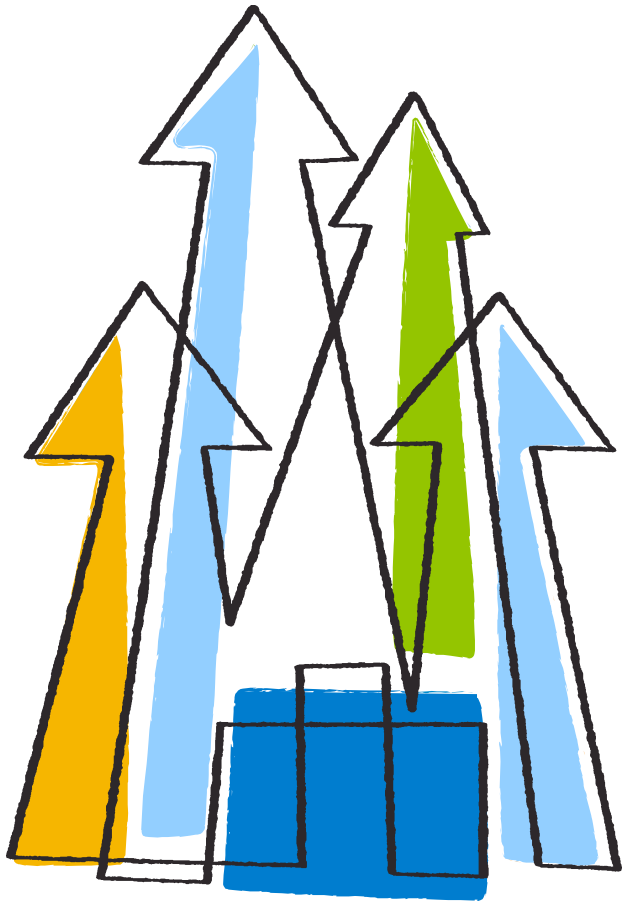
Go further, faster®

Lessons and Observations of a Large F ilesystem

Robert Dilmore

Consulting Engineer II

NetApp



Installation of a 1+TB/sec filesystem

Installation Details-Grove +

- Grove 384 Controller pairs, 23040 3TB Seagate drives, 768 OSS nodes
  - Performance peak 1.3+TB/sec
- 48 RBODs in Vulcan
- 40 RBODs in Porter
- 40 RBODs in Marzen
- 8 RBODs in Stout
- 2 RBODs in Lager

~ Performance well over 2TB/sec ~

Key areas of focus

- Understand and CONTINUALLY monitor:
  - Heat
  - Higher component failures/trending
  - Bandwidth
  - Latency
- Trust no firmware/hardware:

