

Exa-Scale FSIO

Can we get there?

Can we afford to?

07/2010

Gary Grider, LANL

LA-UR 10-04611

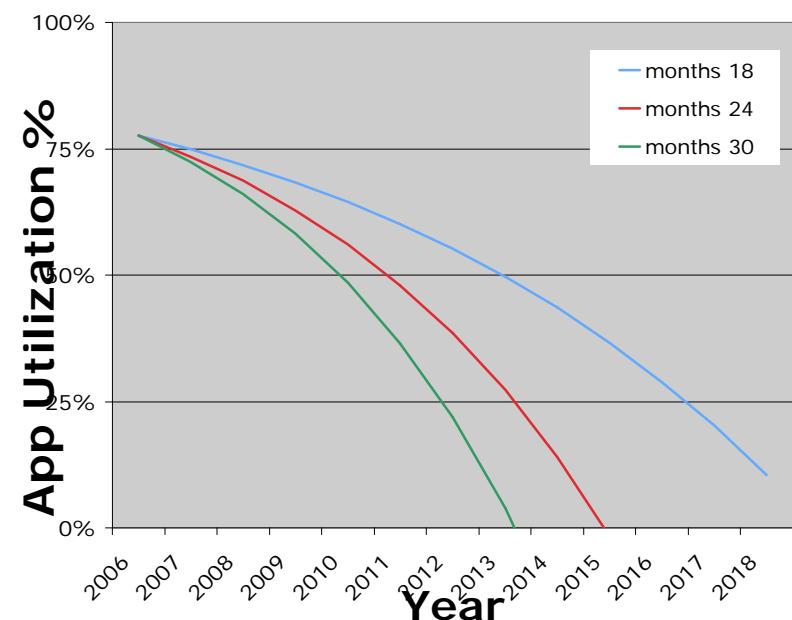
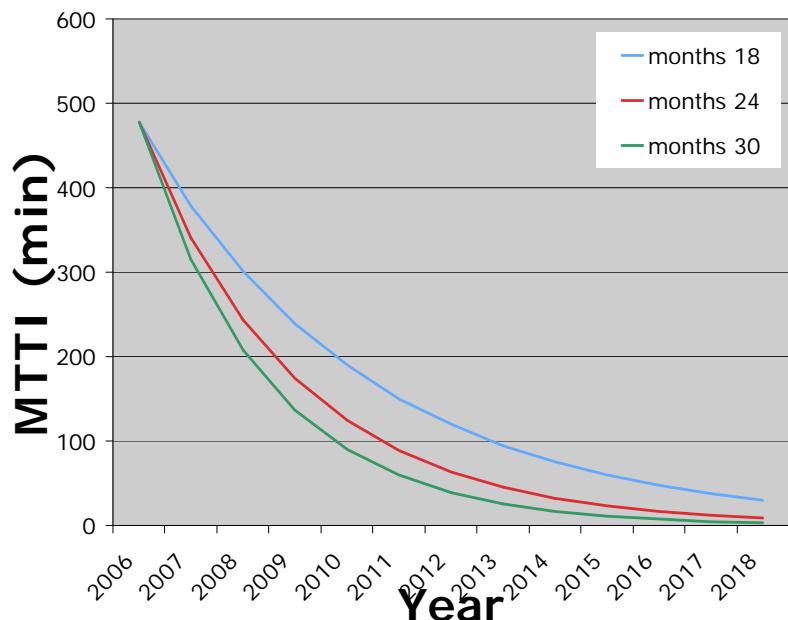


Potential System Architecture Targets

System attributes	2010	“2015”		“2018”	
System peak	2 Peta	200 Petaflop/sec		1 Exaflop/sec	
Power	6 MW	15 MW		20 MW	
System memory	0.3 PB	5 PB		32-64 PB	
Node performance	125 GF	0.5 TF	7 TF	1 TF	10 TF
Node memory BW	25 GB/s	0.1 TB/sec	1 TB/sec	0.4 TB/sec	4 TB/sec
Node concurrency	12	O(100)	O(1,000)	O(1,000)	O(10,000)
System size (nodes)	18,700	50,000	5,000	1,000,000	100,000
Total Node Interconnect BW	1.5 GB/s	20 GB/sec		200 GB/sec	
MTTI	days	O(1day)		O(1 day)	

Gloom and Doom from 2006/

- Petascale computing is coming
 - Orders of magnitude more components
 - **Orders of magnitude more failures**
- Need raw data for better understanding of failures



Past and Future Assumptions

- Past
 - All disk
 - Constant ratio of total \$ to IO infra \$
 - Machines wont accelerate their reliability per flop
- Future
 - Not necessarily all disk
 - Not necessarily same % but close
 - Machines may make accelerate progress on reliability/flop due to integration and industry desire to have constant reliability per socket

Can we do defensive IO at Exascale?

