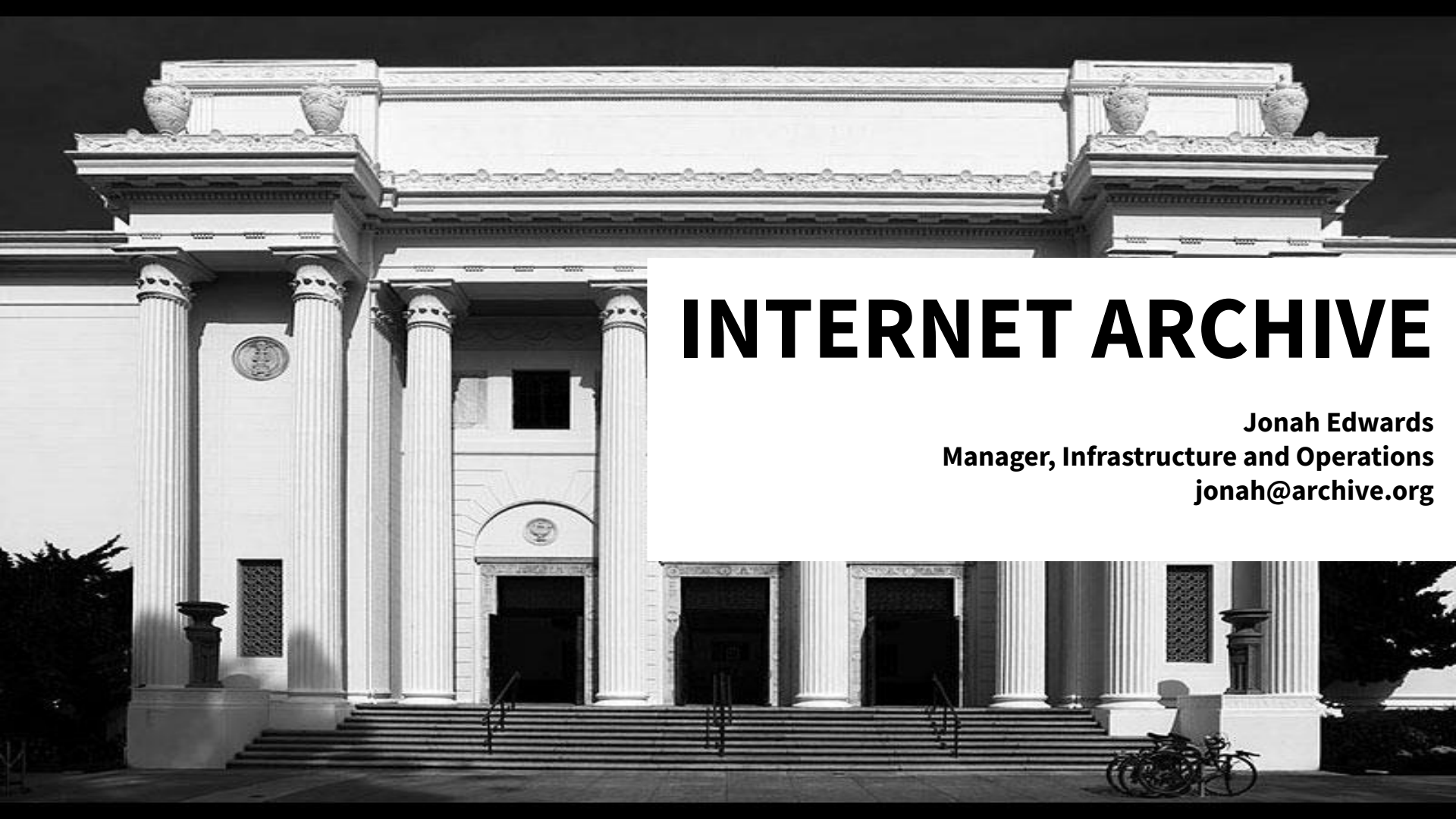


INTERNET ARCHIVE

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

2024 Materials Update

Wayback Machine:
- over 800 billion web pages
- 650 million pages captured per day

