Federated Storage in HPC

Presented by Patrick Fuhrmann with contributions bymany experts.

Contributions and thoughts provided by

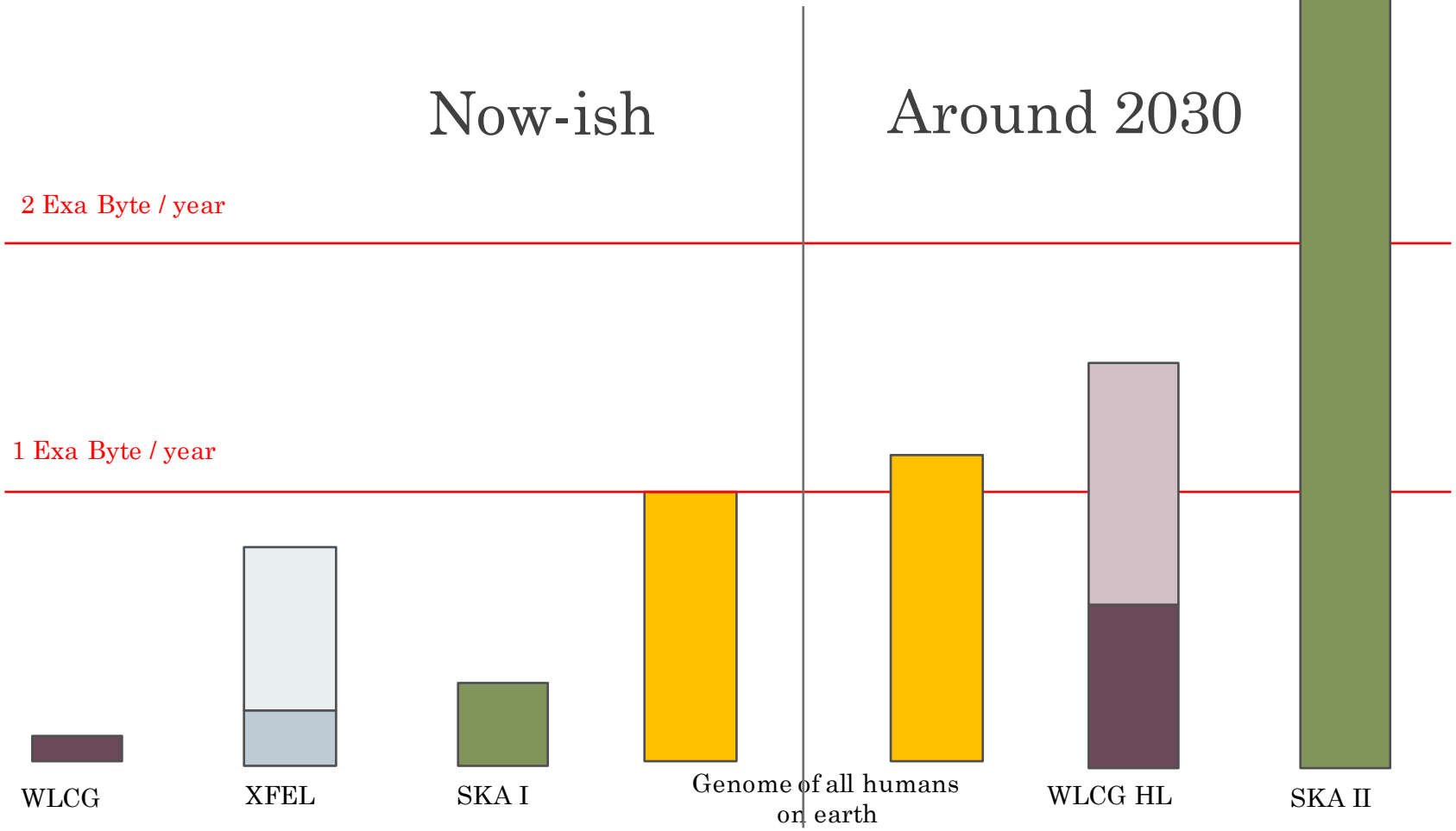
- Daniel Mallmann, JSC Juelich
- Florian Berberich, Prace
- Stephan Kindermann, and
- Michael Lautenschlager, DRKZ Hamburg
- Many more people I cornered and stalked

Kind of : follow up on Olivier's presentation this morning.

One slide on “My Data World”

Antonietta : Male Macho “Showing Off”

