



# Exascale Storage Challenges

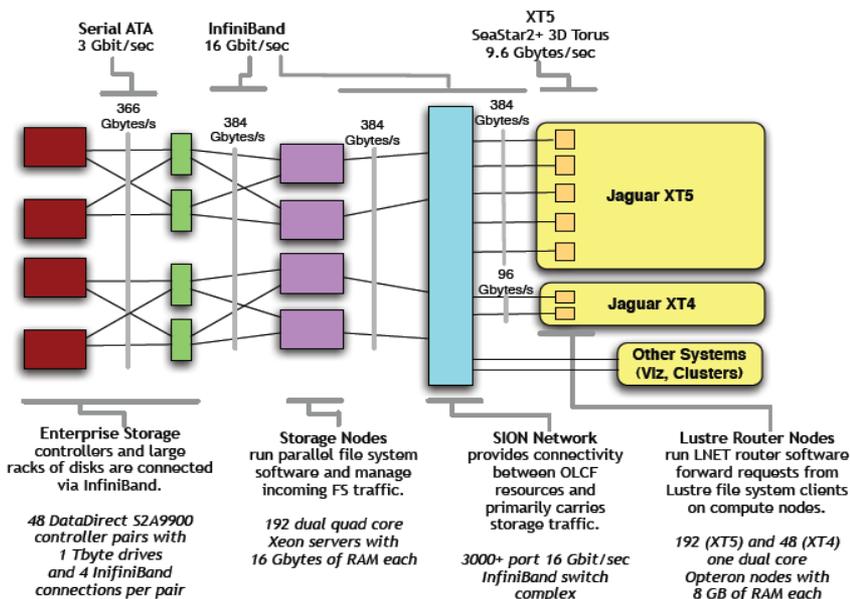
**Roger Haskin**  
**Senior Manager, File Systems**  
**IBM Almaden Research Center**

**Note: the views expressed herein are solely those of the author and not the IBM Corp.**

## Panel questions:

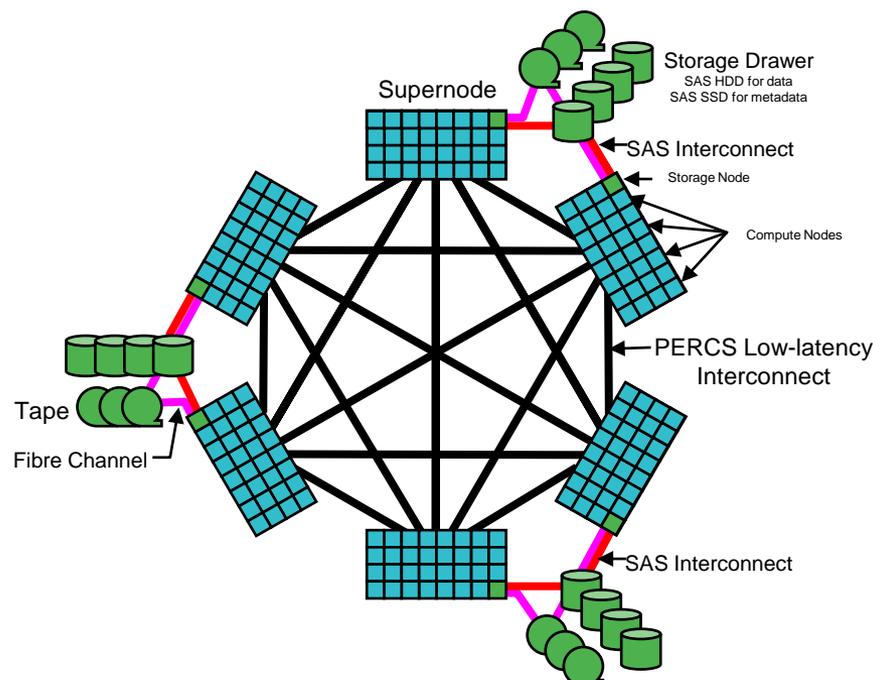
- **What do you *\*think\** parallel FS (hw and sw) will look like in 10 years?**
- **What do you *\*wish\** parallel FS (hw and sw) will look like in 10 years?**
- **What do you think will be the top 3 challenges in the next 10 years to provide IO and associated infrastructure into Exa and enormous data analytics?**
- **Where will HPC be able to find good external tools to leverage (e.g. Google, LSST, etc)?**
- **How does the hype about clouds and virtualization add challenges or opportunities to file system and storage SLAs both at future exa as well as at current more moderate scale?**