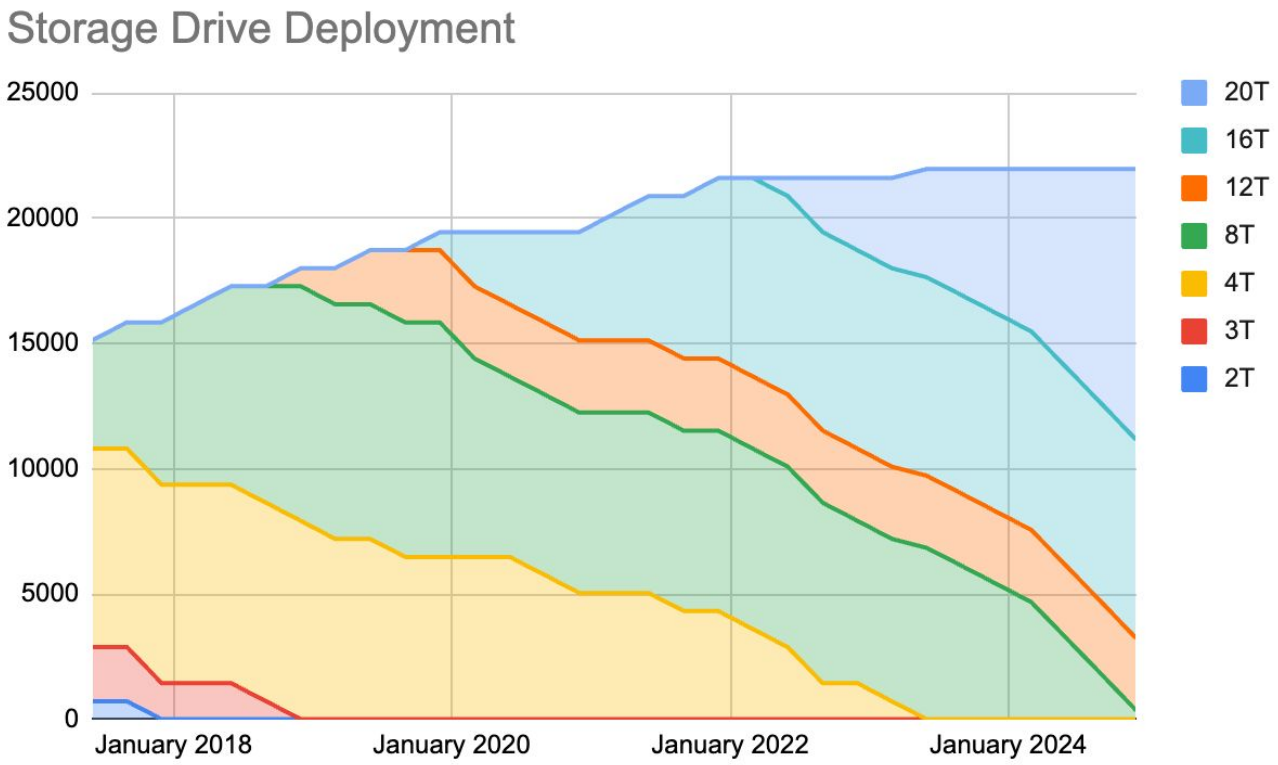
Collections
- 44 million texts
- 7+ million books digitized by us
- 4300 books per day, 20 centers worldwide
- 6 million movies (excluding television)
- 2.4 million broadcast news programs
- 14.7 million audio items
- over 1 million software titles, many emulatable
- 4.4 million images

