Panel: Leveraging FLASH in Integrated, Scalable Systems

Chris Jordan TACC

Mike Vildibill DDN

Brian Van Essen LLNL

Gary Grider LANL

Panel Focus and Questions

- Explain the scalable system you deployed/deploying flash into
- Explain scalable speed / feeds
- Explain scalable operation use cases/software you have added to make the solution useful
- Why flash was chosen?
- What aspect of FLASH helped you achieve your scalability goals - simplified management, performance (latency, bandwidth), other?
- How/why did you make the tradeoff of giving up capacity to add flash (for whatever reason)?
- How do you deal with durability/lifetime mgmt at scale in your application/system
- How is maintenance regarding wear dealt with, replacement or expendable or other

Deeper Storage Hierarchy for Trinity Probably too Deep

Gary Grider

Division Leader
High Performance Computing Division
Los Alamos National Laboratory
ggrider@lanl.gov

Excerpts from LA-UR 14-26443
Oct 2014

HPC at LANL
The Edge of the Computing Envelope for Decades

- Maniac

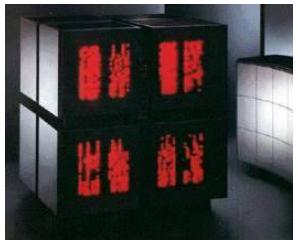
- IBM Stretch

- CDC

- Cray 1

- Cray X/Y

- CM-2



- CM-5

- SGI Blue Mountain

- DEC/HP Q

- IBM Cell Roadrunner

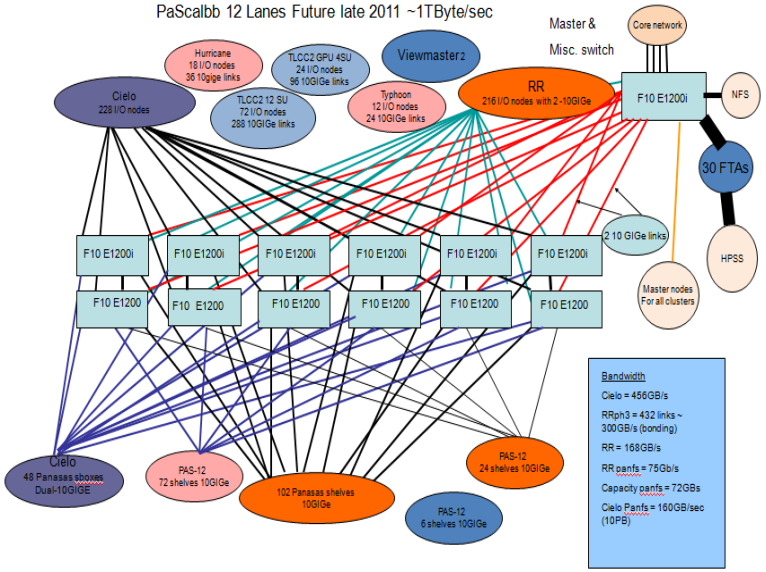
- Cray XE Cielo

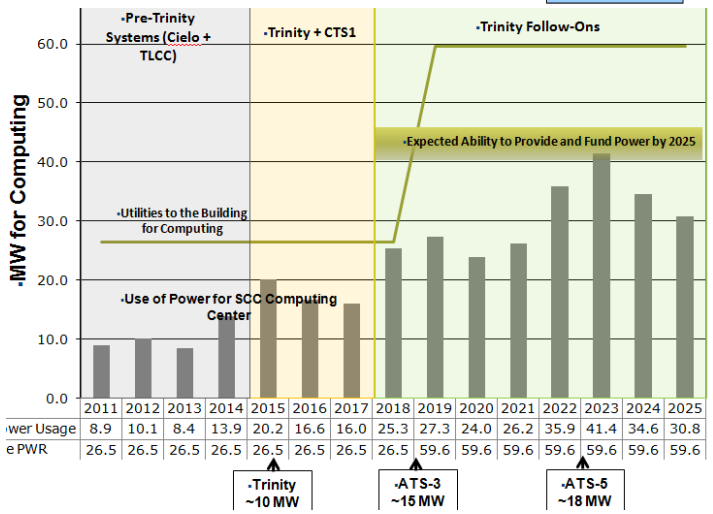
Current Machines

- Cielo (Cray XE6) 1.4 PF
- Luna (Appro) .5 PF
- Typhoon (Appro) 100 TF
- Mustang (Appro) 350 TF
- Moonlight (Appro) 488 TF
- Wolf (Appro) 200 TF
- Pinto (Appro) 50 TF
- Conejo (SGI) 50 TF
- Mapache (SGI) 50 TF
- Hobo (Appro) 300 node Data Intensive
- Helios (Cray XK) Data Intensive

