Hybrid Systems Use Cases

Things are about to get a whole lot messier

Nathan Rutman LAD 2018-09-24

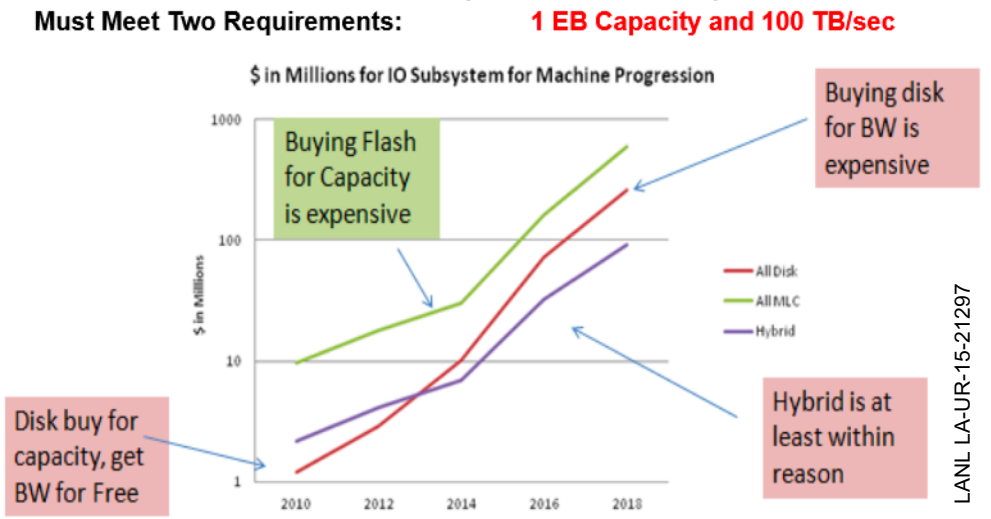
About this preso

- Why hybrid
- How will we use them
- What is my flash for, really
- How big
- Data movement is the answer (?)
- Lustre features that help

Economics

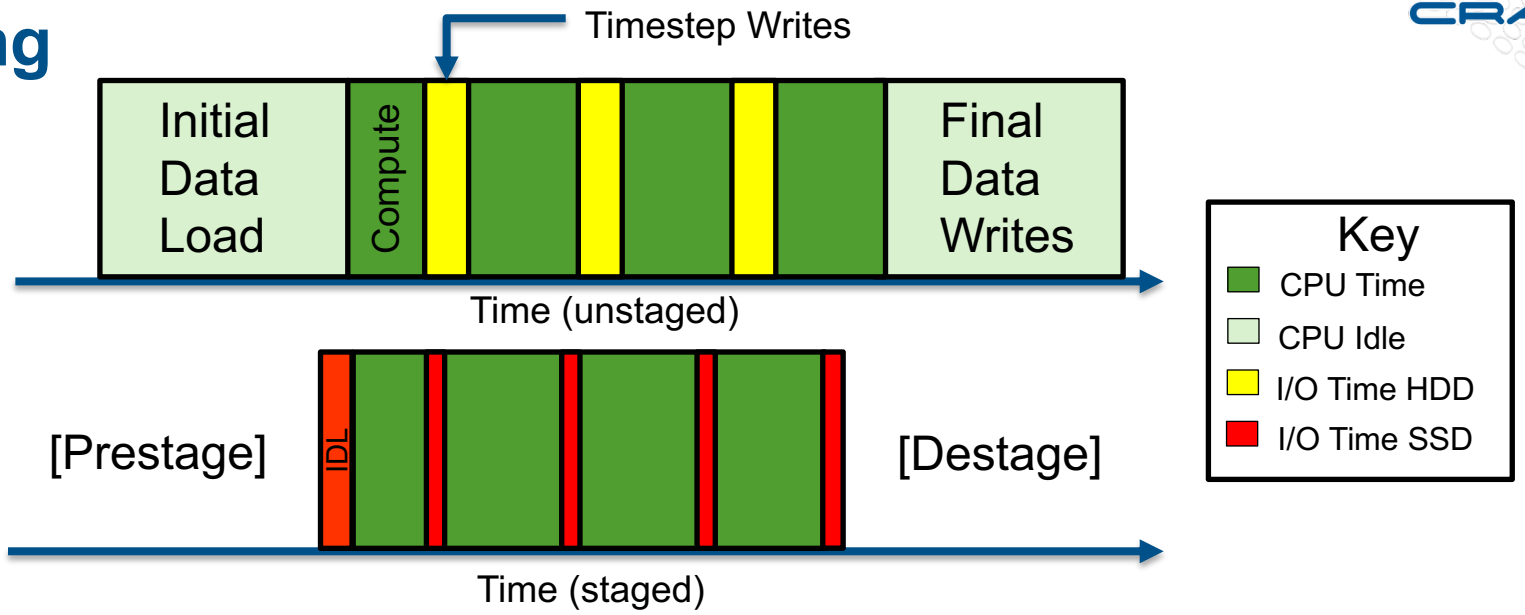
- With two media types, can optimize $ for two constraints (e.g. BW + Capacity)
- Great, buy a bunch of both
- Sum the speeds and sizes
- And we're done, right?