  Trust nothing/Test everything

  Code is tested in a mirrored configuration

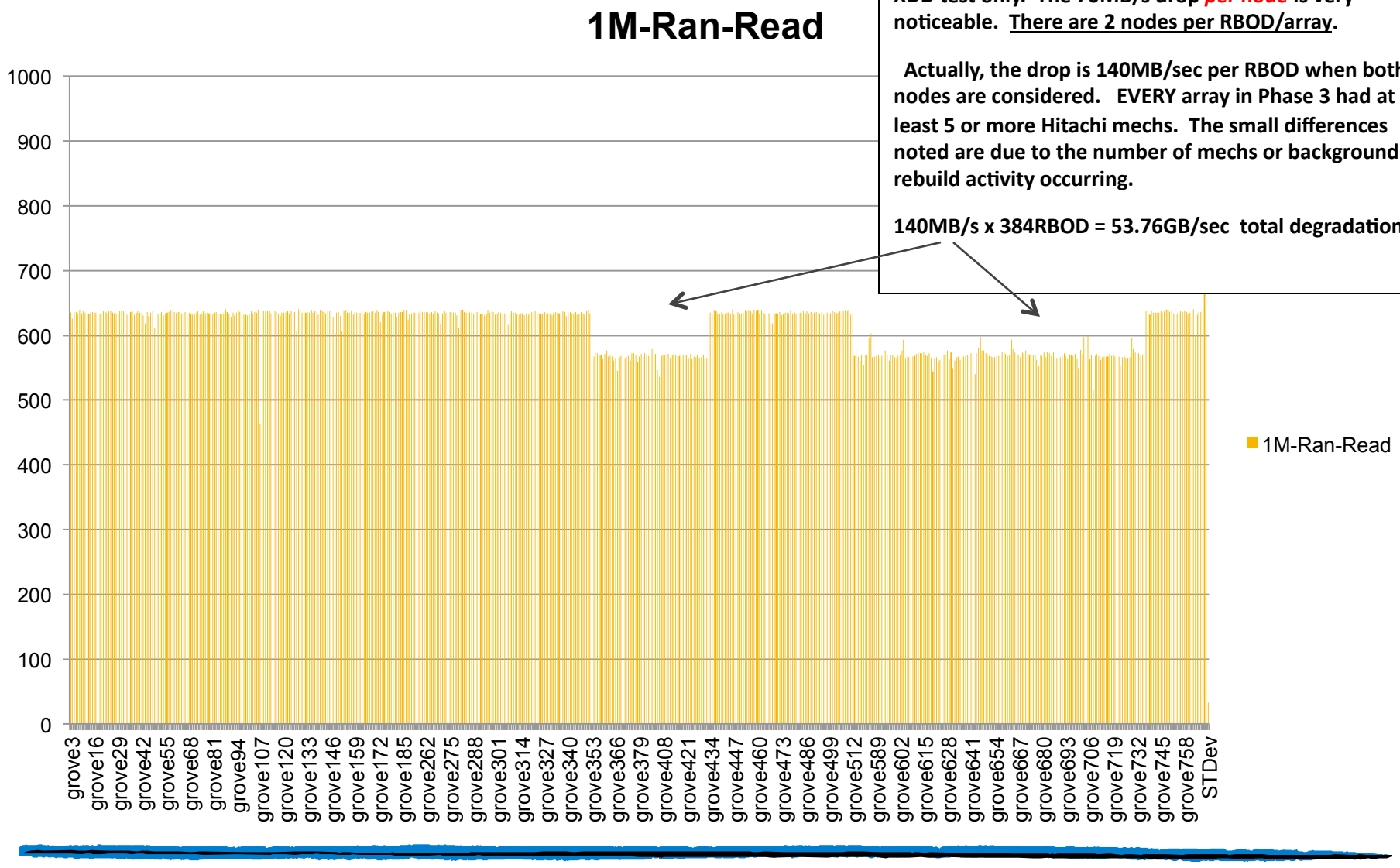
  Spares are cooked and tested

TS Nock-ten and Performance

This is a blowup of the 1M Random Read portion of the XDD test only. The 70MB/s drop *per node* is very noticeable. There are 2 nodes per RBOD/array.

Actually, the drop is 140MB/sec per RBOD when both nodes are considered. EVERY array in Phase 3 had at least 5 or more Hitachi mechs. The small differences noted are due to the number of mechs or background rebuild activity occurring.

