

WISE: Big Data, Little Money - Lessons Learned

TIM CONROW
WISE System Architect

GRITS III

June 17, 2011



The Problem

* Heavy ops processing load

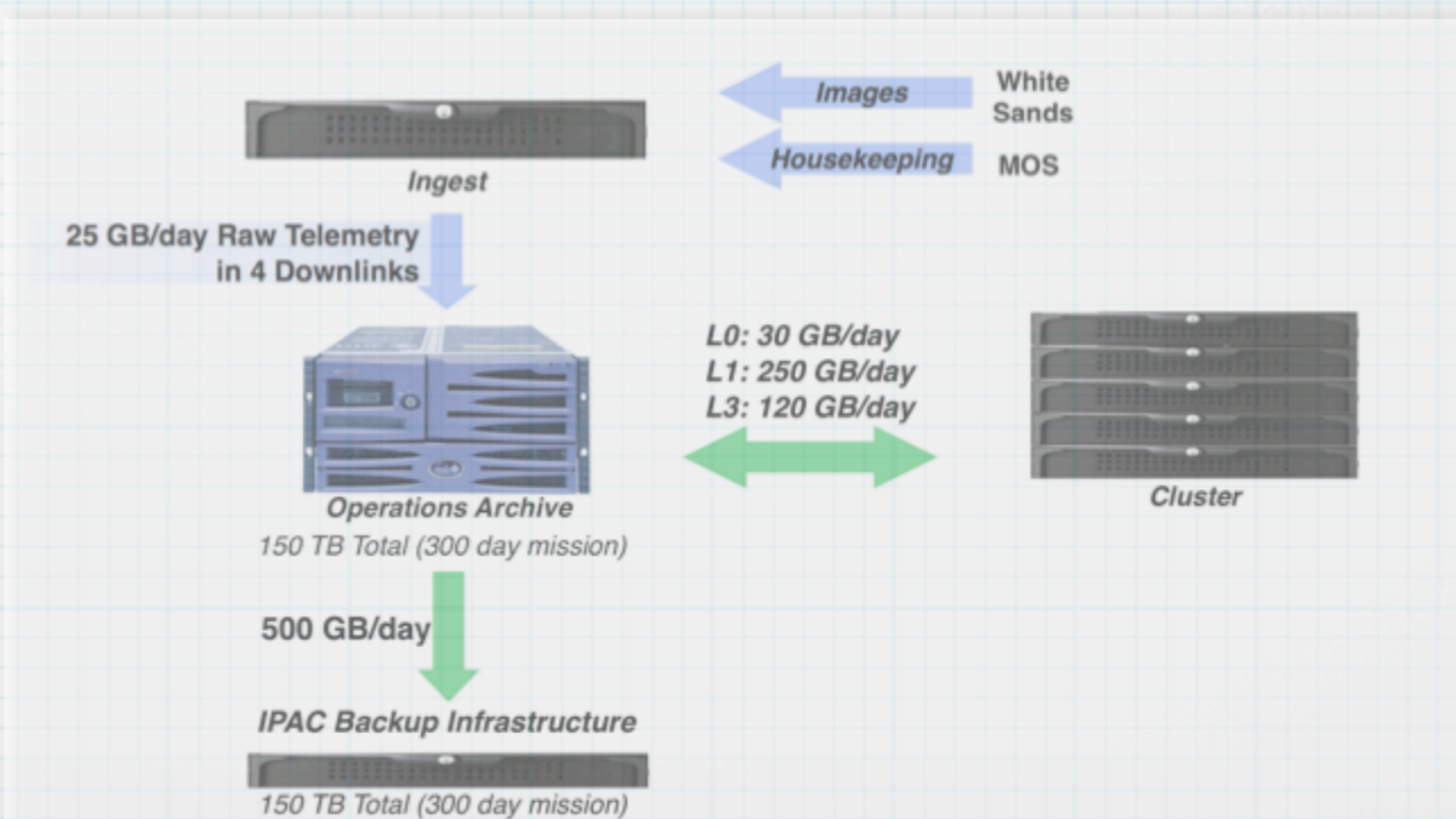
- R/W > 3TB/day, sustained 1Gbit/s network load
- 180 TB ops archive (non-IRSA)

* Compressed budget and schedule

- < \$400k for ops system at launch
- Major hardware purchase at L-6 mo.s
- Simultaneous reprocessing

