A Perspective on Power Management for Hard Disks

Kirill Malkin
Director of Storage Engineering

May 3, 2016

Powering Down Drives

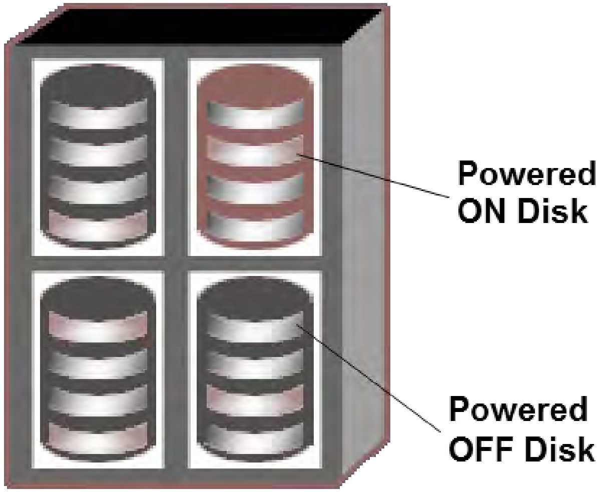
What are the benefits?

- Turning off power may extend drive life
  - Extreme example:
    - Power down drive for 5 years; it should run for 5 years more, doubling drive life
  - Will this hold if we do this more frequently?
  - If we spin up drives only when access is required and then spin them down, will this help extend drive's life?
- Ability to dial in power consumption
  - Limiting total number of drives that can be spinning based on power budget

MAID: Power Managed Disks

- What is MAID?
  - Large number of power-managed disks
  - More than 50% drives powered off
  - Power-cycling by policy
  - Lower management and environmental costs and longer drive life

- COPAN Systems have enhanced MAID
  - Three-Tier Architecture
    - Scales performance with capacity
  - POWER MANAGED RAID Software
    - RAID protection for power-managed disks
    - **Maximum of 25% drives spinning**
    - 1/3 cost of traditional RAID systems
  - DISK AEROBICS Software
    - Disk reliability and data integrity



SNIA Definition

"A storage system comprising an array of disk drives that are powered down individually or in groups when not required. MAID storage systems reduce the power consumed by a storage array."

COPAN Software Features

- POWER MANAGED RAID
  - Drives spin only when necessary to meet application requirements, extending drive life by more than 4x
- DISK AEROBICS
  - Actively monitors and manages drive health
    - Tracks “slow I/O” and timeouts
    - Logs SMART and environmental data
  - Disk Scrubbing
    - Are data sectors readable and consistent?
  - Periodically exercises idle drives
    - Performs self-test
  - Proactive failing of “suspect” drives
    - Evacuate data and request drive replacement

Brief History of COPAN

- COPAN was started in 2003
- First product, Revolution 200 Series, was shipped in 2004
  - Supported MAID LUNs or VTL
  - Intel Xscale based controller
- Performance upgrades resulted in 300 series shipped in 2008
- Hardware redesign and release of 400 series in 2009
  - New disk canisters and backplane
  - COM Express module based controller
- SGI acquired COPAN in 2010
  - Shipped 400 series until 2014