- If we loosen assumptions?
- If we can do it can we afford to do it?

New Assumptions

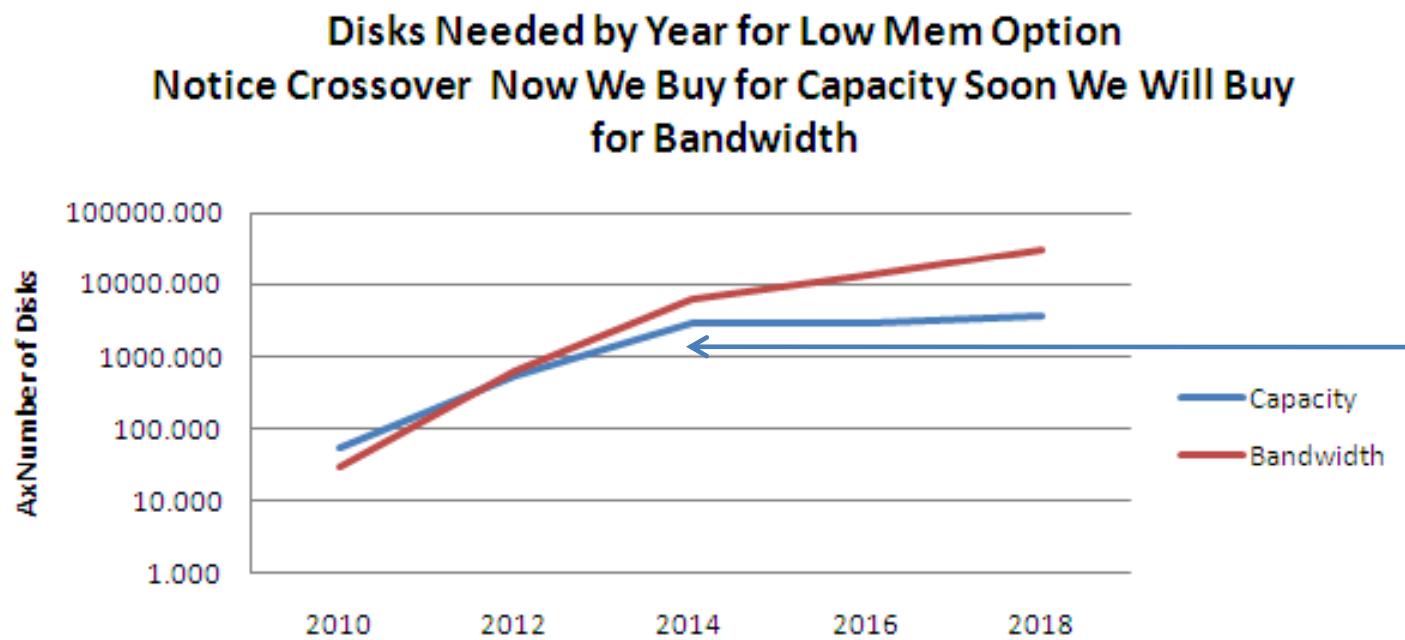
Year	EF	2010	2012	2014	2016	2018
PF		1.000	20.00	200.00	400.00	1000.00
mem low PB		0.004	0.07	0.72	1.44	3.60
mem med PB		0.020	0.40	4.00	8.00	20.00
mem high PB		0.300	6.00	60.00	120.00	300.00
Num Full Mem Cap		30	30	30	30	30
Size Scratch PB low		0.108	2.16	21.60	43.20	108.00
Size Scratch PB med		0.600	12.00	120.00	240.00	600.00
Size Scratch PB high		9.000	180.00	1800.00	3600.00	9000.00
Time to dump Secs		1200.000	800.00	600.00	400.00	300.00
Ckpt BW low TB/s		0.003	0.09	1.20	3.60	12.00
Ckpt BW med TB/s		0.017	0.50	6.67	20.00	66.67
Ckpt BW high TB/s		0.250	7.50	100.00	300.00	1000.00
Disk Capacity TB		2.000	3.92	7.68	15.06	29.52
Disk Speed MB/s	100	100.000	140.00	196.00	274.40	384.16
IO node thrput GB/s	100	1.000	2.000	4.000	8.000	16.000