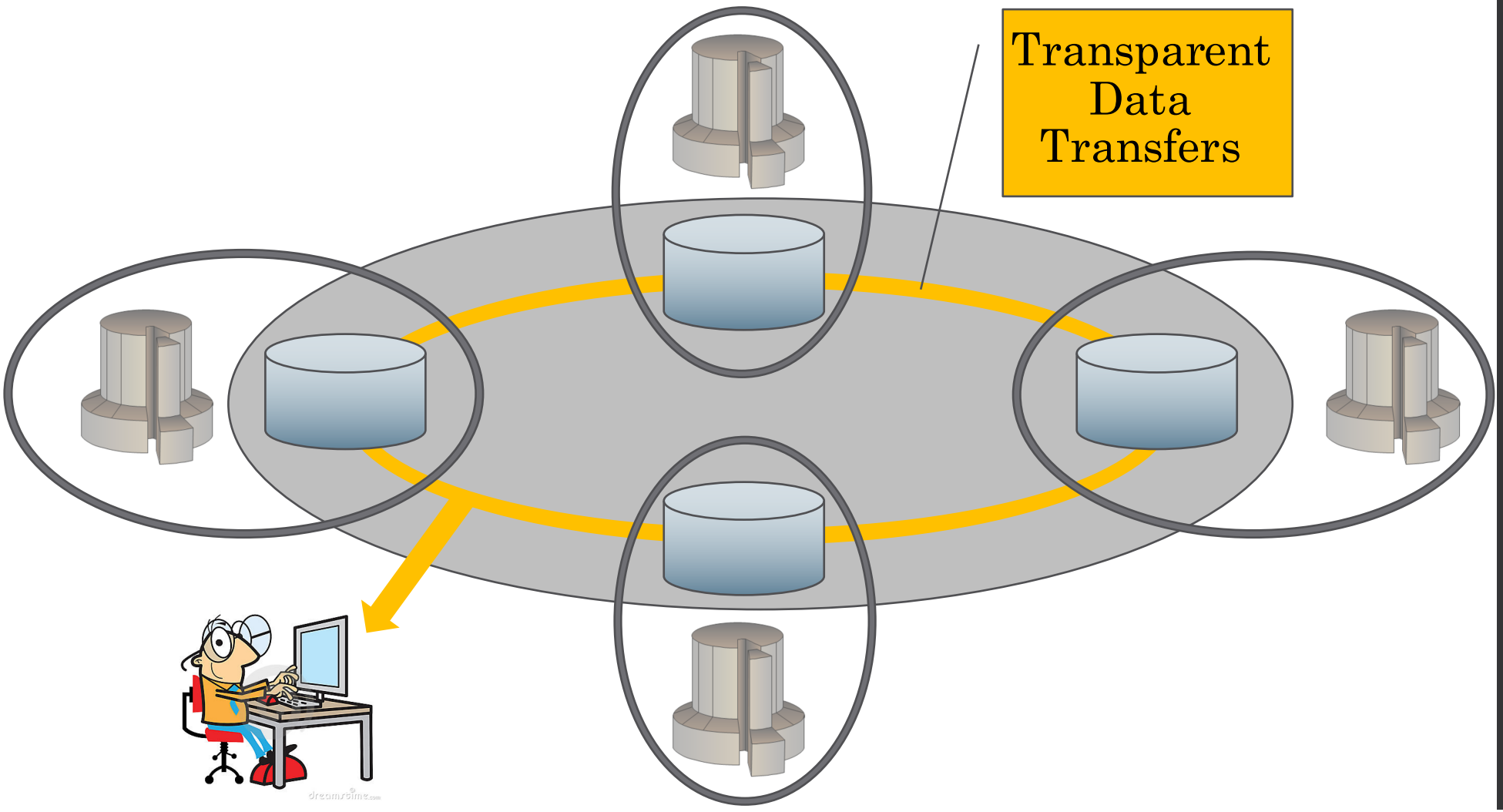
The Data Storage production rates



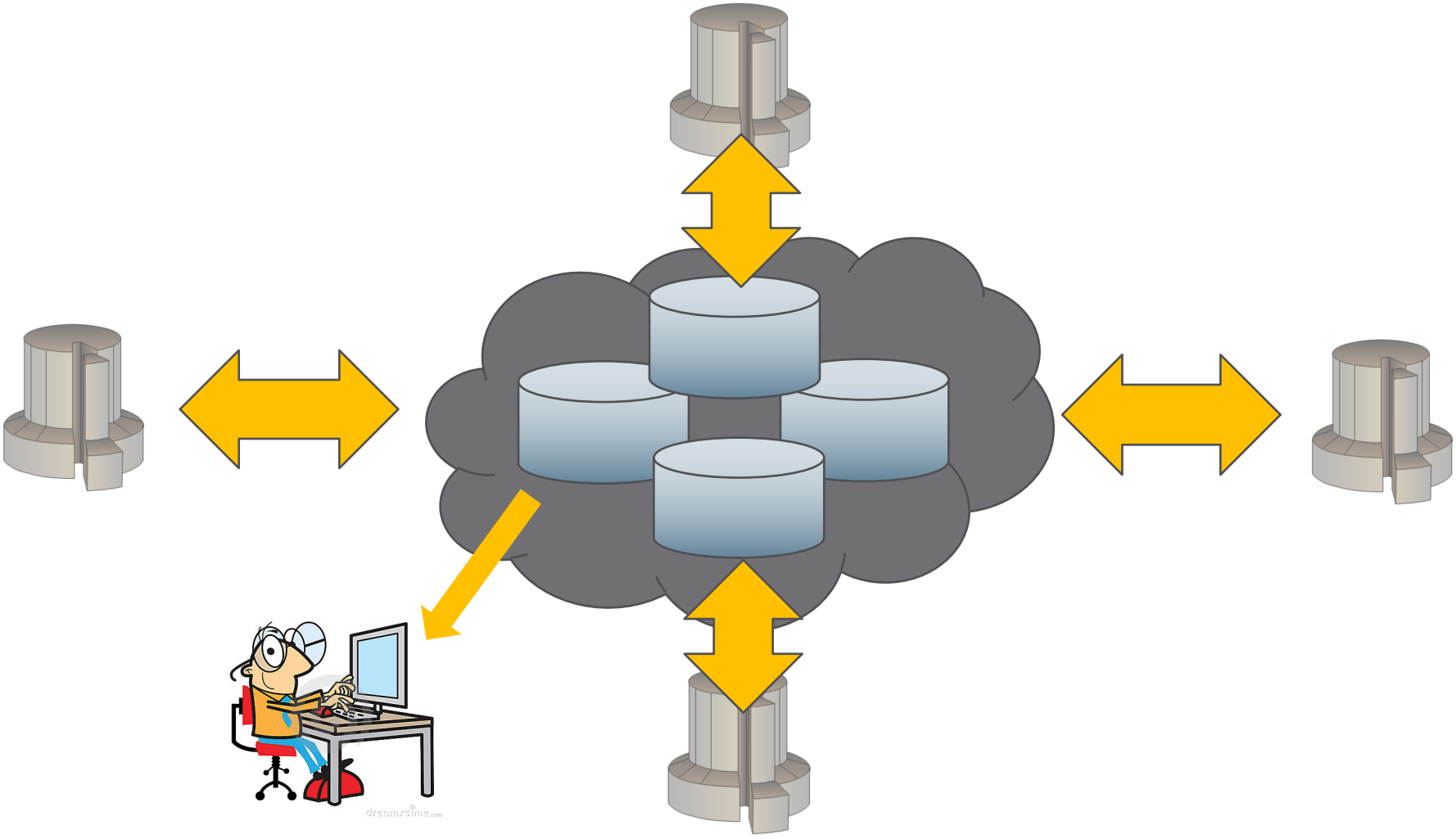
Moving on :

My view of a data federation !

My definition of a data federation !



This is not a data federation !



Why does HPC need data federations ?

- Limited Space at the local site
- Make data available outside of the 'data producer' site
  - Data sharing
  - Reducing network latency for further processing any analysis.
- Data Redundancy
  - Avoids loosing data
  - High availability (primary system could be done, even happens to Amazon)
- Comparing large datasets. (Comparing climate models)
- Moving data to long term storage facilities.
- You can not always move computing to the data (Oliver this morning)
- Funding agencies will ask for a Data Management Plan
- You name it ….

The current E Infrastructure program of H2020 is a perfect match of this workshop.

H2020 EINFRA 21 has essentially two main topics:

HPC

"Support to Public Procurement of innovative HPC systems"

Data Federations

"Universal discoverability of data objects and provenance"

…. **which cuts across geographical**, temporal, disciplinary, cultural, organizational and technological **boundaries**, **without relying on a single centralized system** but rather **federating locally operated systems**

European Open Science Cloud

What do we expect from a data federation ?

Basics

- Should support FAIR principles
- A central catalogue service allowing to find data independently of its location. (findability)
- Users can get access to all connected data services with a single credential.
- Users are consistently mapped to the correct identity at the local and remote site.
- ACLs can be enforced consistently within the federation.
  - Even for temporarily cached data
- Users (sysadmins) can easily trigger bulk transfers.

What do we expect from a data federation ?

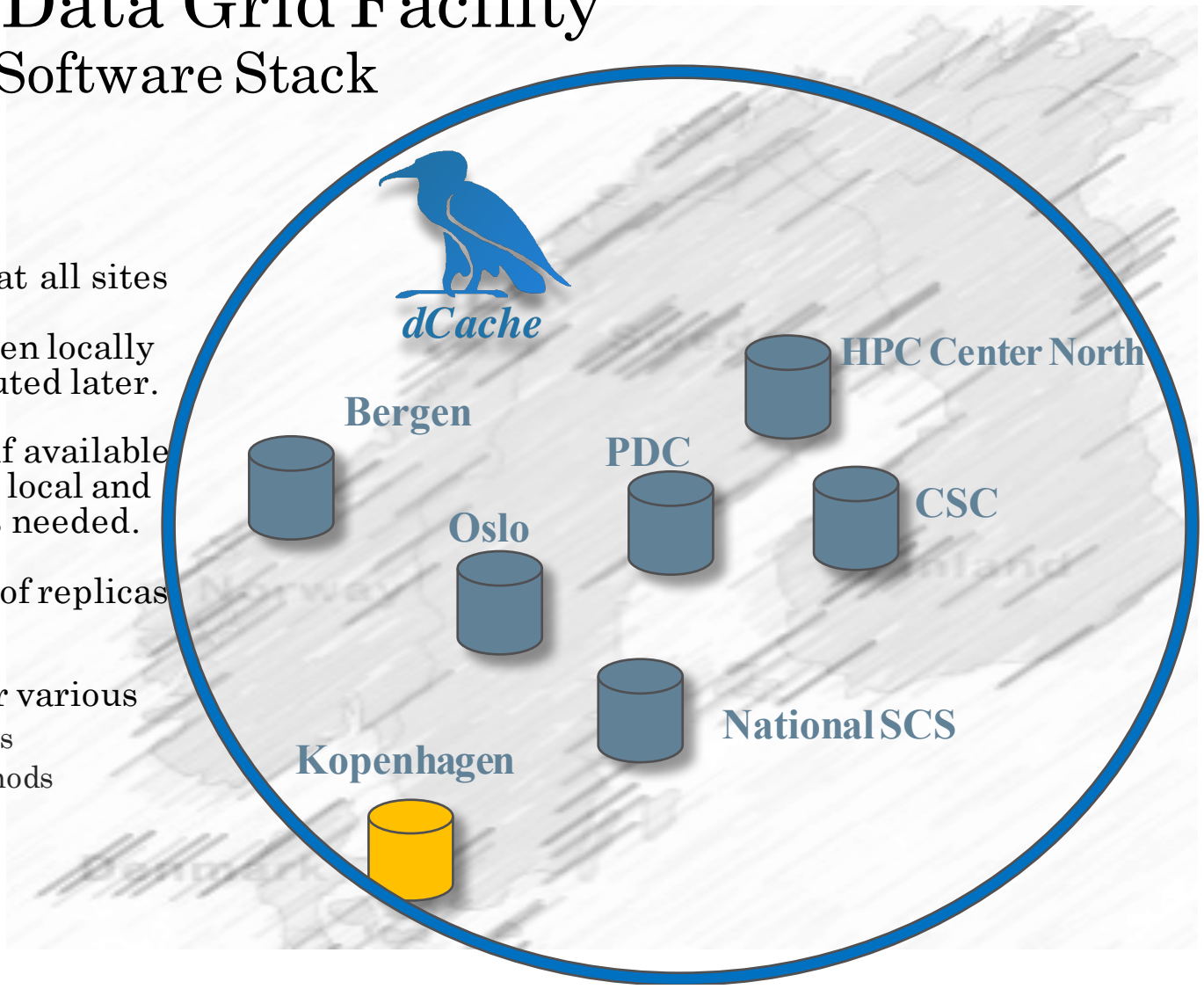
What would be really cool :

- There is only a single namespace, spanning the entire federation.
- Accessing a ’remote dataset’ will automatically trigger a transfer to the local system.
- Local cache copies are created and removed based on the access profile.
- Data Location Orchestrators are migrating data to the most appropriate location based on policies or on detecting data access patterns

Existing Federations

The Nordic Data Grid Facility
Homogeneous Software Stack

- Single name space
- Same User Identity at all sites
- Data is always written locally and possibly distributed later.
- Data is read locally if available otherwise fetched to local and cached for as long as needed.
- Automated creation of replicas for high availability
- Consistent access for various
  - Data access protocols
  - Authentication methods



