Data Movement: Problem Statement & Objectives

For the lone “mid-tail” researcher

Ahmad Hesam, Jan Meijer, Morten Kjeldgaard, Richard Freitag,
Ivan Chorbev GN5-2 WP4 T3

Jakob Tendel - DFN – Work-Package co-leader GN5-2 WP4

TNC’26 – side meeting,

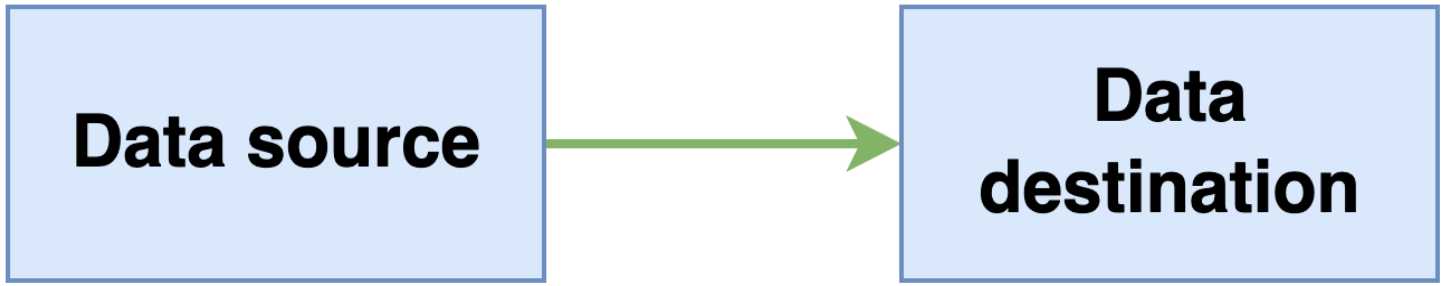
08 Jun 2026

Agenda

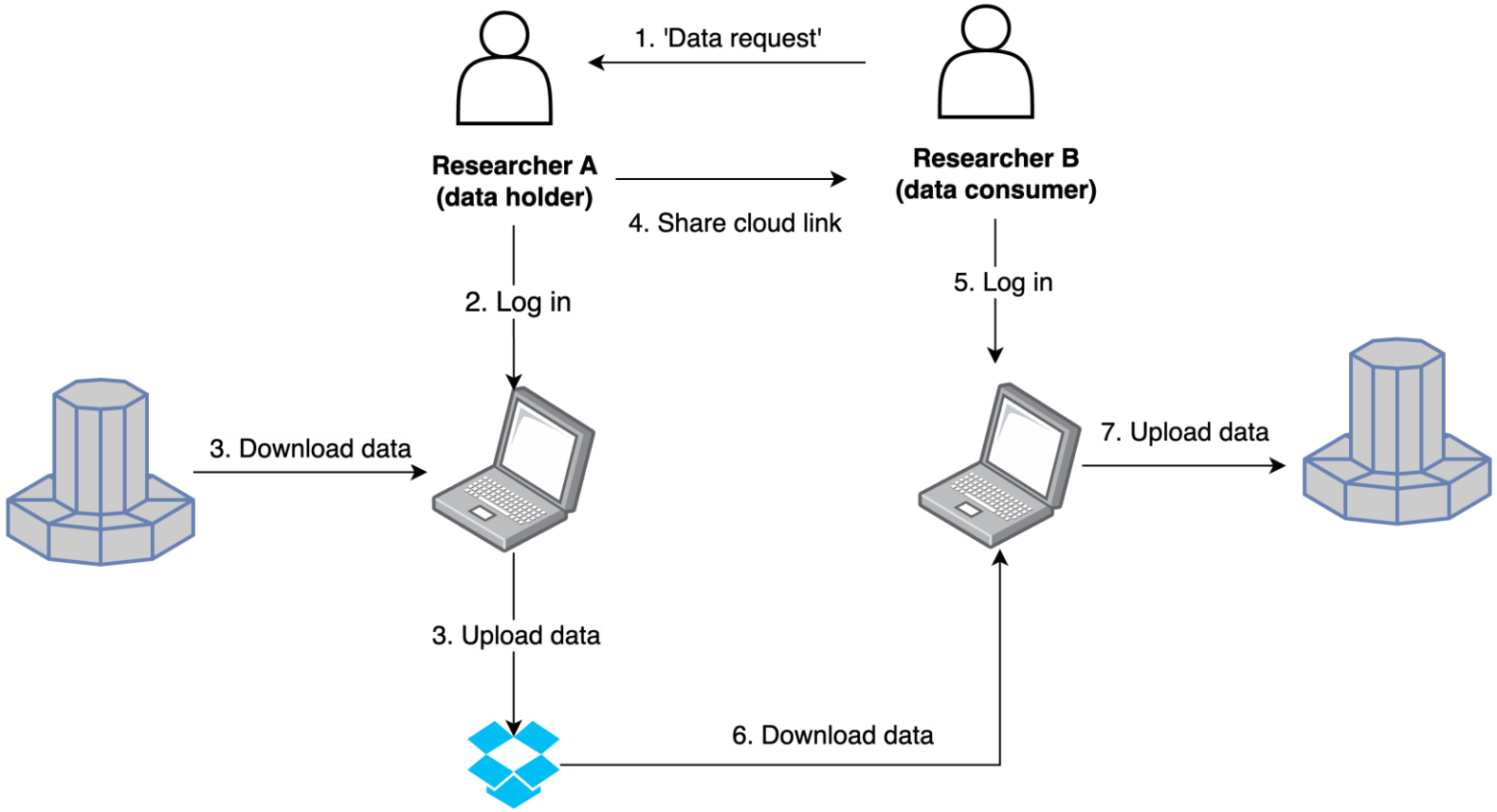
- What do we mean with ‘data movement’?
- The problem as it appears
- Why it might be important for us to address
- Our current questions
  - Is the problem real?
  - Is it in our interest to tackle together?

