WISE: Big Data, Little Money - Lessons Learned

Tim Conrow

WISE System Architect

GRITS III

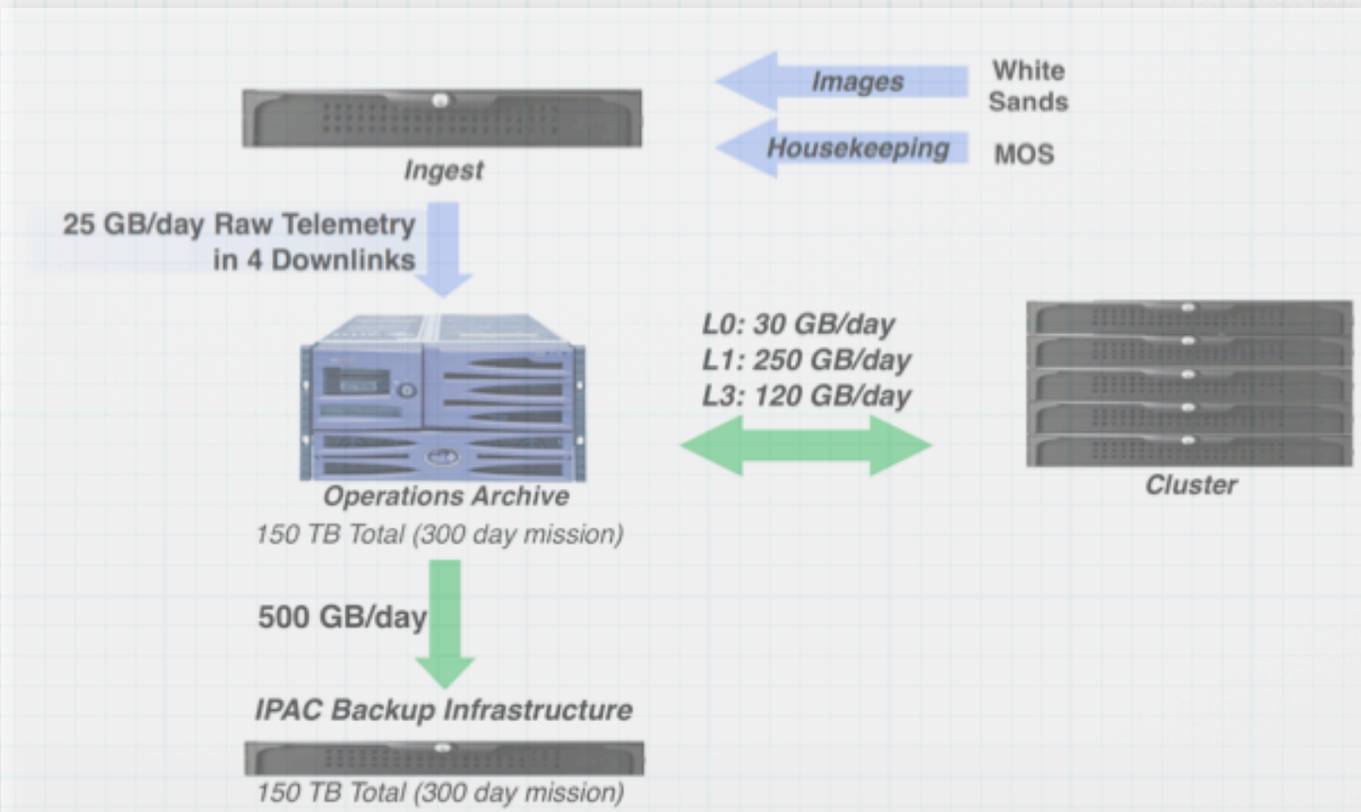
June 17, 2011

The Problem

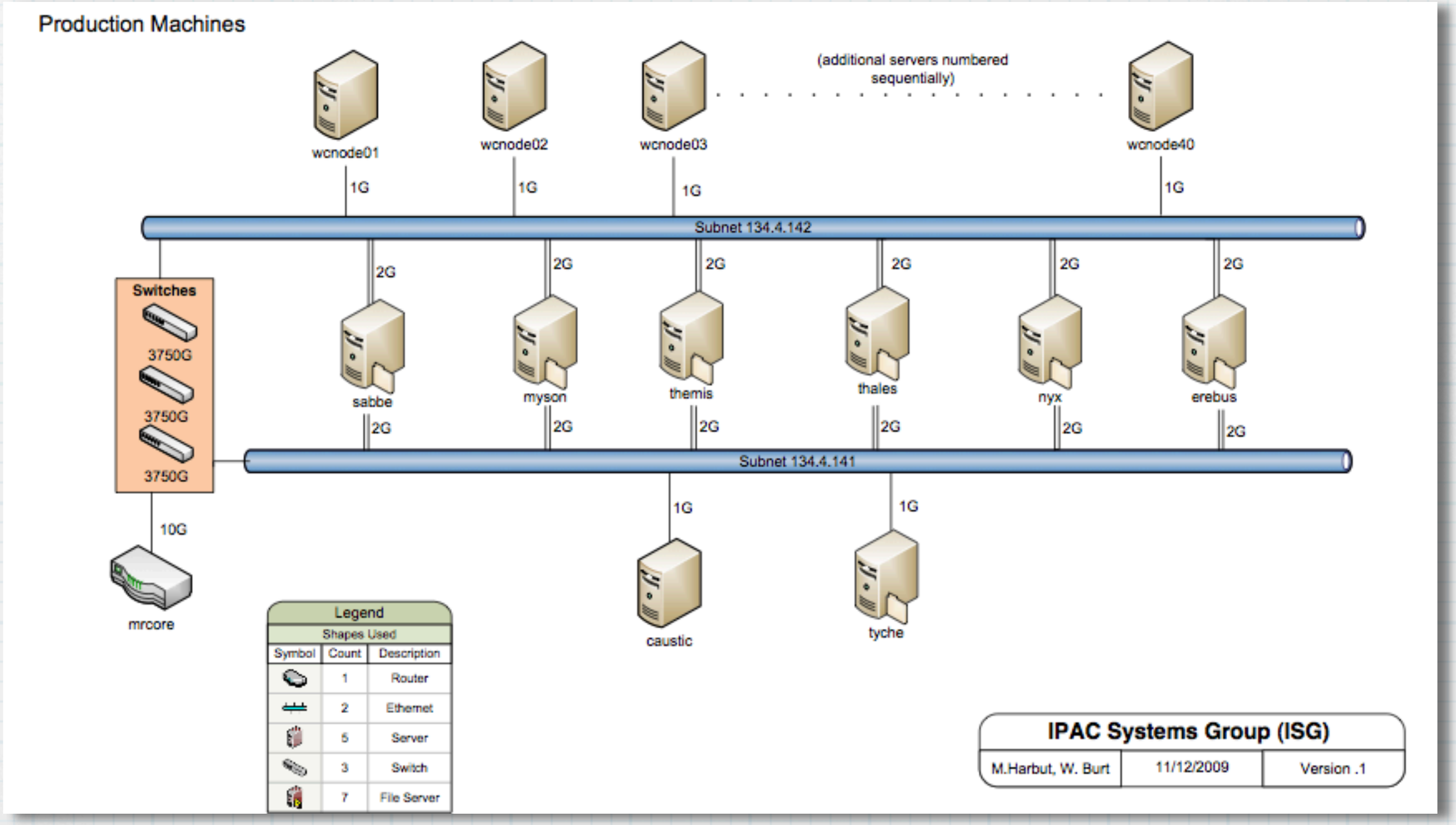
- Heavy ops processing load
  - R/W >3TB/day, Sustained 1Gbit/s network load
  - 180 TB ops archive (non-IRSA)
- Compressed budget and schedule
  - <$400k for ops system at launch
  - Major hardware purchase at L-6 mo.s
  - Simultaneous reprocessing