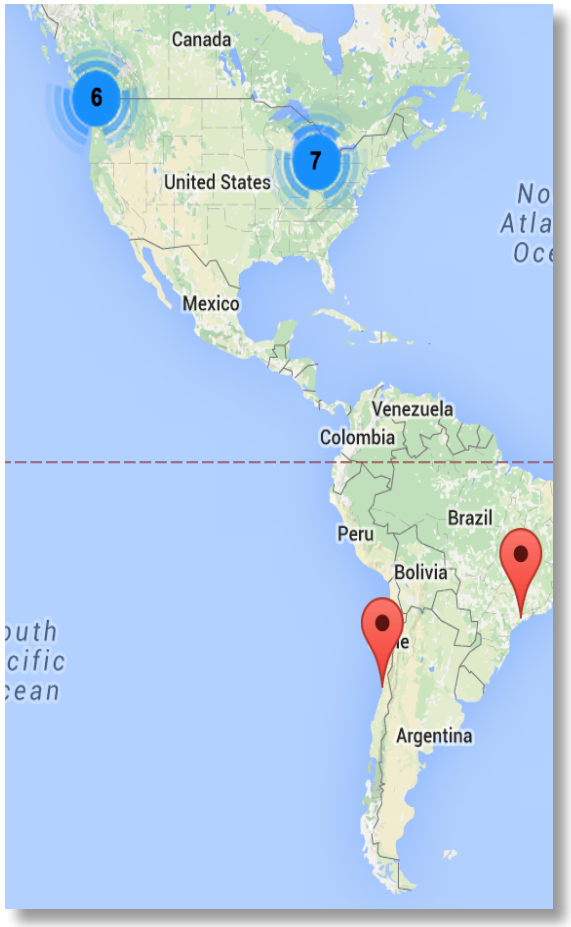
Two slides on dCache

dCache Cheat-sheet

- Combines heterogeneous storage nodes under a common virtual file system tree and scales into 100PB region.
- Provides access to data via a variety of protocol, e.g. NFS4.1, WebDAV, GridFTP, etc. seeing the same name space.
- Provides a variety of authentication mechanisms, like User/Pass, X509 Certificates, Kerberos, in preparation SAML and OpenID Connect, Macaroons.
- Multi Tier support: moves data around between different media types, like Tape, Spinning Disks and SSDs.
  - By user request.
  - Automatically based on the access profile, hot spot.
- Provides resiliency, e.g. through multiple copies.

Worldwide dCache distribution

Worldwide dCache distribution



Other “single system” solutions

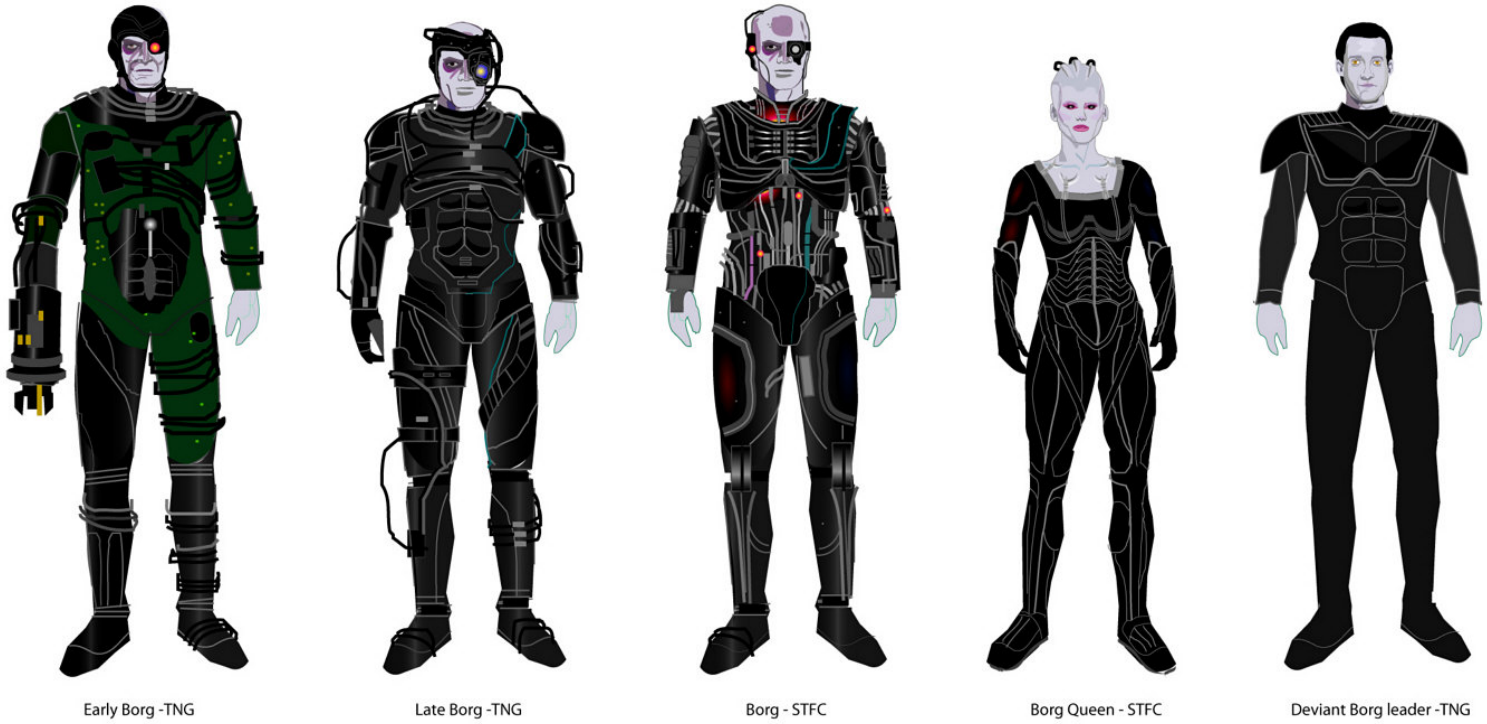
European Infrastructures

- **EUDAT**
  - B2SAFE
  - B2SHARE
  - B2STAGE
  - B2FIND
- **EGI**
  - **Onedata**

IBM : AFM cache

- Read Only
- Local Update
- Single Writer
- Independent Writer

Moving to real heterogeneous Federations

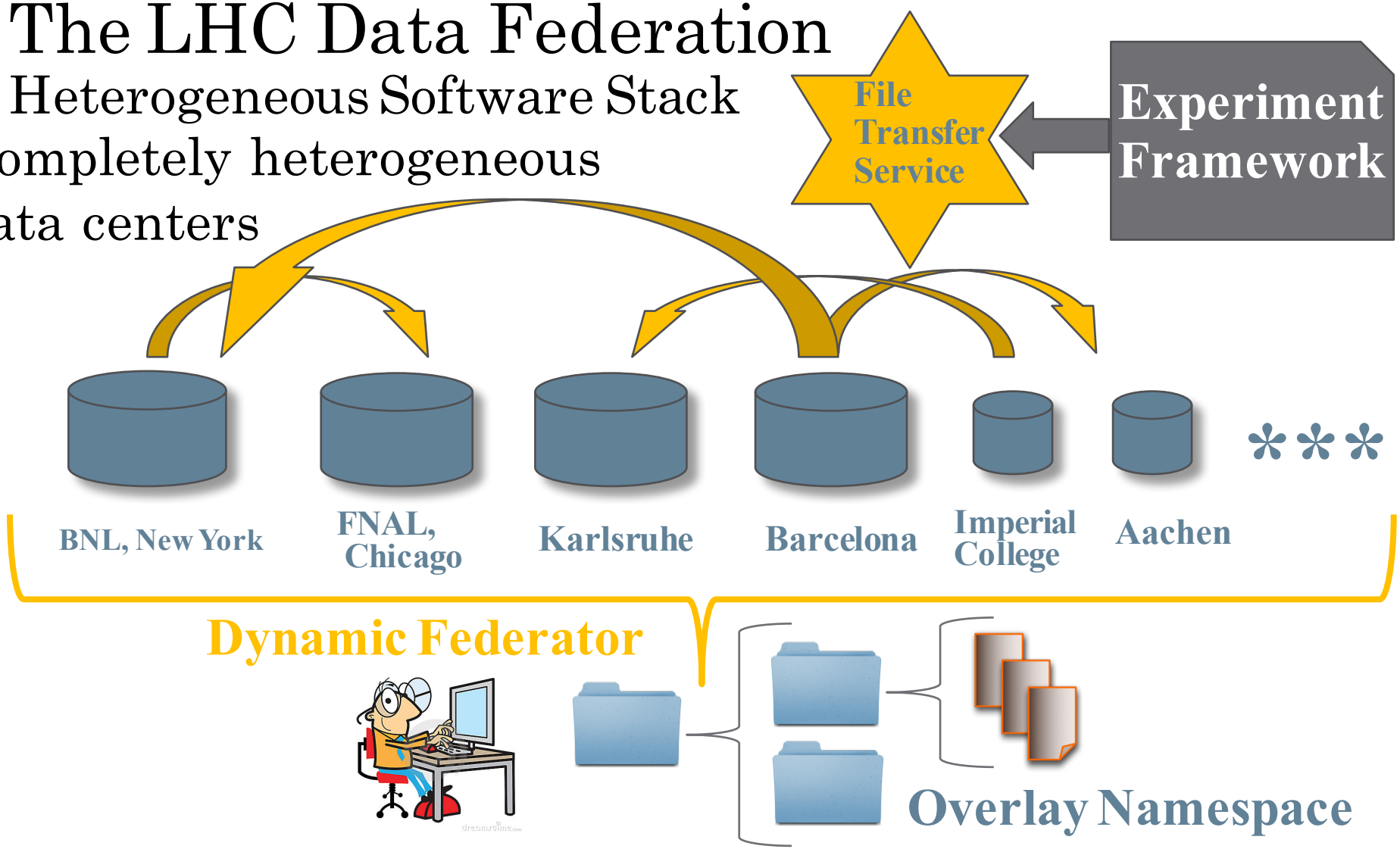


We are WLCG: You will be assimilated

The LHC Data Federation

Heterogeneous Software Stack

Completely heterogeneous data centers



The LHC Data Federation
Heterogeneous Software Stack

- Experiment frameworks orchestrate the data placement.
- Data transfers performed by network bandwidth aware transfer management
- Dynamic Catalogue system generates global namespace on the fly.
- Federator only redirects to the actual location (No data flow through federator)
- Data fetched from closest location or all locations (via meta links)
- Authentication via X509 certificates.
- System transferred One Exabyte in 2016

