INTERNET ARCHIVE: AS THE DISKS TURN

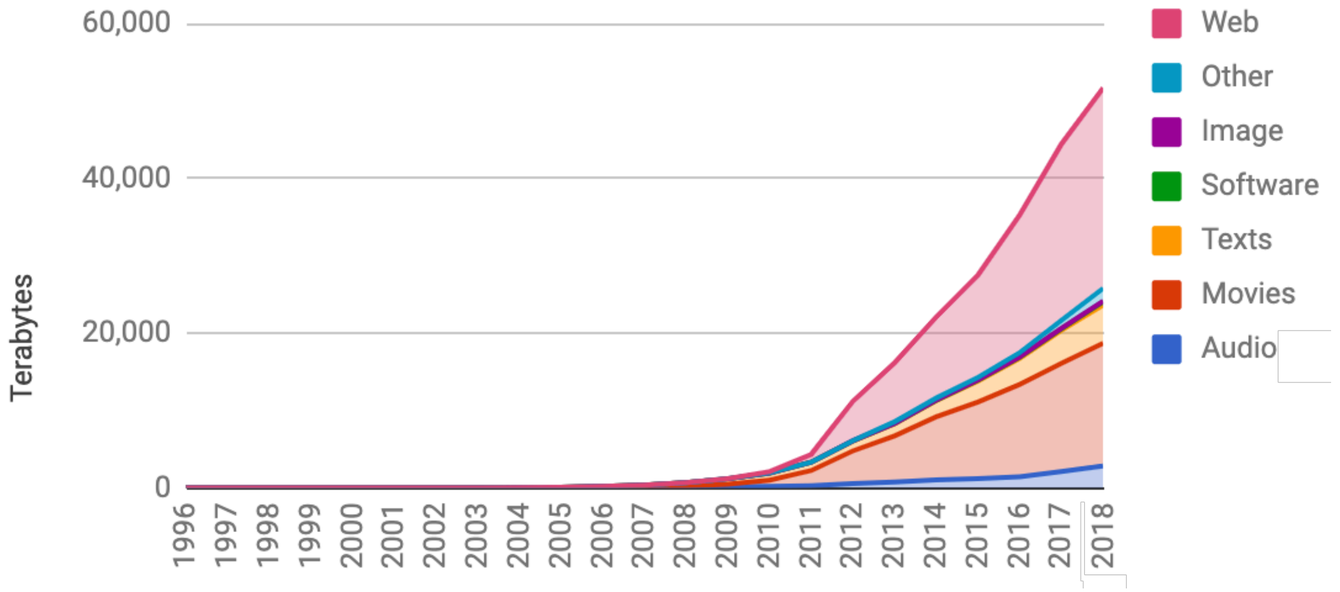
Jonah Edwards
Infrastructure & Operations Manager
jonah@archive.org

THE INTERNET ARCHIVE: NON-PROFIT LIBRARY

Universal Access to All Knowledge

INTERNET ARCHIVE: QUICK OVERVIEW

- 376 Billion Web Pages (totaling over 750 Billion ‘web objects’)
- 23.4 Million Books & Texts (over 1M borrowable books from openlibrary.org)
- 7 Million Audio Recordings (including over 200,000 live concerts)
- 5 Million Videos (including 1.8 Million TV News Programs)
- 3.3 Million Images
- 450,000 Software Programs
- 55 Petabytes of Unique Storage
- ~30,000 Spinning Disks
- ~200 SSDs

