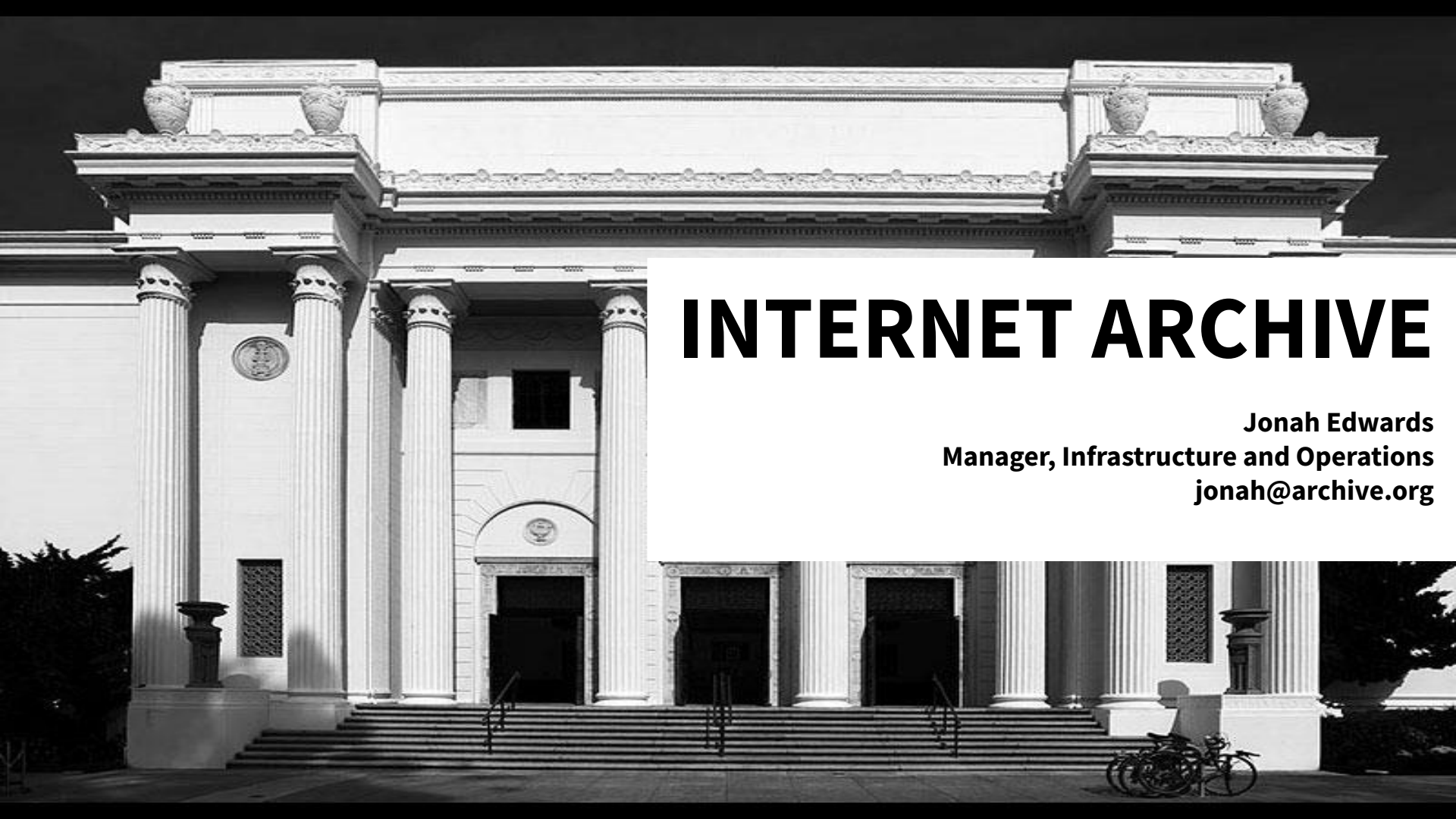


INTERNET ARCHIVE

Jonah Edwards
Manager, Infrastructure and Operations
jonah@archive.org

2025 Materials Update

Wayback Machine:
- 1 trillion web pages celebrated last October
- 800 million pages captured per day