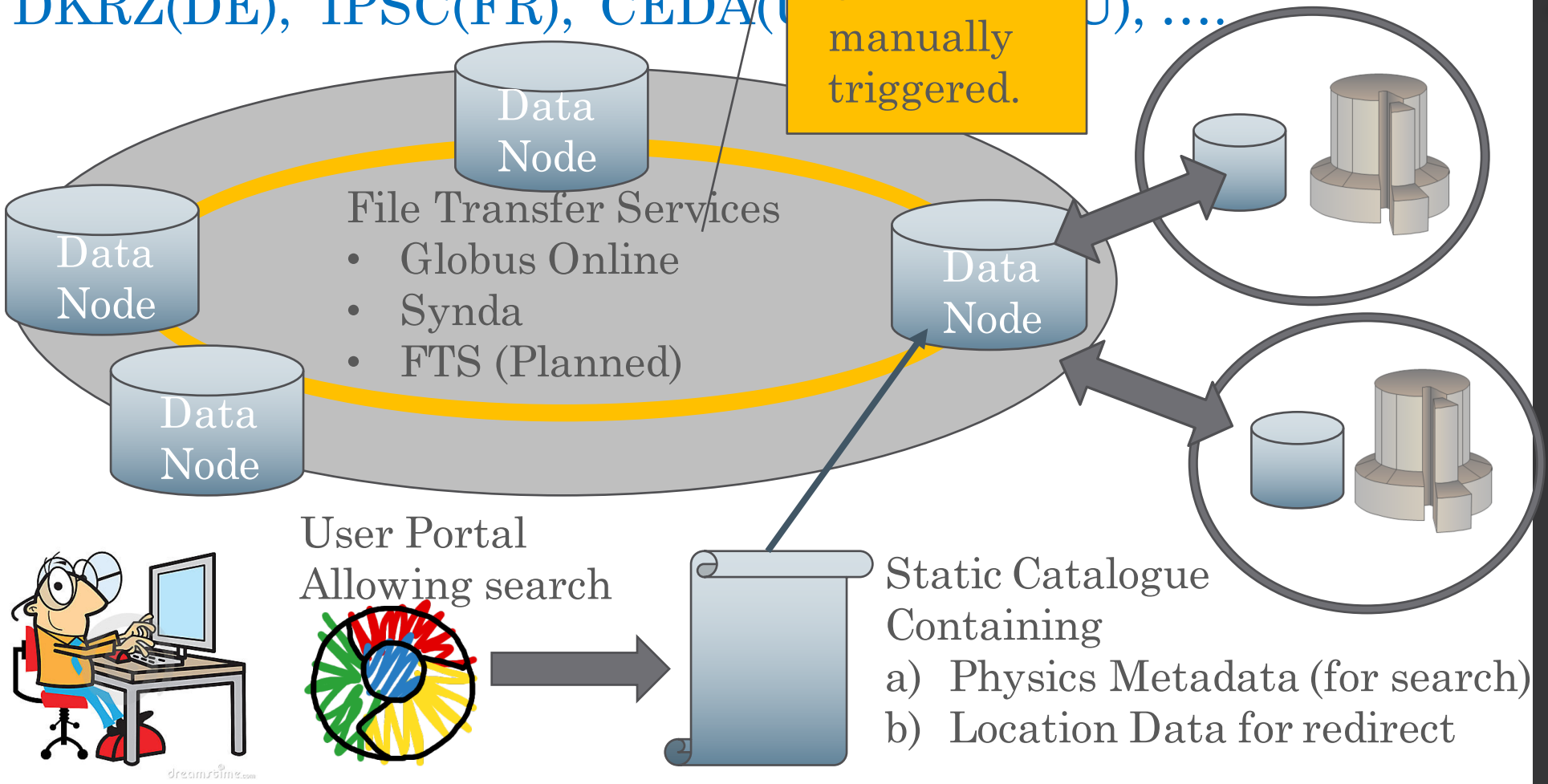
And all the issues ….

- Painful agreement on a “federation control protocol” (SRM)
- Using X509 Certificates, which people find difficult to use.
  - Trying to migrate to OpenID Connect (with Online CA’s)
- No guarantee of the same access control (authorization) at the endpoints.
- Catalogues getting out of sync ;
  - Dark Data
  - Dangling references
- And again the difficulty to predict the future:
  - Lots of unnecessary data transfers
  - Lots of data stays unused and blocks storage space

Moving on to an HPC example

The Earth System Grid Federation

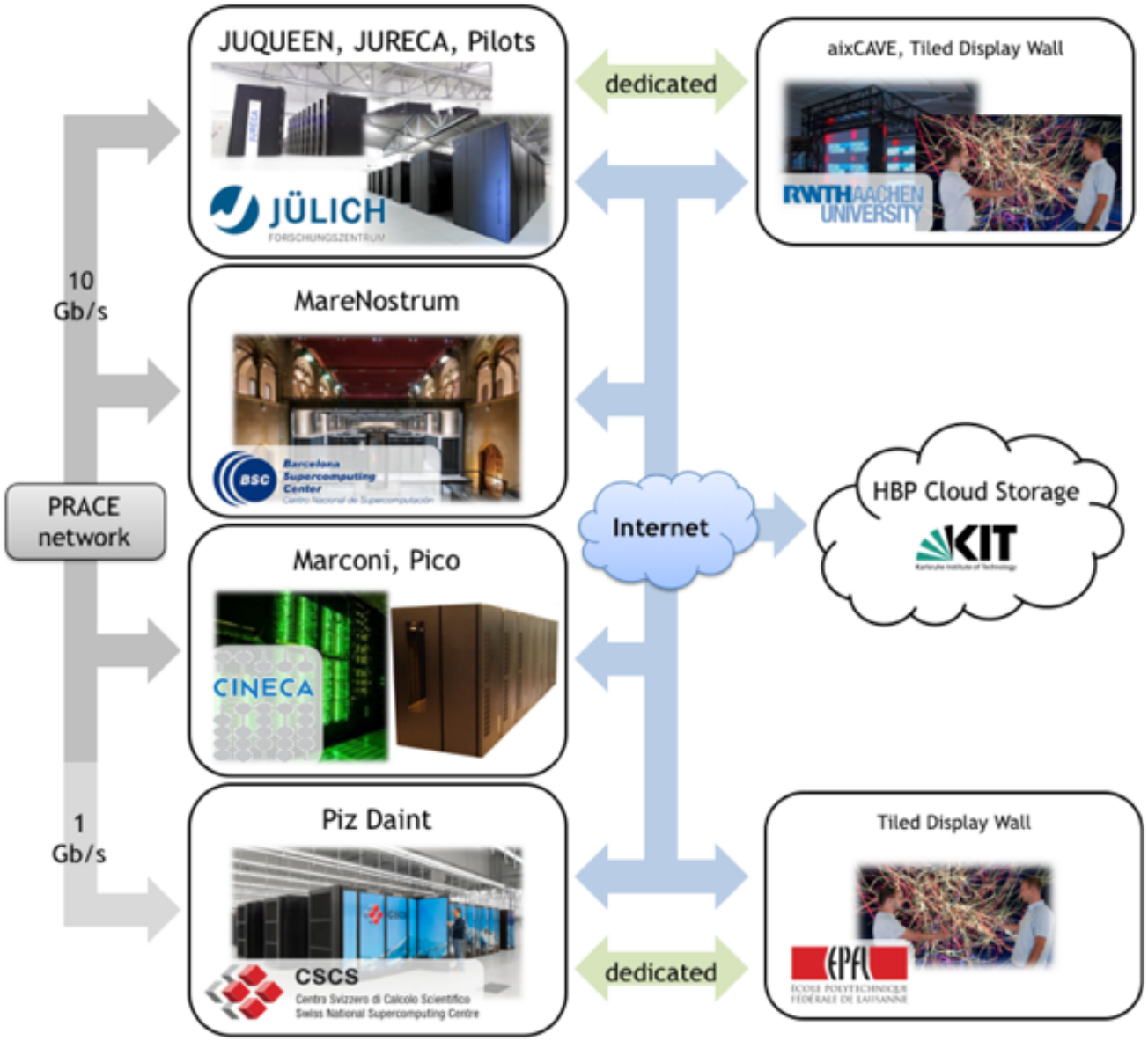
DKRZ(DE), IPSC(FR), CEDA(U…U), ….