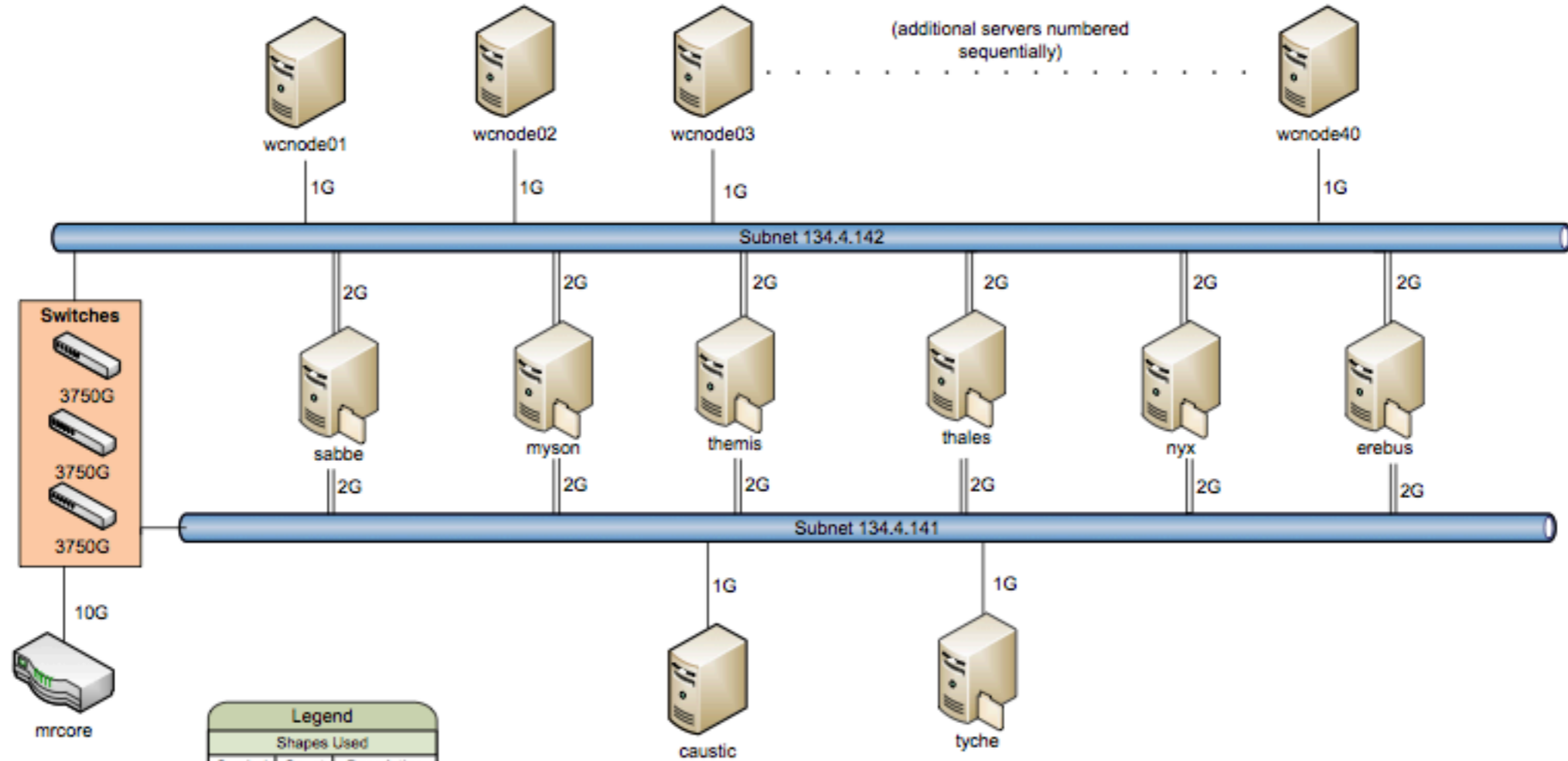
The Solution?

- * Sun servers fronting JBOD's, S/W RAID-Z
- * Local network mediated by a Cisco switch stack (trunked 1Gbit/s to servers)
- * 32 node compute cluster of 8-core Xeon servers
 - use cluster-local storage to off-load network
- * Backup via separate i/f



(The Diagram)

Production Machines



Legend		
Shapes Used		
Symbol	Count	Description
	1	Router
	2	Ethernet
	5	Server
	3	Switch
	7	File Server

IPAC Systems Group (ISG)
 M.Harbut, W. Burt 11/12/2009 Version .1

(The Details)

