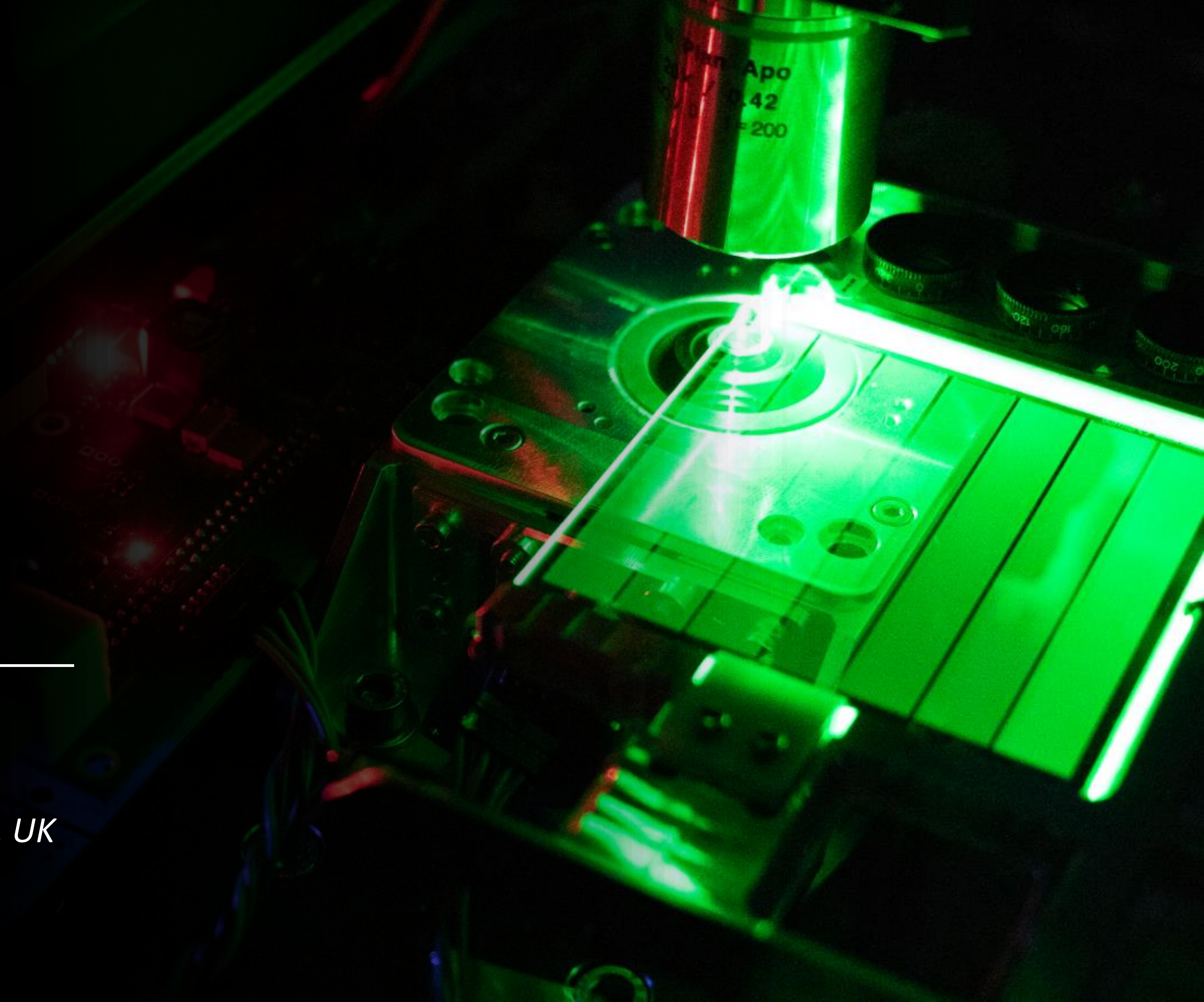
SDC 2024



Project Silica: The Future of Sustainable Archival Data Storage

Ioan Stefanovici

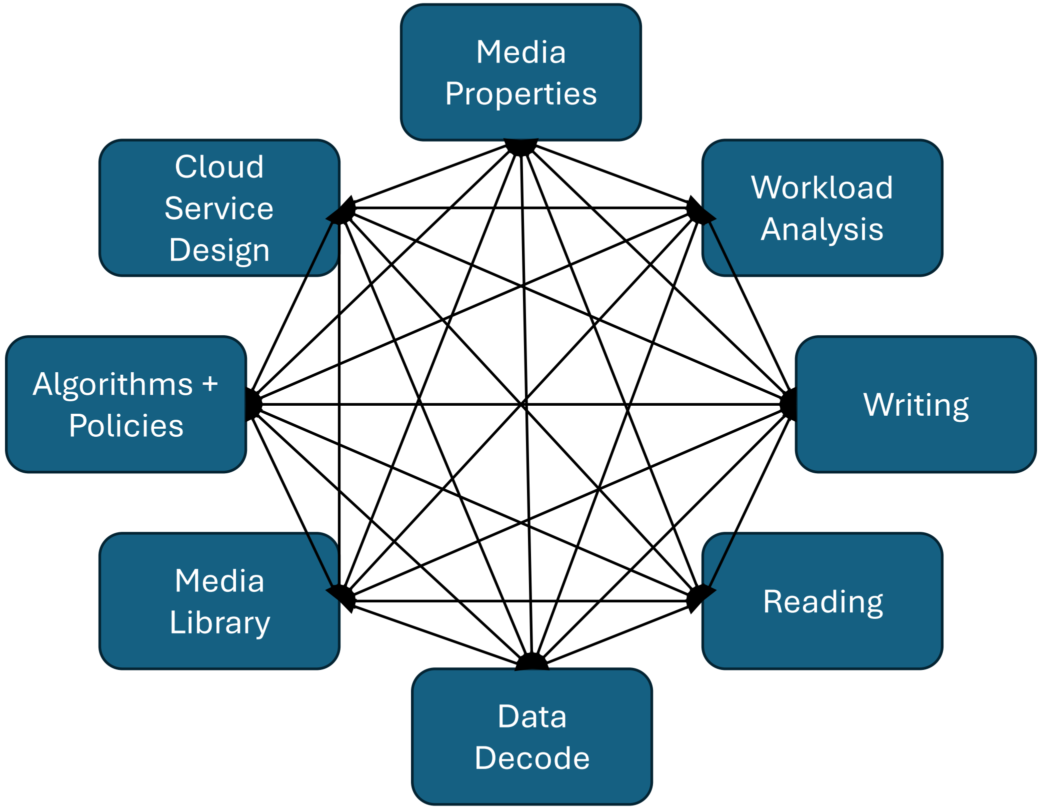
Principal Research Manager

Microsoft Research – Cambridge, UK

Microsoft

Project Silica – A Cloud-First, Multi-Disciplinary Effort

- 8 year ongoing partnership between Azure Storage, Microsoft Research, and more
  - 160+ person-year ongoing effort



- For even more: come talk to me, SOSP’23 paper, [upcoming] Transactions on Storage paper

Archive – The Many Different Kinds

Backup

- Disaster recovery
- Tape’s original market
- Small fraction today

Active Archive

- Scientific computing + measurements (e.g., US National Labs, CERN, Riken, Met Office, etc.)
- Better latency will increase usage and unlock new applications.