# Large-scale parallel FS today – two design points

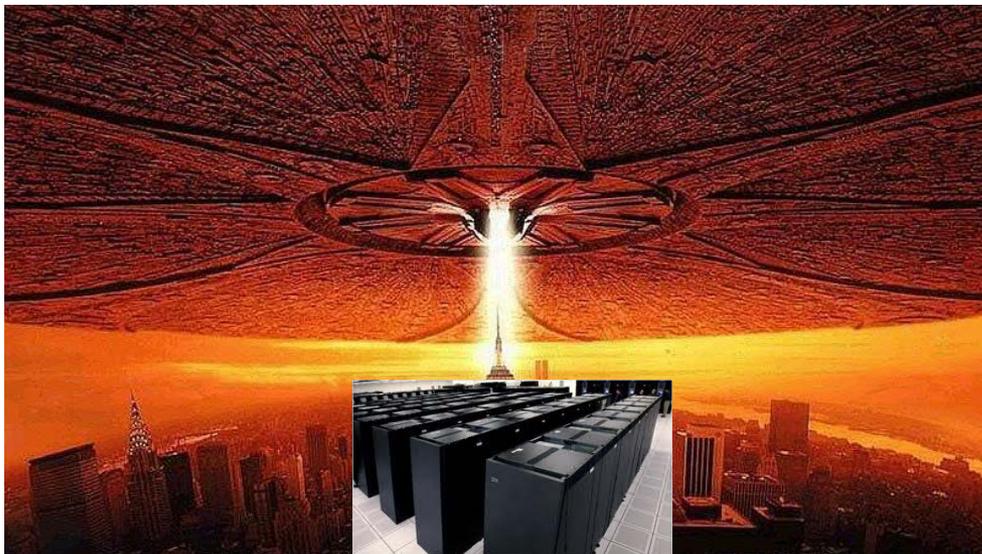


- File system runs directly on the compute nodes
- Storage nodes and associated physical storage are embedded in the compute racks (building blocks)
  - JBOD w/ software RAID in storage nodes
  - Metadata nodes (e.g. lock manager) also embedded
- All communication and I/O use the internal fabric
- IBM PERCS (Blue Waters), GPFS

- File system runs on I/O nodes inside the system
  - Each I/O node serves multiple (up to dozens) of compute nodes
- Storage nodes and associated physical storage are external to the system
- I/O nodes connect ...
  - to compute node via internal fabric (e.g. torus)
  - to storage via separate external fabric (e.g. Infiniband)
- Cray, Blue Gene, Lustre, Panasas, GPFS



# Parallel File Systems at exascale

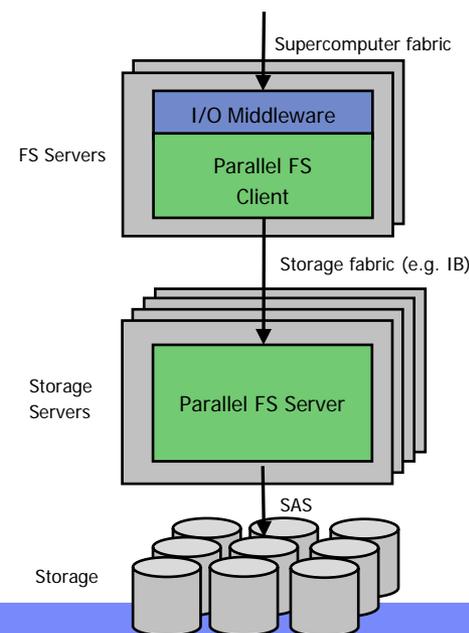


Recently discovered Nostradamus quatrains predict that in 2018...