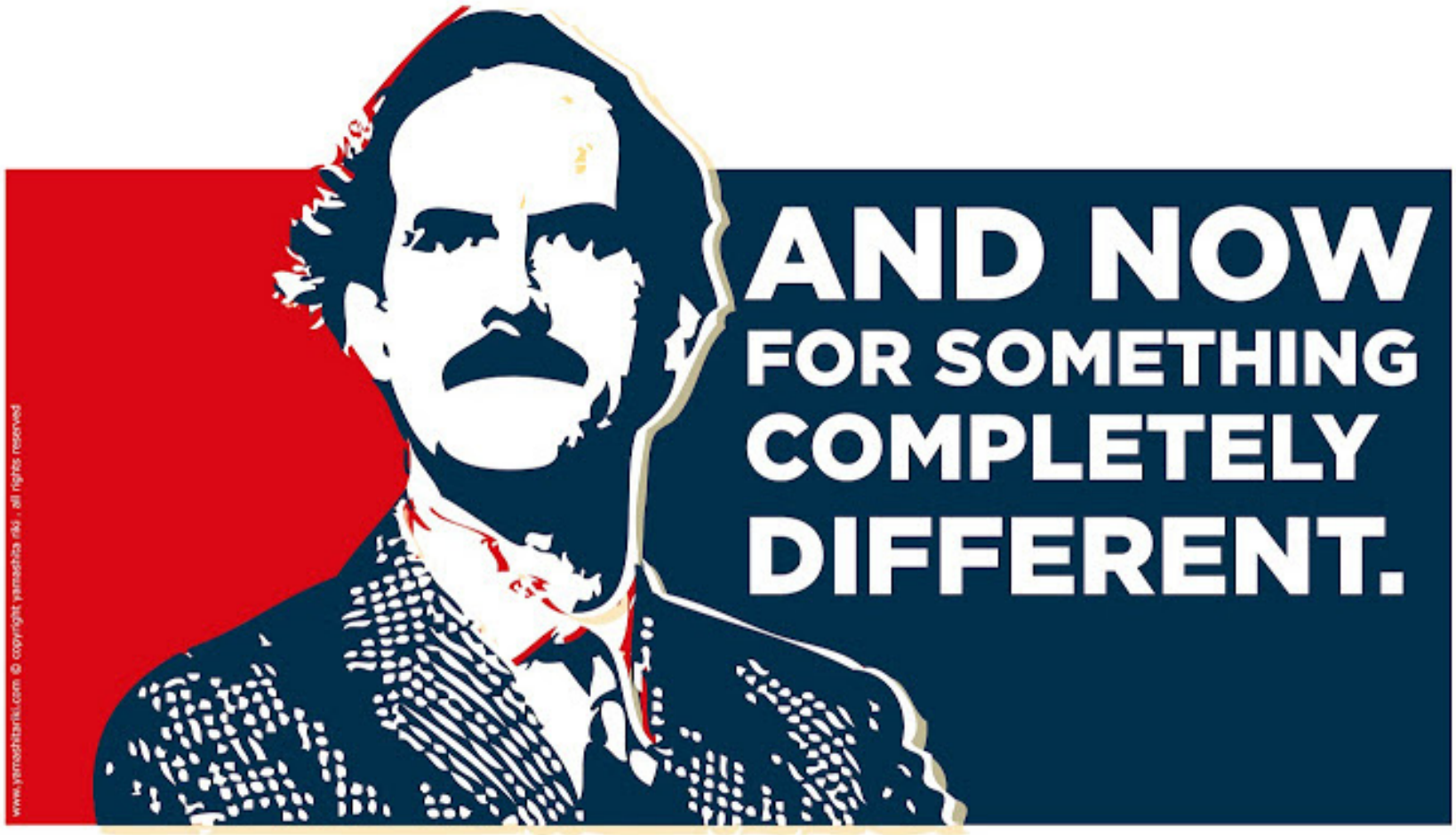


Human Brain Federation

Stolen from:

High-Performance Analytics and Computing Platform specification update (WP7)




AND NOW
FOR SOMETHING
COMPLETELY
DIFFERENT.

German – Russian Federation Project

(Proposal only)

Russian German Federation Proposal

- HGF Project Proposal
- Locations
  - Moscow (Kurtschatow)
  - Dubna
  - St Petersburg
  - Hamburg
  - Munich
- Analysis of Altas and Belle II data

What's so special ?

- Highly distributed physics analysis
- Data is automatically moved to the most optimal location (cached or permanent)
- We intend to use machine learning based on
  - Physics meta data and
  - Access pattern recognition

  To find the best location (caching)

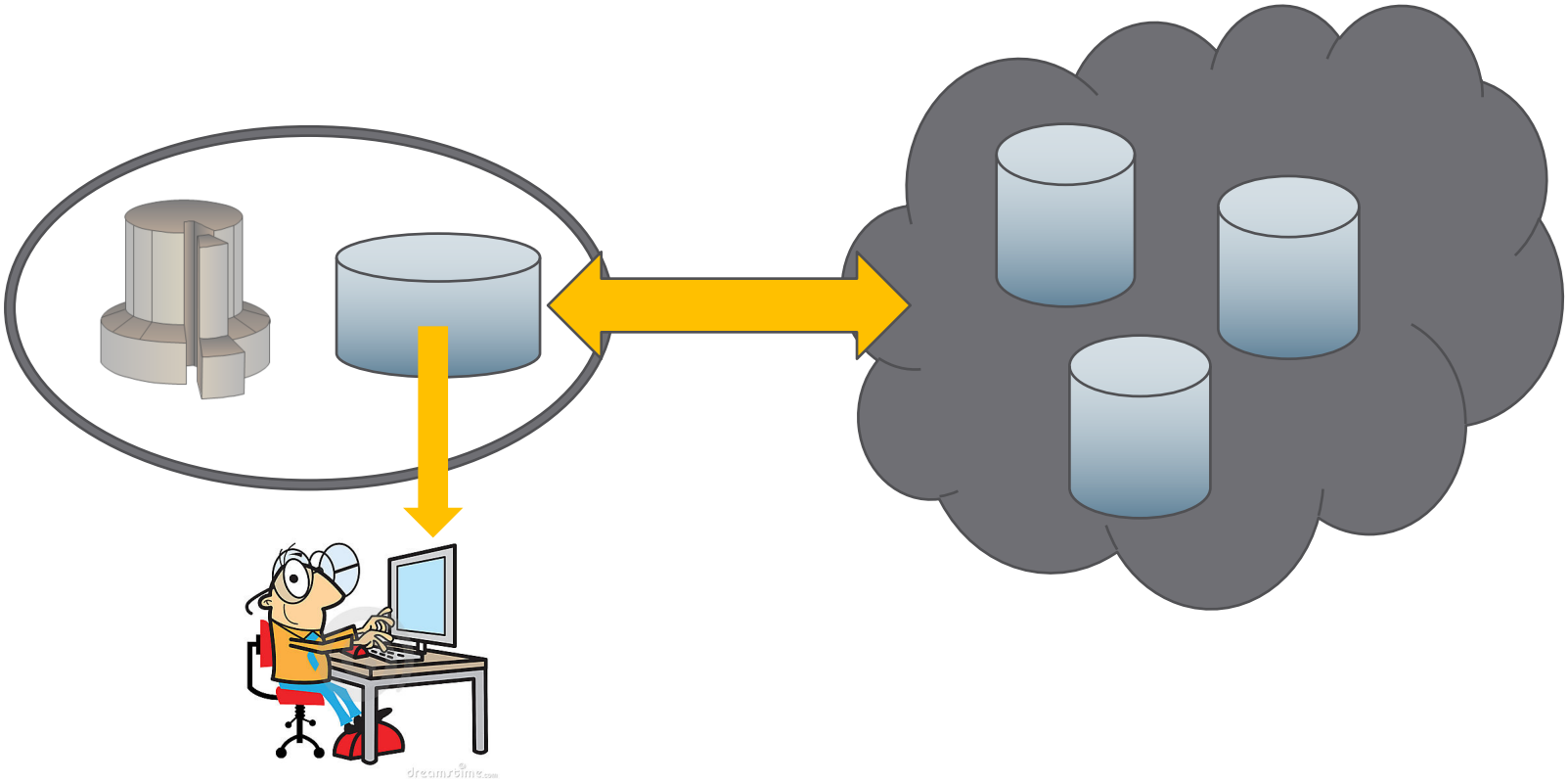
Conclusion

- Data Federations already provide great advantages for some high data volume sciences in terms of
  - Sharing data
  - Data redundancy
  - Data protection
- Although around for more than a decade there are still problems to solve.
  - “Best location” problem
  - Authentication and authorization
- Funding agencies will insist in shareable big data infrastructure.
- Didn’t talk about data privacy issues.
- HPC already makes first steps in that direction.