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

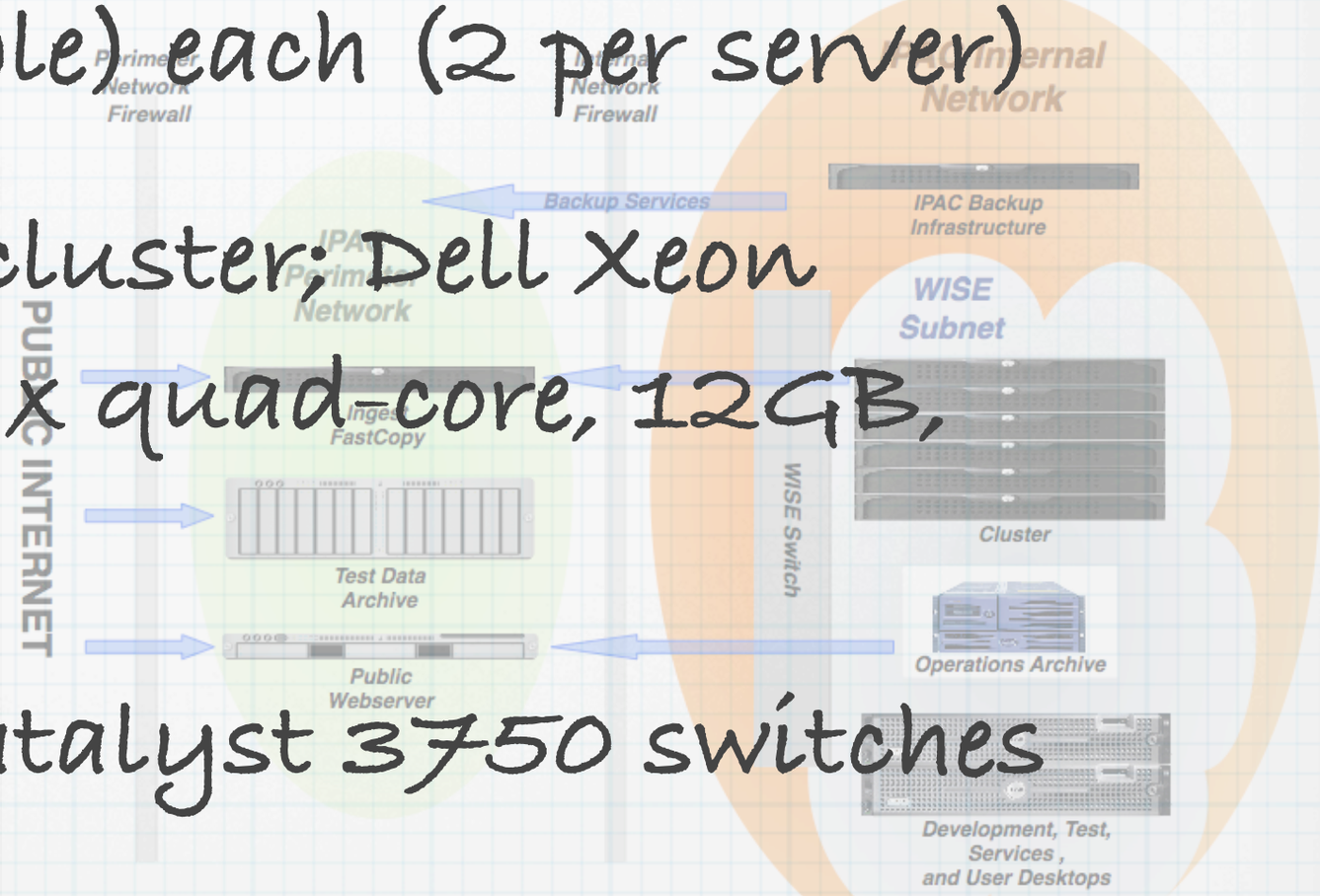


(The Diagram)


Production Machines
wcnode01
wcnode02
wcnode03
(additional servers numbered sequentially)
wcnode40
1G
Subnet 134.4.142
2G
Switches
3750G
3750G
3750G
sabbe
myson
themis
thales
nyx
erebus
Subnet 134.4.141
10G
mrcore
caustic
tyche
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