Based On

- DARPA Exa Study for machine sizes, mtti, etc. except 20 PB med mem machine and 30 dumps in scratch
- Seagate Disk Capacity/Size/Pricing/Power (not shown)
- Micron Flash Capacity/Size/Pricing/Power (not shown)
- 10% of mtti as dump time

I wanted to know – what miracles will we need and to get past what problems.

Status Quo: Use Disk Based Shared Global Parallel File System to Provide Dump Space



Notice that using these modeling parameters, we finally reach the predicted cross over point of buying disk for BW and not Capacity in 2012

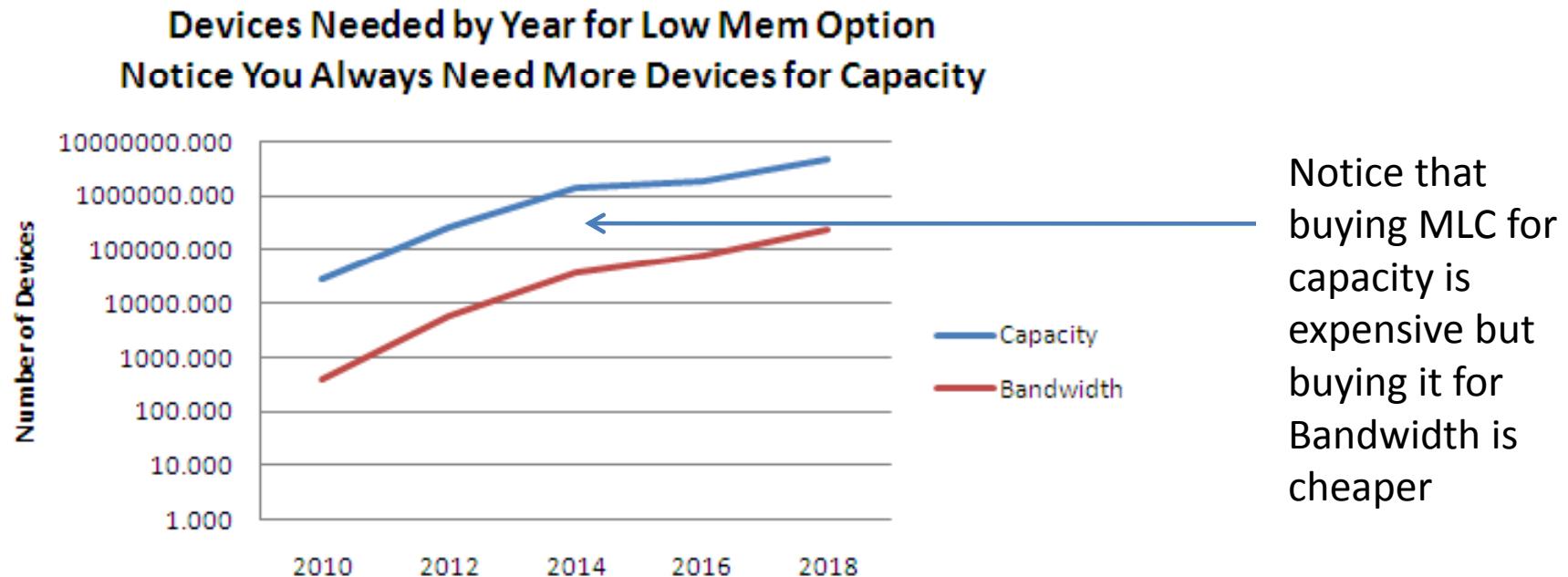
2018 medium memory machine

- 4166 IO nodes, 175k disks
- File System sees 50-100k way parallelism (assumes IOFSL)
- \$225M pessimistic purchase (assumes no technologies pushing disk other than Flash)
- Power 1.5MWatts

Miracle Needed!