2024 Storage Update

Over 175 petabytes of unique data stored

Transitioning storage to new storage model

Over 30PB of data now stored on ZFS

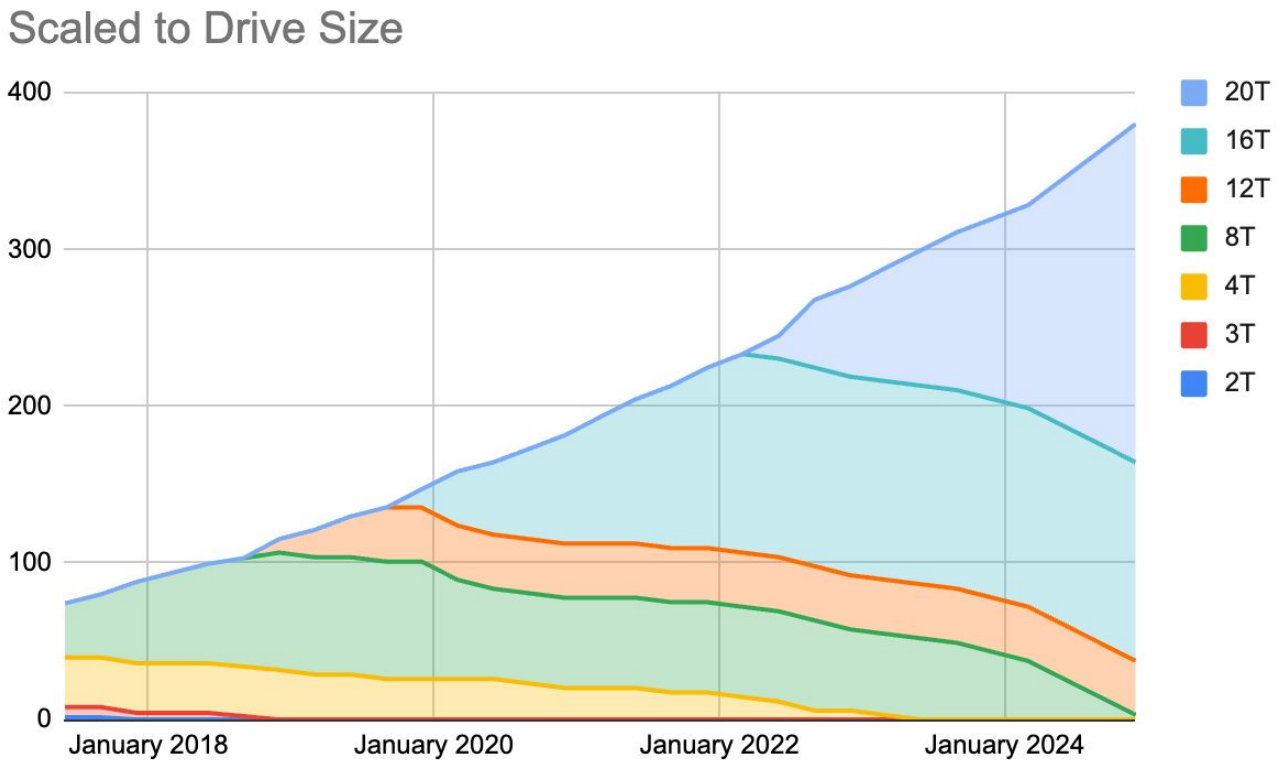


2024 Storage Growth

Averaged 78TB/day of ingested material

YoY growth running around 20%

Total available (not occupied) raw storage volume of 400 petabytes



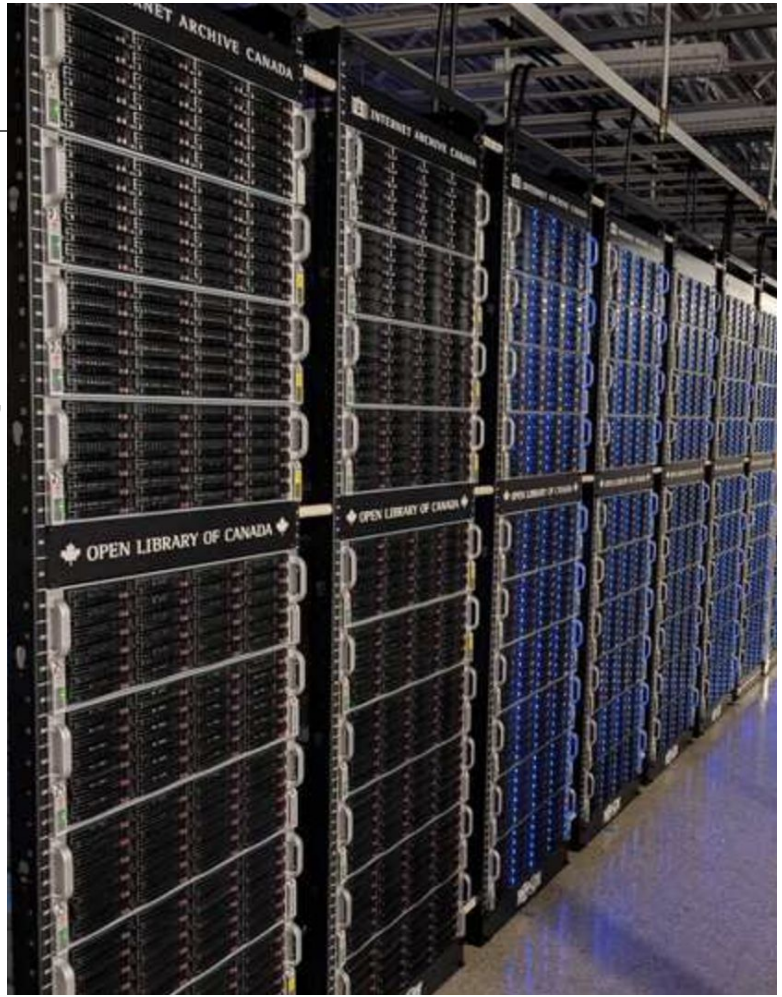
IA Canada Update

- Opened new IAC headquarters in Vancouver, BC in June of 2022

- Ongoing buildout of larger-scale Canadian datacenter, will hold a complete copy of the corpus this year

- Content currently being served from Canadian sources

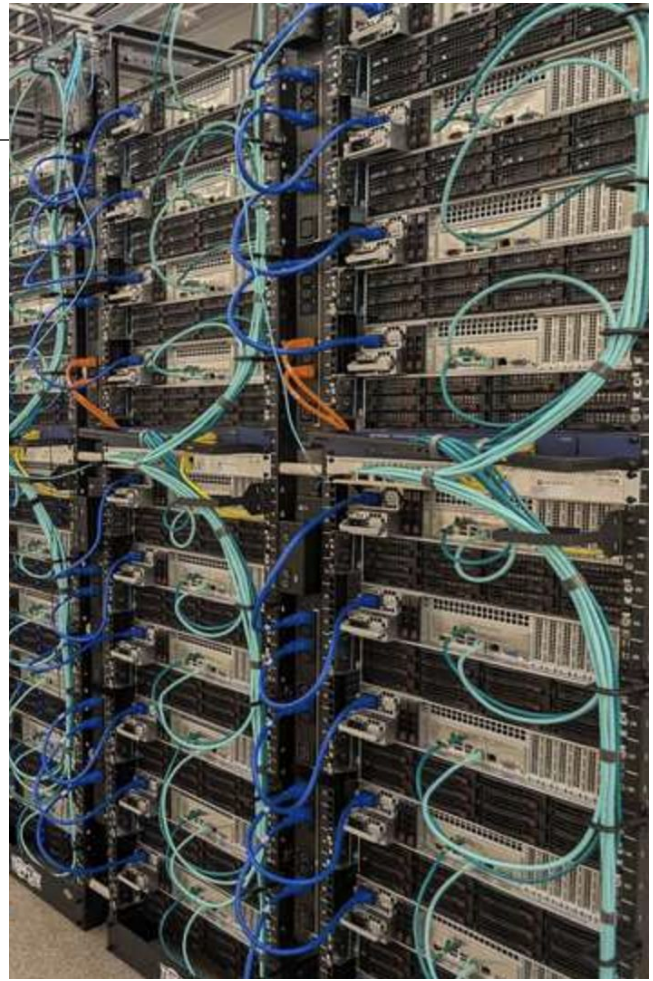
- Moving large volumes of data between facilities -- "Never underestimate the bandwidth of a station wagon full of tapes hurtling down the highway."
  -Andrew Tanenbaum, *Computer Networks*, 3rd ed



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

- New ZFS-based model
-- server-sized storage pools
-- increased redundancy and ease of maintenance
-- primary deploy in Canadian (and other remote) sites, converting systems in primary sites on a rolling basis



Ongoing Challenges

- Increases in scale and scope of operation within existing physical (electrical, environmental) footprint

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

- Maintaining our ability to serve patrons effectively and continue to evolve while taking a much harder line on security and access

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