- Exascale will complete the exile of the file system from the supercomputer
- FS and storage will reside externally, analogous to a file server, with some kind of point-to-point connection to the switch
- The system architect, who at best regards storage as an afterthought, will be liberated to forget about it completely

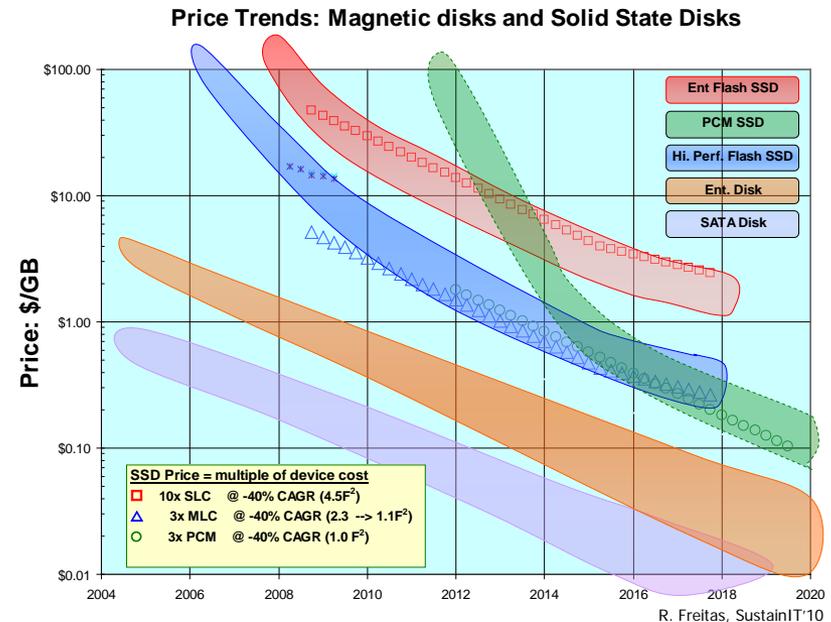
But that's a good thing!

- Exascale has too many compute nodes to run the file system on them directly
- Aggressive exascale packaging will complicate putting storage inside the machine
- Internal I/O nodes aren't usually that suited to their task
  - Insufficient memory, wimpy power efficient processors, ...
- Eliminates the need to port the FS to the supercomputer environment
  - Architecture, OS/distribution software level, ...



# Challenge 1: what to do about solid state?

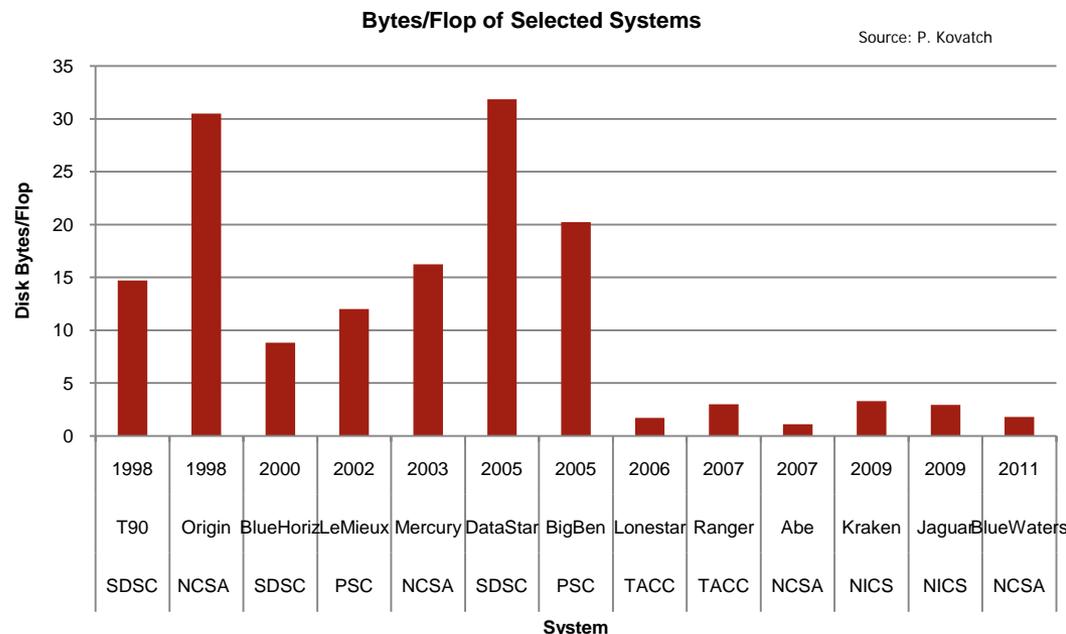
- **Solid state will remain too expensive to use for everything**
- **File system can use it in judicious quantities**
  - To speed metadata operations
  - For hot data, as part of the storage hierarchy
- **What about solid state in the compute nodes?**
  - Could run the file system inside, like ASF, or...
  - PCM can be addressed like DRAM, so persistent GSM model may be better



