frantic uses it.

Sun X4500

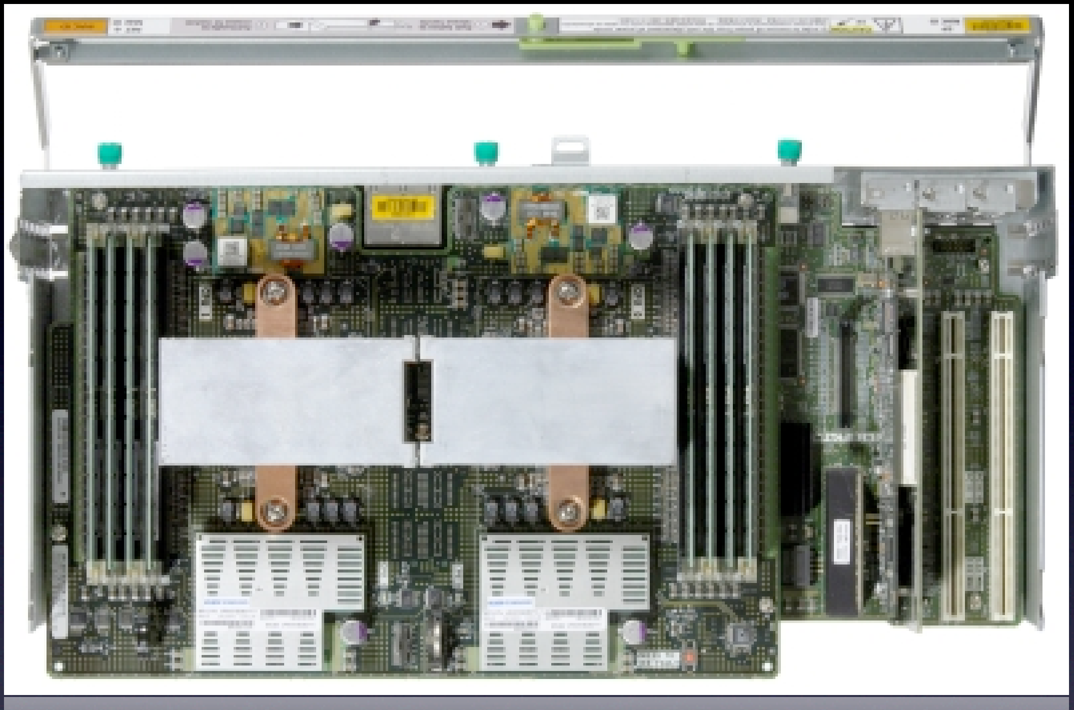
Sun X4500

- Dual Dual Core Opteron 285 (2.6GHz)
- 16GB Ram
- 4 GigE ports
- 6 S-ATA Controllers
- 48 500GB HD's
- 110 lbs

Sun X4500

- Dual Dual Core Opteron 285 (2.6GHz)
- 16GB Ram
- 4 GigE ports
- 6 S-ATA Controllers
- 48 500GB HD’s
- 110 lbs

Sun X4500



Sun X4500

How we use it.

- Frantic will have 72TB of storage on ZFS
- Three Sun X4500's with 24TB each
- Massive amounts of storage for reasonable cost
- Changed the way we looked at storage

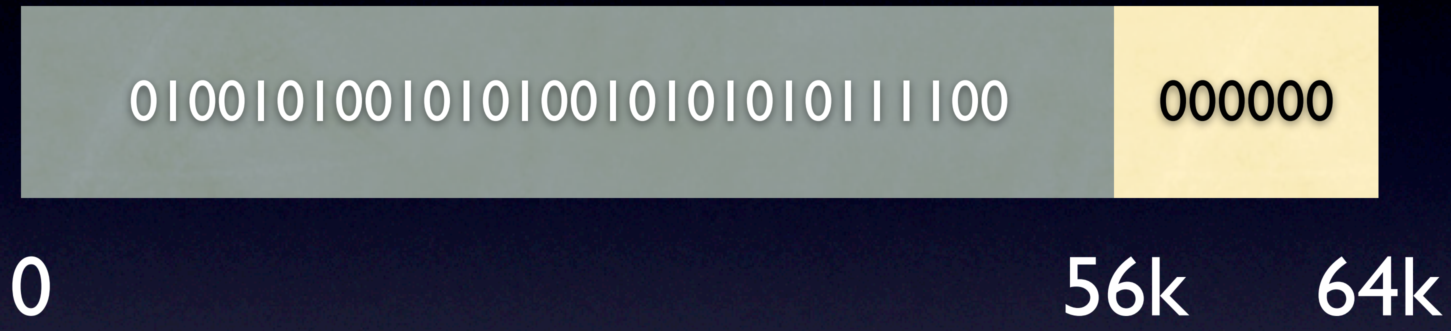
How I (plan) to use it.

