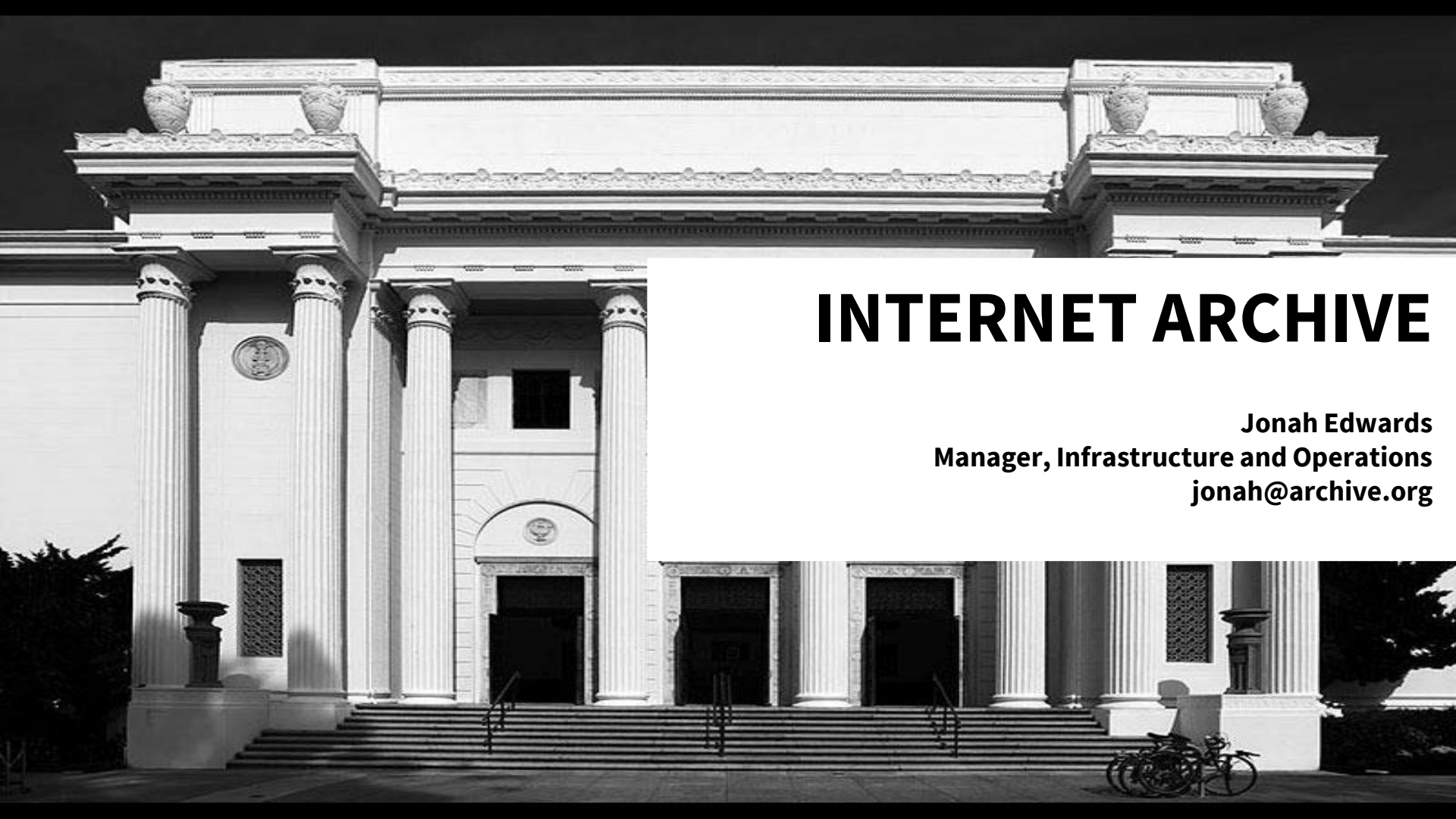


INTERNET ARCHIVE

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

2021 Materials Update

Wayback:
- 625 billion web pages
- 585 million pages captured per day

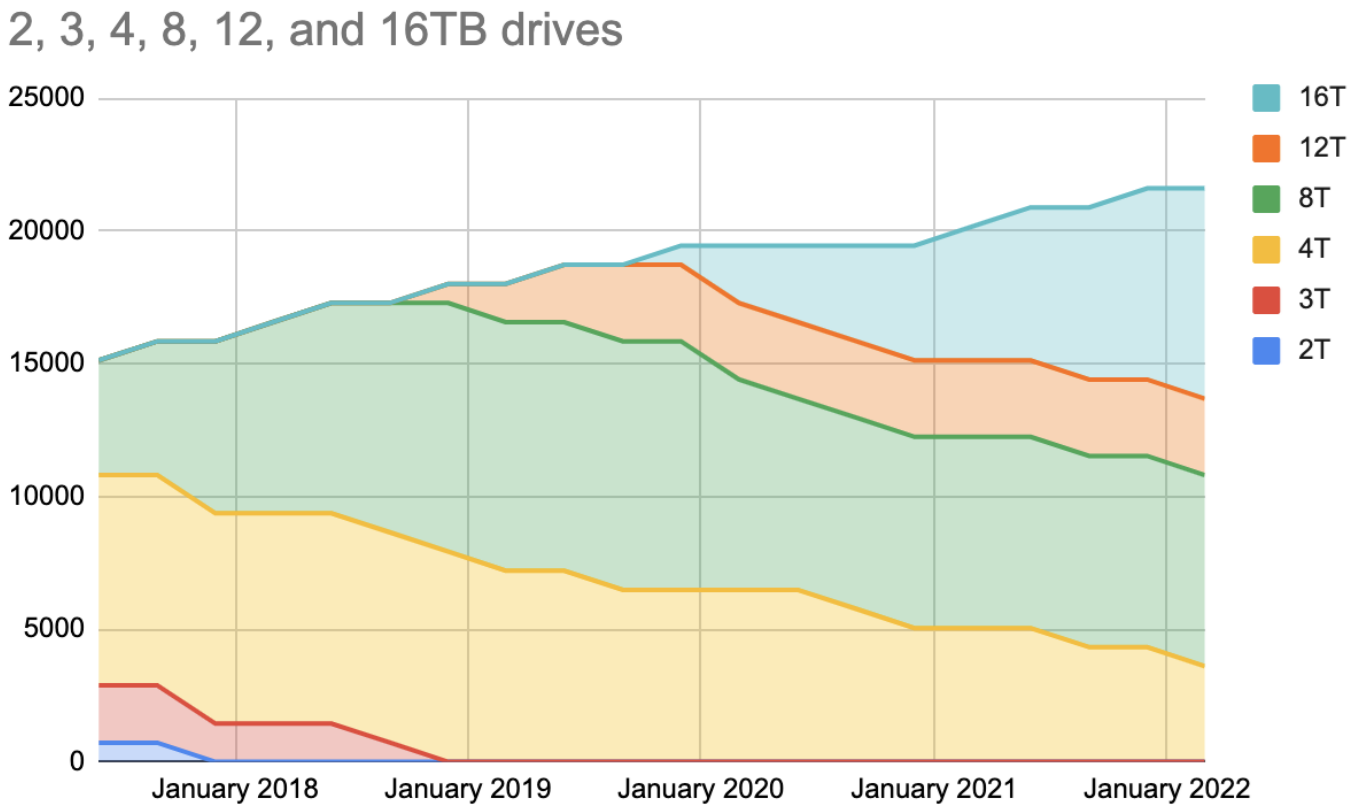
Collections
- 38 million texts
- 6 million books digitized by the Internet Archive
- 4,000 books digitized per day
- 2.2 million news programs
- 5 million movies (not including television)
- 14 million audio items, including 300,000 78rpm sides
- 790,000 software titles, many emulatable
- 4 million images

2021 Storage Update

99+ Petabytes of unique data in the stored corpus

~ 22k spinning disks underlying the paired storage infrastructure, with an additional 10k disks in service in other roles

