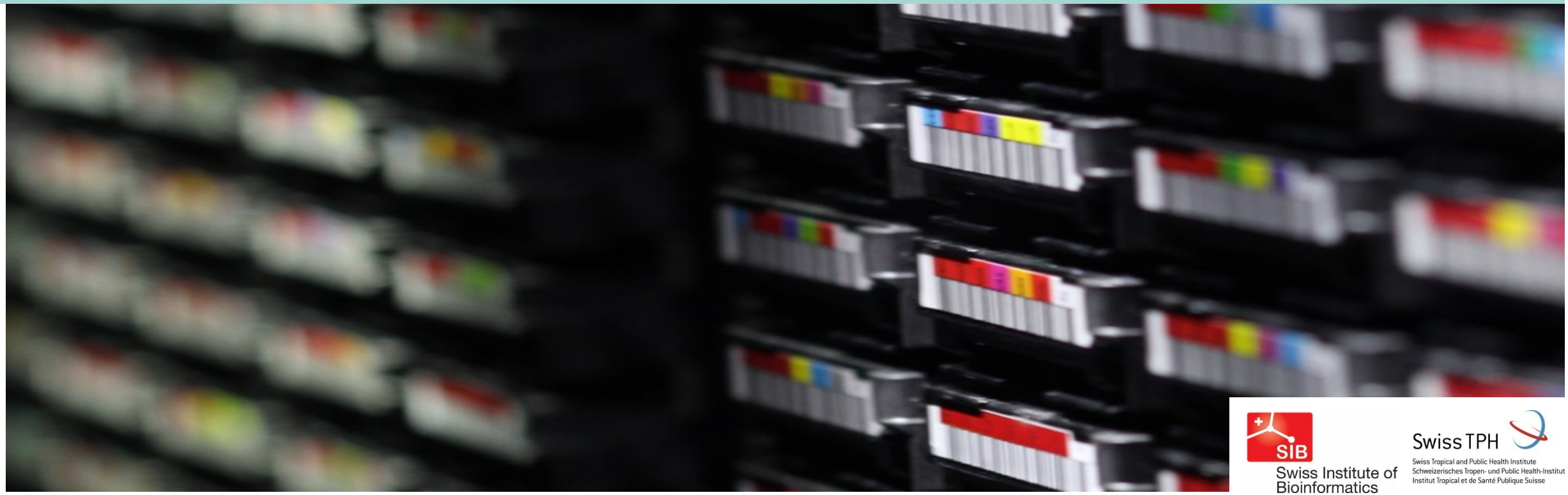
University of Basel

Disaster recovery for large scientific data: what are the options?

HPC-CH May 2024, Lupfig

sciCORE, 2024-05-21



SIB Swiss Institute of Bioinformatics

Swiss TPH
Swiss Tropical and Public Health Institute
Schweizerisches Tropen- und Public Health-Institut
Institut Tropical et de Santé Publique Suisse