What do we mean with ‘data movement’?

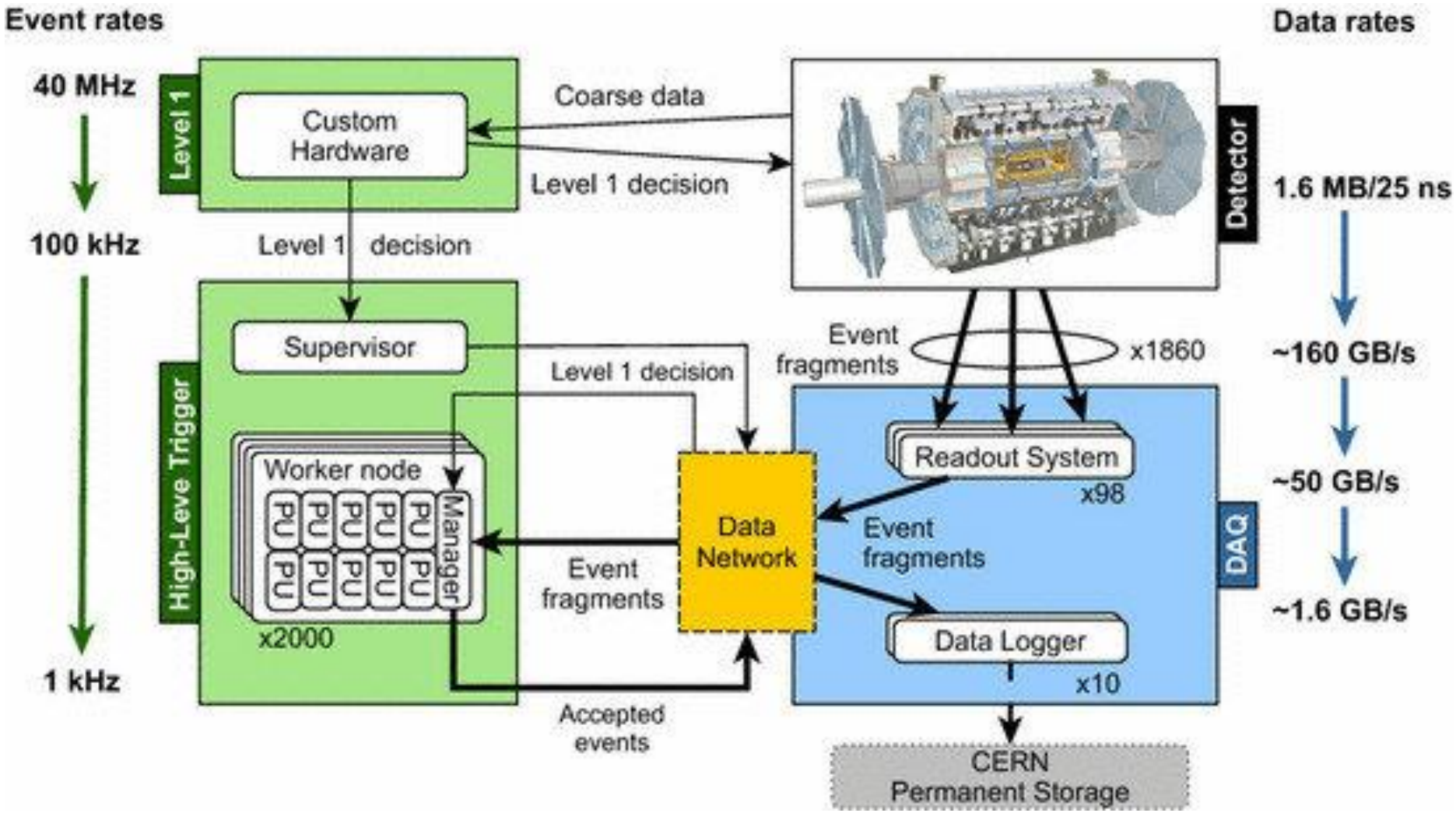
What do we mean with ‘data movement’?