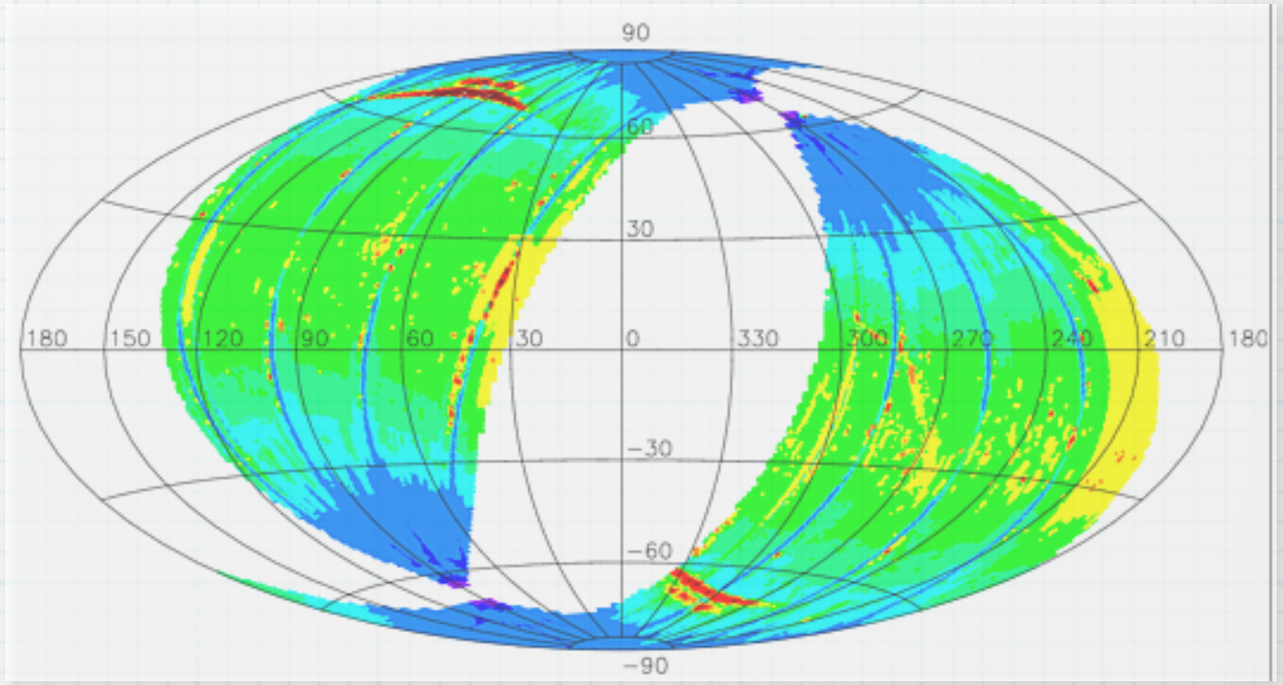
(The Details)

- 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
  - Traced to use of trunked network ports
- Server drops/freezes under load
  - Multiple rounds of patches and work-arounds from Sun (then Oracle)
- Frequent spindle "failures" (really transient SCSI i/f errors)
  - JBOD HBA update + other changes