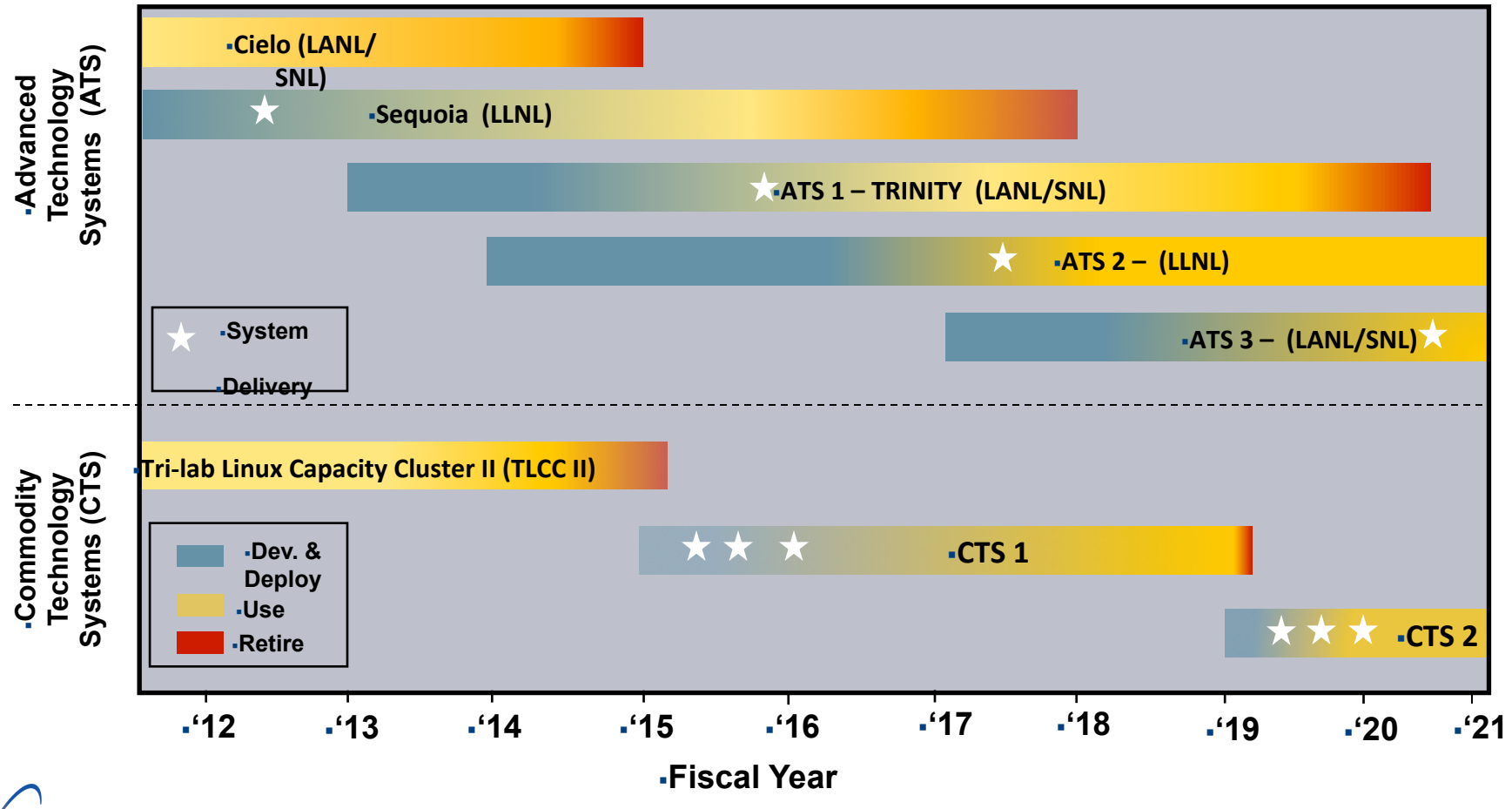
HPC Requires Pretty Big Infrastructure

- ~2 TB/sec SAN -> 10sTB/sec
- 16 PB Scratch File Systems _> .5 EB
- .5 EB Parallel Tape Archives -> 10 EB of Archive
- 20 MW -> 40 MW
- 100-200 M Gallons Water/Yr Evap