140MB/s x 384RBOD = 53.76GB/sec total degradation



Duct Tape and Sand bags and Heat

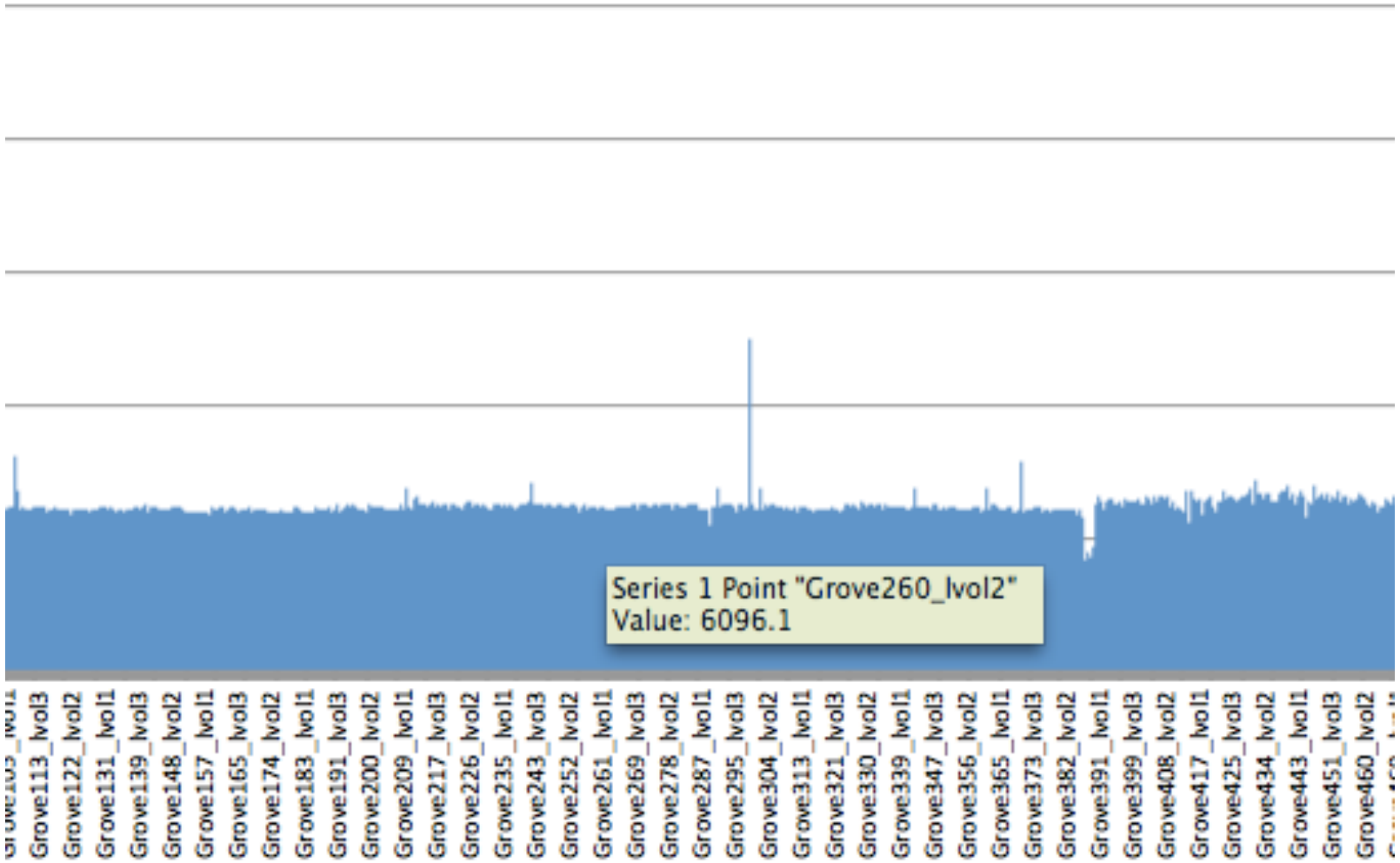
* *The first known proper use of duct tape in history*!

