Exa- to Yotta-scale Data
An Optimistic View

Rob Farber
PNNL

Pacific Northwest
NATIONAL LABORATORY

Optimistic about Storage Bandwidth

- Wide-striping increases bandwidth
  - PNNL MPP2 (~4k disks) achieved through Lustre
    - 86 GB/s sustained write,126 GB/s sustained read
  - PNNL Chinook has 20k new technology disks

Machine	OST/wide-stripe	disk MB/s	Weeks/Exa	Weeks/Zetta
PNNL MPP2	3990	30.00	13.81	13813.19
PNNL Chinook	20000 (5x)	100.00	0.83	826.72
2013	100000 (5x)	300.00		55.11
2018	500000 (5x)	700.00		4.72

Optimistic about Seek Operations

- Current disk drives offer about 200 seek/s per drive
- Current SATA Solid-State drives offer about 10k seek/s per drive (tomshardware.com)
- PCI-E flash offer about 100k seek/s per unit

Concerns

- Need to get wide-striping working in production
  - Currently one slow drive can rate limit the entire system
- Failure and bit-rot in Exa- to Yotta-scale storage
  - Can error correction fix this?
    - I'm hopeful – especially with solid-state
- Space, Power, Cost issues
  - Potential for large-capacity, low-cost solid-state with 3D manufacturing
  - Potential for spin-up/spin-down or power-on/power-off with SSDs
- Limitations of the file-system software
  - Performance CPU appears to be a limitation for path lookup (e.g test by running entirely out of RAM)
- Others
  - Backup!
    - (If for no other reason than everyone mistakenly deletes files)