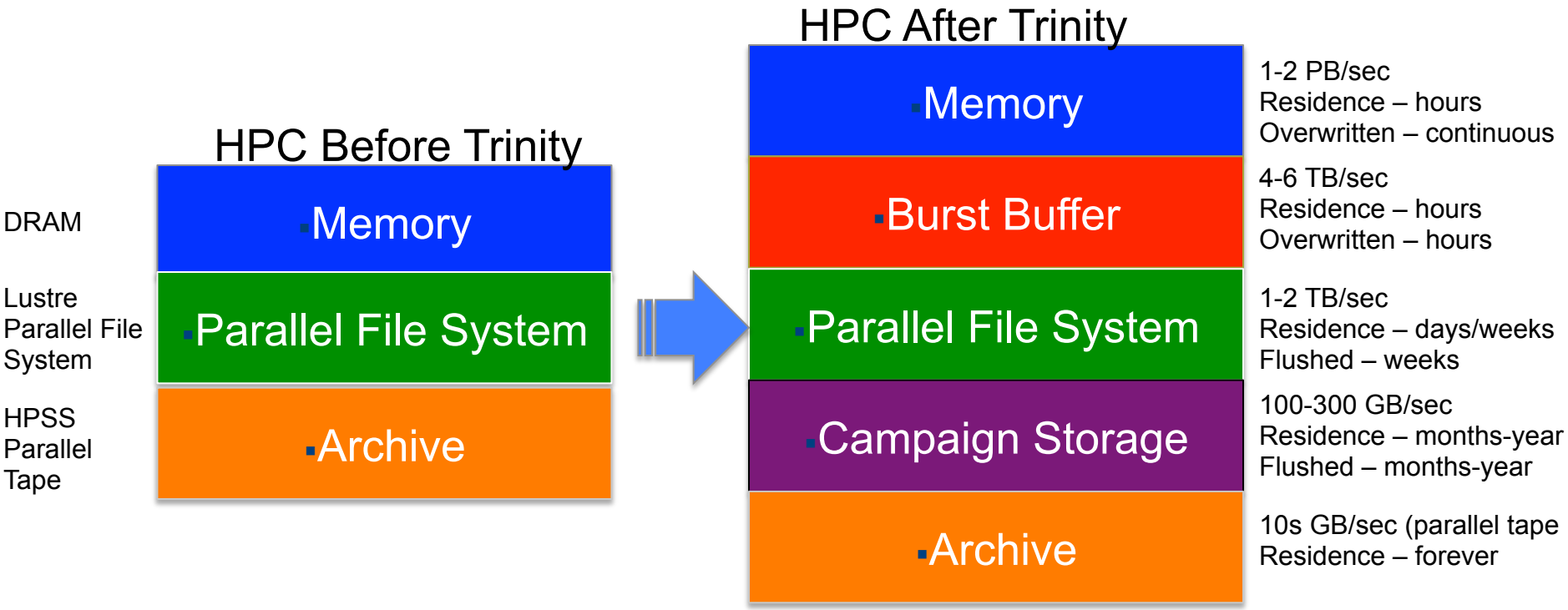
And the need for bigger machines just keeps growing!