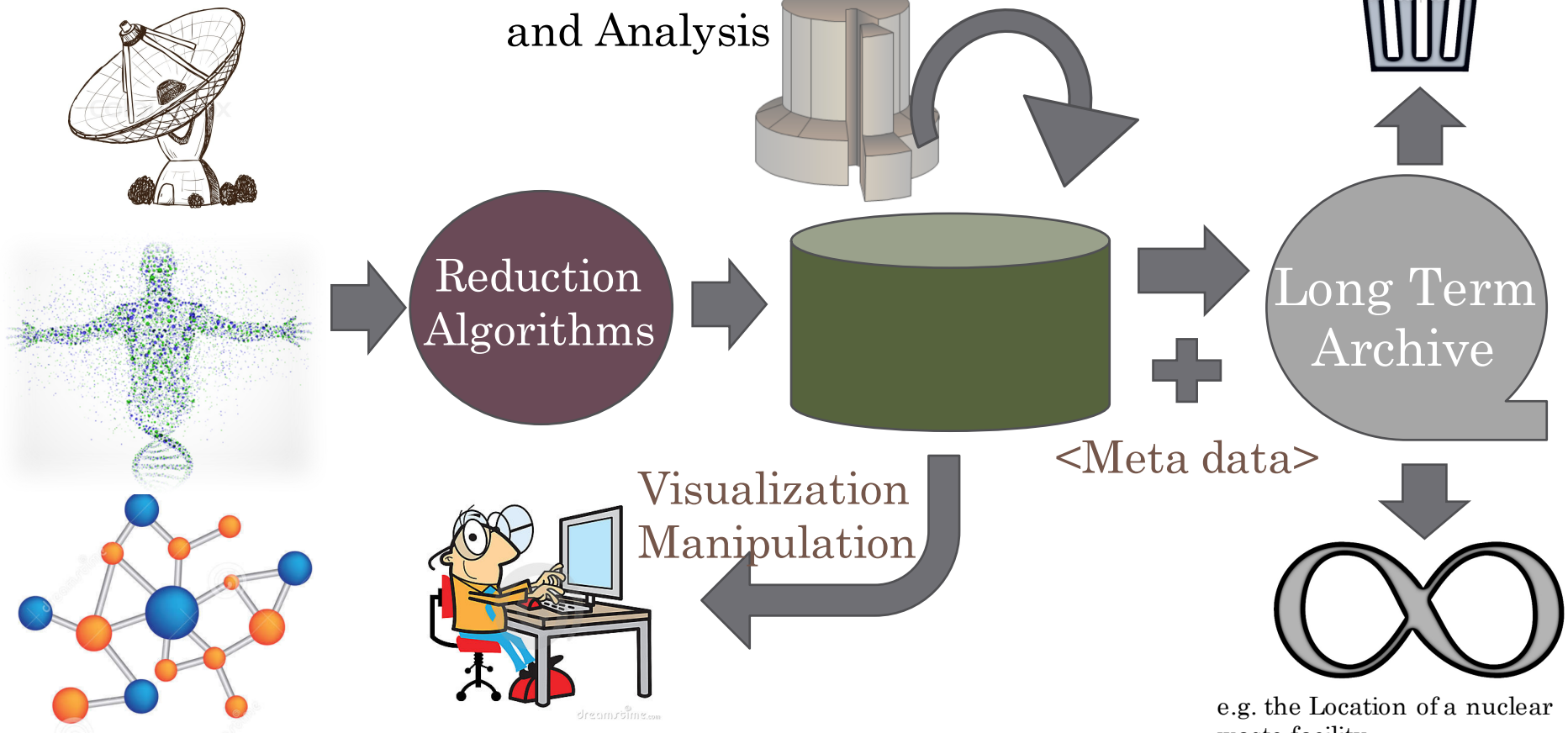
The END

Backup Slides

Which, from the users perspective, is equivalent to :



My simplified idea of data life cycle

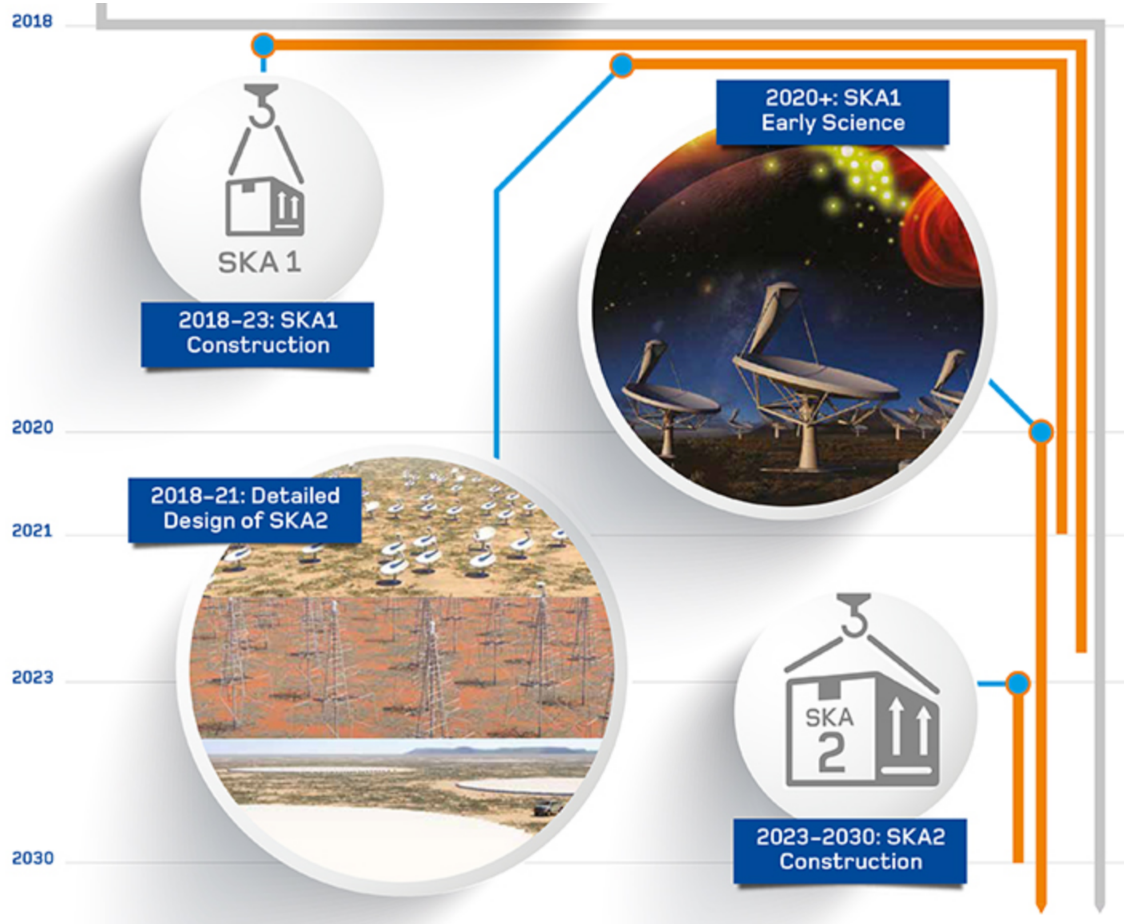


Laying the scene

What is the order of magnitude we are talking about ?