Not so fast…



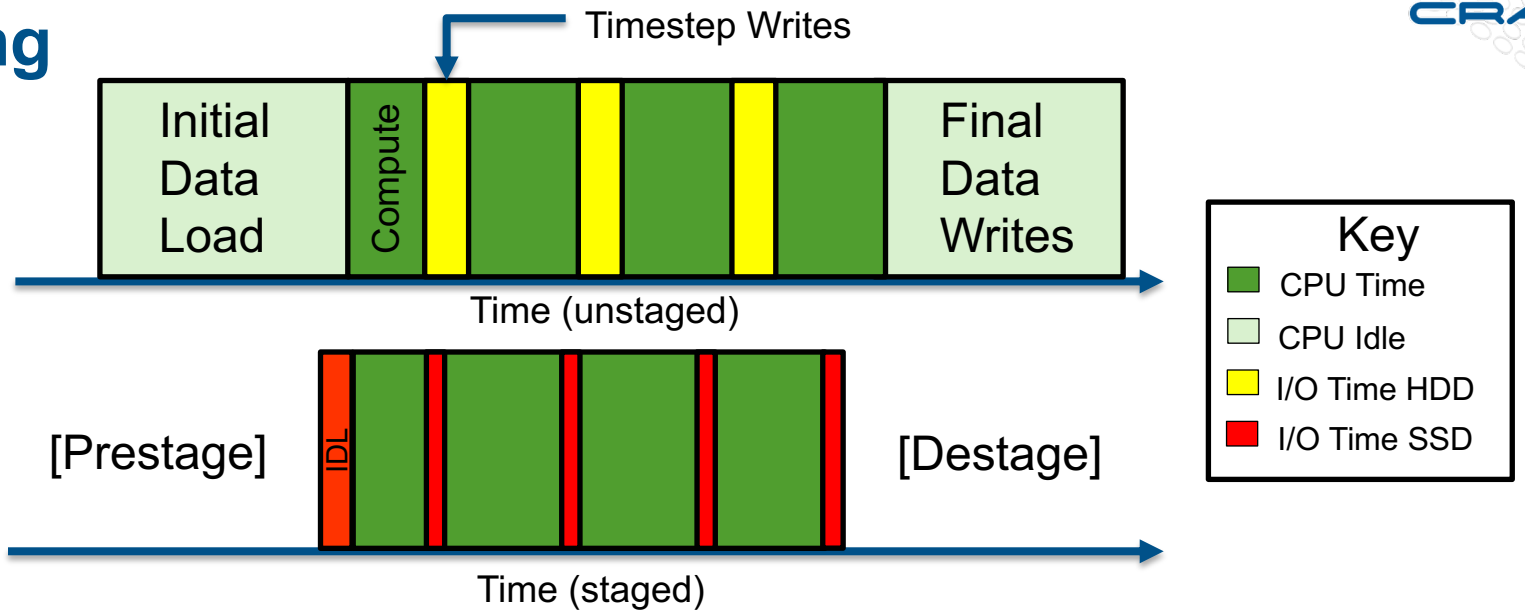
LANL LA-UR-15-21297

Staging



- Compress IDL & timestep writes to flash during “job”
- Reduce job wall time
- Keep CPUs busier