Long-Term Archive

- Scientific measurements
- Medical
- Personal photos & videos
- Financial regulatory
- Extractive industries
- Financial ‘lifeboat’
- Legal, contractual

Preservation

- Nascent business
- Customers who care about media properties.
- 100s of years
- True “air gap”
- Immutability
- Withstands benign neglect.

Sustainable Archival Storage: An Unmet Need

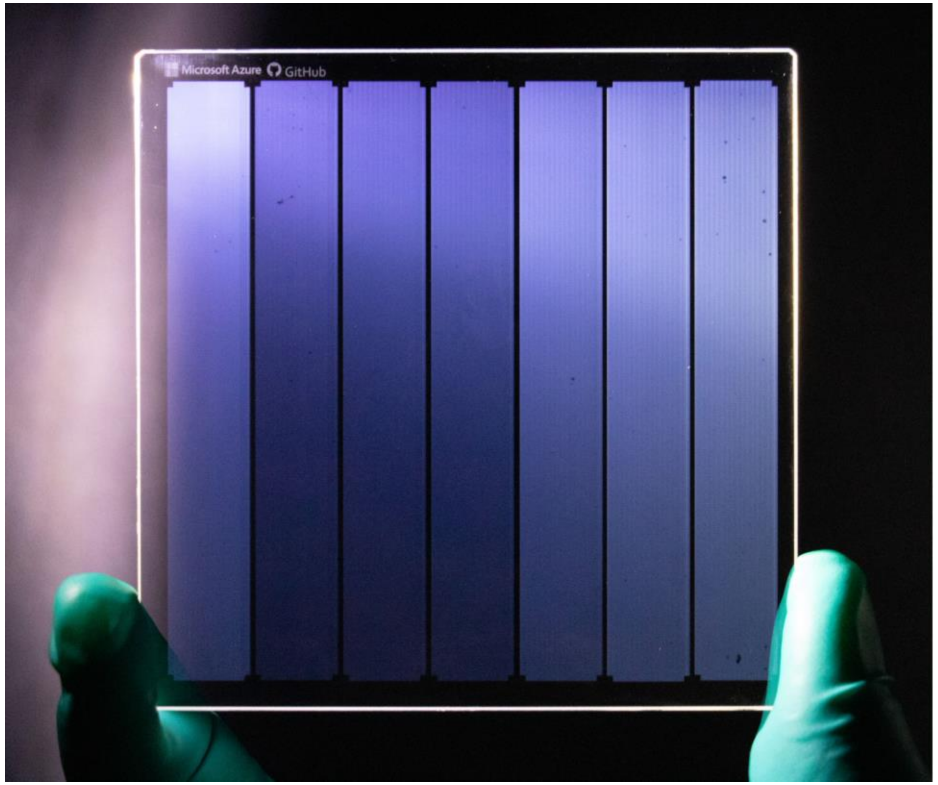
Magnetic media degrades & is susceptible to interference

- Data "refresh" to new media every few years
- Energy required to "scrub" media

Cost (emissions, energy, $) scales with the **lifetime** of the stored data!

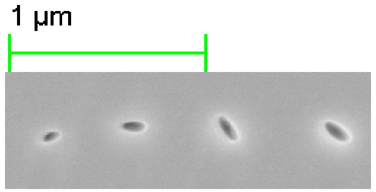
A Disruptive Archival Media: Glass

- Low-cost material: fused silica (quartz glass)
- Durable media
- Electromagnetic field-proof
- Write Once Read Many (WORM) media
- No bit/media rot
  - Data lifetimes > 1,000s years
  - No scrubbing required!
- Data can be left in situ **forever**!



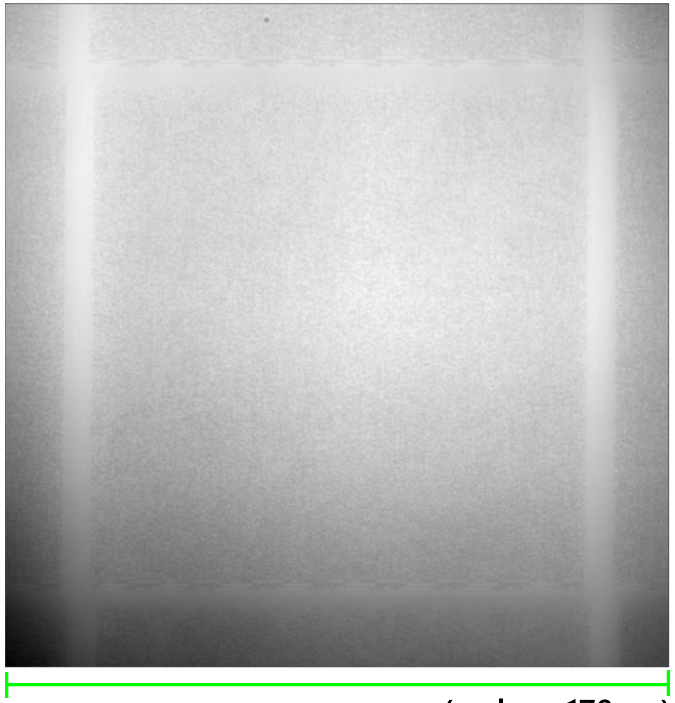
- **Operational proportionality**:
  - Cost (emissions, energy, $) now scales with the **operations** performed on data!

Writing Data in Glass



SEM image of 4 multi-bit “voxels”