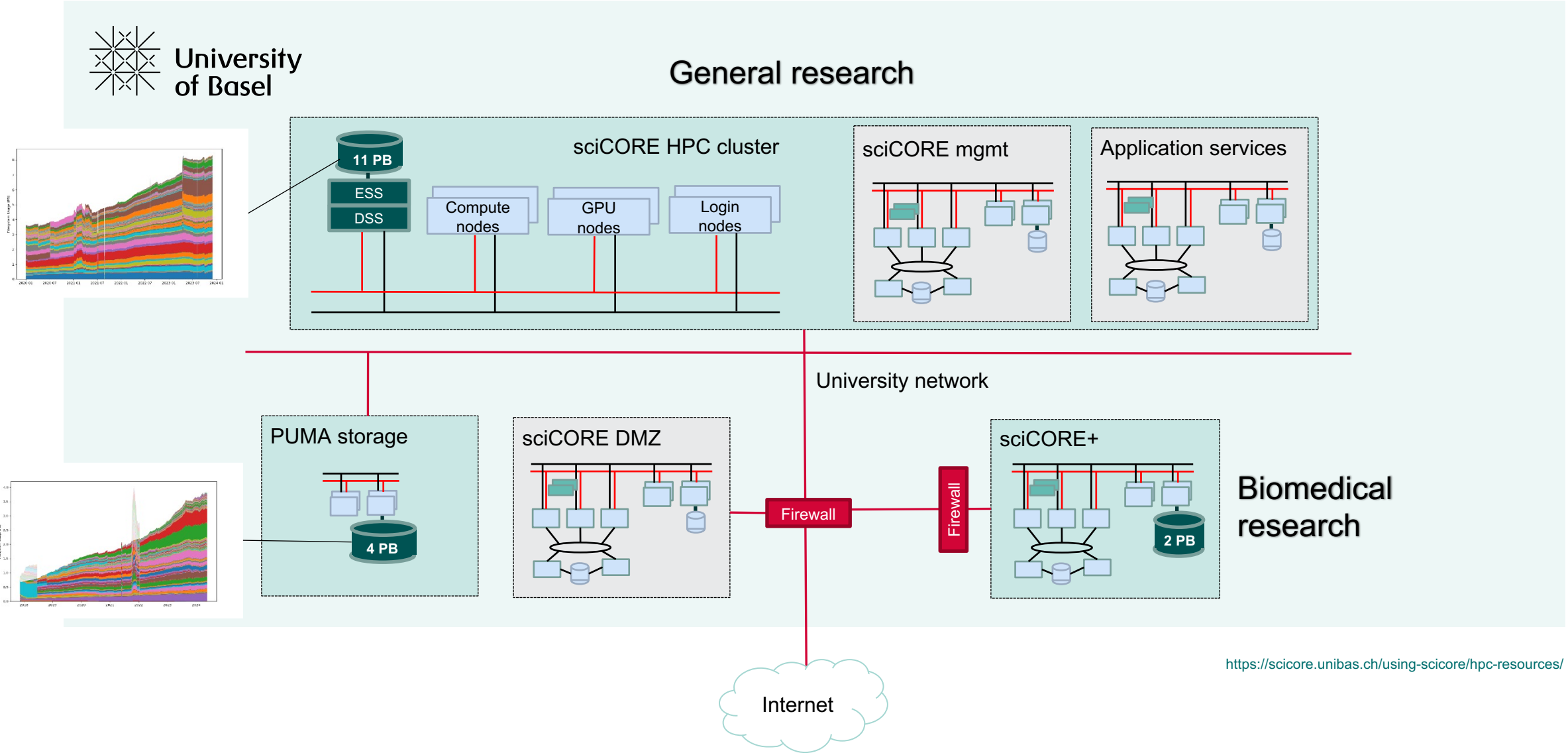
Abstract

Ensuring academic research continuity in case of major disaster is a challenge when dealing with data volumes in the petabyte range.

In this presentation we'll share the state of our thought process concerning disaster recovery at University of Basel.

We want to open the discussion, and possibly seed synergies across organizations to address this challenge.

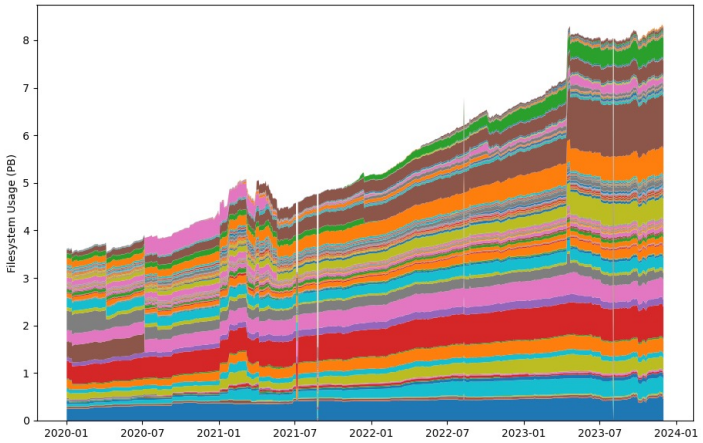
sciCORE evolution: from computing facility to data hosting service