Drive Life and Reliability Promise

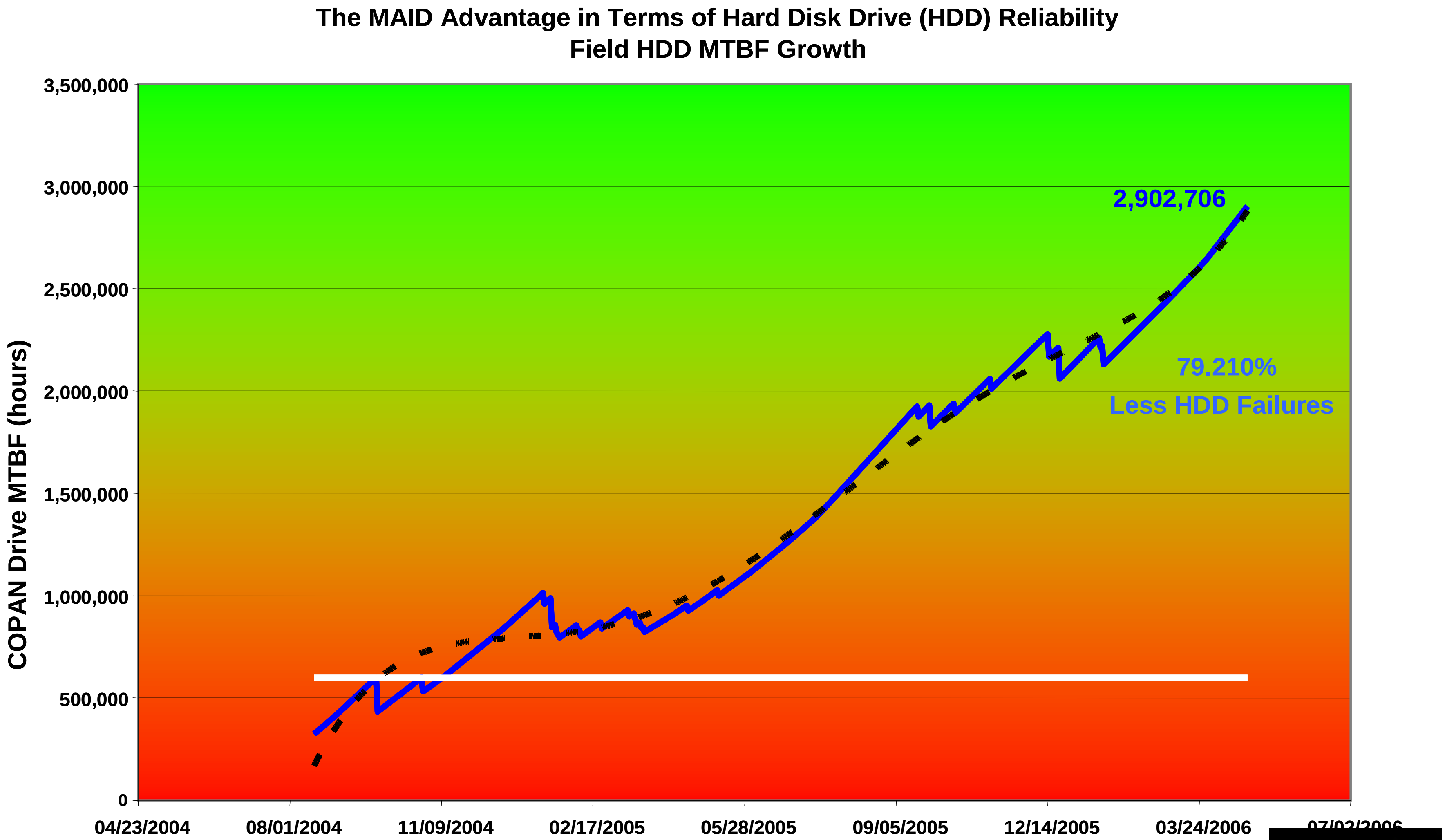
- Extended drive life and reliability
  - Compared to standard SATA disks, COPAN has less than ¼ the failure rate
  - Field MTBF is more than 4x SATA disks, more than 2x FC disks
  - Service Life: expect more than 4x
- Disk Reliability and TCO benefits
  - Assuming 1,000 drives, expect:
    - COPAN: 3 failures per year
    - SATA: 15 failures per year
  - Standard SATA platforms have
    - ~5X drive replacements
    - 17 touches versus 1 touch for COPAN

MTBF (hrs)	AFR (%)	Disk Specification
8,000,000	0.11%	
5,000,000	0.18%	
3,000,000	0.29%	
2,902,706	0.30%	COPAN - Apr 2006
2,400,000	0.36%	
2,000,000	0.44%	
1,200,000	0.73%	Fibre Channel
1,000,000	0.87%	Fibre Channel
800,000	1.09%	
600,000	1.45%	SATA
400,000	2.17%	SATA
200,000	4.29%	
100,000	8.39%	

600K hrs = 68 yrs
2.64M hrs = 331 yrs

Early Drive Reliability Field Data

Suggests 4x improvement



Did the promise hold up?