- Femtosecond ($\sim 10^{-15}$s) pulsed-laser writing

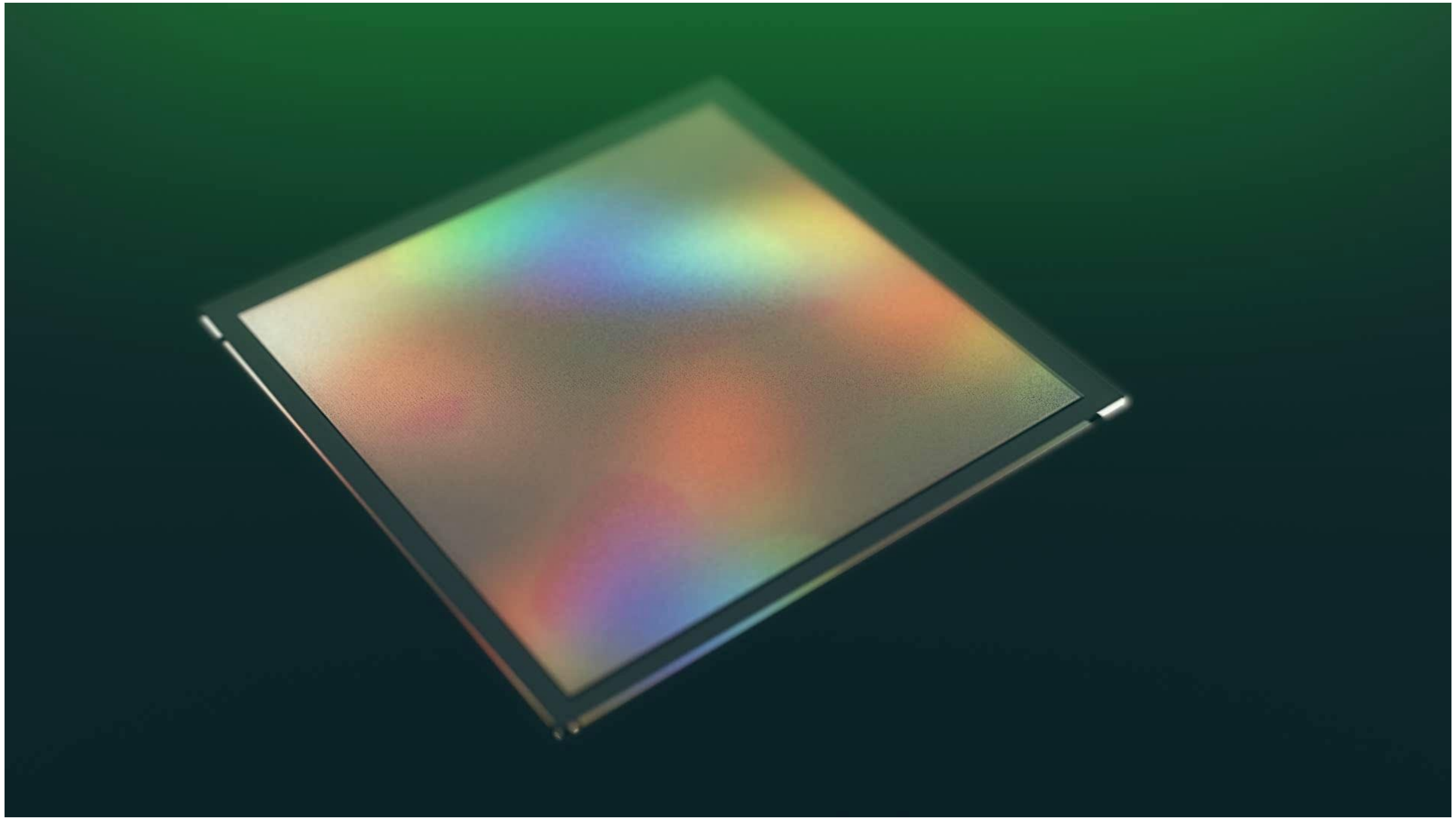


(scale: ~ 170µm)

Data sector from prototype hardware

Reading Data from Glass

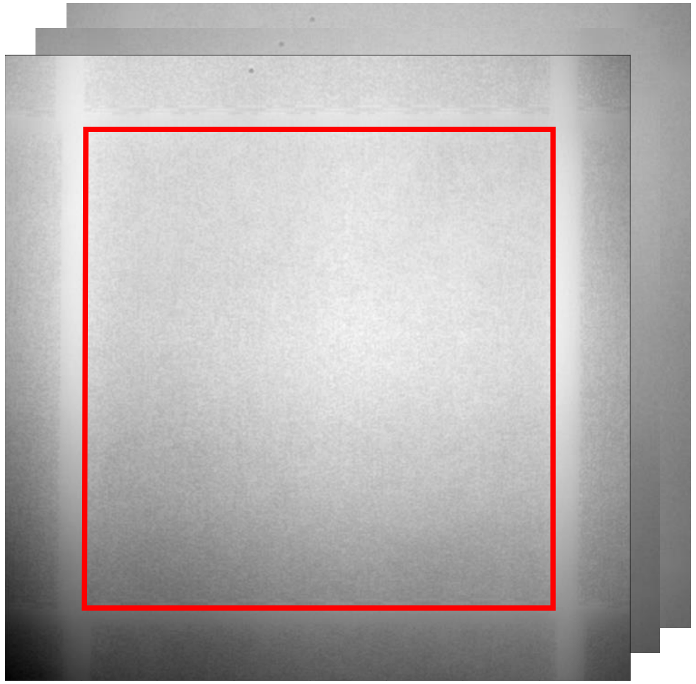
- Polarization microscopy
- Low-powered illumination
- Data overwrite is fundamentally impossible!
- 2D sectors at a time
- Random access media
    - Quick random IOs