Buying disk for capacity is reasonably priced but buying disk for bandwidth gets expensive fast!

Use MLC Based Shared Global Parallel File System to Provide Dump Space

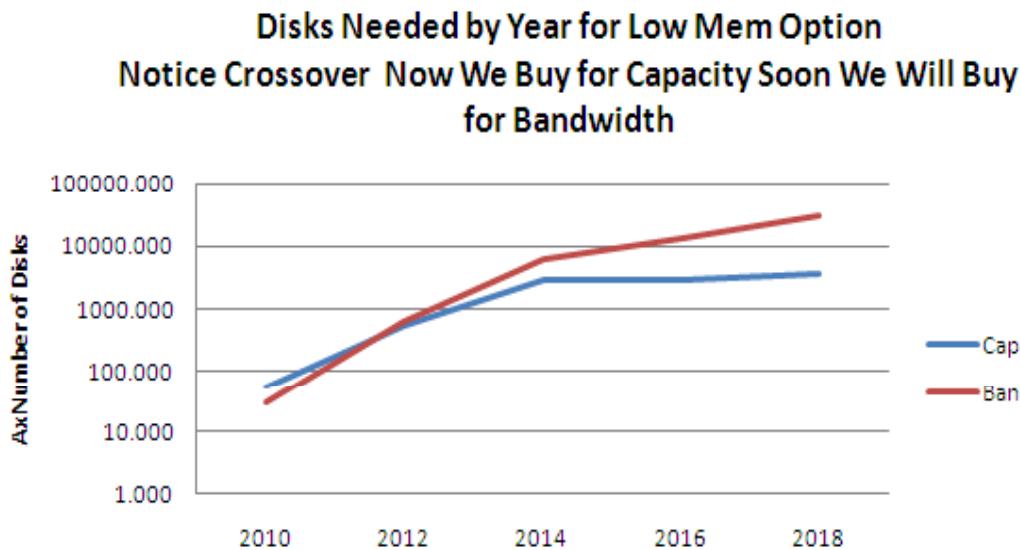
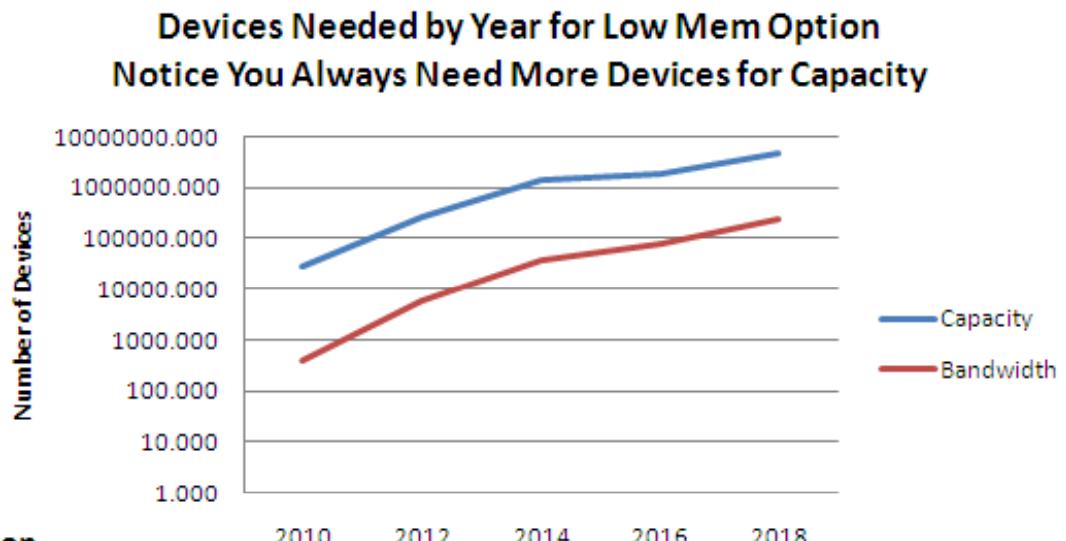


2018 medium memory machine

- 4166 IO nodes
- File System sees 50-100k way parallelism (assumes IOFSL)
- \$625M pessimistic purchase (assumes no technologies pushing disk other than Flash)
Miracle Needed!
- Power 2.5MWatts (have to buy so much to get capacity)