Staging



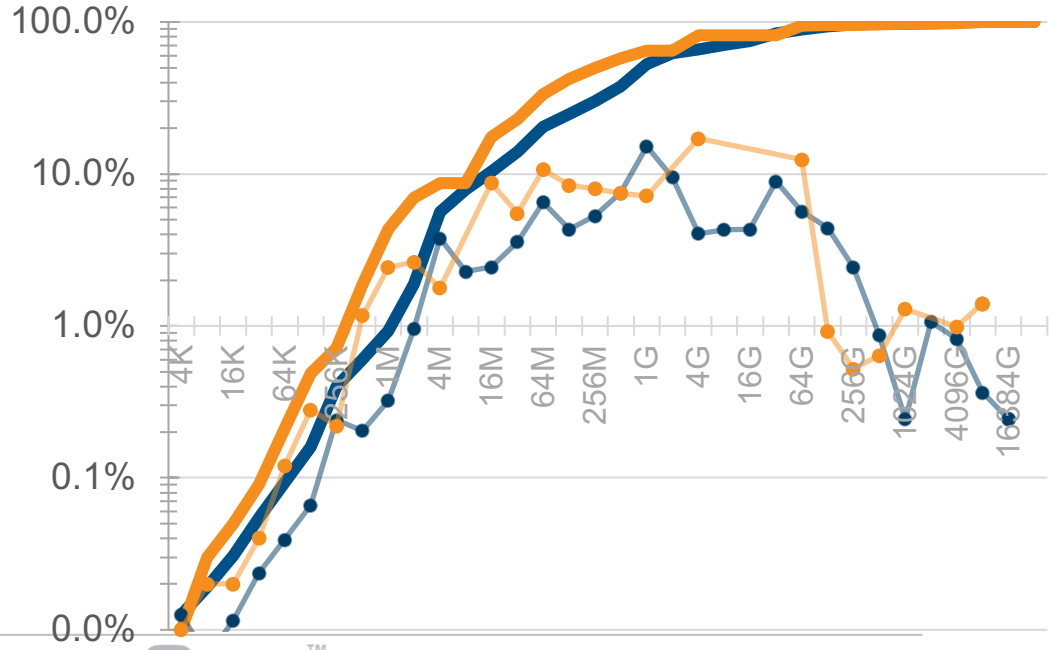
- Pipelining issue - requires intelligent scheduler
- Data movement requires bandwidth in HDD + SSD - twice
- No permanent flash files (need space)

Initial Placement

- Place (and leave) your data in the “right” place
- Stream to HDD OSTs, random to SSD OSTs
- So – flash not as a burst buffer, but as a random-IO tier
- Is that how you sized your flash?

Tier Sizing

- **We initially sized our flash for peak bandwidth**
- **But if we're going to leave files there, we really care about capacity**
  - SSD capacity for IOPS files
  - HDD capacity for streaming files
- **How big?**
  - Small files as a proxy for random
  - File size distributions



DoM

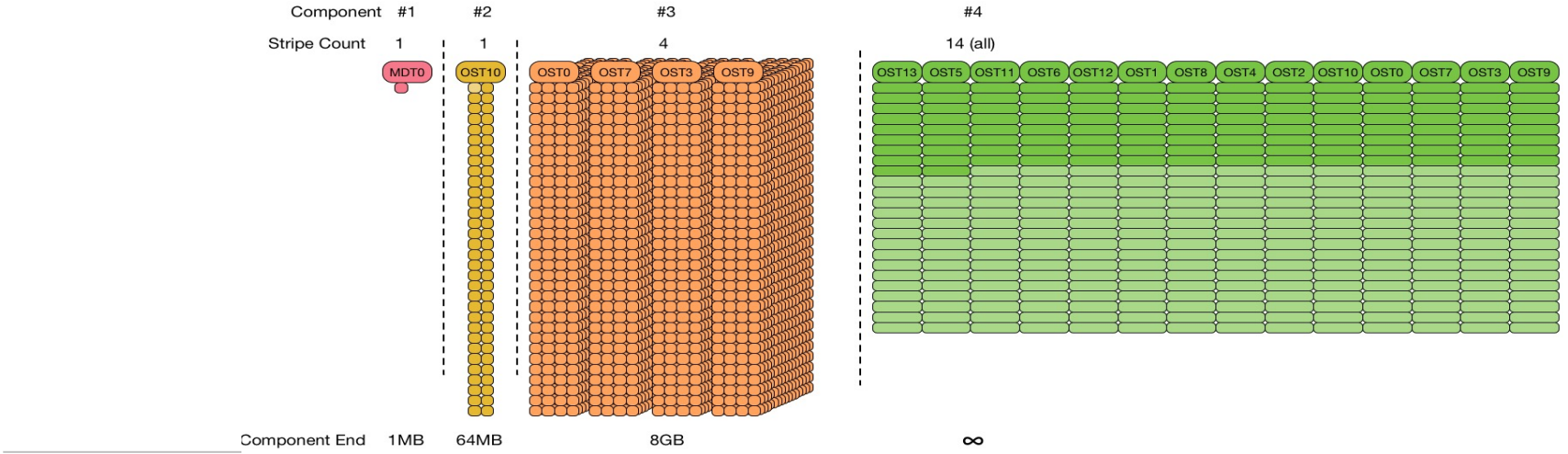
- If we really mean small files, flash DoM is better than flash OSTs
  - DoM for small files
  - Flash OSTs for large-but-random files
  - Disk for sequential files
- Beware new load on MDS's

Placement controls

- **Directory defaults for known apps**
  - Pools, striping params
- **PFL for unknowns**
  - Want “as much as possible” in flash, but no more
  - Thresholds based on file size distributions
- **Enforcement**
  - Default FS pool = HDD (or PFL)
  - Pool quotas ≠ project quotas!
    - LU-11023 Pool (not project) quotas

Two notes on PFL

- **Assume we want PFL to fill all tiers at the same % rate**
  - But this means flash is empty most of the FS life ☹
  - Can increase it to fill fast, but then we will have to move it ☹
- **Don't consider individual PFL files as "mixed media"**



Performance: does 5+3 = 8?