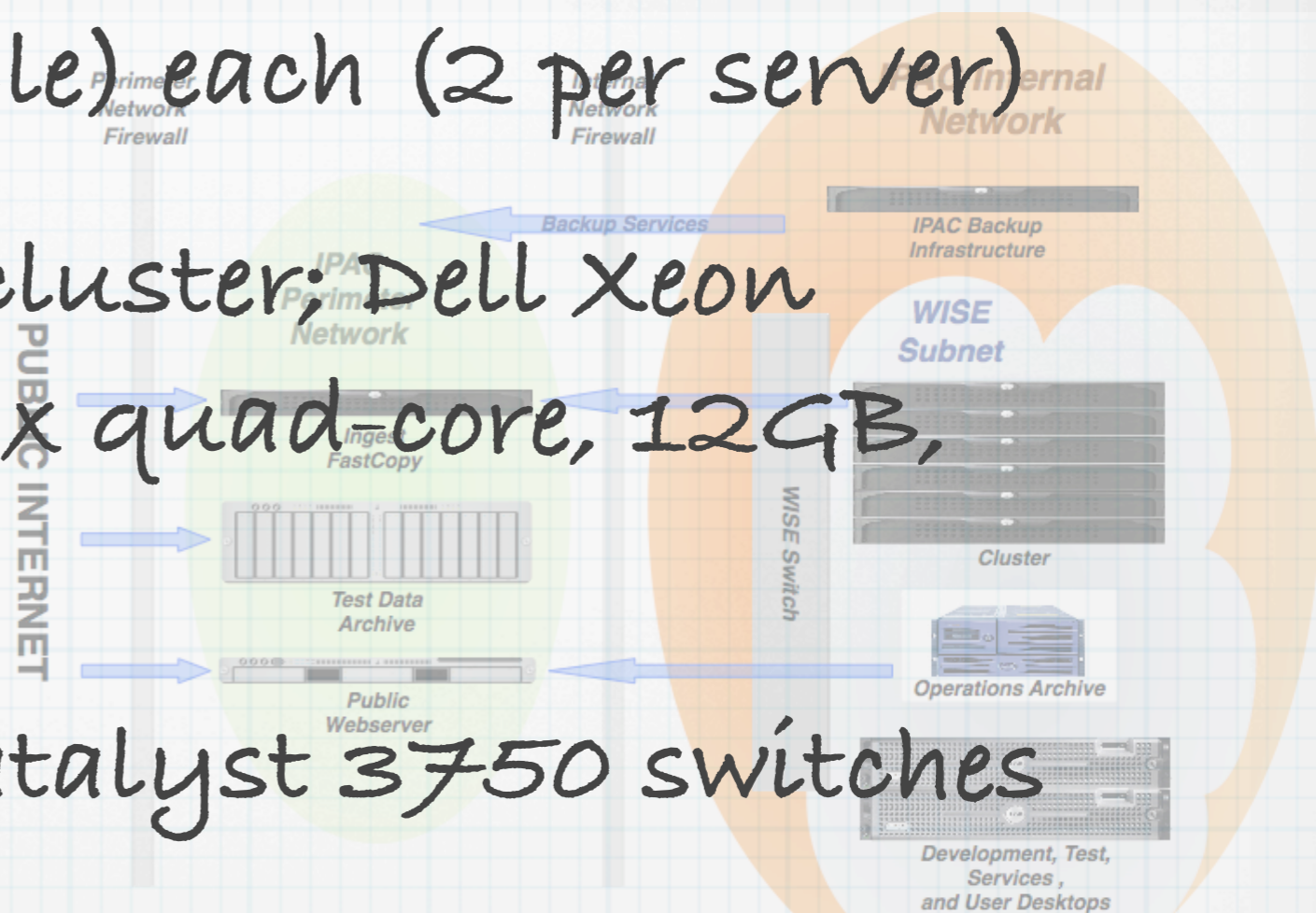
* 5 SUN X4150 Servers (4 ops + spare)

* 10 SUN J4400 JBOD's, SAS+RAID-Z, 24TB (18TB usable) each (2 per server)

* 32 node compute cluster; Dell Xeon PE1950, R610 (2 x quad-core, 12GB, 500GB)

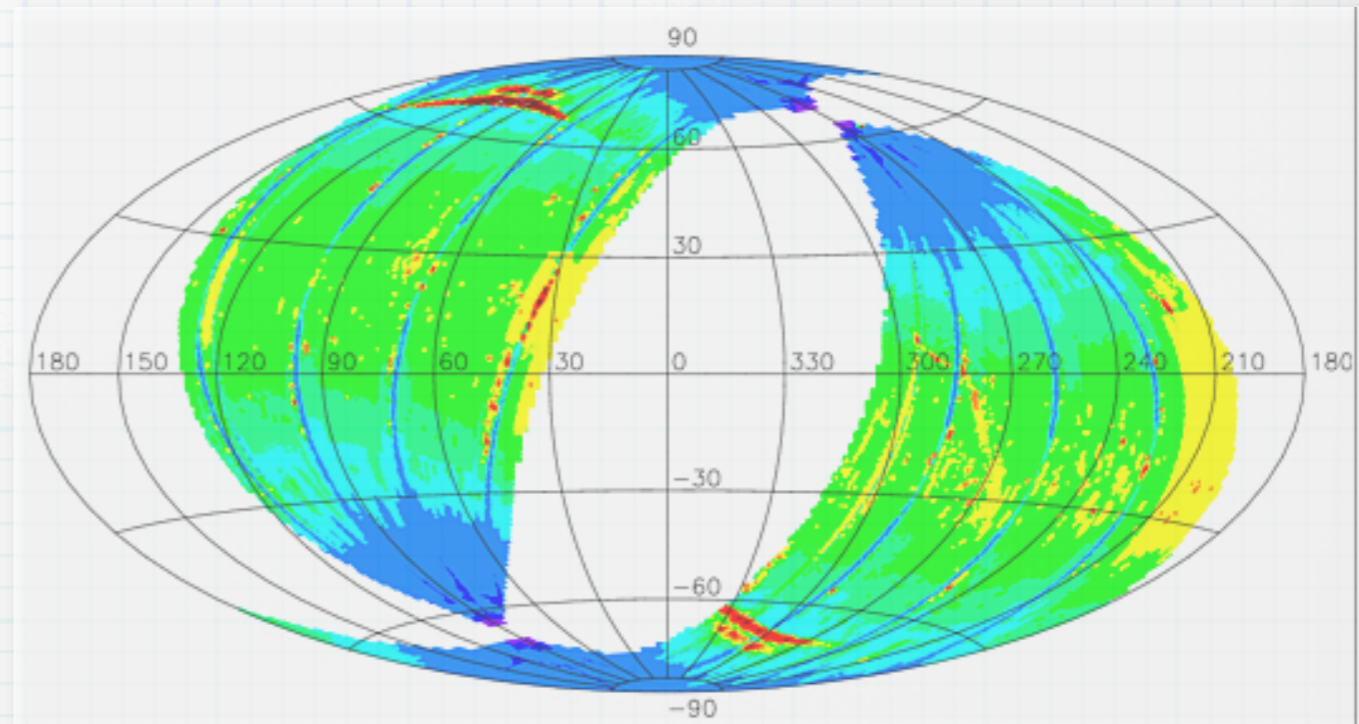
* 2 Cisco 48-port Catalyst 3750 switches