	Flash*	PCM*
Cell Size	2 F <sup>2</sup>	5.8 F <sup>2</sup>
Read	20 us.	1 us.
Write	200 us	3-5 us
Erase	2ms	n/a
Endurance	low	high
Access	Page	Bit

## Challenge 2: Cost of storage

- **\$/byte not scaling with \$/flop**
- **Acceptable cost of storage is around 5-15% of the system**
- **As a result the traditional scaling ratios are out the window (see graph)**
- **Things will be much worse for exascale (see table)**
  - Can we live with a million disks?
  - Can we live without them?
  - Will there even be such things?

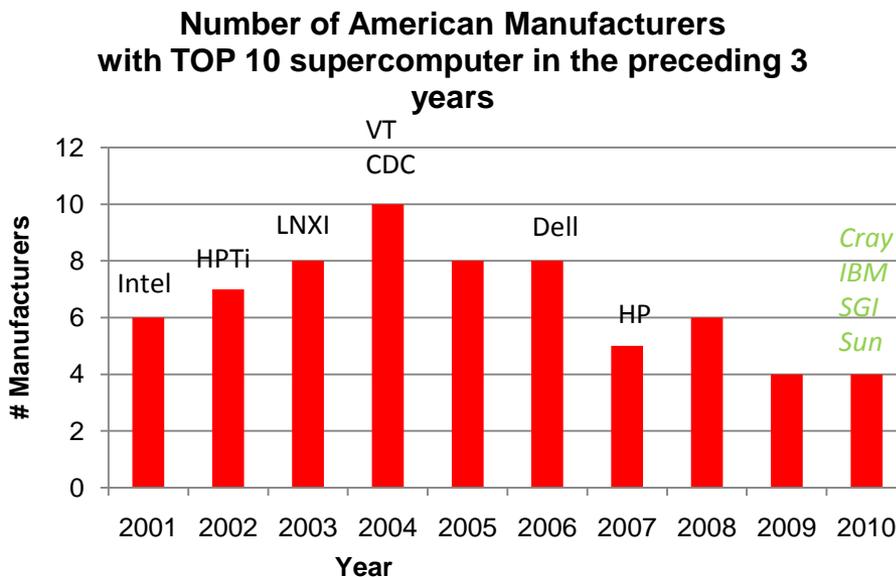


B/F	0.1	0.2	0.5	1	2	5	10	20	50
B	1.00E+17	2.00E+17	5.00E+17	1.00E+18	2.00E+18	5.00E+18	1.00E+19	2.00E+19	5.00E+19
Drives	11844	23688	59219	118438	236875	592188	1184375	2368751	5921877
TB/S	1.473553	2.947105	7.367763	14.73553	29.47105	73.677629	147.3553	294.7105	736.7763
B/S/F	1.47E-06	2.95E-06	7.37E-06	1.47E-05	2.95E-05	7.37E-05	1.47E-04	2.95E-04	7.37E-04
x Memory	3	20	50	100	200	500	1000	2000	5000
Ckpt time	20359	10179	4072	2036	1018	407	204	102	41

## Challenge 3: Who's going to build it?

### ■ Who's going to build it?

- Is exascale storage a viable market to interest the major vendors?
  - The existing parallel file systems each represent a major investment that the respective vendors are increasingly trying to monetize
  - Major software vendors (Oracle, MS) have revenue 20x R&D cost
- Can research consortia and/or startups build, test, and support a reliable system at exascale?
  - Typically 5 years to maturity
  - Test infrastructure represents a substantial capital investment



Source: [www.top500.org](http://www.top500.org)

# Questions?