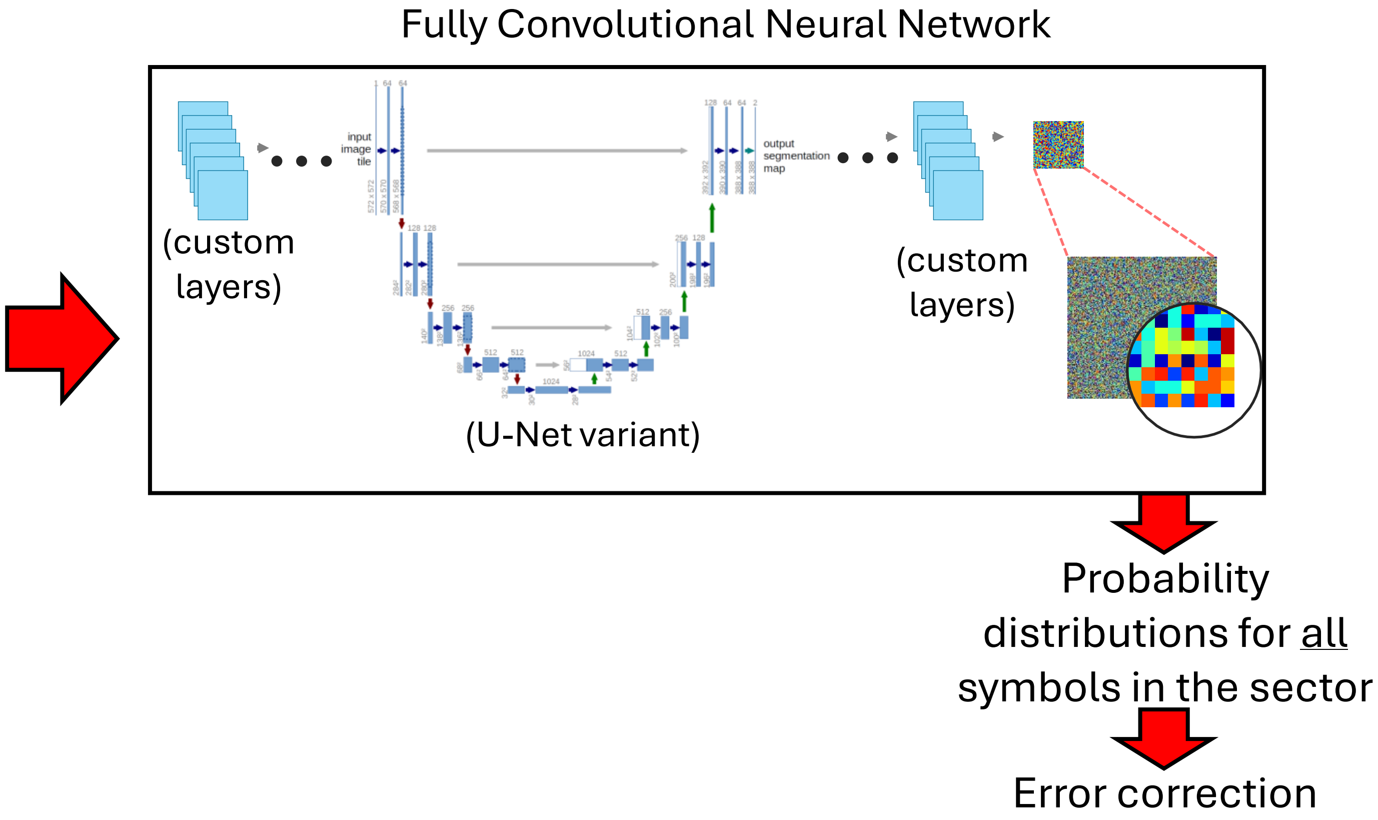


Microsoft

Machine Learning Data Decode

- Most accurate analog-to-digital signal conversion





Fully Convolutional Neural Network

Probability distributions for <u>all</u> symbols in the sector

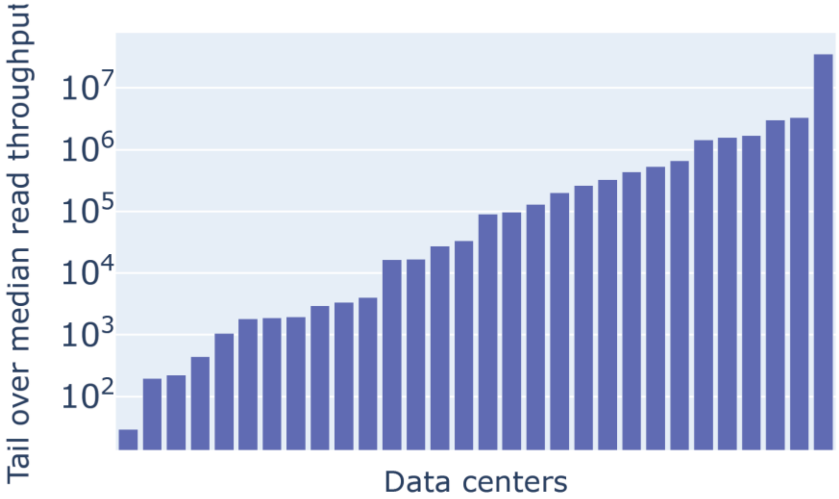
Error correction

- ML decode stack: fault-tolerant elastic microservice

Archival Workloads Require Flexibility

- Studied tape-backed cloud archival service with 10+ hour SLOs
- Long-term average 47x more MB and 174x more IOs Write than Read

Read Variability



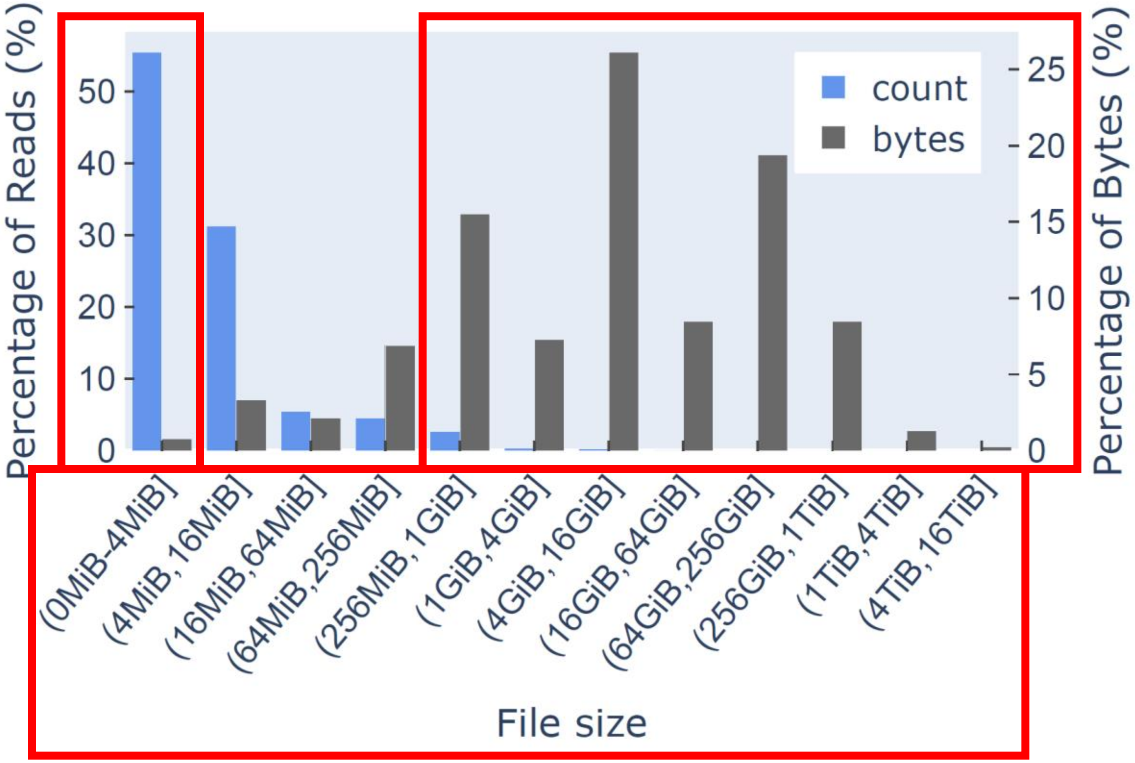
- Up to 7 OoM mean-to-tail within-DC
- Variability across DCs: over 5 OoM diff

➢ Write and Read bandwidths need to be provisioned & scaled independently in the system
➢ Modular & flexible library design needed to support variability across deployments

Archival Workloads are IOPS-Dominated

- 58% of Reads < 4MiB
  - But only 1.2% of total read bytes
- 85% of bytes read > 256 MiB
  - But < 2% of total read requests
- ~ 10 OoM range in file sizes