* RHE4, Solaris, ZFS, Condor, Ganglia, NFS3

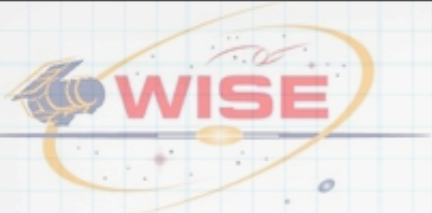


The Result

- * Supported required throughput
- * No data loss (Yay!)
- * Met processing goals



- * **BUT**, considerable time and effort spent resolving H/W issues, which we would rather have spent elsewhere



The Blowback

* Collision/Retran cascades => network stalls

The HBA firmware is the first thing to update; the specific bug (663190) is identified with the following error:

```
Apr 6 03:14:02 myx local0: [ID 101823] kern.warning: WARNING: /pci00.0/pci8086.25a
202/p2.000,350000/pci1000,01800/pci1000,31000/pci090,0 (sc221)
Apr 6 03:14:02 myx      Command failed to complete...Device is gone
```

- Traced to use of trunked network ports

The work-around is to reboot, and the disks come back online. The fix is to update the HBA firmware to 1.1a.03; your current HBA firmware is at 1.1a.00 (sysystemid 0x150 and the output of the HBA "prometheus", then find it's firmware version).

* Server drops/freezes under load

You can download 1.1a.03 (1.26.03 decimal) via:

http://www.lsillogic.com/support/sun/sa_xpci8sas_e_sPoHS.htm
http://www.lsillogic.com/support/sun/files/LSISAS1_update_14.2.2.zip

- Multiple rounds of patches and work-arounds from Sun (then Oracle)

If that does not resolve the problem, then the following related patches are out of date:

Storage	P2 mpt Patch	143129 missing (current -03): SunOS 5.10_x86: mpt patch
Storage	P2 mpt Patch	144018 missing (current -01): SunOS 5.10_x86: mpt.so.1 patch
Storage	P2 Load Patch	142258 missing (current -03): SunOS 5.10_x86: load patch
Storage	P2 scsi patch	144022 missing (current -01): SunOS 5.10_x86: kernel/mach/amd64/scsi patch

* Frequent spindle "failures" (really transient SCSI i/f errors)

...along with several kernel patches:

SunOS	P2 Kernel Patch	142901 missing (current -08): SunOS 5.10_x86: kernel patch
SunOS	P2 Kernel Patch	144111 missing (current -01): SunOS 5.10_x86: kernel/mach/amd64/pcplusmp patch
SunOS	P2 Kernel Patch	141535 missing (current -01): SunOS 5.10_x86: poll patch
SunOS	P2 Kernel Patch	137124 downrev (inst. -01, current -02): SunOS 5.10_x86: devfsadm, devalloc, and libbsm patch

- JBOD HBA update + other changes

If you want to update the mpt patches, find the mpt driver that is the driver used by this HBA.

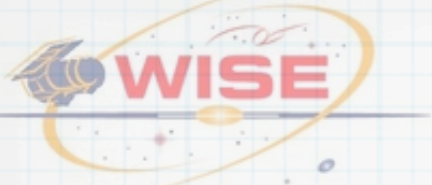


The Problem 2

Now, do it all again, 3X faster!

No time for
H/W issues.





