Exascale Failure Modeling with CoFaCTOR

*Co*rrelated *Fa*ilure *C*onsultation *T*ool for *O*perational *R*eliability

PIs:

Dave Bonnie,

Dominic Manno,

Wendy Poole,

Brad Settlemyer

May 21, 2019

NNSA
National Nuclear Security Administration

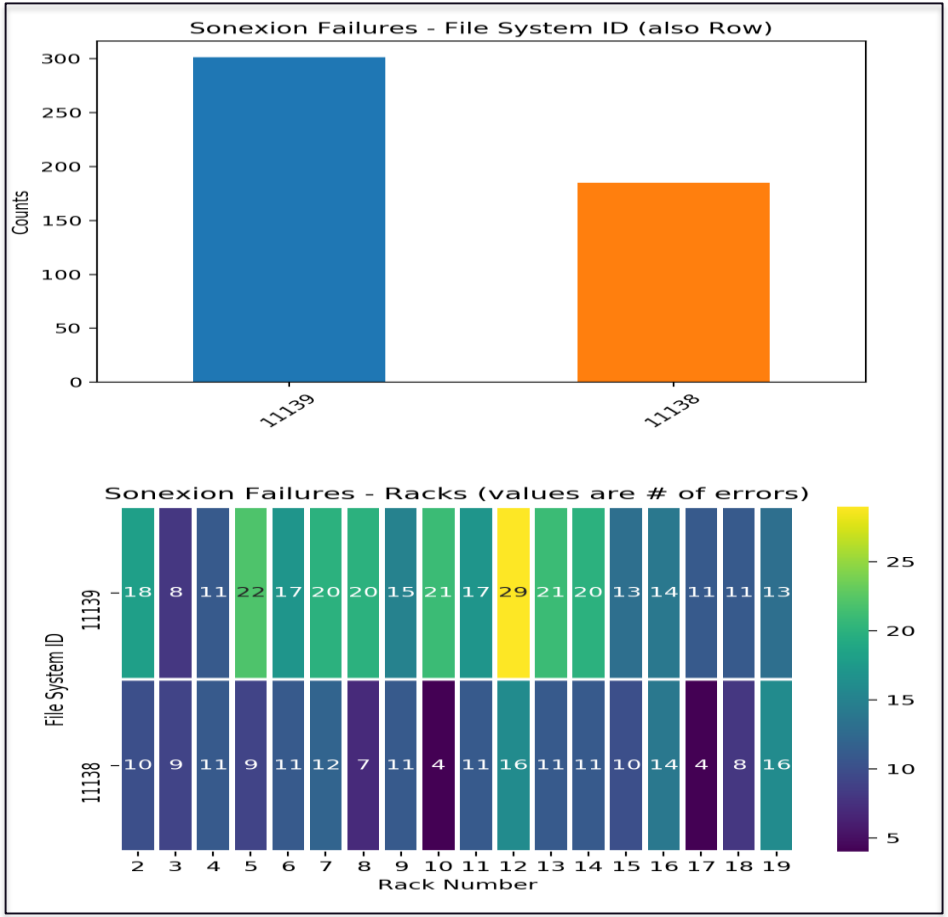
Managed by Triad National Security, LLC for the U.S. Department of Energy's NNSA

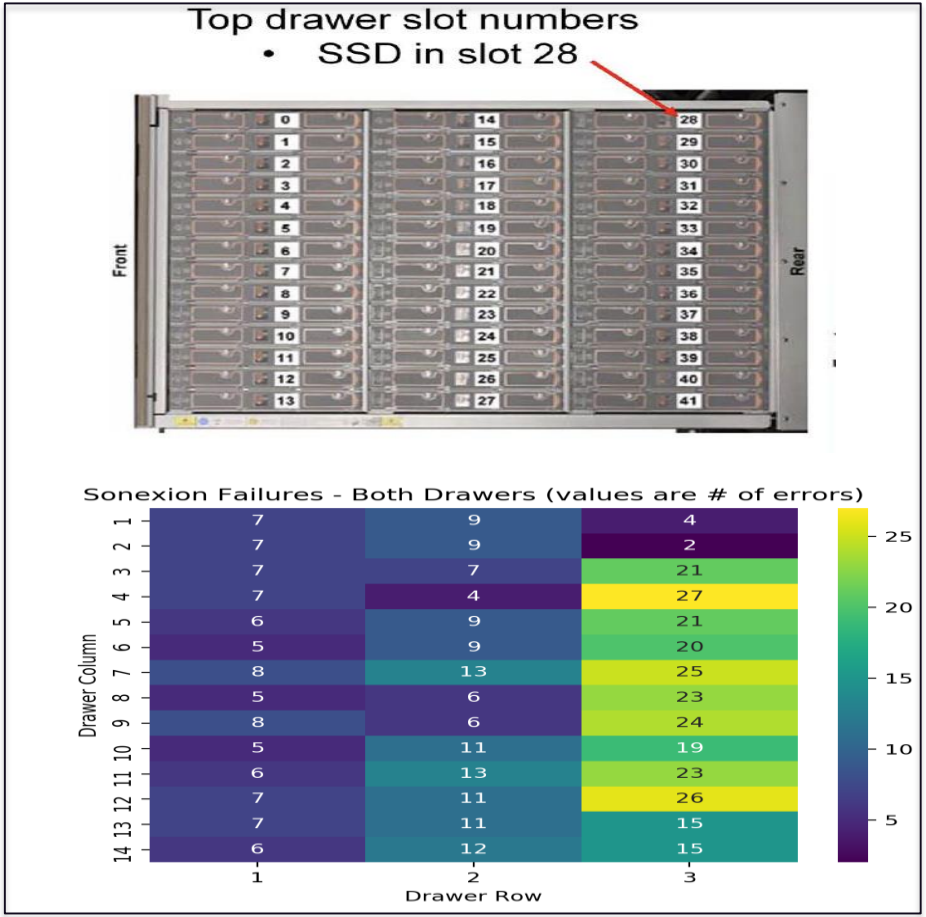
LA-UR-19-24681

BLUF (Bottom line up front)