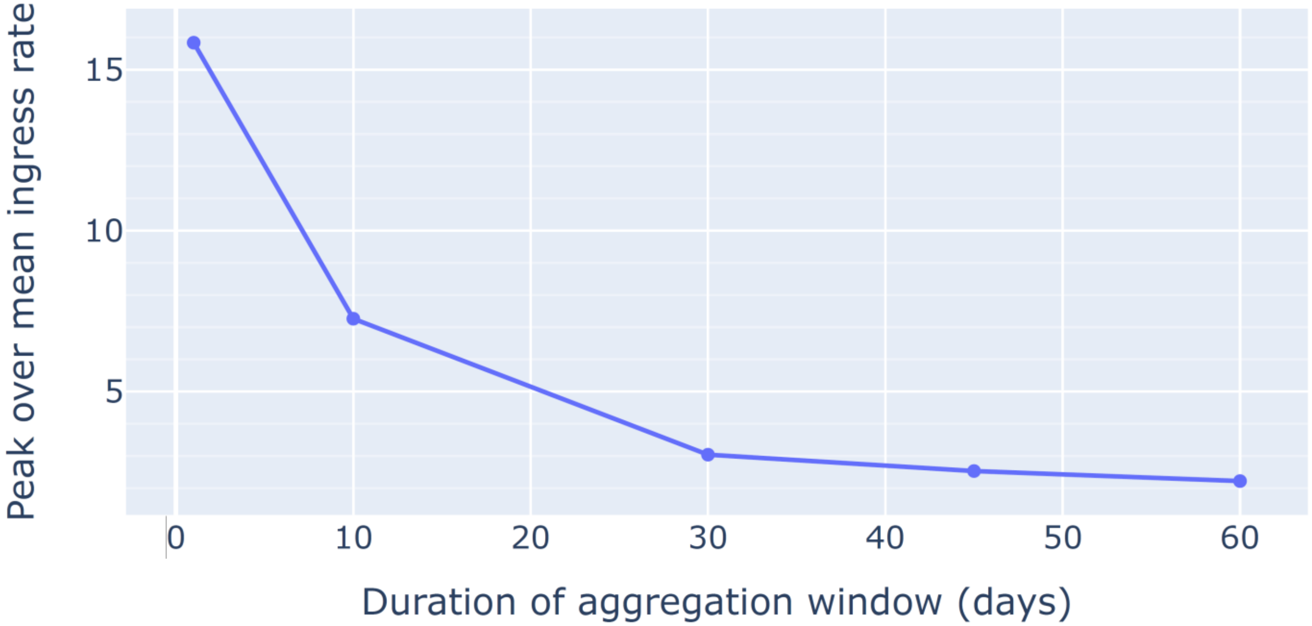
Read Characteristics



- ➢ Per-drive read throughput less important; sharding for small fraction of large files
- ➢ Minimize mechanical overheads (media transfer, mounting, seeking delays)

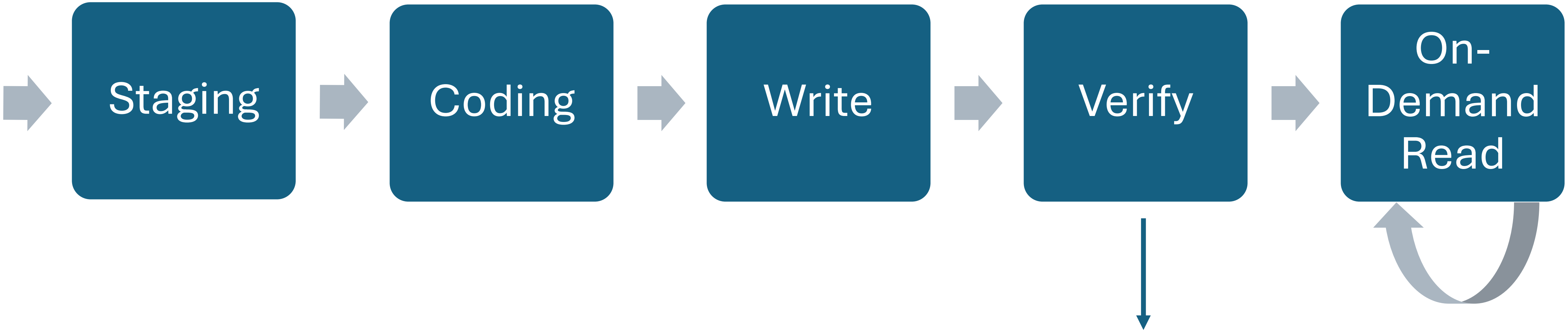
Archival Ingress Benefits from Staging

- Archival ingress rate is highly bursty: > 15x peak-to-mean ratios
- Cost of Silica system dominated by writing
  - Maximizing write drive utilization is important