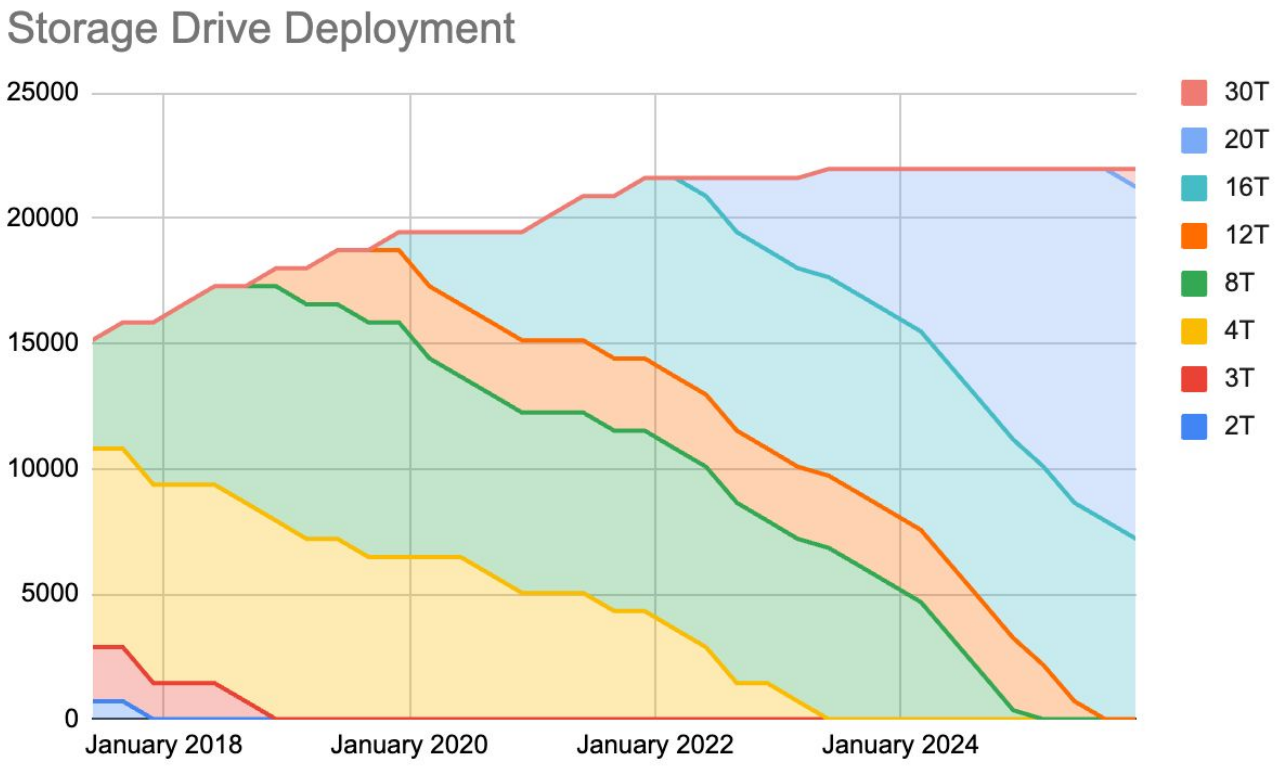
Collections
- 56 million texts
- 8.4 million books digitized in 22 centers worldwide
- 11.5 million movies (excluding television)
- 3.3 million broadcast news programs
- 13 million audio items
- - 280,000 live concerts from 7,600 bands
- - 187,000 78 RPM recordings
- 1.3 million software titles, many emulatable
- 5.3 million images

2025 Storage Update

22,000 spinning disks in Bay Area datacenters

Ongoing move from "paired storage" model of matching disks to ZFS-based single-server pools