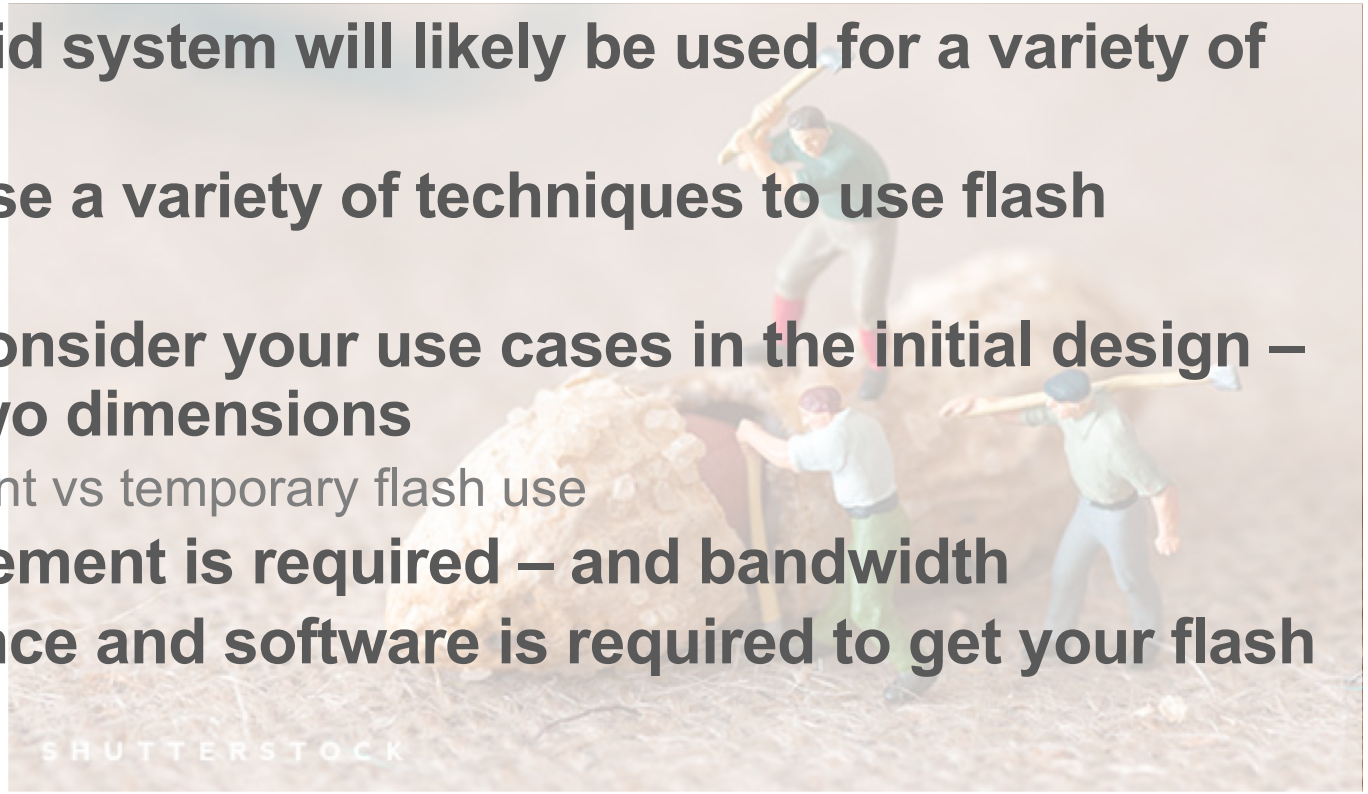
- If my flash tier goes at 5 TB/s, and my disk tier at 3 TB/s, can I get 8 TB/s for my app?
- Not with PFL - wrong SSD:HDD ratio
- FPP job with 5 nodes writing to SSD for every 3 nodes writing to HDD
  - Non-trivial setup in app and/or Lustre
- Is this how you sized your system?

When we get initial striping wrong

- **Can ENOSPC on small flash OSTs**
  - Spillover space – delayed allocation built on PFL
- **Or move/migrate files**
  - Requires policies and efficient copytools
  - Turning into an HSM problem
- **hsm migrate** LU-6081
- **hsm mirror sync**

All Together Now

- Your hybrid system will likely be used for a variety of purposes
- Need to use a variety of techniques to use flash optimally
- Need to consider your use cases in the initial design – not just two dimensions
  - Permanent vs temporary flash use
- Data movement is required – and bandwidth
- Maintenance and software is required to get your flash benefits



Low-Cost Hybrid Flash/Spinning System

Thank you