The Problem 2

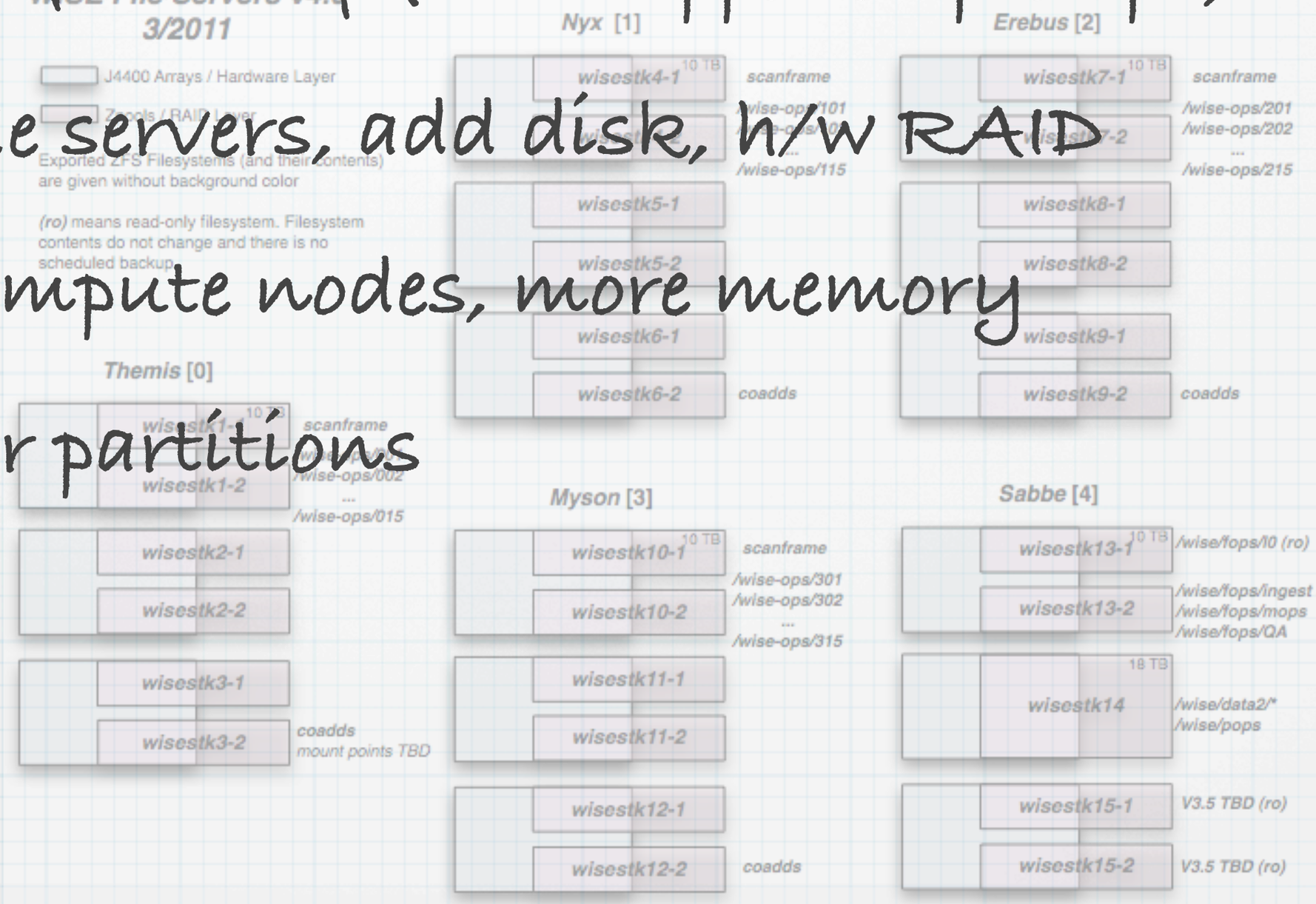
Now, do it all again, 3X faster!