To answer that we needed failure analysis

- Goal was to determine observed AFR
  - Ideally, for each year, needed number of disks under support and number of disks replaced
  - Detailed service data was a bit difficult to obtain, used indirect or incomplete data
    - Interesting statistical exercise in forensics
- Decided to organize data by disk capacity to enable validation checkpoints
  - This proved to be quite useful

Failure Rate Analysis

Based on Service and Sales Data

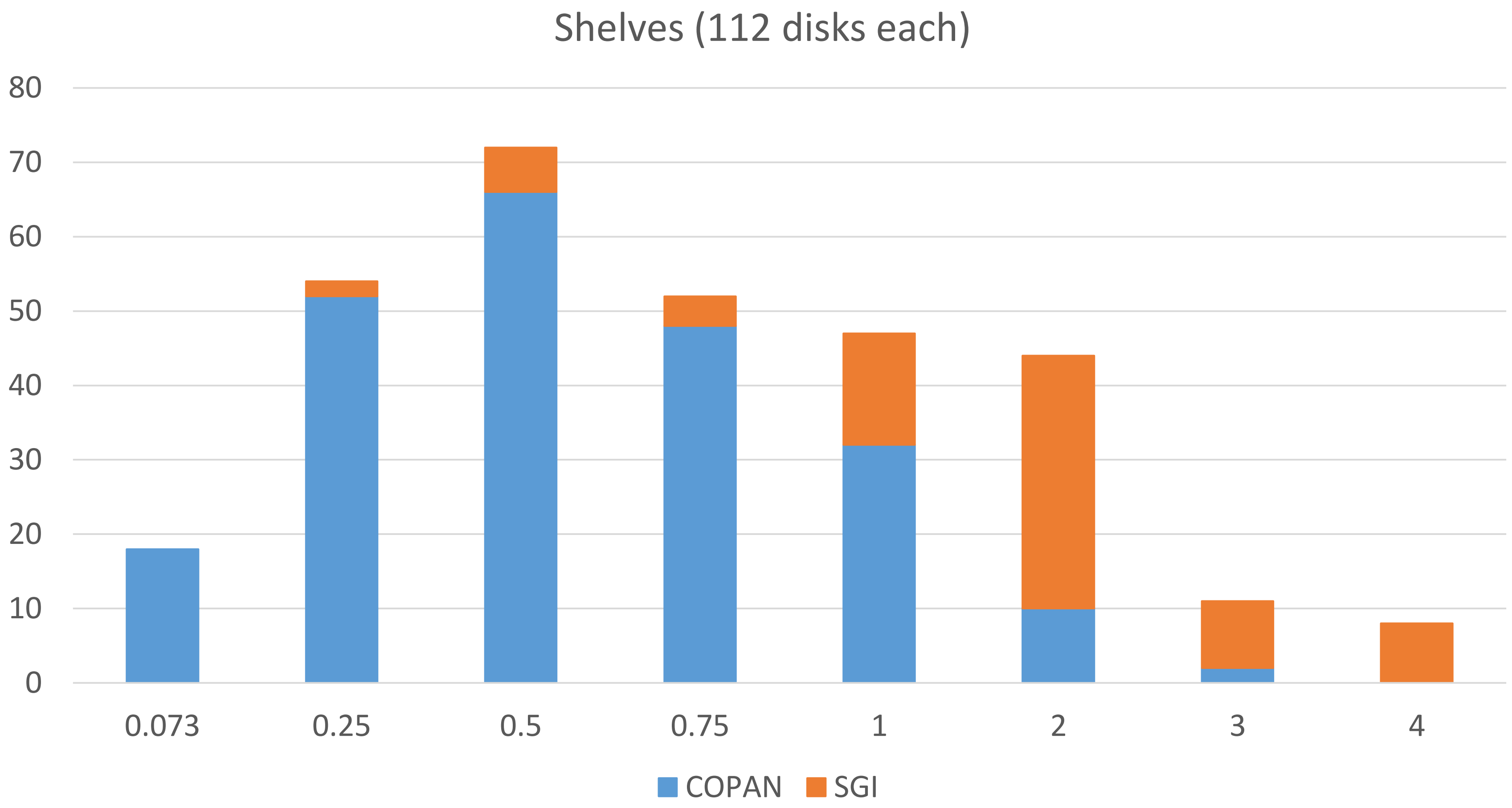
- Determining number of failed disks per year
  - Located post-2010 disk FRU codes and matching system data
    - Part number (indicating COPAN or SGI part)
    - How many disks replaced
    - In how many shelves
    - Year first replaced
  - Assumed even failure distribution after year first replaced
  - Pre-acquisition service data was not available
    - No data available for 2007-2010 ☹
    - Analyzed post-2010 data, including existing and new installations
- Determine overall disk count
  - Obtained post-2010 sales data, i.e. sales by SGI
  - Extrapolated system count sold by COPAN based on number of installations under support contracts