Densification within existing physical footprint



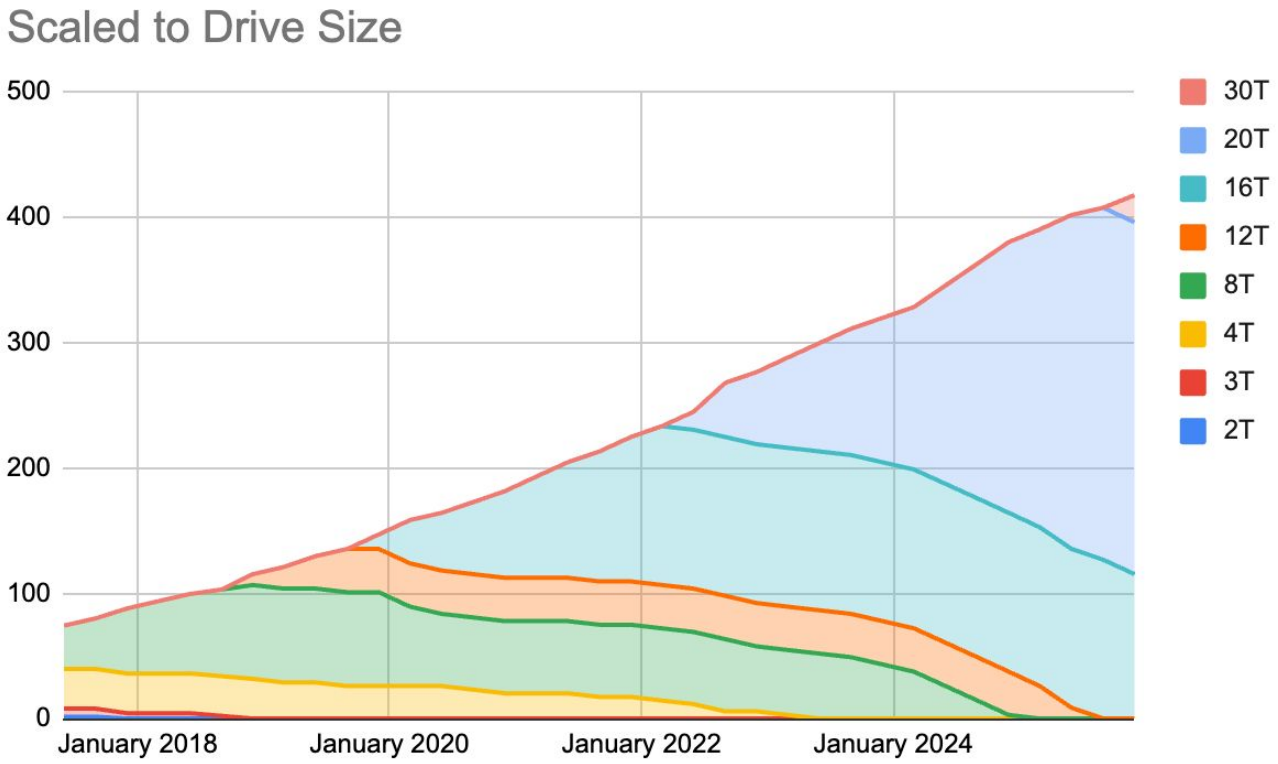
2025 Storage Growth

Over 210 petabytes of unique data

Averaged 123TB/day of ingested material

YoY growth running around 20%

Total available raw storage volume of over 400 petabytes in Bay Area datacenter



International IA

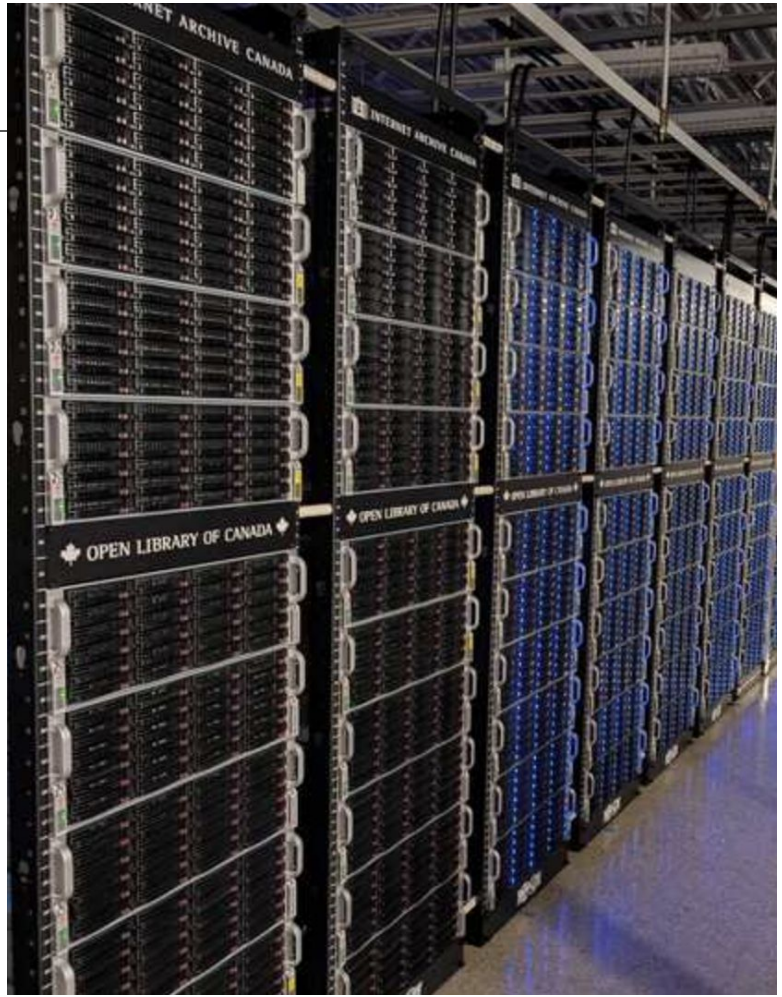
- Opened new IAC headquarters in Vancouver, BC in June of 2022 and full datacenter space in 2024