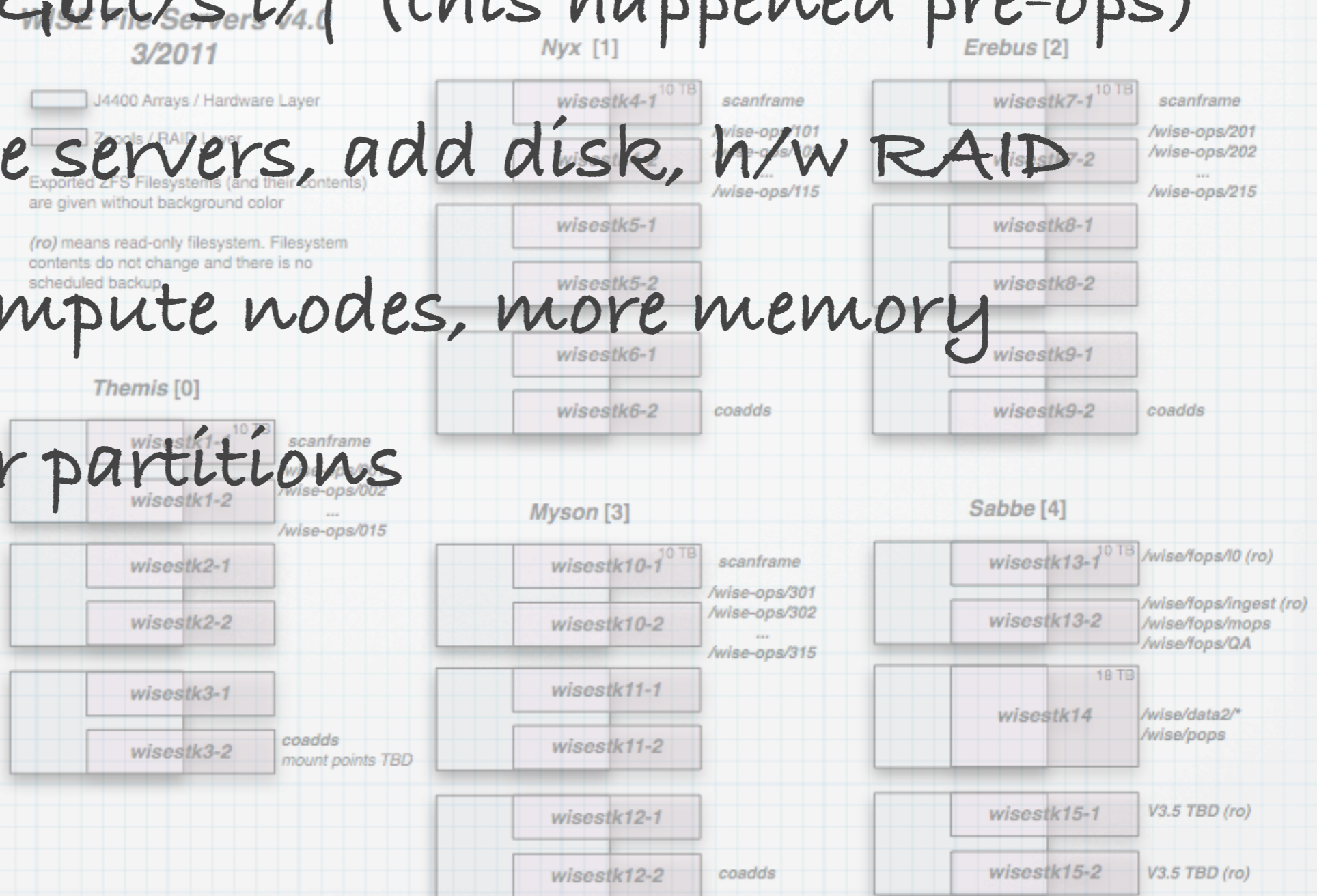
The Solution 2

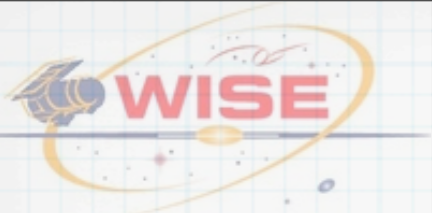
* Upgrade switches to replace trunked ports with 10Gbit/s i/f (this happened pre-ops)

* Upgrade servers, add disk, h/w RAID

* More compute nodes, more memory

* Smaller partitions





The Result 2

* Running at 3.3 ops days/day

- 7 mo. cryo mission done in 2 mo.s

* 15TB/day network I/O

* W/R 8TB, 15M files/day

* No problems

- no downtime
- no stress

* Yay!

* Thanks ops team!

* Thanks ISG!