Trinity

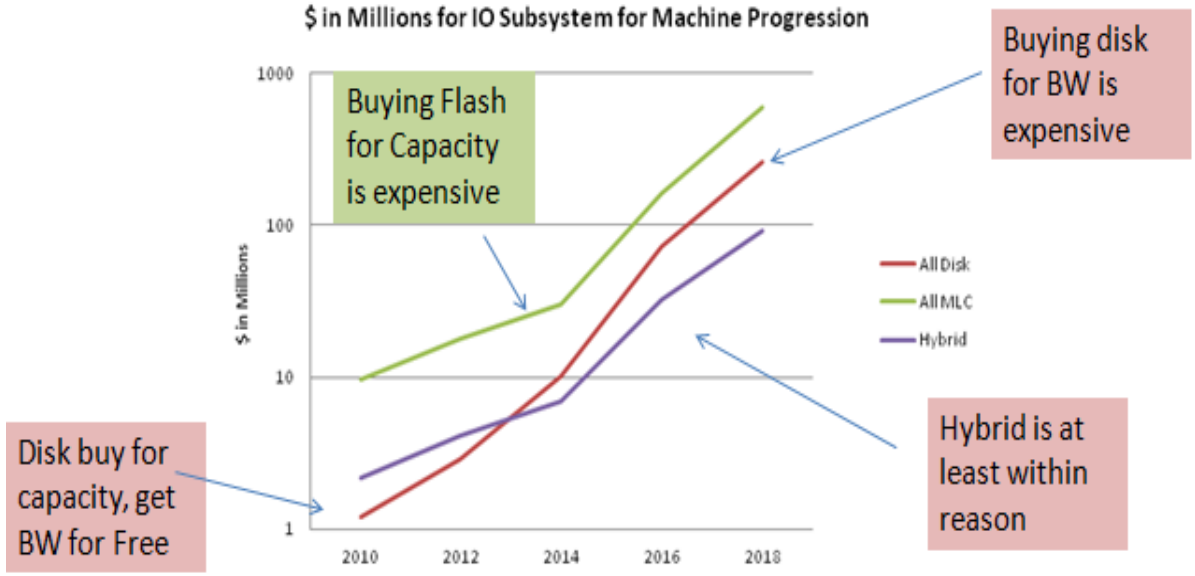
- ~21,000 nodes
- 1-2 M cores
- ~3 PB dram
- 6-8 PB flash burst buffer (4-6 TB/sec)
- 80-100 PB parallel file system (1-2 TB/sec)
- 300-500 PB campaign storage (50-100 GB/sec) growing to EB
- 8-12 Mwatts of power
- Begin install summer 2015
- **Typical 3D run might be 1 PB DRAM over ~1M cores for 6 months to 1 year!**

What are all these storage layers? Burst Buffers? Campaign Storage?



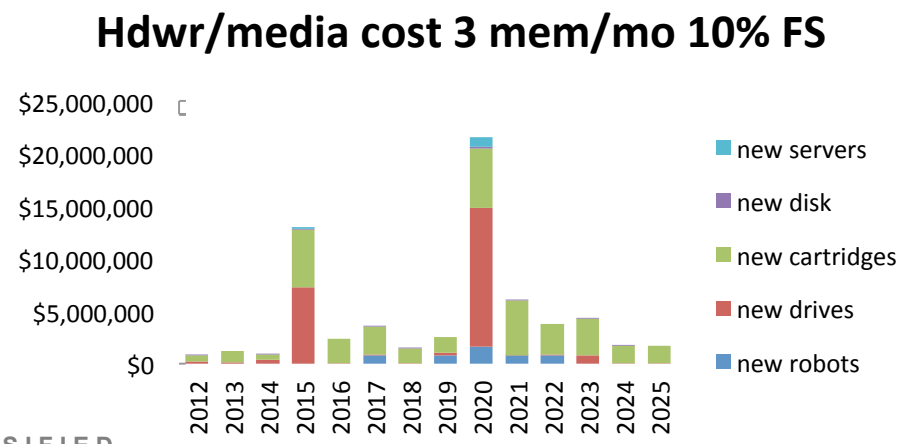
- **Why do we need all those layers?**
- **Economics and maturity**

Why Burst Buffers and Campaign



- **Economic modeling for large burst of data from memory shows bandwidth / capacity better matched for solid state storage near the compute nodes**

- **Economic modeling for archive shows bandwidth / capacity better matched for disk**



What about this campaign storage thing?

- Campaign storage will grow to Exabytes in a few years
- Bandwidth needs too high for parallel tape
- Number of disks implies the need for erasure based systems
- Why not borrow from the cloud storage community, object erasure systems. After all we are the same as Dropbox except our single images are a little bigger (say 5 orders of magnitude).
- Very parallel use of object erasure systems has promise for this need

What did you mean by maturity?