Lets Try to Buy Disks for Capacity and MLC for Bandwidth == Hybrid Model

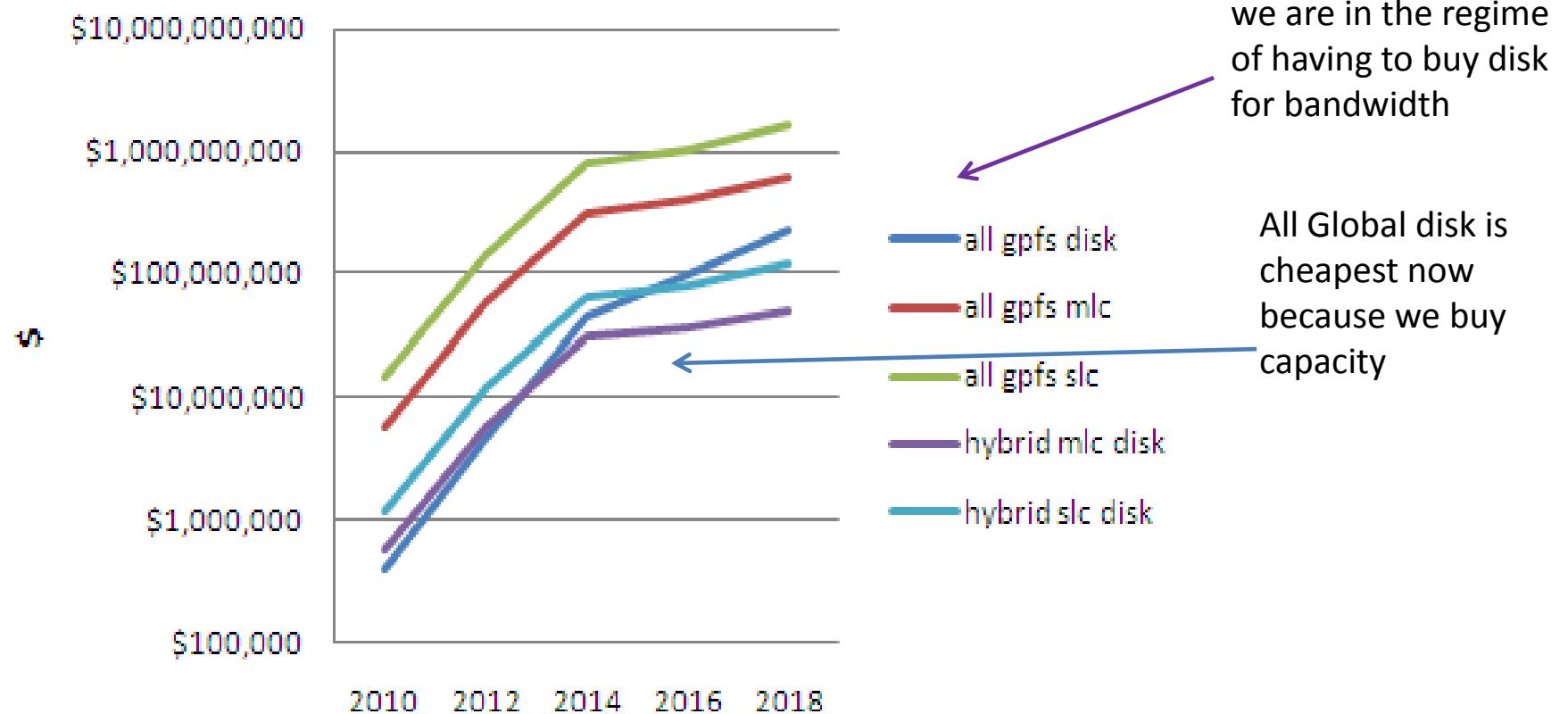
- Twin tailed non global MLC connected to NN compute nodes N I/O Nodes, compute nodes dump to MLC at 10% MTI time and IO nodes bleed to global disk without causing jitter at 1/10th the dump-burst data rate or less