The Square Kilometer Array

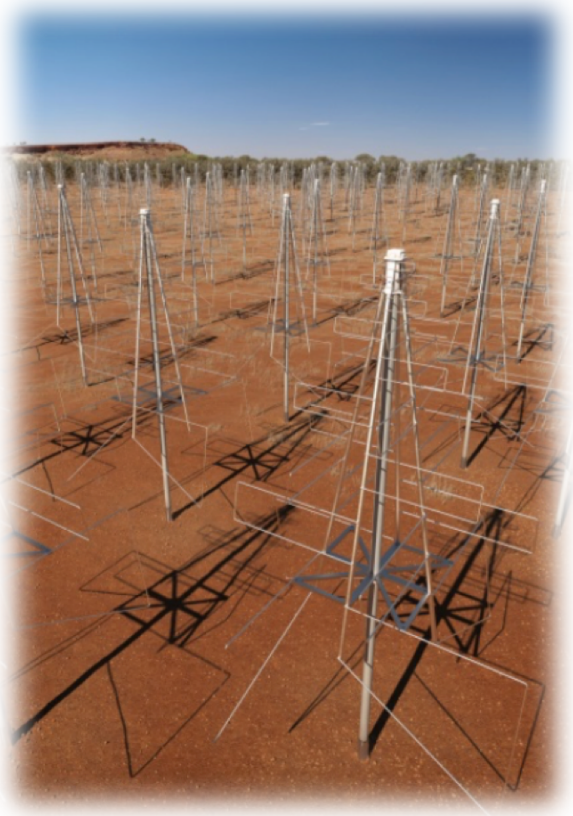
SKA I, 2018



SKA II 2030

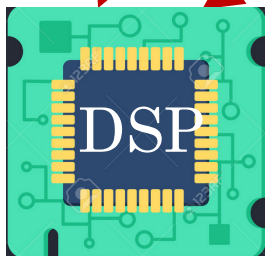
Stolen from the SKA Homepage

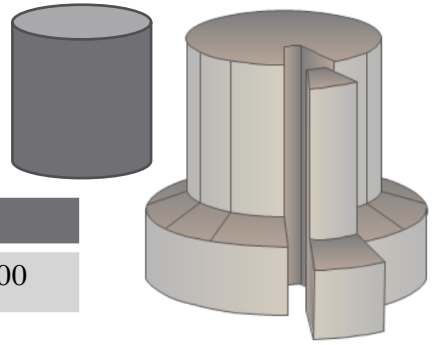
The Square Kilometer Array



2020	2030
5.000	100.000

Number in PBytes/Day

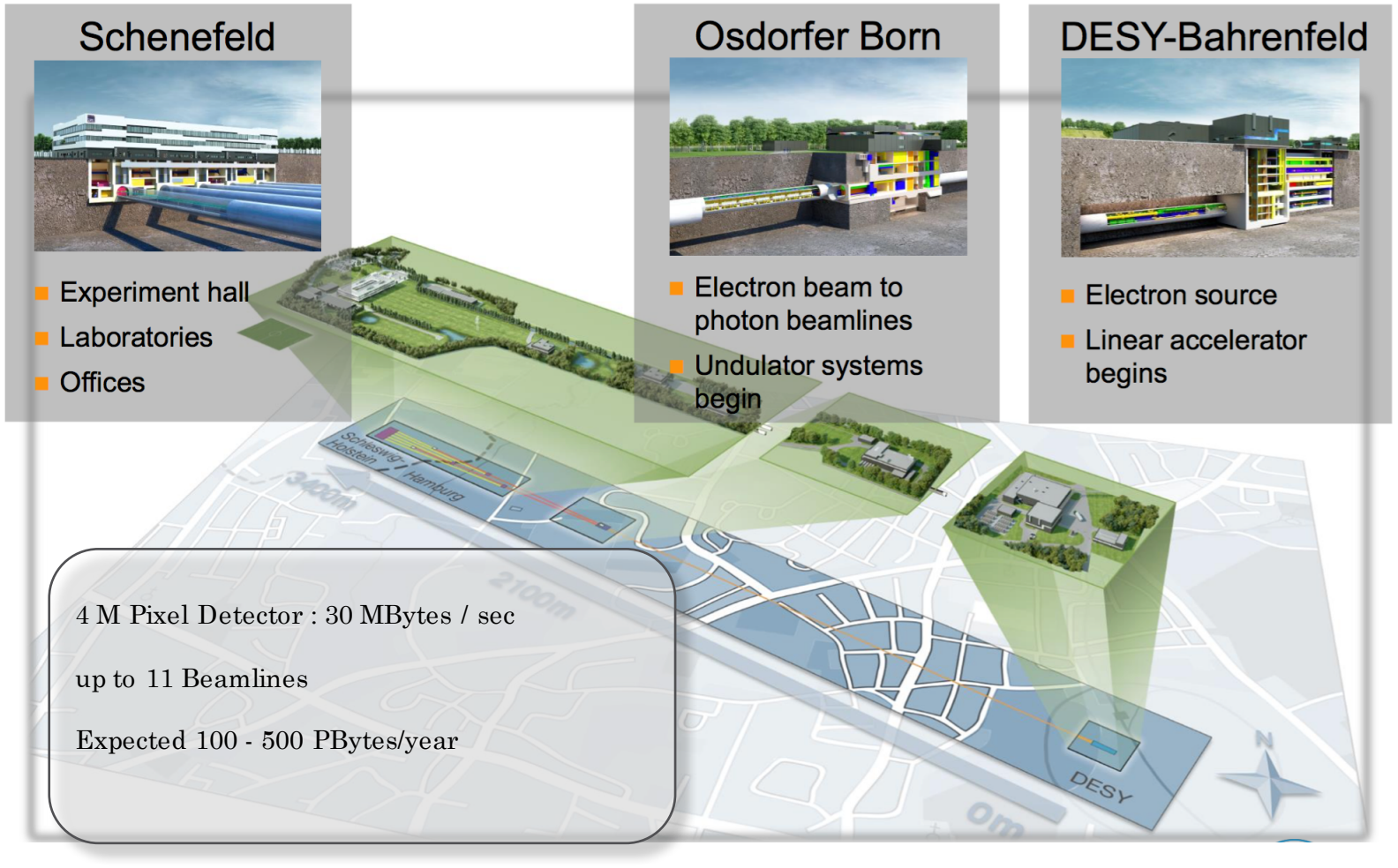




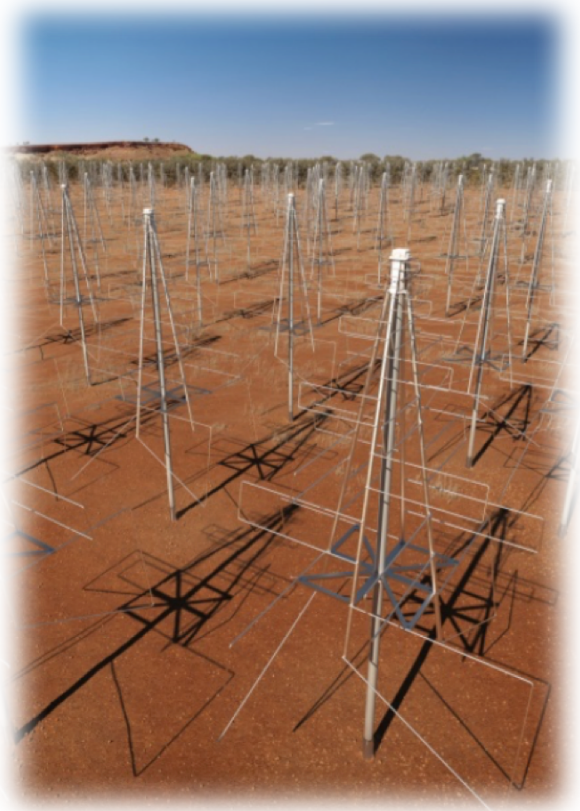
2020	2030
1 - 10	3.000 - 10.000

Stolen from the SKA P. Alexander

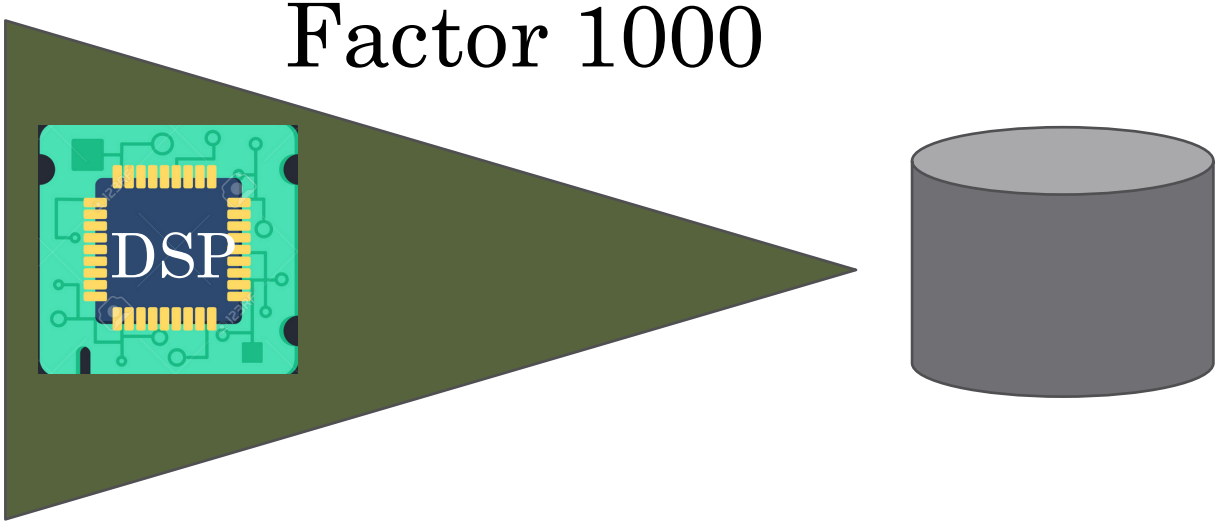
The European XFEL



Data Injection



Factor 1000



Experiment Specific :

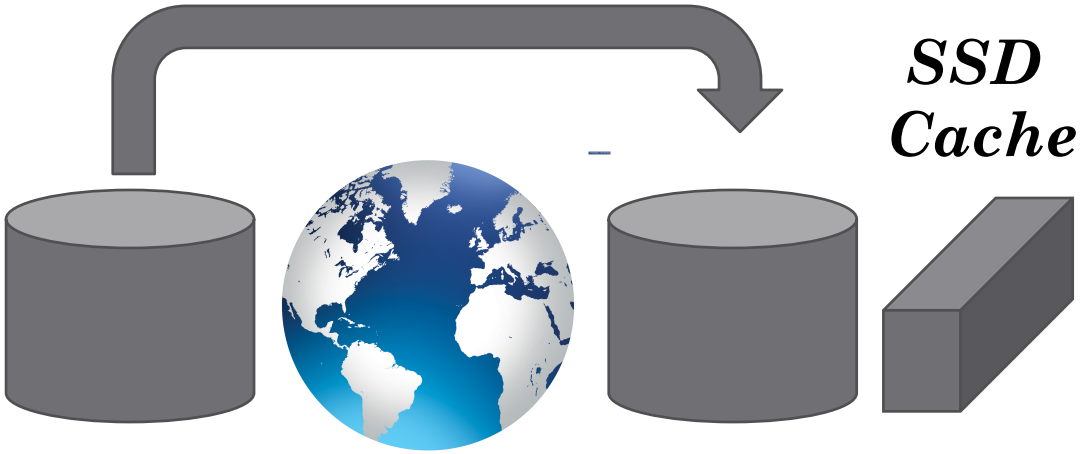
- Here the key-phrase is ”Smart Algorithm”
- In some fields you should be very careful rejecting events.
- It’s a difficult issue but not really our problem.