- all ZFS-based, mirroring content from Bay Area

- increasing functional operations from remote sites

- additional sites being brought online this year in EU and Caribbean

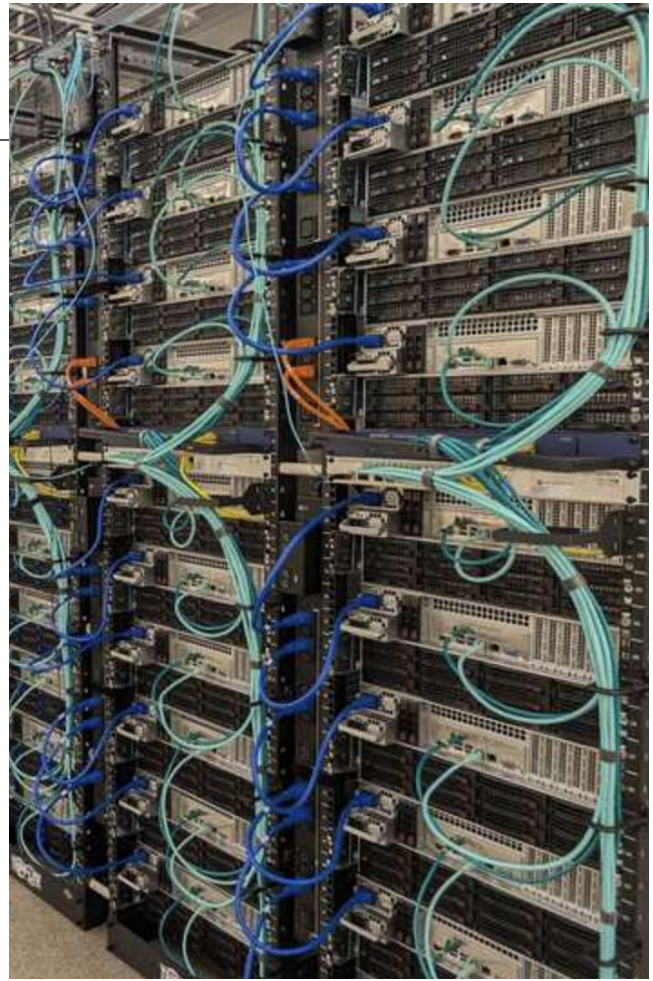
- building a model for running a site of arbitrary size in alternate facilities



Next-gen Storage Model

- old storage model (still in use): Paired Storage
-- relies on full-disk replication mediated by our internal catalog system
-- advantages include simplicity and transparency
-- replication time for (rare) full-disk failures scales with drive size, requires frequent hands-on work

- New ZFS-based model
-- server-sized storage pools
-- increased redundancy and ease of maintenance
-- conversion of paired systems results in significant capacity increased in same footprint, but relies on remote mirrors for redundancy and recovery



Ongoing Challenges

- Balancing longevity of equipment with need for increased capacity and computational power

- Component pricing and availability timelines adding significant challenges to capacity planning

- Continued pace of growth, both in primary and remote sites, in face of logistical and other factors

- and for AI: used heavily internally for OCR, transcription, etc, but impact of AI culture both in terms of our mission and impact on our services largely negative