➢ Staging data maximizes write drive utilization + minimizes system costs

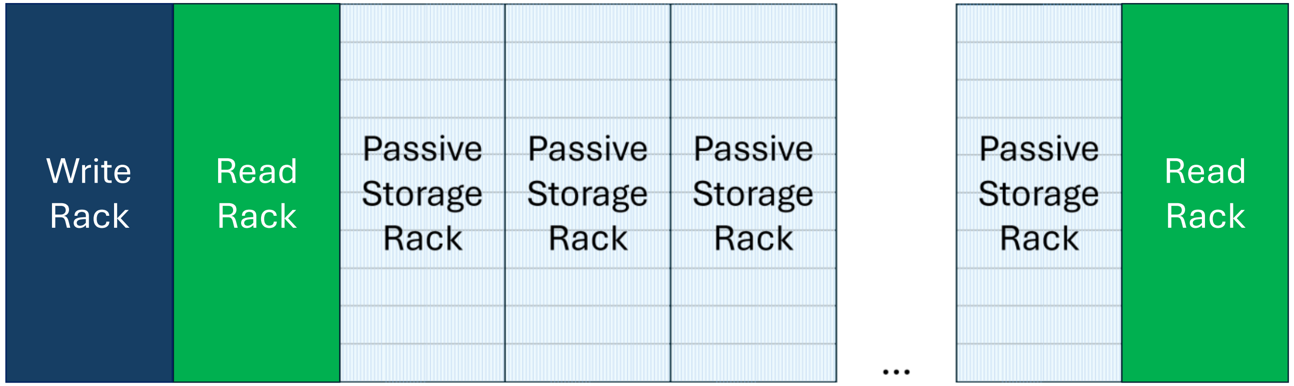
Data Lifecycle in a Silica System



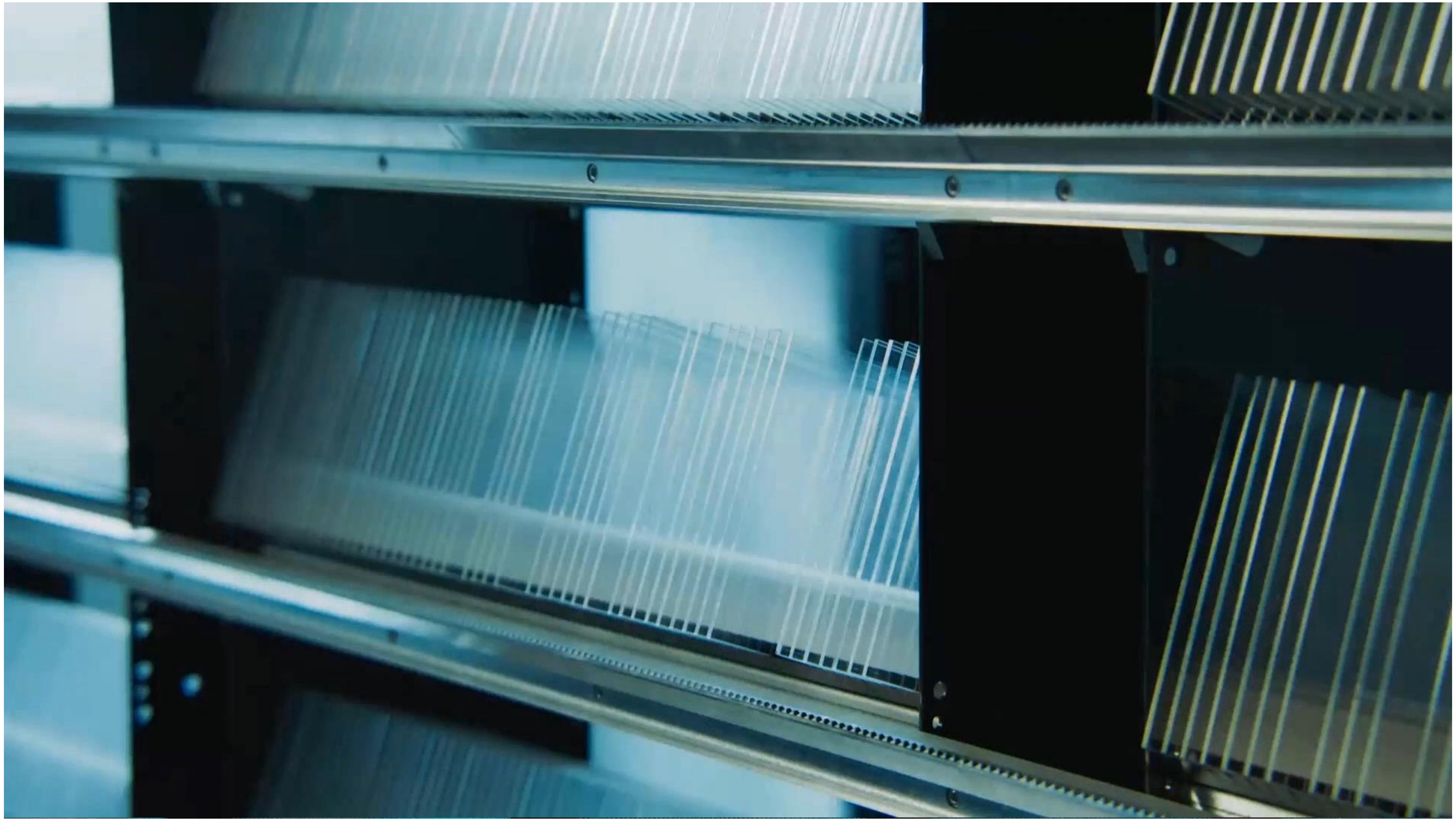
Only needed once!

The Silica Library

- Design principles
  - Modular Write + Read bandwidth: provisioned + scaled independently
  - Minimize mechanical overheads
  - Data at rest should not consume any resources
  - Library lifetime should match DC lifetime



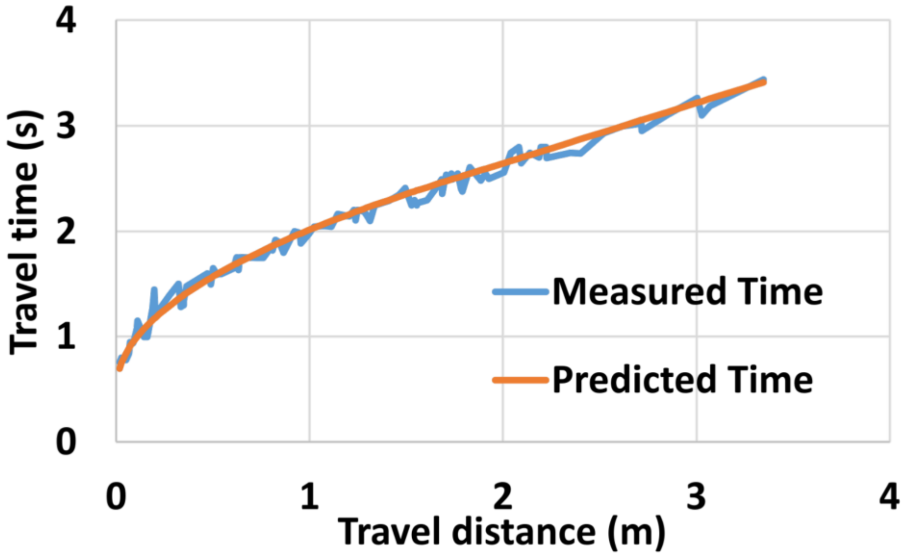
Media Handling in the Silica Library Prototype



See aka.ms/Silica for more!

Silica Archival System Evaluation

- Digital Twin of Silica library
    - Discrete event simulator
    - Cross-validated w/ prototype
- Workload traces:
    - Geo-distributed multi-tenant archival service
    - 12-hour intervals across 6 months
    - Typical, IOPS, Volume

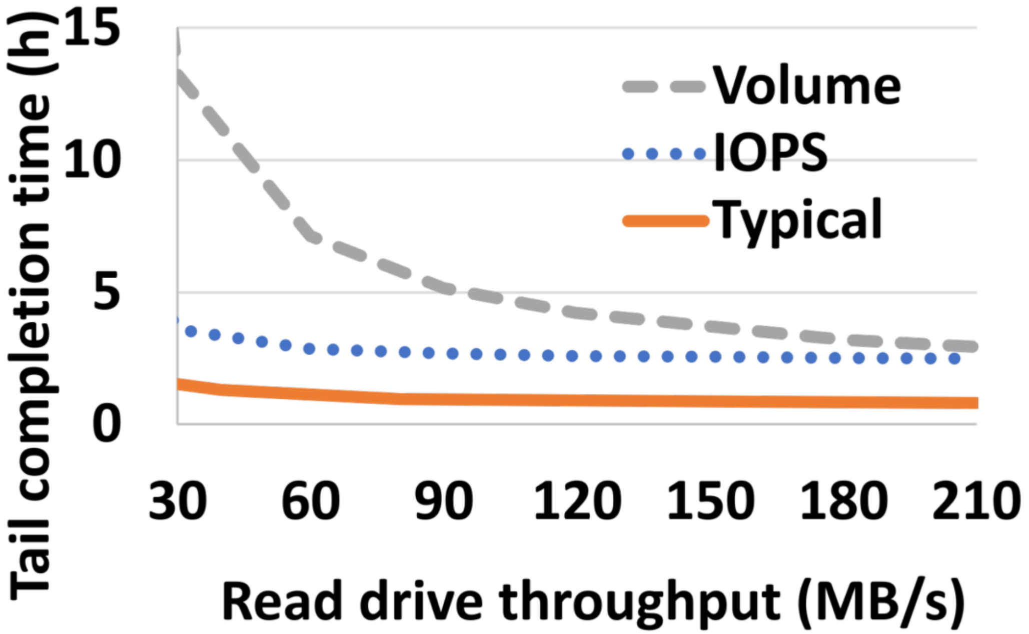


Horizontal shuttle travel

- Evaluation Questions:
    - What is the impact of read drive throughput on system SLO?
    - How much overhead do shuttles add?
    - Can we achieve high read drive utilization with our design?
    - Is the library design future-proof?
    - Many more in our papers (and come chat to me)!