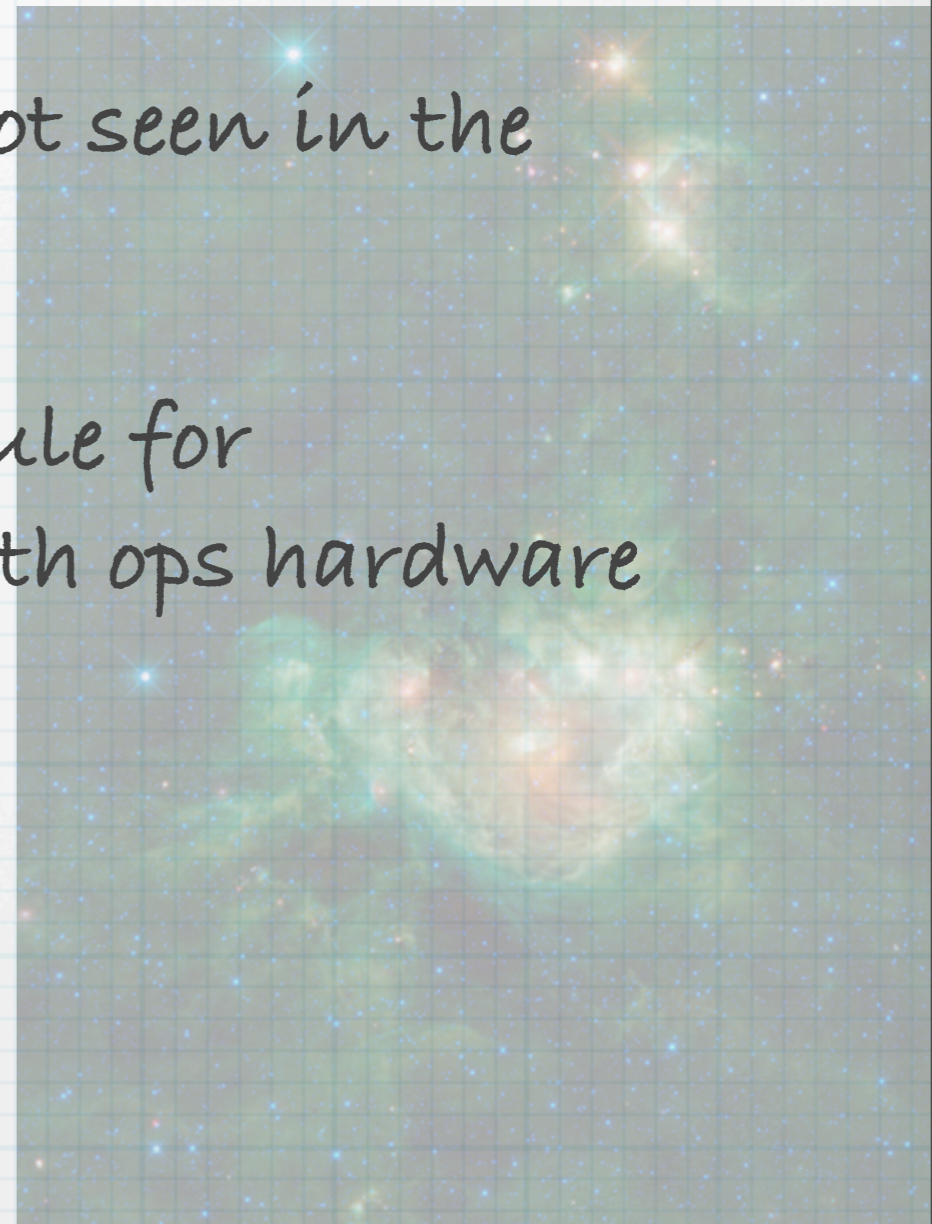




The Lessons

* Mind the server performance envelope

- Performance and configurations not seen in the wild could be risky
- Have a large test dataset and schedule for prolonged ops-like run scenarios with ops hardware
- Have a backup server!



The Lessons

* Keep it local and under control

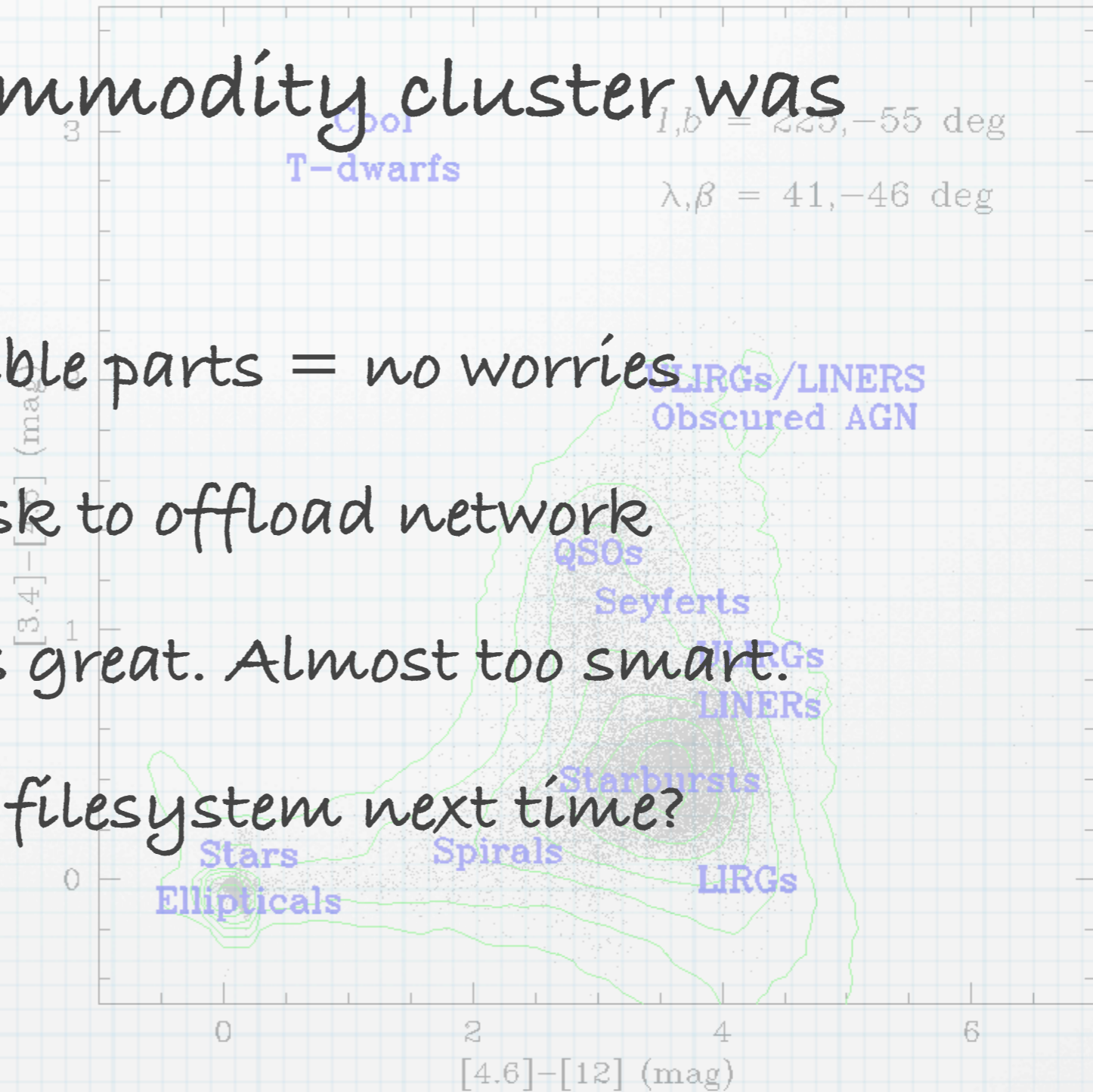
- Segregate from core router
- Local LDAP server or local ops accounts
- Restrict non-ops access
- Have good network and cluster monitoring and debugging tools