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

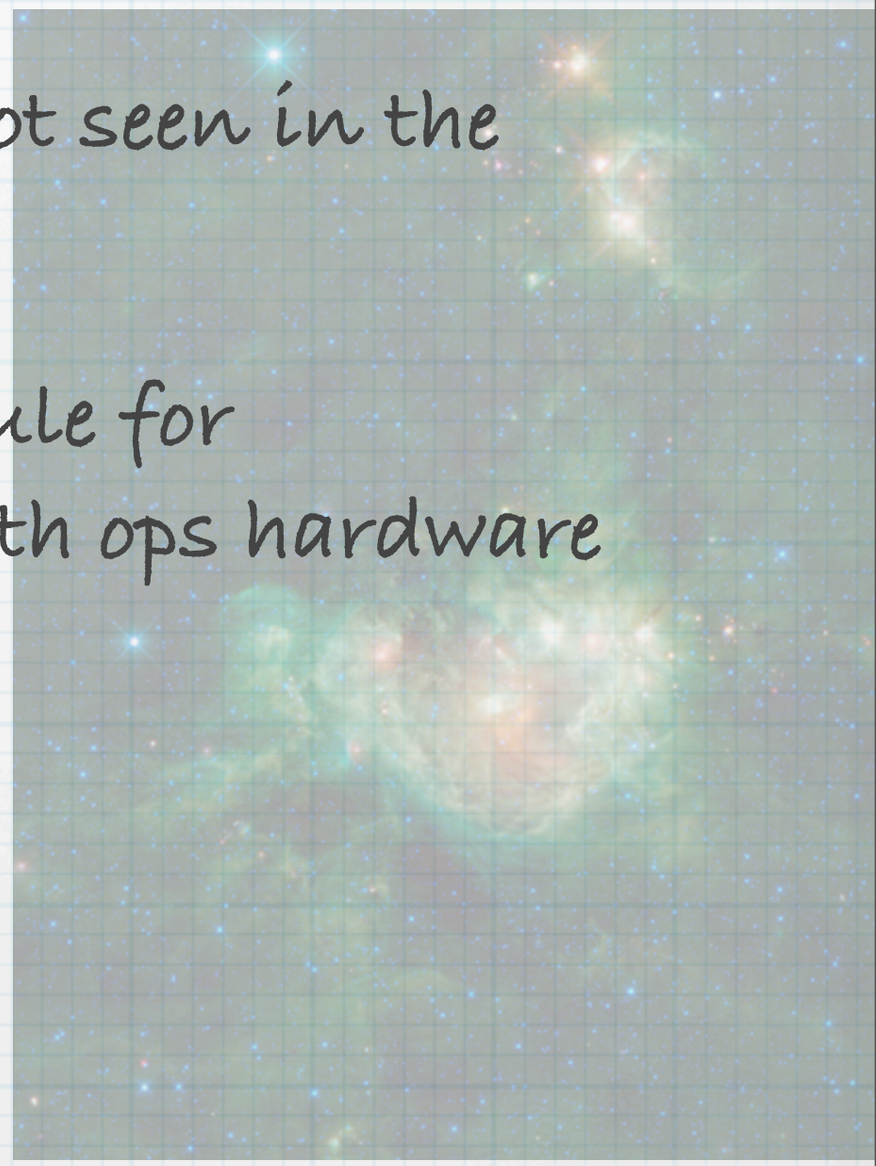


The Result 2

- 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!