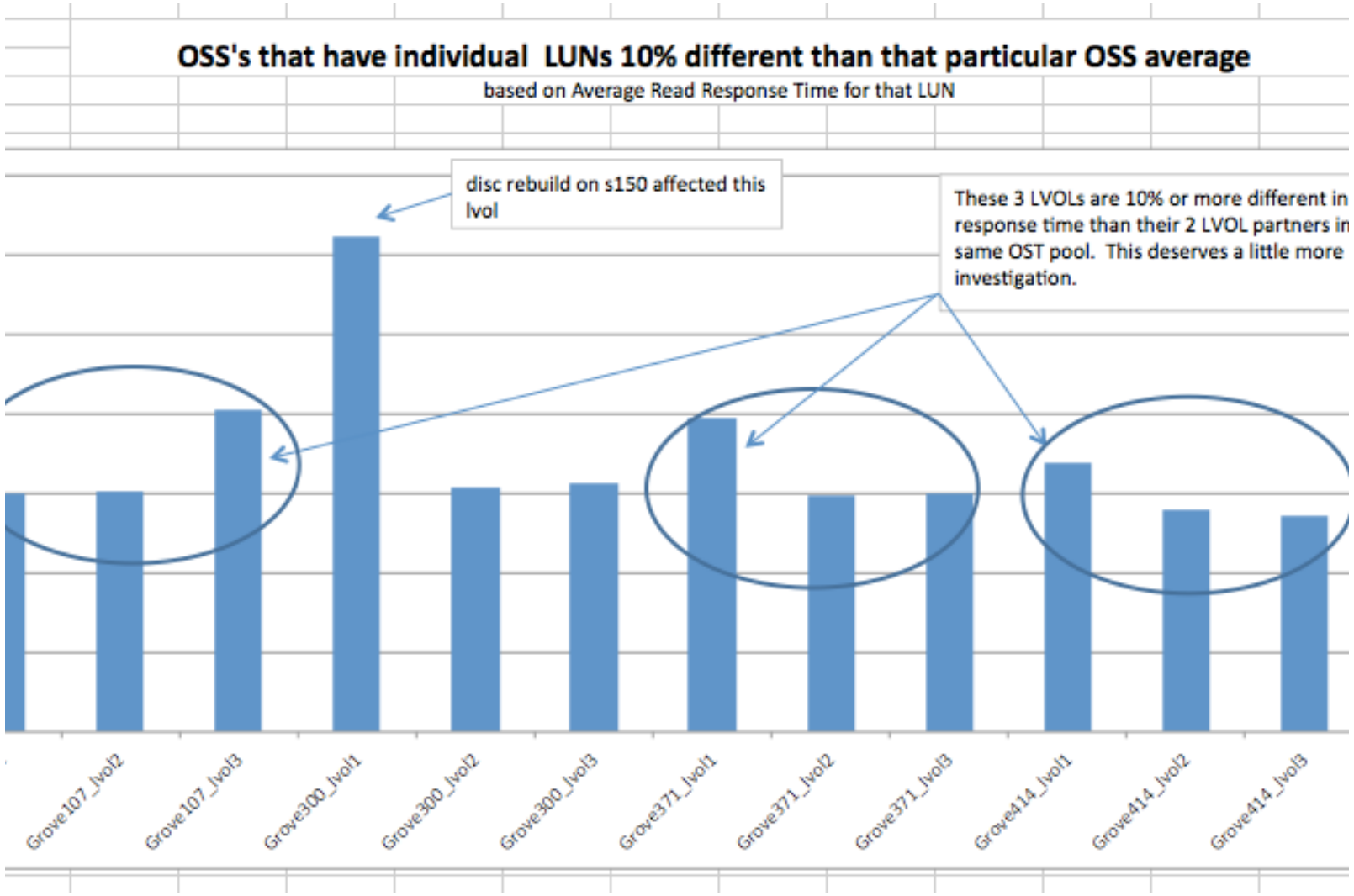
Many places to monitor heat and flow

- All disk drives have temperature sensors
- All trays have 2 sensors
- All other components have AT LEAST 2 sensors
- 100+ temperature monitoring points

- Most installations use one-the light on the front of the box

Tracking Average response time




OSS's that have individual LUNs 10% different than that particular OSS average
based on Average Read Response Time for that LUN
disc rebuild on s150 affected this lvol
These 3 LVOLs are 10% or more different in response time than their 2 LVOL partners in same OST pool. This deserves a little more investigation.
Grove107_lvol2
Grove107_lvol3
Grove300_lvol1
Grove300_lvol2
Grove300_lvol3
Grove371_lvol1
Grove371_lvol2
Grove371_lvol3
Grove414_lvol1
Grove414_lvol2
Grove414_lvol3