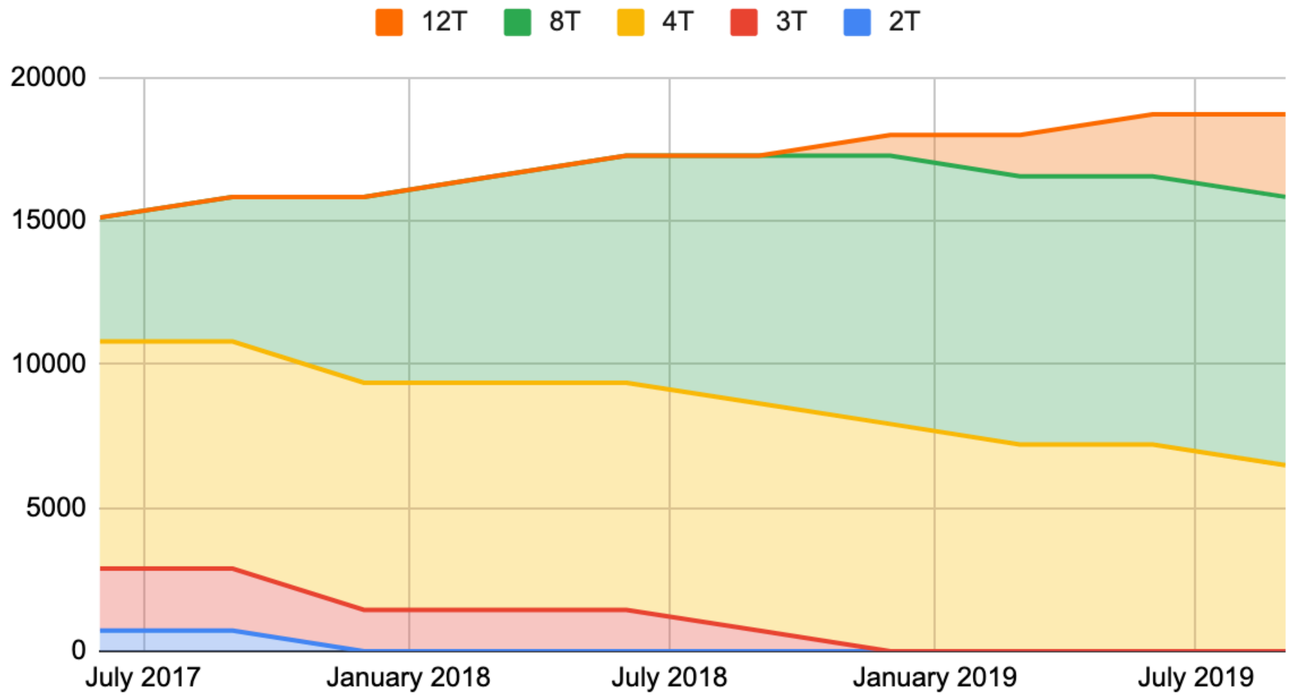


HOW? PRINCIPLES OF OPERATION

- Transparency: Items in the archive are directories on disks
- Simplicity: The basic unit of storage is the disk
- Durability: Disks are replicated across datacenters
- Performance: Content is served from all copies
- Longevity: Formats evolve as needed (and old content is re-derived)

- For more details:
  https://blog.archive.org/2011/03/31/how-archive-org-items-are-structured/
  https://archive.org/services/docs/api/internetarchive/cli.html

STORAGE IN 2018/2019



Drives by size in data storage pool

- Higher density of storage allows us best use of limited space and minimizes overhead costs of storage
- Move away from SMR disks means some gains in cross-datacenter replication times, but trend is still bad for this model (mean stat on full drives):
  - 4T replication time ~12hr
  - 8T (SMR) replication time ~72hr
  - 12T (CMR) replication time ~36hr
- Implies an endpoint for this model dictated by individual disk size

CONSERVATIVE APPROACH

- Battle-tested suite of data validation
- Example: 2018 excursion into next-generation SMR disks
  - Caught via 12-year-old code originally designed to compensate for disk controller without ECC RAM
  - After data is written to disk, caches are flushed and data is immediately read back
  - Infrequently (on a newly installed disk, avg one incident per 12 hours of writing… but would stop once disk was 5-10% full), newly written blocks would read back all zero for 10-30 minutes, then later read back correct data.
  - Hard to catch in action, tons of help from vendor
  - However, timeline for resolution meant we skipped that generation of storage

SHORT-TERM OPTION: REDUCE THE REPLICABLE UNIT TO THE ITEM

- Instead of replicating whole disks, allow individual items on the disk to replicate out to multiple destination endpoints
- Already have the metadata system for associating items with specific storage endpoints
- Allows us to remove the disk-write bottleneck for storage transfer speed
- Allows us to initiate replication at the time of issue detection, before redundancy is lost
- Leverages existing mechanisms for verification and integrity checking

LONG-TERM OPTION: ABSTRACT THE STORAGE LAYER (SAFELY)

- Looking further ahead, how could we take advantage of modern clustered storage systems without compromising durability?
  - Potential for expanding underlying technology pool to diversify risk
  - First pass: intermediate abstraction layer using user-space filesystems — make our catalog system think it's still operating on standard block devices
- Problems…
  - Standard clustered filesystems are primarily eventually consistent
  - Want to avoid changing meaning of operational verbs (e.g. "flush cache") in abstraction
  - Loss of operational simplicity and recoverability

POSSIBLE FUTURES

- Cost of on-prem storage still seems impossible to beat, particularly when availability requirements are below the mode
- Expanding geographic diversity requires review of bandwidth and latency requirements for a platform which has always existed in a single region
- Never going to reduce redundancy below 2x hot/warm access (have >2x including cold storage), but could have a parity-based 1.3x as the primary online store with synchronization to the other 1+x in alternate facilities, services, clouds…

INTERNET ARCHIVE: AS THE DISKS TURN

Jonah Edwards
Infrastructure & Operations Manager
jonah@archive.org