Read Drive Throughput & System SLO

- Each library:
  - 20 read drives
  - 20 shuttles (1 per drive)
  - # platters ∝ total data at trace time
- Metric: p99.9 tail latency
  - Archival service SLO: 15 hours TTLB

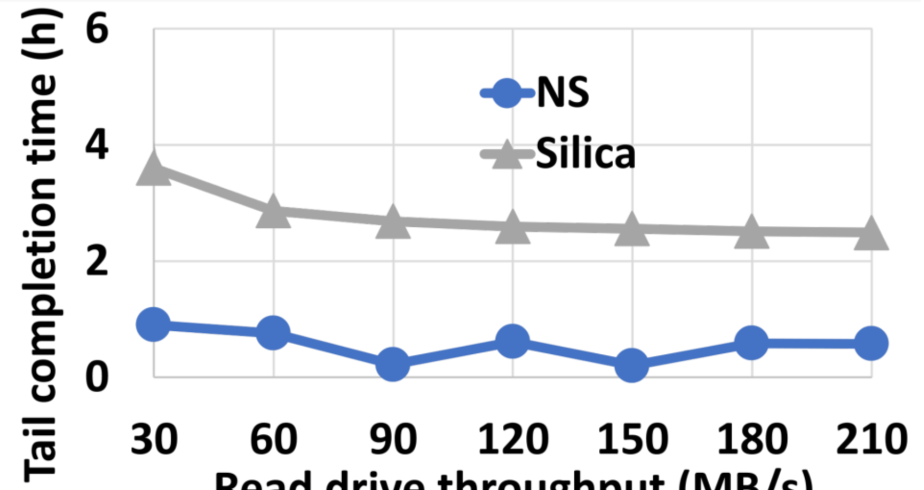


- High-throughput Read drives are not necessary!
  - Diminishing returns past 60MB/s even for Volume workload
  - Minimizing mechanical overhead is crucial!
- How much overhead do shuttles add?

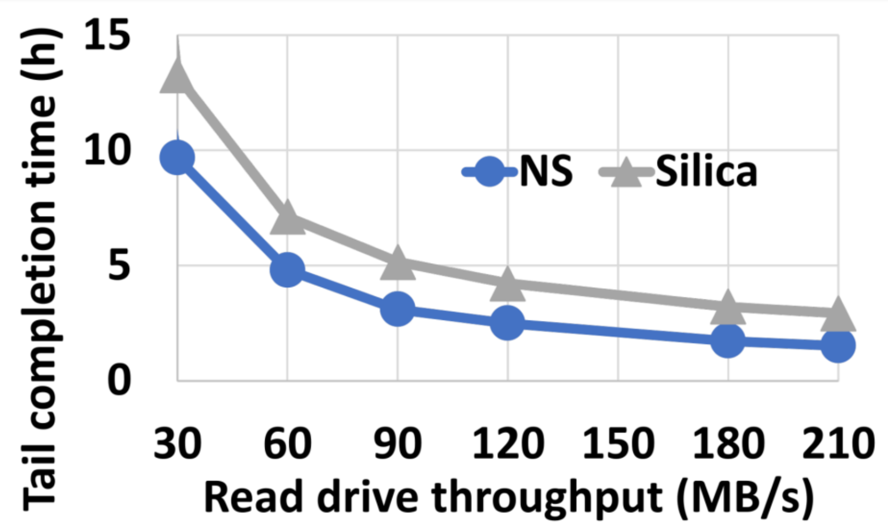
Quantifying Shuttle Overheads

- NS baseline: platter is loaded as soon as read drive is available
  - i.e., “infinitely fast shuttles”

IOPS



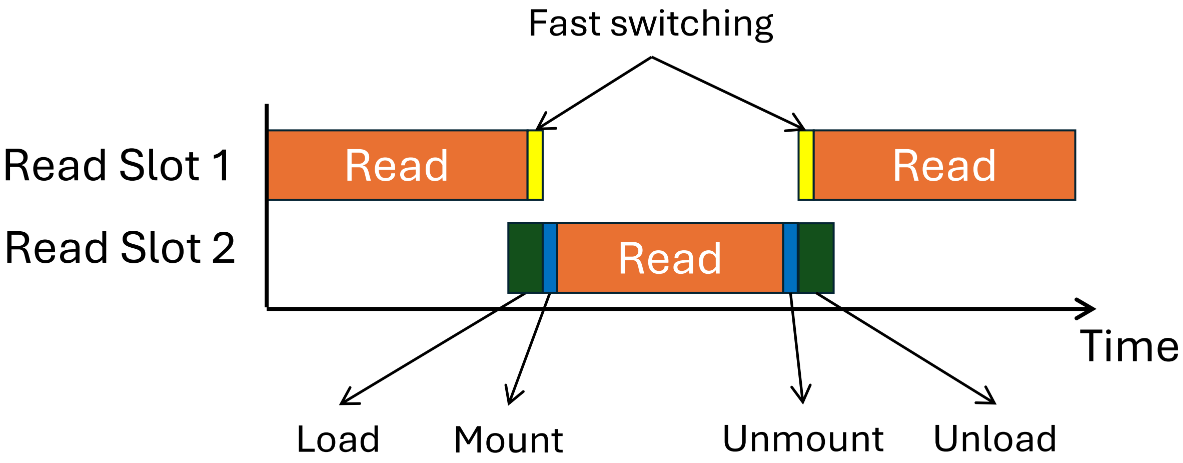
Volume

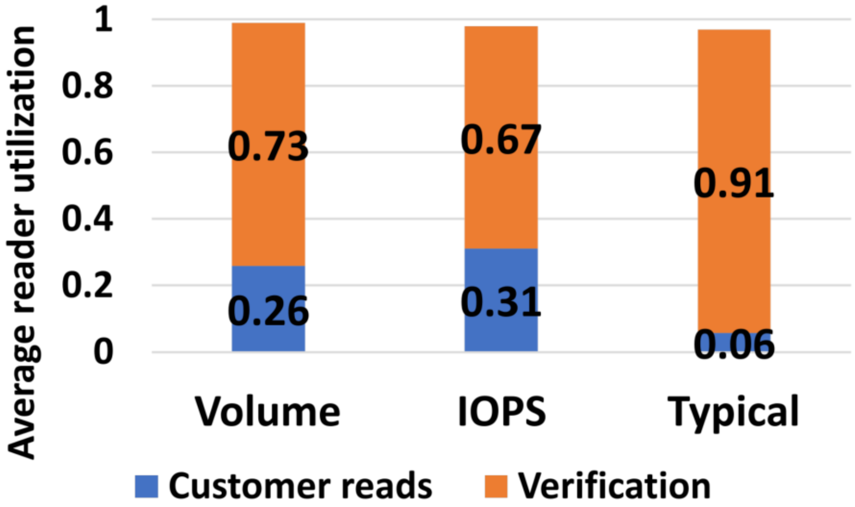


- Current Silica shuttles + policies enable great performance

Read Drive Utilization

- Read drives have 2 slots
  - “Fast switching” enables pipelining between 2 Read streams
  - Physical realization of double buffering