Tools developed to monitor-DDRT

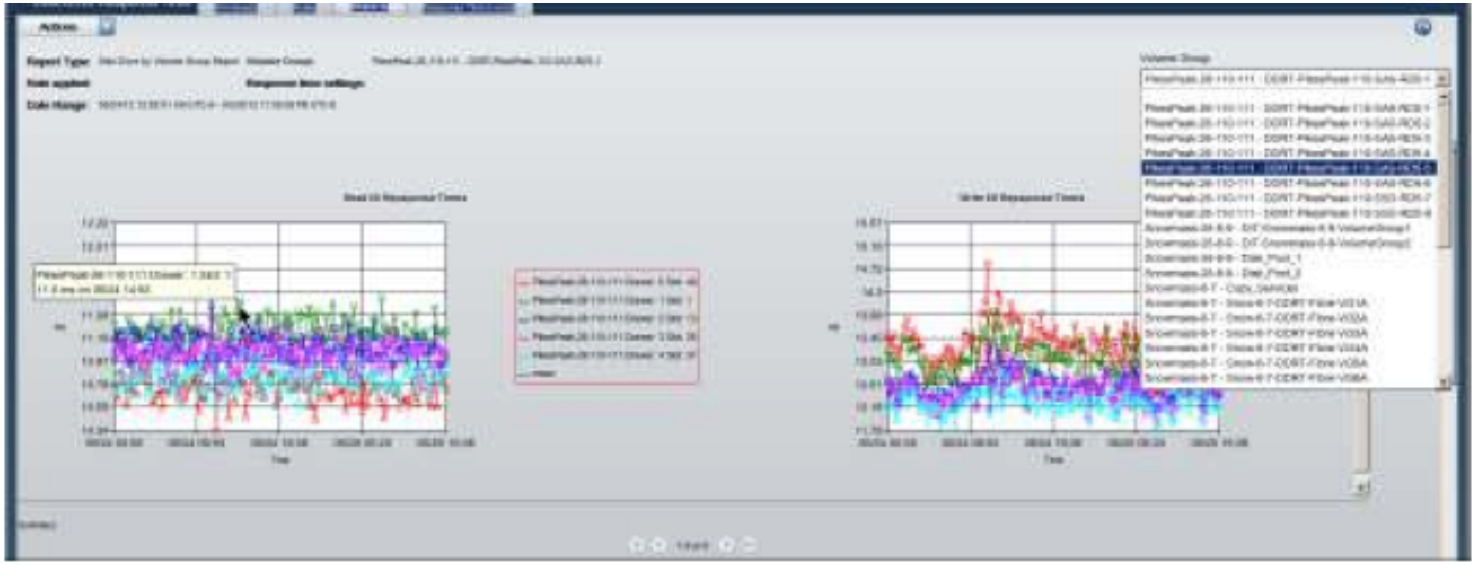