https://scicore.unibas.ch/using-scicore/hpc-resources/

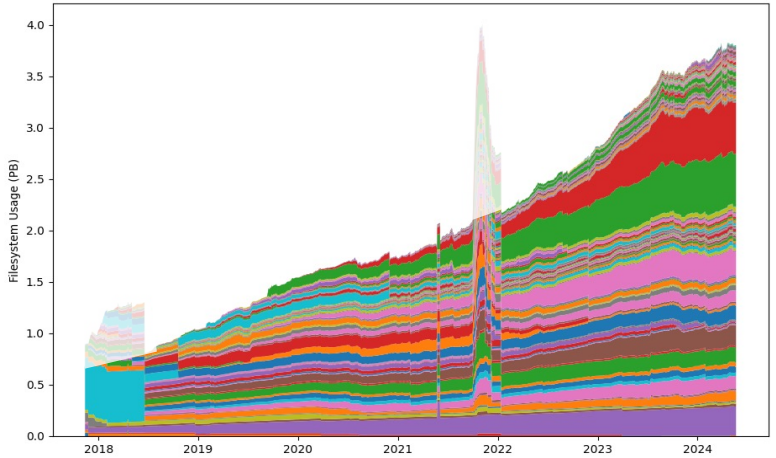
sciCORE storage evolution

HPC-cluster attached storage

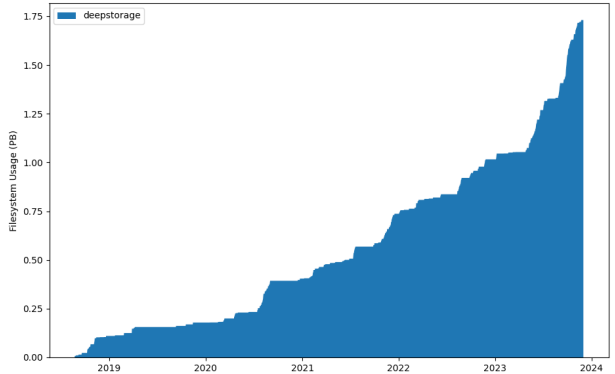


On disk	
2014	0.2 PB
2024	15 PB
2030	42 PB

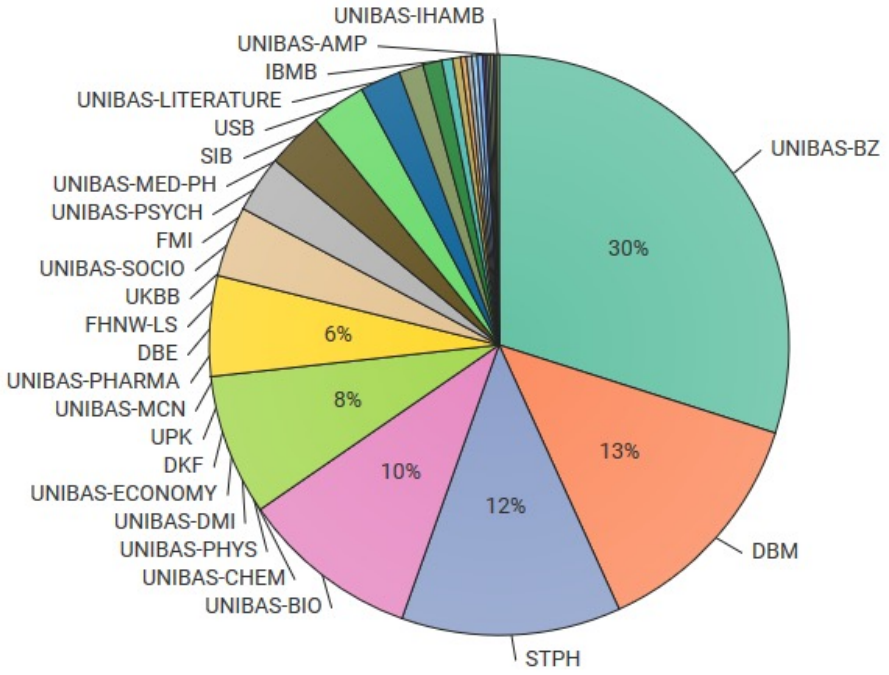
Workstation storage (NAS)