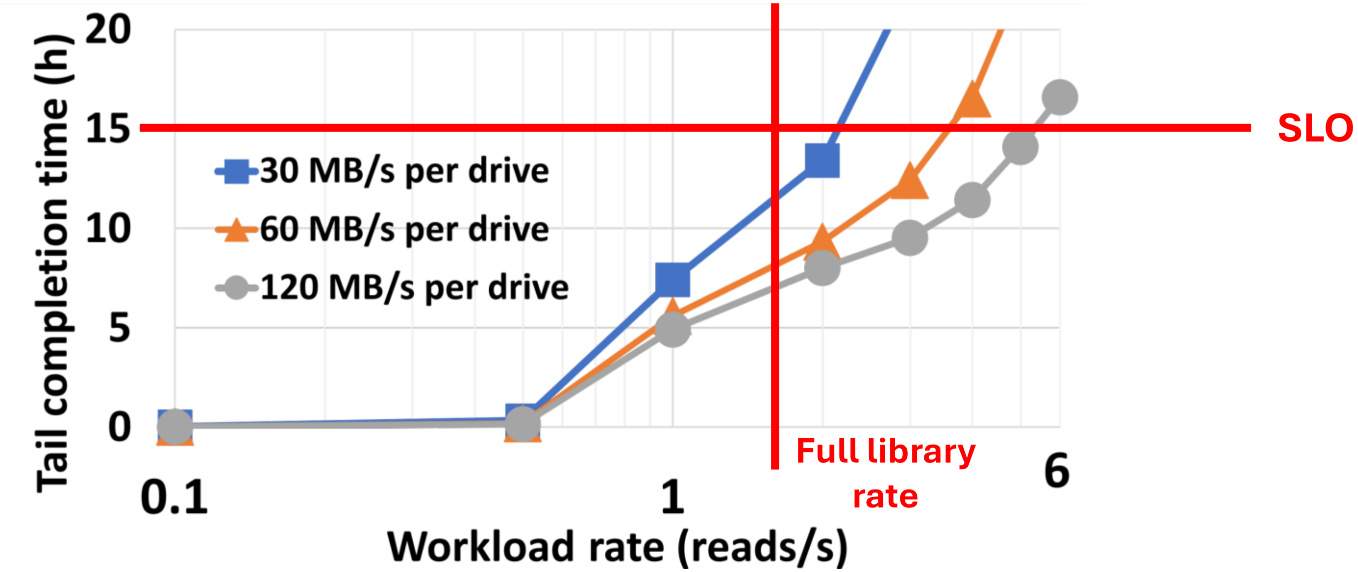




- > 96% drive utilization for all workloads!
- Fast switching: efficient mechanism to achieve high drive utilization

Performance of a Full Library

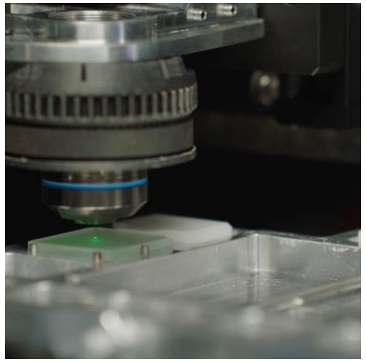
- Synthetic workload:
  - 100MB files, 5% yearly deletion rate, 10% cooldown rate
  - 0.3 reads/s → 1.6 reads/s (full library)

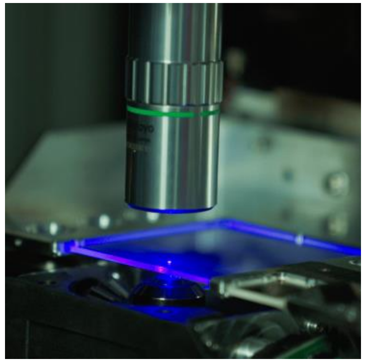


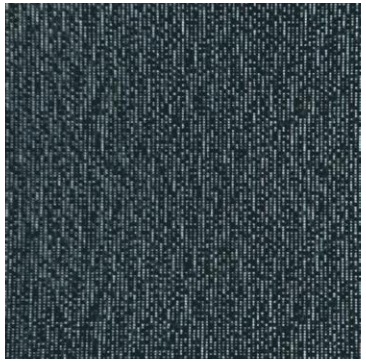
- Library design can comfortably meet SLOs for full libraries
- For more aggressive workloads: add Read rack + shuttles to library

Thank You!

- **Extremely** multi-disciplinary team
  - Computer systems, machine learning, materials-processing physics, free-space optics & microscopy, chemistry, electrical & mechanical engineering, industrial design, and more!







Patrick Anderson, Erika Blancada Aranas, Youssef Assaf, Raphael Behrendt, Richard Black, Marco Caballero, Pashmina Cameron, Burcu Canakci, Thales de Carvalho, Andromachi Chatzieleftheriou, James Clegg, Rebekah Storan Clarke, Daniel Cletheroe, Bridgette Cooper, Tim Deegan, Austin Donnelly, Rokas Drevinskas, Alexander Gaunt, Christos Gkantsidis, Ariel Gomez Diaz, Istvan Haller, Freddie Hong, Teodora Ilieva, Shashidhar Joshi, Russell Joyce, Mint Kunkel, David Lara, Sergey Legtchenko, Fanglin Linda Liu, Bruno Magalhaes, Alana Marzoev, Marvin McNett, Jayashree Mohan, Michael Myrah, Truong Nguyen, Sebastian Nowozin, Aaron Ogus, Hiske Overweg, Ant Rowstron, Maneesh Sah, Masaaki Sakakura, Peter Scholtz, Nina Schreiner, Omer Sella, Adam Smith, David Sweeney, Benn Thomsen, Govert Verkes, Phil Wainman, Jonathan Westcott, Luke Weston, Charles Whittaker, Pablo Wilke Berenguer, Hugh Williams, Thomas Winkler, Stefan Winzeck, **and many more!**

- What's next?
  - Looking beyond archival! Join us!