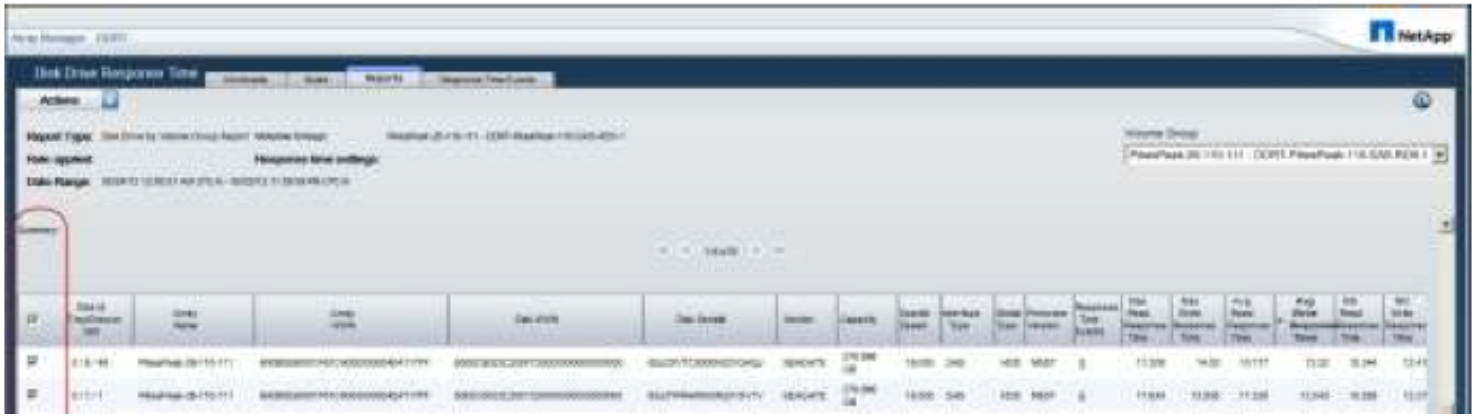
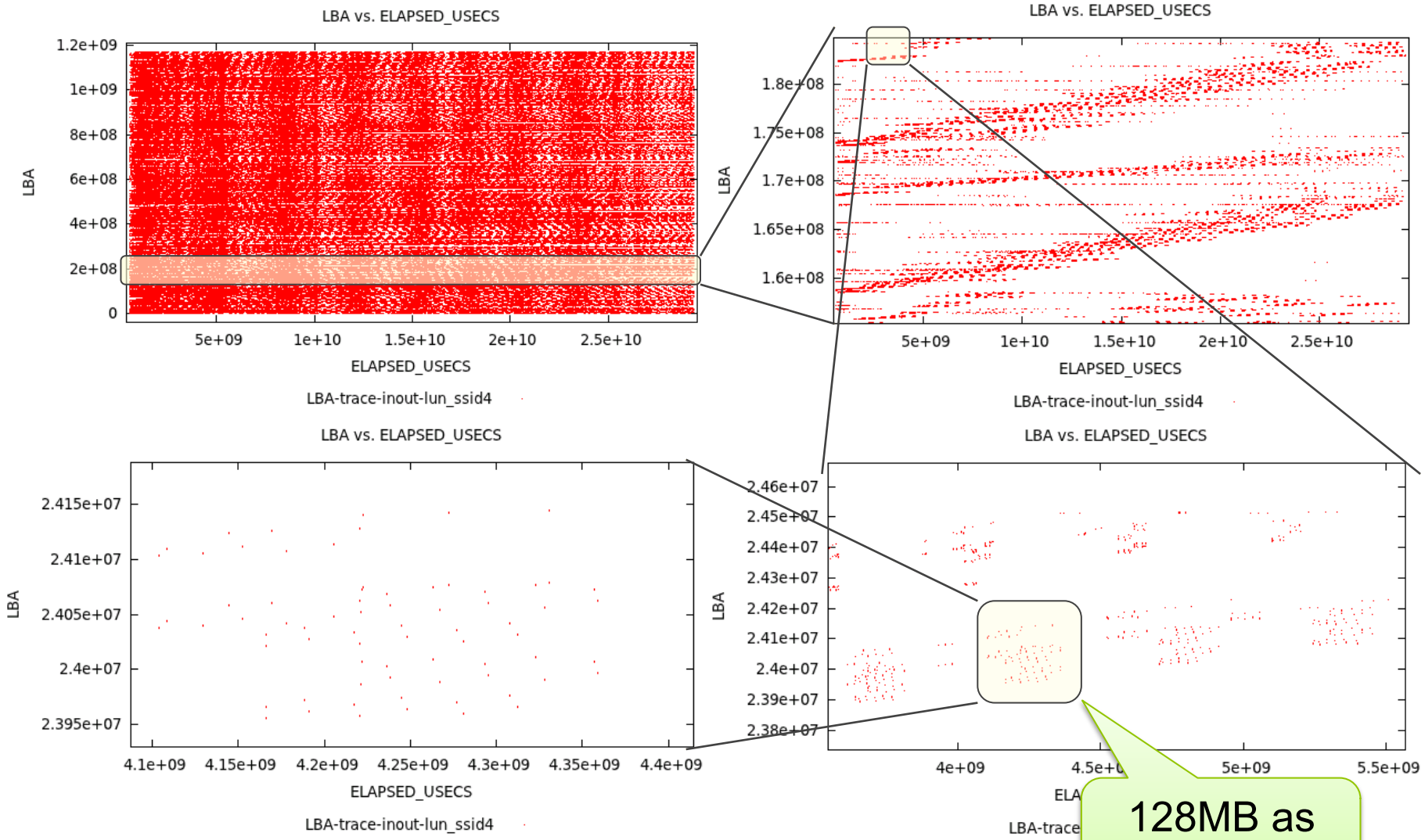