- 3 memory dumps in MLC
- 30 dumps in global disk

Hybrid MLC burst / Disk Global

Med Mem

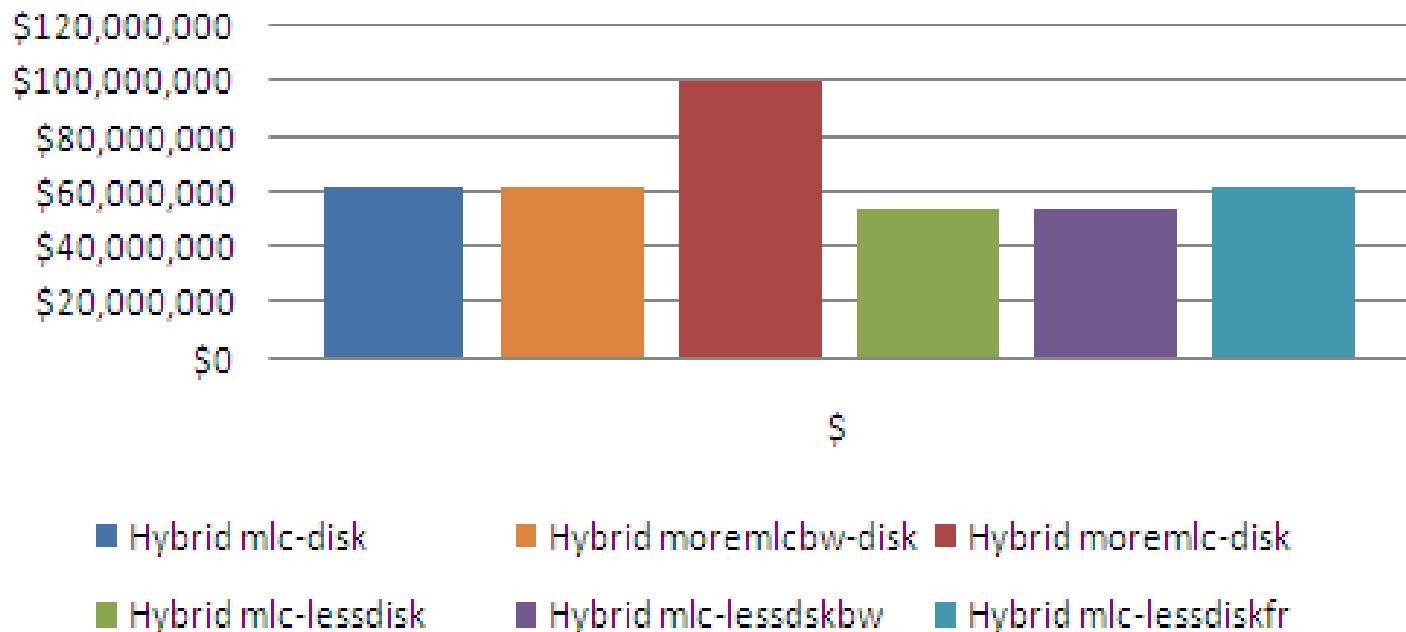


2018 med mem mach

- 416 IO Nodes, 20k disks not much of a stretch
- Disk FS sees modest parallelism assumes IOFSL/burstbuffer etc.)
- \$60M pessimistic purchase - worst case (all migrated to disk and tech price)
- Power 2.2MWatts