Analysis Summary

- Total storage analyzed
  - 34,200 total disks – good size sample
  - 8 different disk capacities
    - 73GB, 250GB, 500GB, 750GB
    - 1TB, 2TB, 3TB, 4TB
  - 31.6PB total capacity
- No post-failure analysis
  - Any replaced disk is considered failed
  - No consideration for non-disk related failures (e.g. backplane issues)

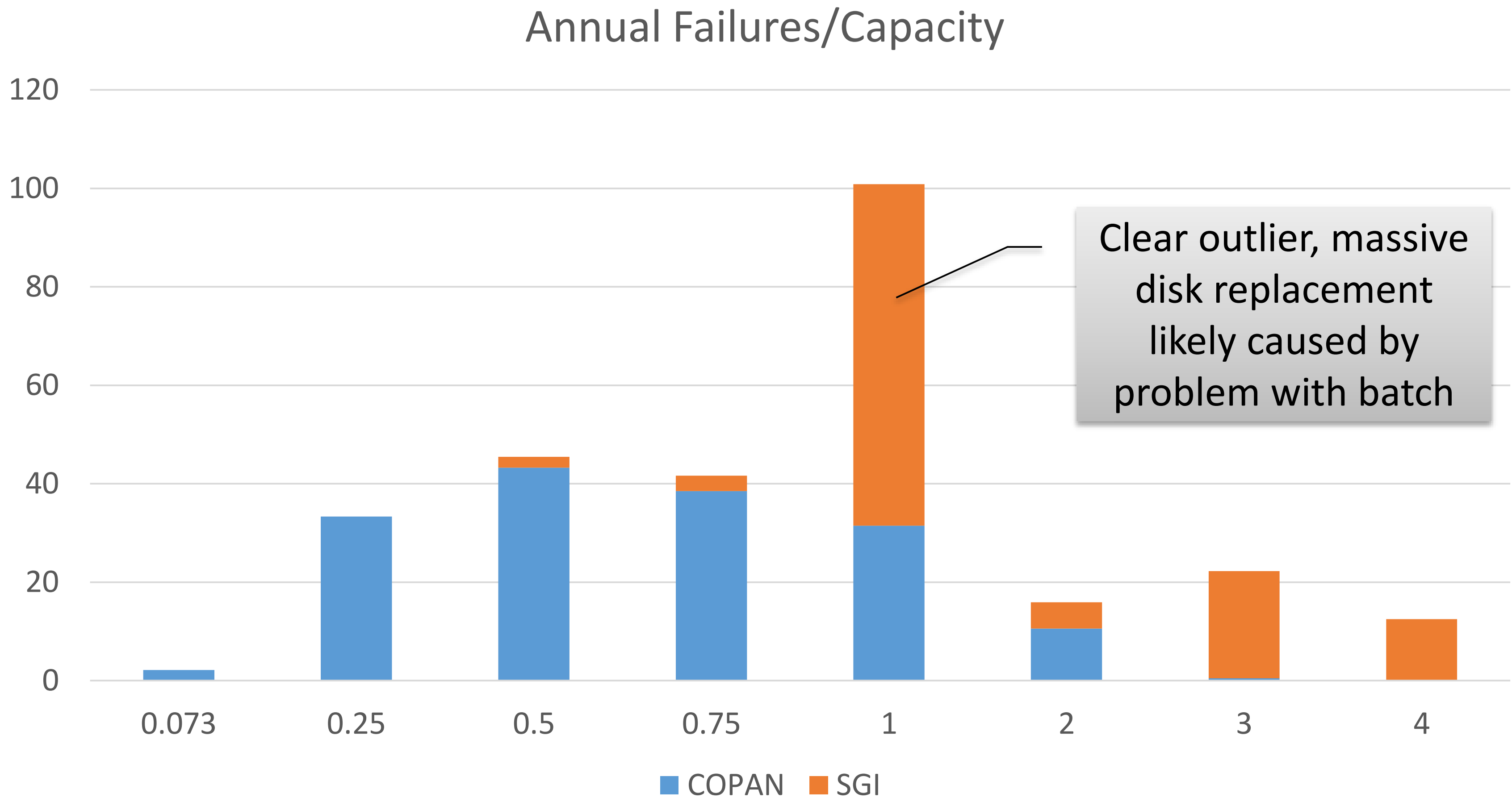
MAID Shelves Analyzed

Separately for COPAN and SGI