Figure 60 Disk Drive Response Time by Volume Group Report—Summary Table



Workload in Pictures–clustered FS



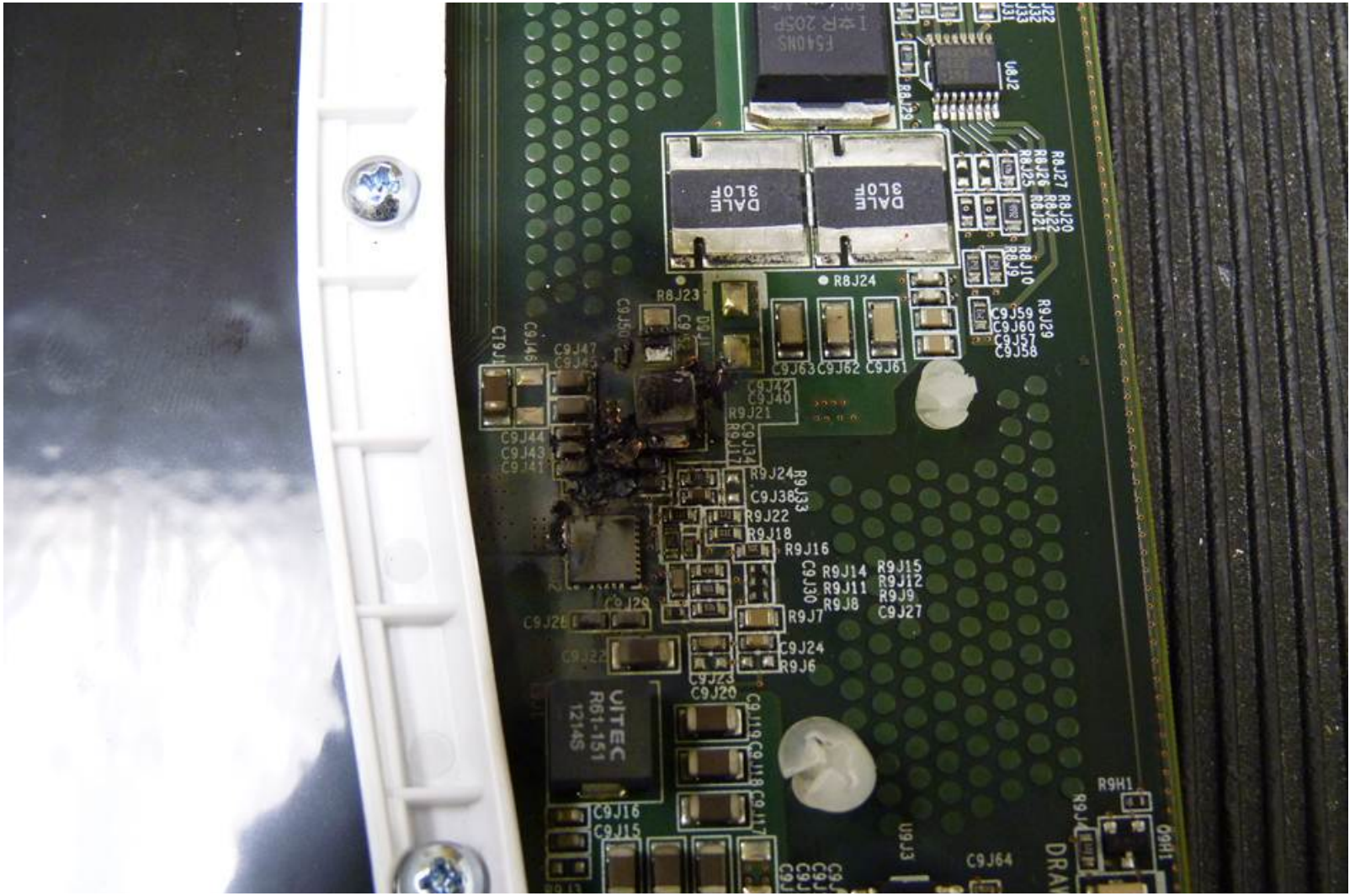
Drive are not failing!

- 57 drives in 342 days (49 weeks) = 1.16 drive failures per week

- 57 drive failures out of 23,040 drives
- .247% total disc failure in a year (not even 1%)

- By the way, we just racked our 500th E5460 over in Bldg 451 on Tuesday. By the time we rack these next 7 RSSU's, we will have a total of 546 E-Series enclosures out here at LLNL.

The case for RAID6



Test-Everything

- Test code @ Hyperion

-found a firmware issue, reboot avoided

- Test spares at site Hyperion

-cold spares CANNOT be trusted

-still have not used initial 100 spare drives

-previous slide show where magic smoke escaped from component

NetApp

Thank you