- I have about 4TB of FC disks over 100 spindles (it’s loud)
- 9 large UltraSPARC machines on a SAN with NetApp, Dell, and Sun FC Arrays
- GIS and large database projects

Technical details

- Metadata allocated dynamically
- Uses 256bit checksums to ensure data integrity
- Adaptive Endianness
- Copy-on-Write
- NFSv4/NT Style ACLs
- Variable Stripe Size - No RAID5 ‘write-hole’

RAID 5 ‘write-hole’



010010100101010010101010111100

000000

0

56k

64k

- Write 56k of data
- Read 8k of old data
- Calculate parity
- Write parity

Limitations

- Booting from ZFS is not supported, yet
- No file-system level encryption, yet
- No per user/group quotas, yet

Getting Started

- Get Solaris or OpenSolaris
- You can use files instead of disks
- ‘Disks’ or files must be 128MB
- *zfs*(1M) and *zpool*(1M)

zpool

- Used to create/manage pools
- *zpool create $pool_name [mirror|raidz|raidz2] c1t0d0 c2t0d0 ...*

zfs

- Used to create/manage filesystems
- *zfs create $poolname/$filesystem*

DEMO

ZFS and Zones

- Zones can inherit a ZFS filesystem

ZFS and Zones

- Zones can inherit a ZFS filesystem
- Clones are ideal for Zones

ZFS Resources

- Solaris ZFS Administrators Guide

  http://docs.sun.com/app/docs/doc/819-5461

- OpenSolaris ZFS Community

  http://www.opensolaris.org/os/community/zfs/

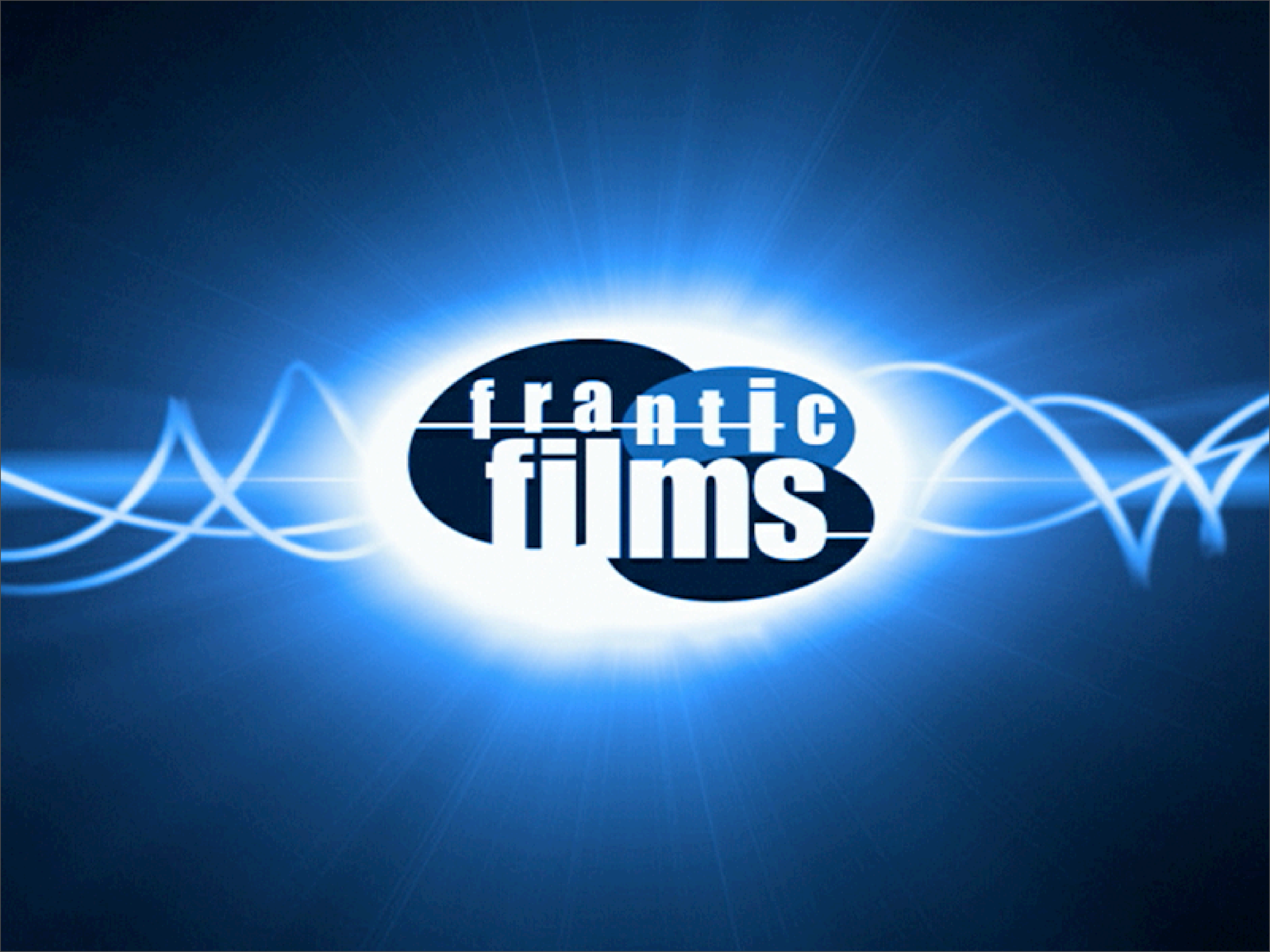
- OpenSolaris ZFS-Discuss mailing list

  zfs-discuss-subscribe@opensolaris.org

- Solaris Internals ZFS Best Practices Guide

  http://www.solarisinternals.com/wiki/index.php/ZFS_Best_Practices_Guide

Questions?


frantic
films

frantic films

NOW HIRING