Everpure

Towards Archival Flash Storage

Ethan L. Miller
Everpure

Why use flash for archival storage?

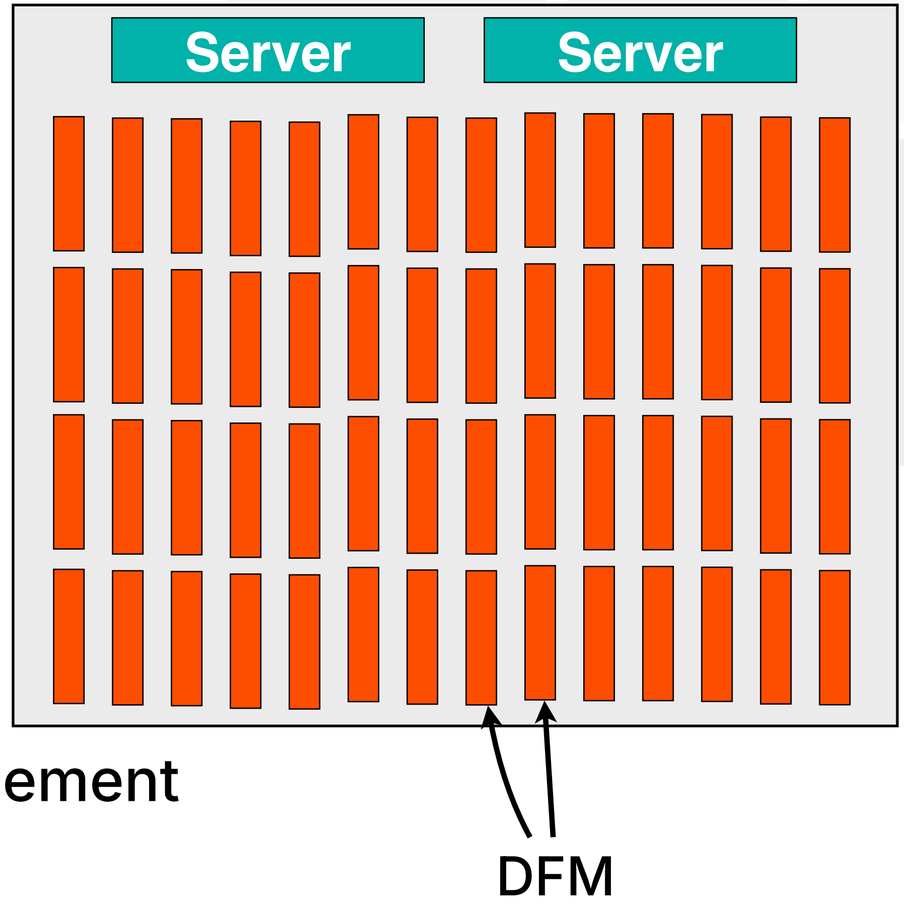
- **Goals of archival storage are**
  - **Preserve data for a long time at a low cost**
    - Low capital cost: inexpensive to purchase
    - Low operating cost: low power, high density
  - **Make the data easily and rapidly available when needed**
- **Tape and optical disk mostly satisfy the first goal**
  - **But they're very slow ➜ not so good at the second goal**
- **Flash is becoming a viable option!**
  - **Total Cost of Ownership (TCO) is decreasing**
    - Relatively high capital cost (we'll discuss the recent cost bump later)
    - *Very* low operating cost
    - Zero migration cost: replace drives as they fail
  - **Extremely fast, especially for reads**
  - **Extremely dense: could store 500PB in a single 19″ rack using *currently-available* drives (300TB)**

DISCLAIMER

These views are mine alone, and may not represent those of Everpure

What would a flash-based archive look like?

- **Pack DFMs vertically into a shelf**
  - Similar to Backblaze pods
  - 216 DFMs fit in a single 5U shelf (24✖9)
  - NVMe backplane on the floor of the shelf
  - Result: 1700+ DFMs in 40U
- **Servers at the back of each shelf**
  - Coordination for accessing drives
  - Cross-drive issues: redundancy, power management
  - Network access to the collection of drives
- **Top-of-rack switch for access to shelves**



(view from above)

Issues: power management & data retention

- **DFMs consume 25W at peak, but 1–3W at low IO rate**
  - Supplying 45kW to a rack (full power) will be difficult
  - Cooling such a rack will be *very* difficult
- **Solution: limit DFM performance**
  - Run ~10% of DFMs at any time
  - Cap IO rate for each DFM
  - Combined approach limits power to <10kW per 500PB rack

- **Flash cells leak electrons: data on flash doesn't persist forever**
  - Rate depends on feature size and materials: set at manufacture
  - Older cells typically leakier
  - Need to refresh about every 1–3 months
- **Solution: periodically read all of the data, correct errors, and rewrite it**
  - Might require multiple days!
    - Not necessarily uninterrupted…
  - Can be done internal to the DFM
- **Combine refresh with power-on time for each drive**

How will this be cost-efficient?

- **Cost efficiency comes from multiple sources, primarily operating cost**
  - Much less server room space and power
  - Essentially no mechanical failures
    - Flash chips are inherently longer-lasting than mechanical disk and tape systems ← technically capital cost
    - Flash drives are more likely to partially (rather than fully) fail, causing a smaller blast radius (and easier recovery)
  - Much higher read and write parallelism: increased bandwidth
  - Random access allows more efficient failure handling: erasure codes can leverage high IOPS

- **But isn't flash *more* expensive today?**
  - Storage demand from AI is driving cost up
  - Higher capital cost, but lower operating cost
- **Flash prices will likely drop in the future**
  - **Flash is reaching the point of diminishing technological returns**
    - Feature size can't get much smaller: already at ~100 electrons in a flash cell
    - Can't fabricate multi-layer flash much higher: already at 300+ layers
  - **New flash fabs are likely to last much longer**
    - Fabs are **expensive**: $10+ billion
    - If they can be productive longer, companies will be more willing to build them
- **Flash cost goes down if fabs last longer and companies are willing to build more fabs**

Benefits of archival flash

- **Archive becomes *active*: all data is read several times per year**
- **Opportunistically run analyses when the data is read for refresh**
  - Declarative I/O (current research) lets users specify operations to be done on data “whenever you read it”
  - Also supports on-demand reads
- **Data for analysis can be read directly from archival flash**
  - Archival flash can support reads at “regular” flash rates (just not all drives simultaneously)
- **Flash is more reliable than disk or tape**
  - Lasts longer and much denser: minimal need for migration
  - Augment capacity over time with additional DFMs
  - “Fail” drives that get too old and need to be replaced: same mechanism as handling drives that *actually* fail

What's the status of archival flash?

- **Archival flash is (currently) a research concept**
  - Nobody (to my knowledge) is building such a system — yet
- **Several issues need to be successfully addressed**
  - Power management and scheduling
  - Declarative I/O (or something else) to allow "background" applications to run during refresh
  - Packaging to make archival flash volumetrically dense
    - Doesn't require tech advances
- **Flash costs may need to drop to make this more viable**
  - But that's going to happen: flash prices are cyclical
  - Flash allows archival data to be better-utilized
- **If archival flash were available, would it be useful for users?**
  - This is the key question: please give me feedback!

PURESTORAGE®

Questions?

elm@purestorage.com