Deep storage (tape based)



Diversity of users and usage patterns



sciCORE user community (HPC)

Over 1100 HPC user accounts (~200 active users, shifting)
70% life sciences, long tail in all university domains
Overall over 1500 researchers rely on sciCORE

Data diversity

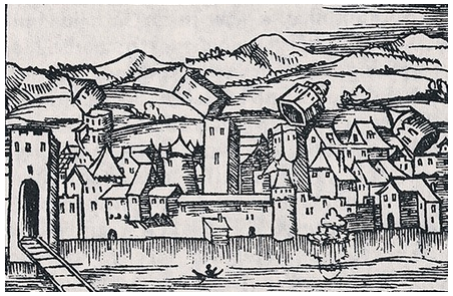
- Analysis codes
- Facilities data
  - -omics data
  - Microscopy
- Clinical study meta-data
- Scans of ancient manuscripts
- Microeconomic data
- Chemistry and physics simulations
- …

Value of data at sciCORE (in risk analysis)

- Value of data is in the **collective work** to create, curate, analyze and publish them
  - Includes work of PIs, wetlab technicians preparing samples, admin personal, etc.…
- Data have value for 5 years (duration of a grant)
- Value = (#FTE involved) x (mean salary) x (5 years)
- For sciCORE: data value > 100 MCHF

Risk analysis for scientific data: loss of availability

Major hardware failure



Cyberattack

Both risks:

- High probability
- High impact

Impact: Potentially large and lasting slow-down of research at university

Disaster recovery (and business continuity)

At sciCORE

- Data backups on tape
- Depending on disaster scenario, restoring work environment can take several months
    - Recall from tape (slow, prioritization difficult) (est. 9 months for a full recall)
    - Time to provisioning hardware (possibly)

Question from our rectorate:

> What are the options to shorten scientific IT downtime in case of major disaster?

Disaster recovery (and business continuity)

At sciCORE

- Formal disaster recovery framework planned (based on outcome of consulting with company)

- Main current gaps:
  - Identify technical options
  - Understand costs

Disaster recovery (and business continuity)

Three scenarios under consideration

- **Full DR**: full copy of data, close to an infrastructure which allows fast access to data (research continuity), failover of management services and core compute capacity
  - Offsite volume: 15-20 PB
  - High cost
- **Essential data**: researcher-curated subset of data, close to an infrastructure allowing fast access to data, failover of management services and core compute
  - Offsite volume: 1 PB
  - Mid cost
- **No additional infrastructure**: data are restored at slow speed from backups
  - No additional cost

Option 1 for offsite DR infrastructure: cloud

Advantage:

- Technologies in place
- Flexible
- Vendors have solutions in place for fast initial ingest
- Collocation of flexible compute and data

Disadvantage:

- Vendor lock-in
- Pricing never entirely clear
- Expensive (arguably)

Pricing discussion with cloud vendors (allegory)

Option 2 for offsite DR infrastructure: build own infrastructure

DR hosting at commercial rack space provider

Advantage:

- Control on hardware costs (competitive acquisitions)
- Regional solutions available

Disadvantage:

- Duplication of effort
- On-site support to be clarified
- Full-cost on storage hosting (rack space and energy)

Option 3 for offsite DR infrastructure: bilateral university partnership

DR hosting at partner university

Advantage:

- Control on hardware costs (competitive acquisitions)
- Lowering costs:
  - Use of existing partner data center
  - Agreement on energy costs TBD

Disadvantage:

- Duplication of effort initial
- Need agreement on “small scale” maintenance (SLA)

Uni Basel data center at Biozentrum

Option 3 bis: partnership through consortium

Discussion points

Is DR for scientific data a topic in your organization?

- Who is responsible for data loss in case of disaster?

Is there an interest in investigating a bilateral approach to DR?

- Feasibility study, PoC, alignment on technologies
- “Exchange” of data center space (a few racks), energy, (lightweight) local support

What about consortium for very large scale storage as DR solution for multiple universities?

- Hundreds of PBs
- Can we leverage infrastructure grants?

University of Basel

Thank you
for your attention.

Axes of DR strategy