- We need a tool for Decision Support
  - Why?
    - We need to be able to do predictive analytics for system failures as well as potential catastrophic disk or filesystem failures.
    - Within the Exascale environment, the filesystems and network will be so complex determining failure causes without the assistance of a tool will be nearly impossible
- We don't believe the failures within the storage environment have been well-identified
  - Traditional assumptions on drive mix and failures no longer hold
  - multi-factor positional considerations (shelf position, rack, row) (vibration…)
- LANL's filesystem environment are very different from the cloud env.
  - A loss of 1MB stripe may invalidate the entire 1PB file
    - We save memory state, not cat pictures and memes

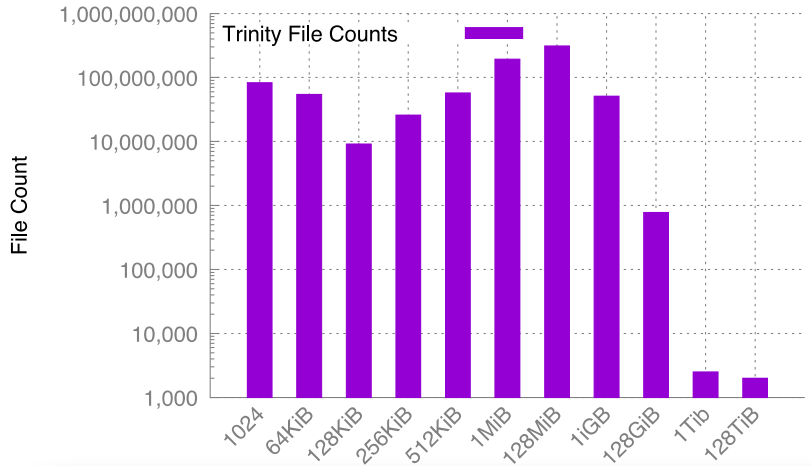
Problem: Correlated disk failures

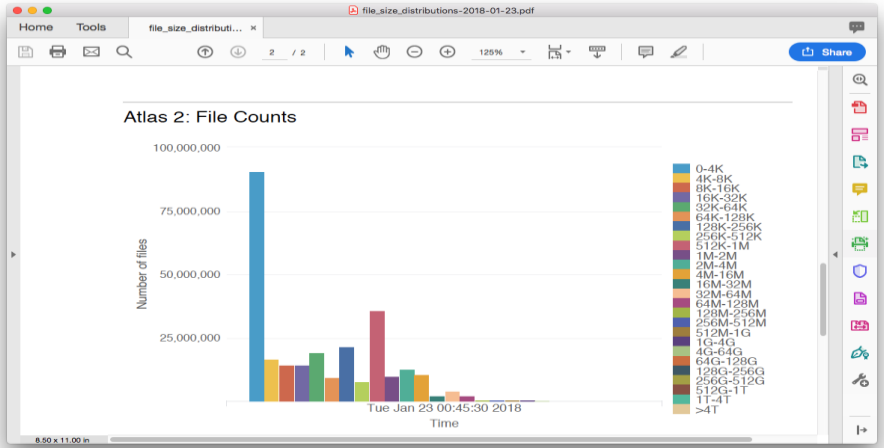




Problem: Why is LANL hitting this now?

- Large differences between hyperscalers and LANL
  - 98.3% of Youtube videos are less than 25MB*
- LANL has very large files, tens of TB is not uncommon
  - Data loss event from failure is orders of magnitude different





* "*Statistics and Social Network of Youtube Videos*", Cheng, Dale, Liu

Components Available Today

- Data already collected
  - LANL File distributions
  - LANL Failure events
  - Industry failure events also published (e.g. BackBLAZE)
- Industry standard data protection schemes
  - Parity-based data protection (GridRAID, ISA-L, RAIDZ3)
- Existing Statistical Techniques appropriate
  - Monte Carlo simulation
  - Failure modes well studied for disks
- Existing simulation toolkits are sufficient
  - PySIM, LANL's Simian, OmNet++

Components Available Today

These techniques are well understood in OR, statistics, reliability community.

We simply need to apply them correctly to our data!

CoFaCTOR Overview

- Model inputs:
  - LANL's empirical file distribution
  - Storage system characteristics
  - Protection strategies
  - Do we need more error protection and at what levels/complexity?

- Evaluate failures via Monte Carlo
  - Generate realistic failure traces
  - Identify probabilities of loss
  - Analyze data loss scenarios
  - Evaluate and analyze distributions of effected files

CoFaCTOR Usage

- Using existing methods will enable:
  - Understanding catastrophic data loss scenarios
    - LANL HPC field's approximately 11 different file systems – each with different data protection schemes and data retention times
  - Evaluating new technologies (with respect to data protection capabilities)
  - Improving future storage procurements

- Additional opportunities:
  - Provide further collaboration for environments with similar predictive requirements (anyone here want to share?) :)
  - Provide further collaboration in this space across other federal agencies - DOD