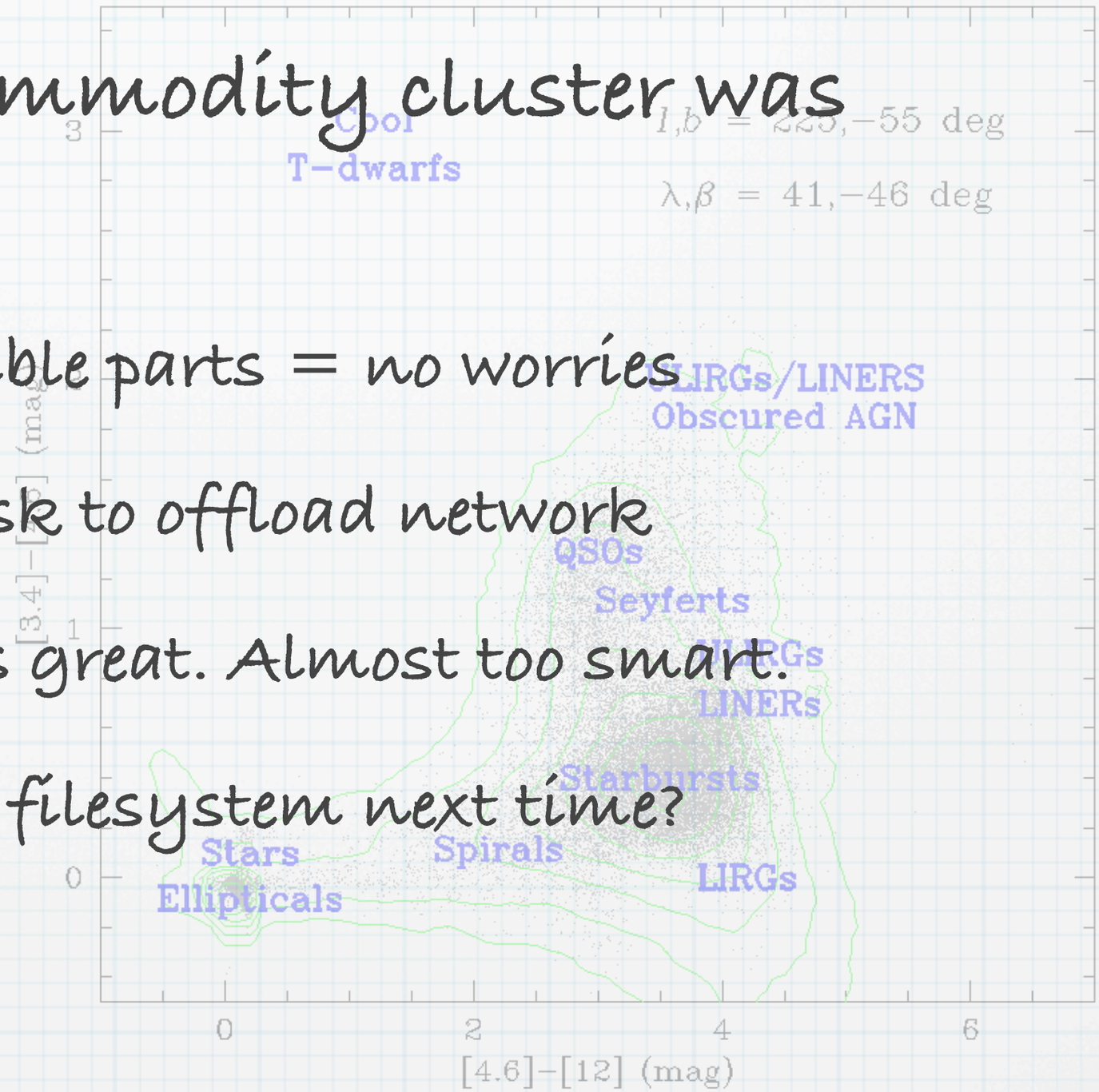


The Lessons

- 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

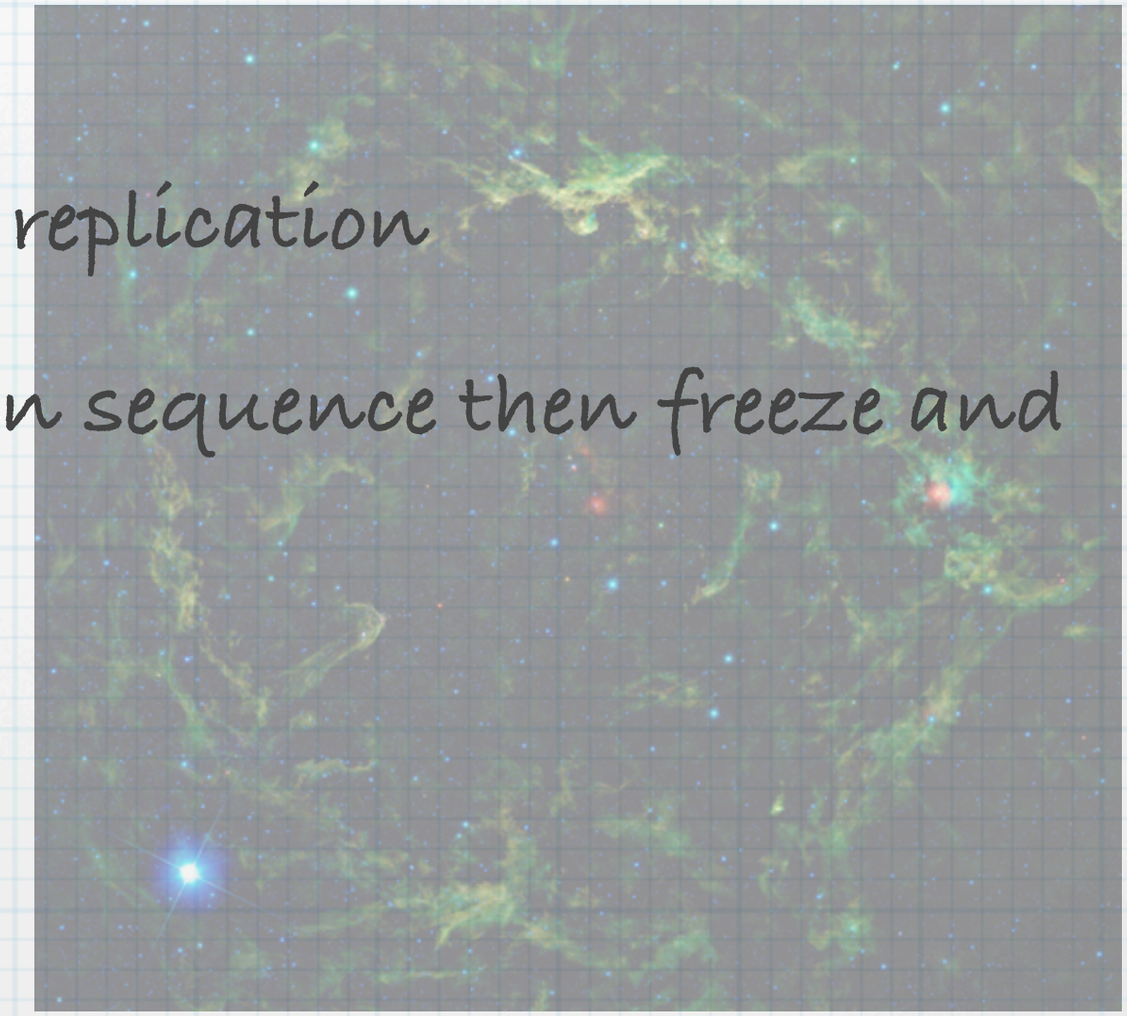
The Lessons