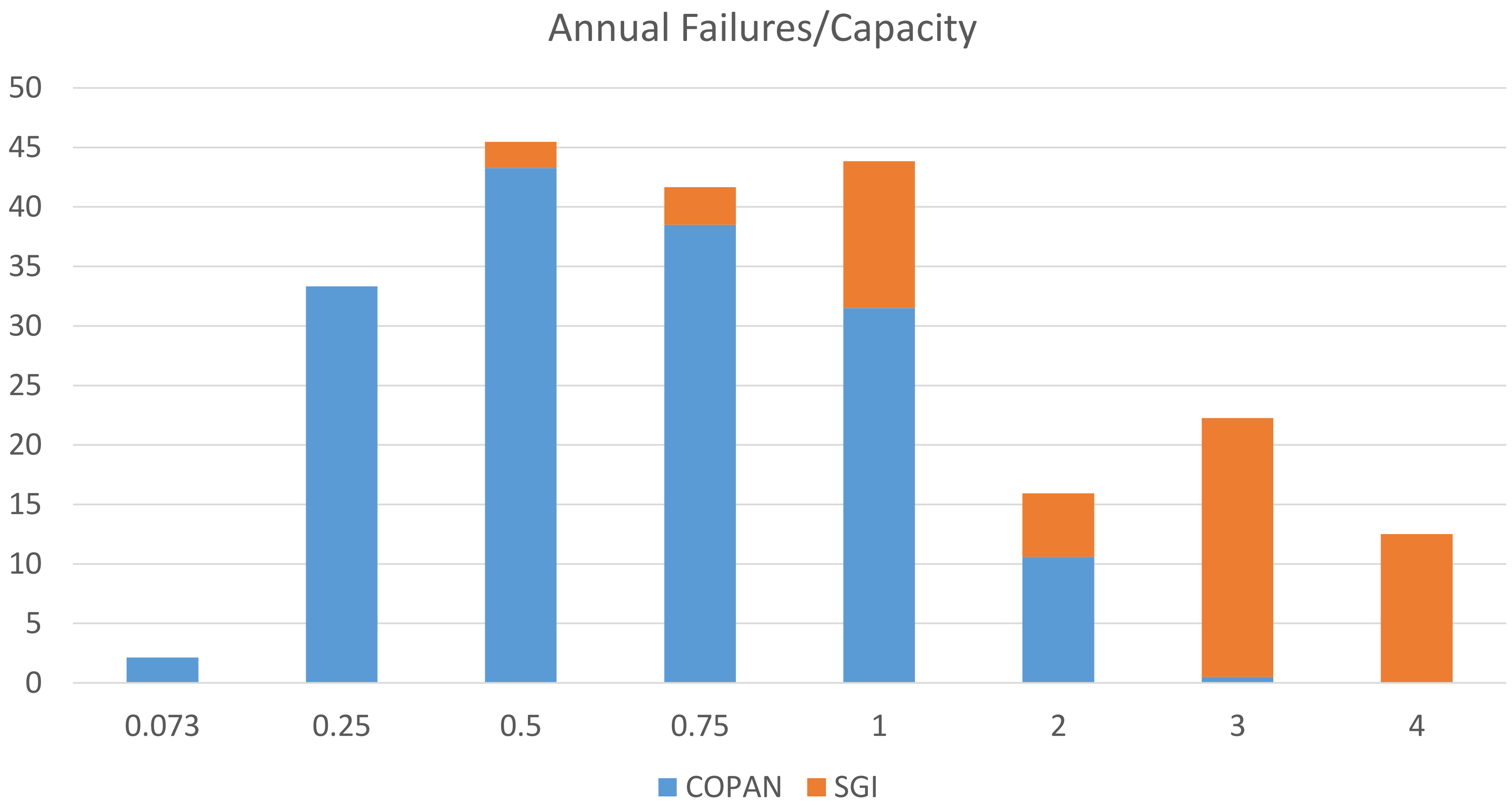
Unscrubbed Failure Data

Organizing by capacity helped find outlier



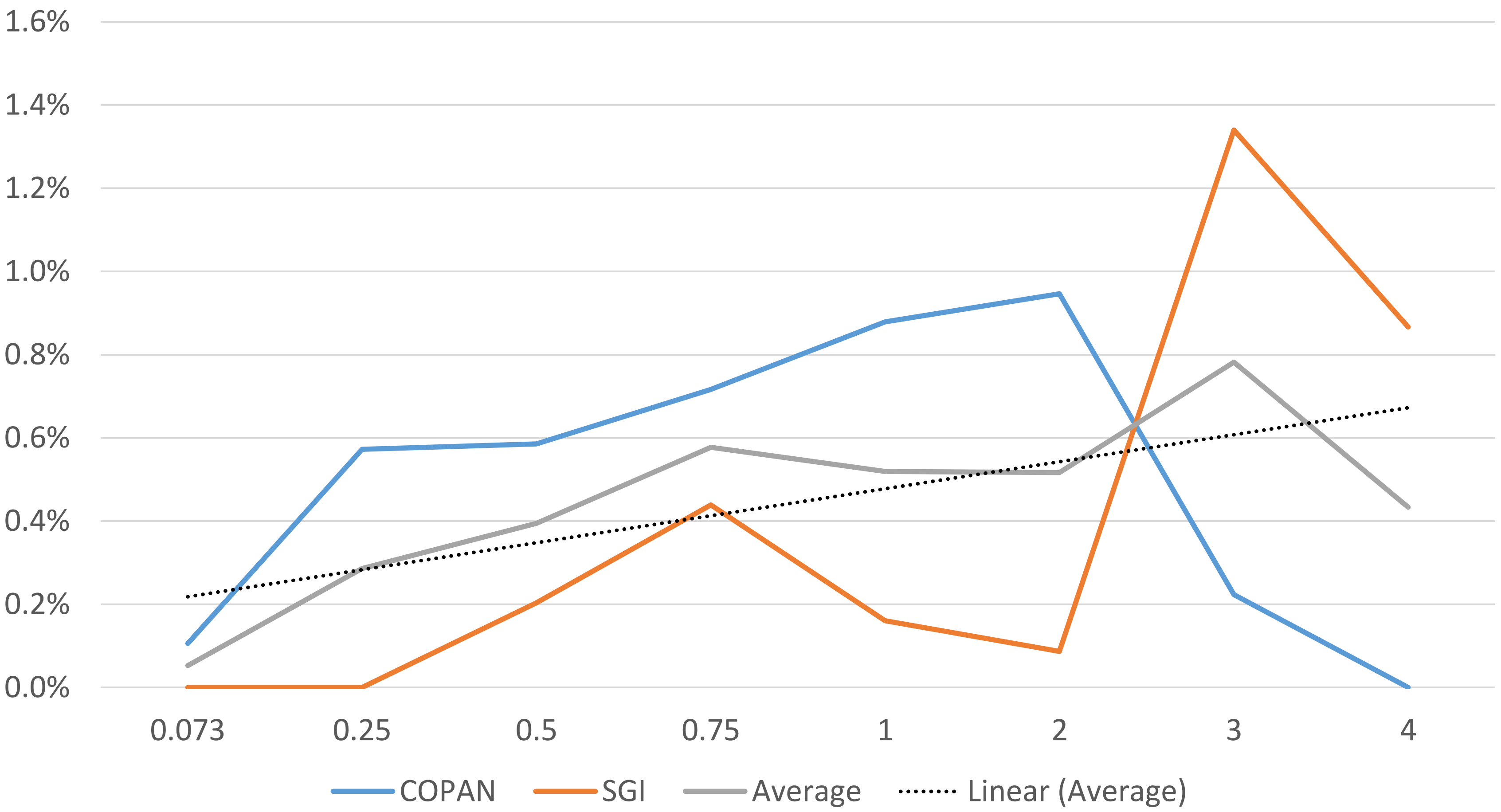
Clear outlier, massive disk replacement likely caused by problem with batch