The ad-hoc data transfer



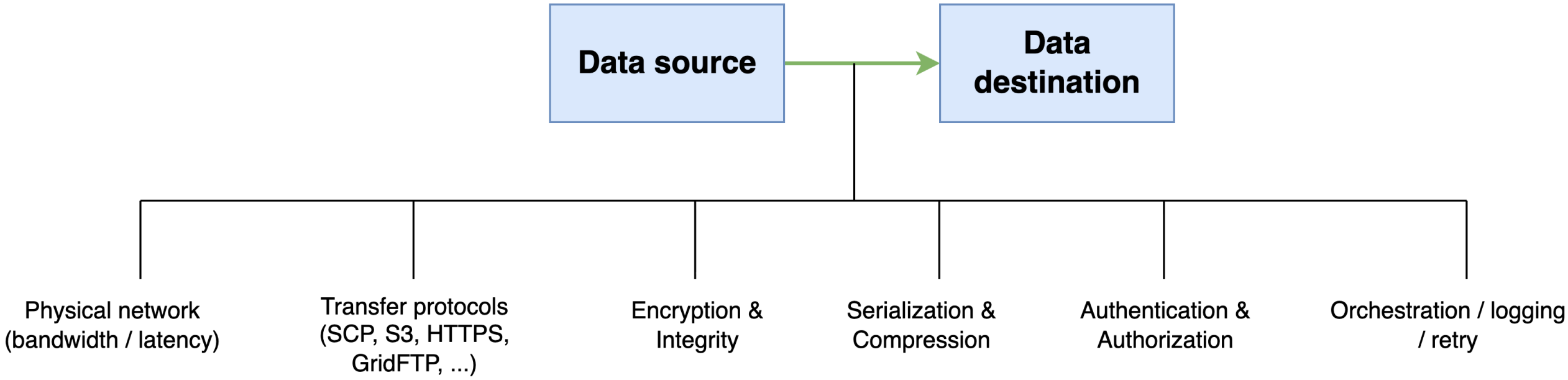
What we do **not** mean with ‘data movement’



Colombo, T., Fröning, H., Garcìa, P., & Vandelli, W. (2016). Optimizing the data-collection time of a large-scale data-acquisition system through a simulation framework. The Journal of Supercomputing, 72.