20T disks currently in testing

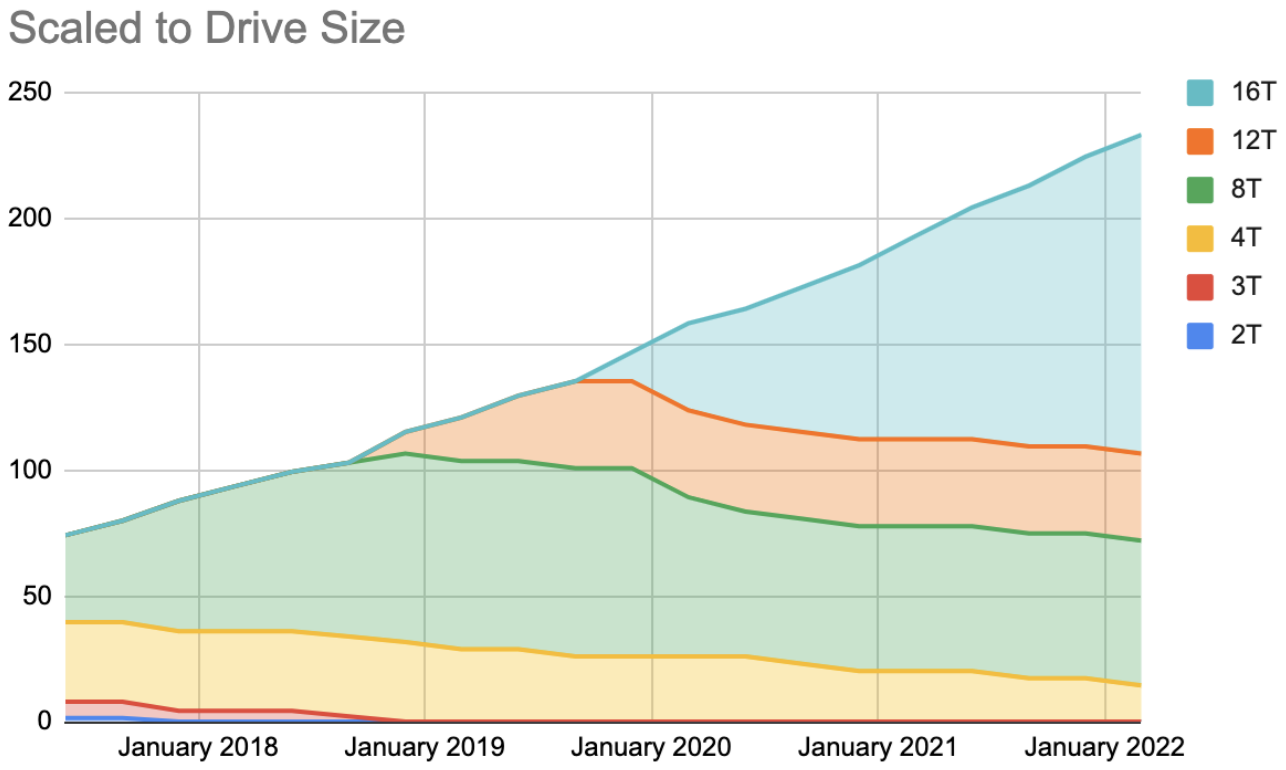


2021 Storage Growth

Anticipate adding 20-25 Petabytes in 2022

Currently averaging 60+ TB/day of data ingest

Stored corpus continues to grow by approximately 25% per year



Non-HDD Storage Components

- Continue to increase use of solid state storage, primarily using SATA SSDs as backing for virtual machines running on shared underlying commodity hardware, and integrated NVMe as backing for large-scale search indices and high-performance cache layers

- Anticipate significant increase in NVMe use for cache fronting and as underlying storage for search

- Still using HDD in non-storage roles for processing scratch

Next-Generation Storage Model

- Paired storage model has been reliable and highly introspectable
  - Able to diagnose and pinpoint issues to specific disk components
  - Underlying storage issues raised directly in fixity checks

- Next-generation storage platform currently in testing
  - Based on medium-sized ZFS pools
  - Relies on high-performance integrated NVMe as write caching layer
  - Tooling being built to allow continued use of commodity systems

Next-Generation Internet Archive

- Continued use of commodity systems, but potential for break between compute and storage
- Geographic expansion requiring significant overhaul of internal catalog systems
- Building further ahead -- mitigating ongoing supply chain problems
- Moving from a "library that's sometimes closed" stance to a more highly available platform