Hybrid MLC burst / Disk Global

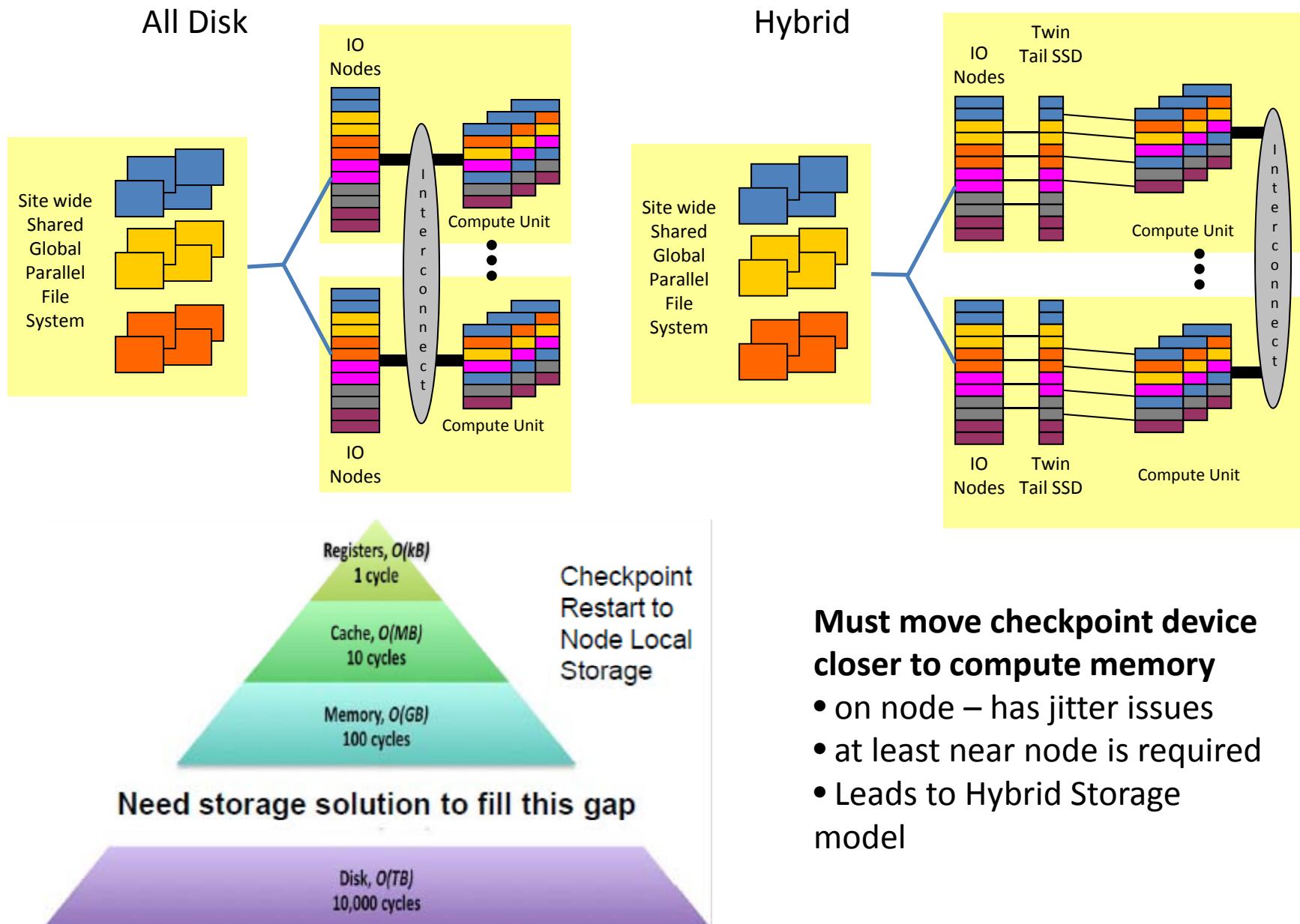
First Order 2018 Med Memory Sensitivity Analysis



Cost Driver Sensitivity

- More MLC BW (free – capacity driver)
- More MLC Cap (costly – capacity driver)
- Less Disk Cap (small savings (MLC capacity driver))
- Less Disk BW (small savings controllers/ION etc. (MLC capacity driver))
- Less Frequent MLC to Disk (no savings, Disk Capacity Driver)

Hybrid Disk/SSD



A Feasible Evolutionary Approach?

Summary: Issue	Action
Probably pretty close on storage densities, bandwidths, and costs, in fact it may be a bit conservative (maybe more than a bit)	Continue to update model
Based heavily on MTTI assumptions in the DARPA study and that study indicates a pretty large per socket improvement in MTTI without good substantiation	Get serious about measuring and predicting this!
Assumes that existing techniques like RAID or other redundant techniques will keep the burst buffer working often enough to not have issues without substantiation	Keep our eye on Flash reliability – prospects are good given wide use
Assumes existing RAS techniques for file systems will be able to keep up without substantiation	Keep our eye on this
Have to have burst buffer so we will need software to manage MLC burst buffer, with bleed to global disk	SCR LLNL / PLFS LANL / ADIOS ORNL / MPI-IO ANL. Zest PSC, ...
Assumes flattening to get high % of peak on disks (like log structure)	PLFS LANL / ADIOS ORNL / MPI-IO ANL, Zest PSC, ...
Need a way to deal with large numbers of files	Giga+, etc.

Maybe we can get to Exascale with evolution only, but it would be pretty sad if we didn't also attempt some more fundamental revolutionary approaches!