The problem as it appears

Stack friction



Condensed overview

Use Case	Domain	Computing Type	Sensitive Data	Data Transfer Protocols	Data Transfer Sizes	Number
LISA	Astrophysics / Space-based GW detect	HPC / Distributed (ESA Science Data Cen	No	CCSDS telemetry; TCP/IP for ground distribution	~0.3–0.6 GB/day (~100–200 GB/year)	≈1,500
KM3NeT	Astroparticle Physics (Neutrinos)	Hybrid (onshore HPC + distributed HTC g	No	UDP (offshore→onshore); TCP/IP; GridFTP/xrootd	~25 Gbit/s raw per BB → ~500 TB/year proces	≈360
Einstein Telescope	Gravitational Wave Observatory	Distributed HPC/HTC (multi-site)	No	Rucio / GridFTP / high-speed research networks	Multi-PB/year (similar to LIGO scale)	≈1,700+
LHC CERN	High-Energy Physics	Worldwide LHC Grid (HTC)	No	gridFTP / xrootd / FTS / TCP/IP	~10 PB/day → multiple EB total	≈12,000
Gaia	Astronomy (Astrometry)	ESA Cloud / Data Centers	No	HTTPS / FTP / TAP (Virtual Observatory protocols)	~100 TB per release (~1 TB/day during peak)	7,000+
LENSS (LOFAR)	Radio Astronomy	HPC / Distributed archive	No	GridFTP / LOFAR LTA API / TCP/IP	~55 PB total archive; TB-level per project	≈2,000
SKA	Radio Astronomy	Global HPC (SRCs) / Cloud hybrid	No	Dedicated optical networks / Rucio / GridFTP	~700 PB/year	1,000+
Federated Self-Driving Labs	Materials Science / Automation	Federated / Local + Cloud AI	No	Federated learning protocols (no raw data transfer)	Model weights (~MBs) exchanged	<100 (earl
CAROLINE	Geoscience / Infrastructure monitorin	Cloud + HPC (InSAR processing)	No	HTTPS / APIs to Copernicus hubs	Tens of GB per site/year	10–50 org
EMPOWER	Bioinformatics	Cloud / HPC	Yes (human dat	HTTPS / REST APIs / Nextflow pipelines	10s TB per workflow cohort	~80 labs ('
ADES	Genomics (Alzheimer's)	Cloud / Secure cluster	Yes	SFTP / EGA / Cloud API	~10–100 TB	Dozens (E
MinE	Genomics (ALS)	HPC / Cloud hybrid	Yes	Secure file transfer / HTTPS / Cloud pipelines	~22,500 genomes → 500 TB+	~100 rese
vantage6	Federated Health Analytics	Federated (local + cloud orchestrator)	Yes	HTTPS / REST / Federated learning APIs	Only model results (~MBs per job)	~50–200 a
SUPERGPS 2	Positioning / Metrology	Local / HPC for simulation	No	Optical fibre time transfer / custom sync protocols	GBs experimental data	<50 (R&D
ODISSEI	Social Sciences	Local / Cloud / HPC	Yes	HTTPS / REST / Secure file transfer	10GB - 1TB	~ 10,000 r
Health-RI	Life Sciences	Cloud / HPC	Yes	SFTP, HTTPS, rsync, Globus, Aspera, S3	10GB - 10TB	40,000 - 5
ScilifeLab Sweden	Life Sciences	HPC & Cloud	Yes	HTTPS / S3 APIs, SFTP / SCP, Globus	50 GB – 2 TB per project	~4,000 res
Georeferenced Data Collection (Swed	Earth Observation / Remote Sensing	Cloud	No	S3 API, HTTPS	70 GB - 200 TB	50 - 200

- Metadata enriched via public documentation / reports / papers
- 90% obtained by roughly normalizing the 'Data Transfer Sizes' by the 'Number of Active Users'
- Excluded CERN ffrom the equation

Why it might be important for us to address.

Not our fault. But maybe our problem?

- Average researchers
- User experience, network data transfer
- Frustrating / ineffective
- Our networks don’t feel useful 

-> Our problem!

Our current viewpoint

The current problem statement

- Researchers generate and share **vastly larger datasets** (10 GB → 1 PB), but still move them with **manual tools** (SCP, SFTP, FTP, rsync).
- **90 % of users** work with data in the 10 GB–10 TB range yet lack reliable, resumable, or automated transfer methods.
- **Protocols and workflows don’t scale —** they were designed for single files, not multi-terabyte, multi-party transfers.
- **Ad-hoc project setups** dominate; every institute reinvents its own solution, leading to high overhead and inconsistent security.
- Even where networks are fast, **usability and interoperability are the real bottlenecks —** not bandwidth.

The current objectives

- Make research data transfers **reliable and effortless** in the 10 GB – 10 TB range
- **Bridge the gap** between manual tools and facility-grade systems
- **Guide users though a ‘Data Transfer Wizard’** for optimal, compliant routing
  - **Automate protocols** like SCP, S3, GridFTP, and rsync
  - **Leverage existing high-speed networks** through simple interfaces
- **Deliver a consistent experience** across research domains

Questions for audience validation

- Have we described the correct core problem?
  - Now? Foreseeable future?
- Do we have the right order of magnitude?
  - Is 10 TB a barrier for ~1 million researchers
- Are NRENs in a position to “fix it once” centrally?
  - Collective action at the GÉANT level and partners
- Is it important enough for us to do so?

Discussion

SURF's Future Network Needs - Findings

- Gather long-term views on data traffic in research
- Interviews with 51 SURF employees
  - Mostly project leads and community managers
- 23 external contacts
  - Researchers, (IT) support staff, professors that have research collaborations with SURF
- Various domains
  - Astronomy, physics, chemistry, geoscience, healthcare, extended reality, humanities, social sciences, IT security
  - 40+ use cases gathered (of which about 15 were relevant)

Data Movement Layer

Data movement layer

- Protocols (S3, rsync, SCP, GridFTP)
- Identity & Policy (AAI, access rules)
- Automation (brokers, orchestration)