The Lessons

* The cheap commodity cluster was awesome

- Lots of fungible parts = no worries
- Used local disk to offload network
- Condor works great. Almost too smart.
- Try a cluster filesystem next time?



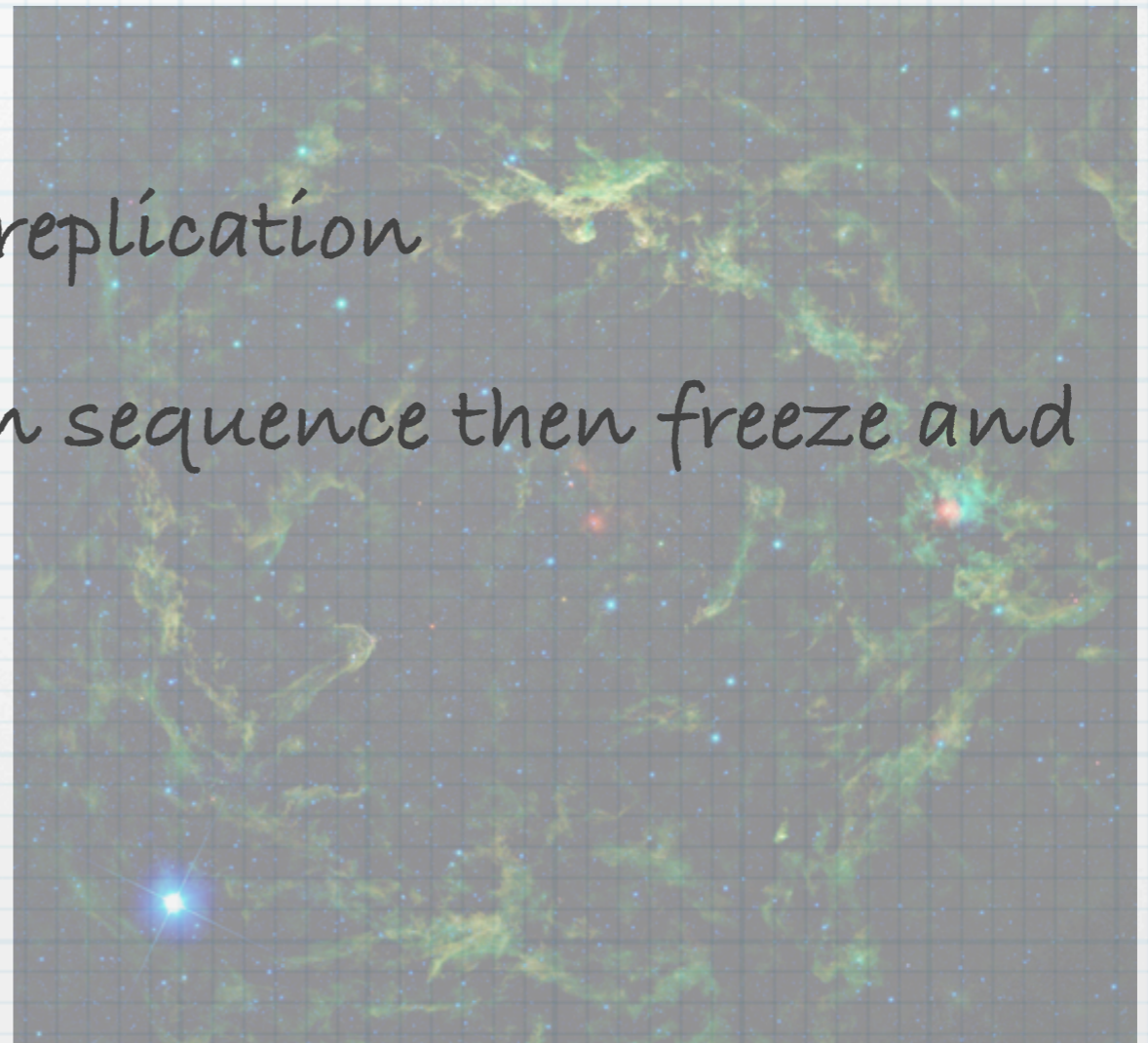
The Lessons

* Massive backups are massively painful

- Do you really want conventional tape backups?
Really?