- 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

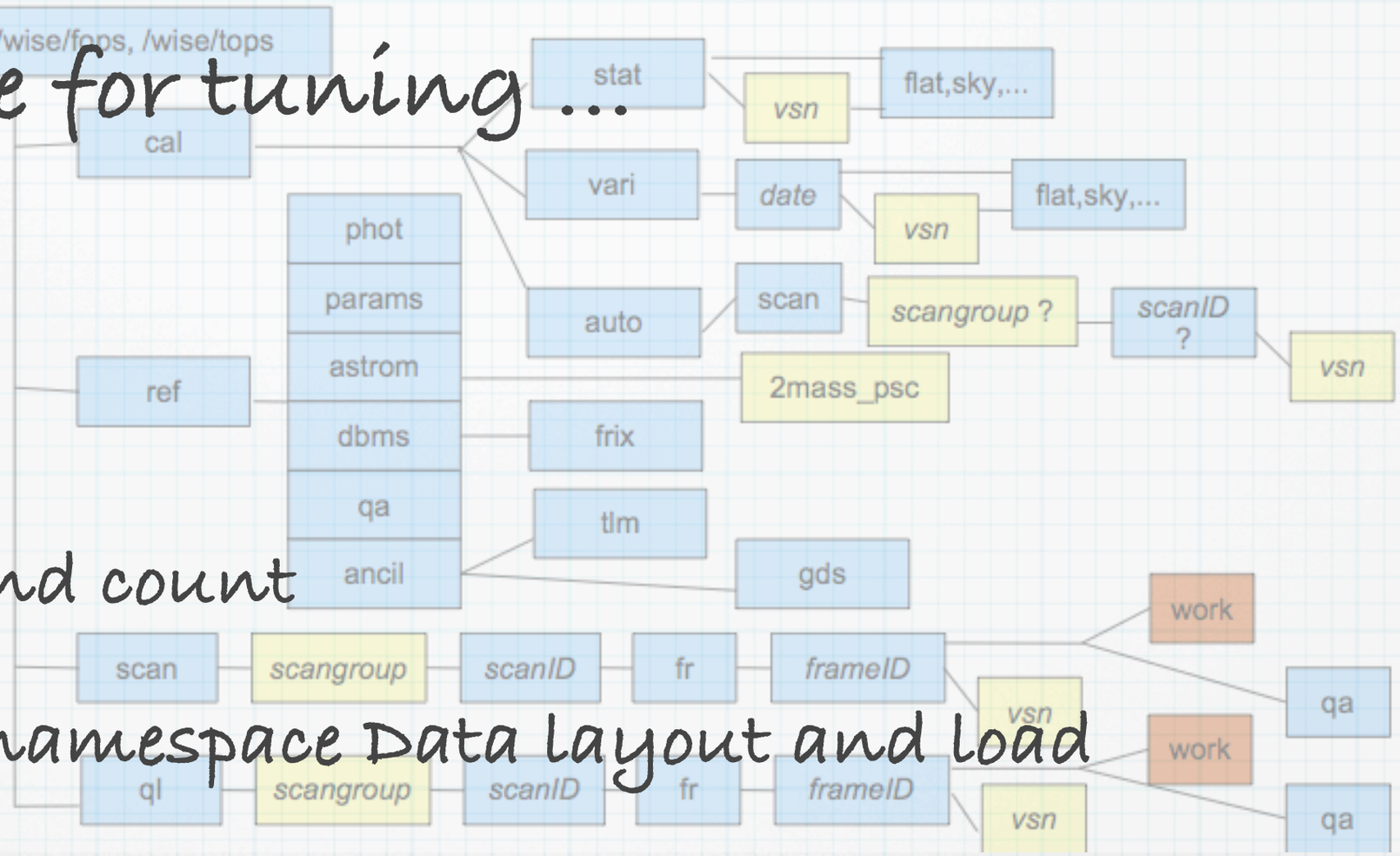


The Lessons

- 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?