Scrubbed Failure Data



Observed Annual Failure Rate

Organized by disk capacity



©2016 SGI

Summary & Conclusions

- **YES** – spinning disks down appears to improve AFR compared to manufacturer’s, extending drive’s life
  - COPAN reported 4x and better
  - Manufacturer’s AFR is 0.73-1.4%
  - **Observed AFR is 0.3-0.6%**
  - AFR tends to increase with disk capacity
- **Observed improvement is about 2x**

Beyond COPAN: SGI JBFS

High-Performance & Cost Optimized for SGI DMF

- JBFS is an acronym for JBOD File System
- SGI JBFS provides mounting services and serial access to disk media
- SGI JBFS enables rich capabilities for power management on any JBOD hardware supporting per-drive power control
- Leverages significant data management software IP that came with SGI's acquisition of COPAN
- Ability to deliver data access and I/O performance significantly beyond alternatives

SGI JBFS

- Any Number of LUNs or Devices
- Full Power Control
- Recoverability
- High-Performance
- Flexible to Many Media Types

SGI DMF7

Data Management for HPC

PBS Professional® slurm workload manager

HPC and HPDA Compute Nodes

sgi Multi-Rail Lustre intel

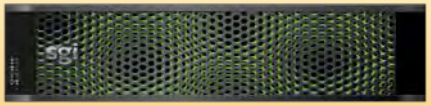
IME® Burst Buffer

High-Performance Parallel File Systems

NetApp DDN® STORAGE intel SEAGATE

TierZero™: Dynamic POSIX Namespaces

On-demand Flash-based filesystems collocated with compute and managed by DMF



SGI DMF v7

Scalable Data Management Platform

- 100s of billions of objects
- Dynamic Namespace
- Job Scheduler Integration

Tape Libraries

SPECTRA ORACLE IBM

Lowest Cost & High Durability

Cold Storage with SGI JBFS

sgi

Low Cost & High Performance

Cloud / Object Storage via S3 API

S3

High Scalability and Geo-Distribution

sgi®