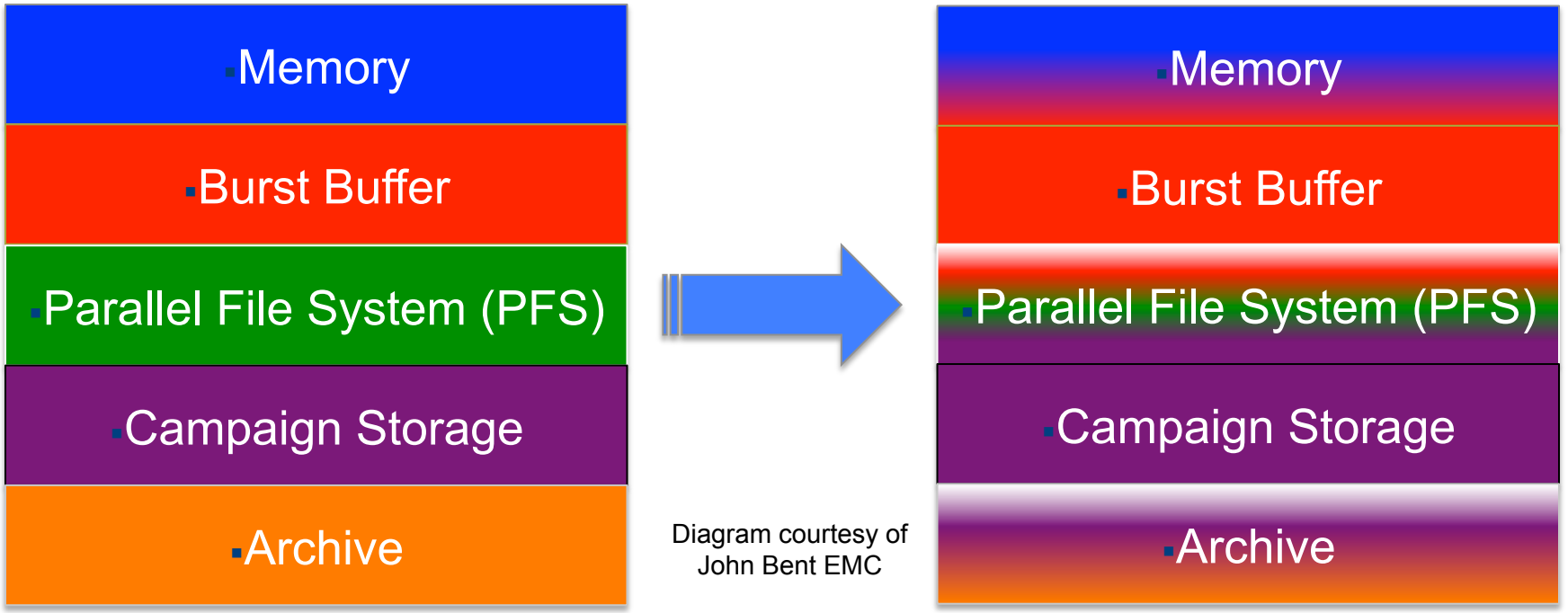


Diagram courtesy of John Bent EMC

- **If the Burst Buffer does its job very well and campaign storage is works out well, do we need a parallel file system anymore, or an archive? Maybe just a bw/iops tier and a capacity tier.**
- **Too soon to say, but this seems feasible longer term.**

Panel Focus and Questions

- **Explain use cases/software you have added to make the solution useful**
  - Checkpoint, Out of core, In Transit Analysis
  - Looks like a parallel file system, prejob stage, postjob destage (even on job failure)
- **Why flash was chosen**
  - Hybrid solution, Flash cheapest for BW and Disk cheapest for Capacity
  - Both procurement costs and power costs due to idle.
- **The tradeoff of giving up capacity to add flash (for whatever reason)?**
  - We bought what we needed of both
- **Durability/lifetime mgmt at scale in your application/system**
  - Write limited per job/flash allocation to rate of 10 overwrites/day
- **How is maintenance regarding wear dealt with**
  - In maintenance contract but only with rate limiter turned on

Thank You and RIPPFS