- Regeneration and/or replication

- Fill smaller partitions in sequence then freeze and dump





The Lessons

* Who you gonna call?

- ISG, purchasing, vendor contacts on speed dial
- Management may need to elevate issue priority with vendor
- Local (or not so local) experts



The Lessons

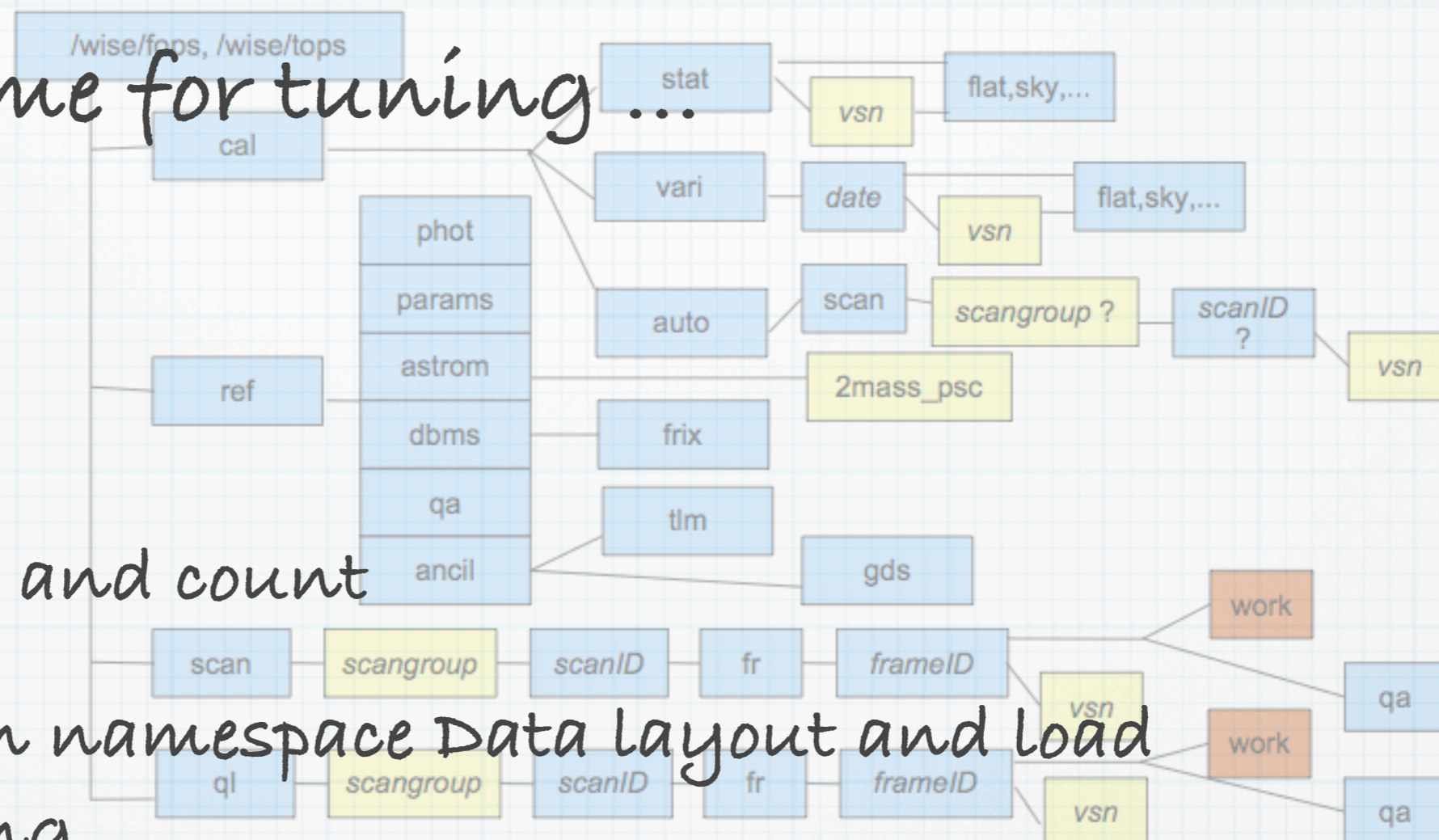
* Make time for tuning ...

- NFS
- LDAP

- File size and count

- Common namespace Data layout and load balancing

- Subnet layout; ops, dev, backup, desktop isolation





(The Future?)

- * Could a WISE-like mission work with computing resources shared among projects? The cloud?
- * Do missions with high/low data rates, operational intensity mix well?
- * What would have happened if our problems occurred on a multi-mission platform?
- * QOS contracts at what cost?