Next up …

Get data back for processing (analysis)

Current Data Access Structure

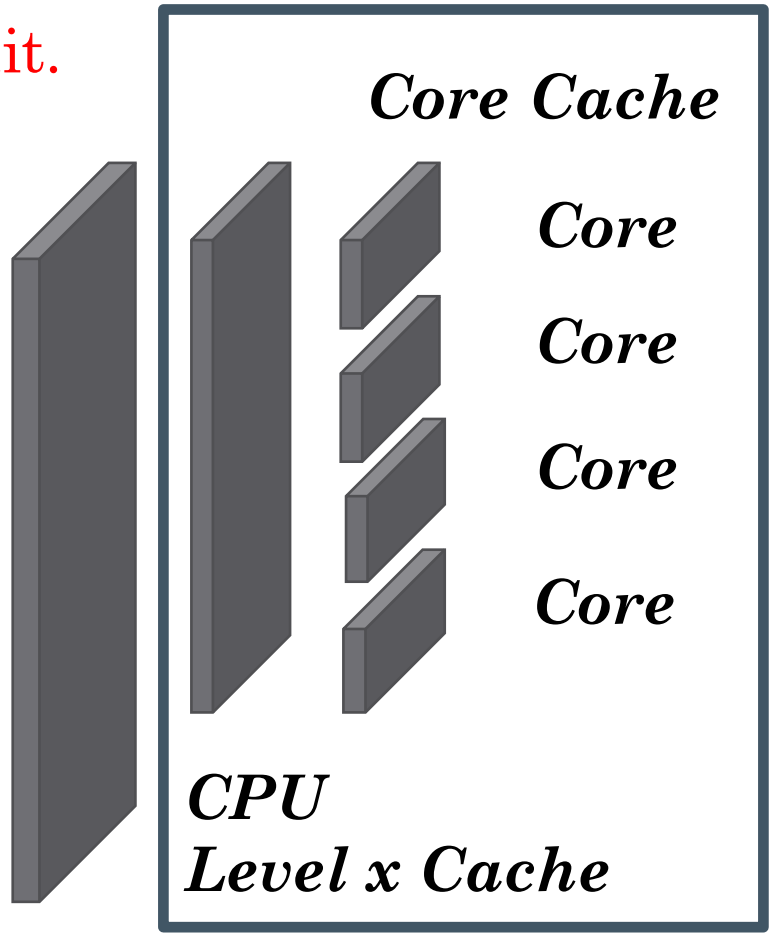
Super Bottleneck: Process power is expensive, don't let your CPU/core wait.



Distributed Storage

SSD Cache

Files System Cache



Things to look into

- Predictive engine for smart data placement (caching)
  - Allow platform layer (experiment framework) to steer data location before processing.
  - Use deep learning to predict access pattern.
- Use vendor provided API's for SSD
  - Move data from spinning devices to SSD (Flash), as you application knows when data is needed.
- Consider circumvent file system cache.
  - I you know what you are doing.
- Consider HADOOOP/Sparc approach

Things to look into (continued)

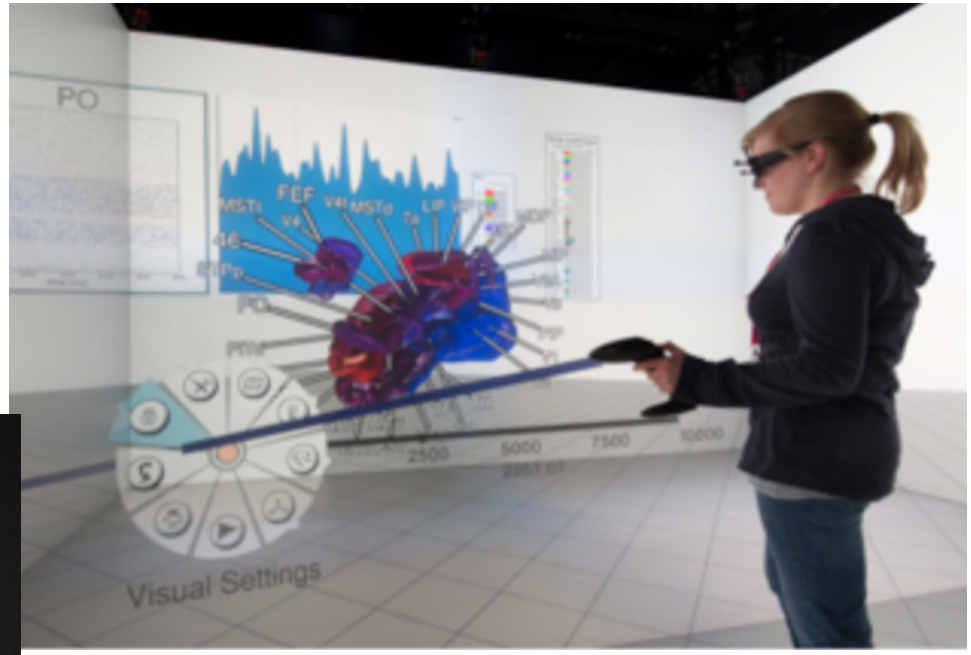
- Simplify access layer (avoid name space lookups)
  - Option : Object stores.
  - Skips name space lookups, client calculates the location of the data itself.
  - Experiment frameworks store IDs anyway.
- Application Software
  - Improve algorithms
  - Teach 'best practice programming'
  - Learn to 'parallel programming'
  - Port old applications properly.

Sensitive Application: Visualization

Latency become a major issue.

After milliseconds delay, viewer gets nervous or runs against a wall.

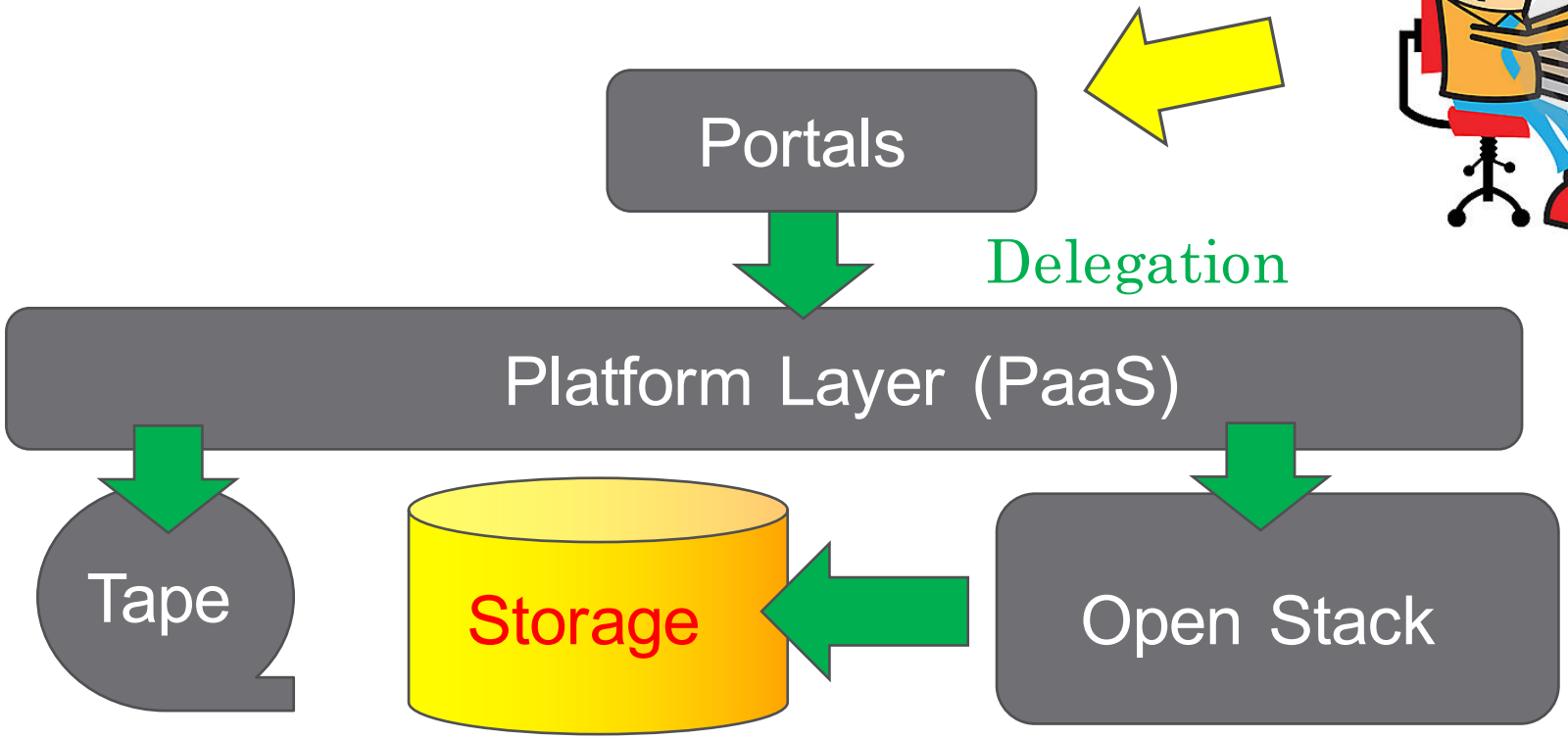
Oculus Rift



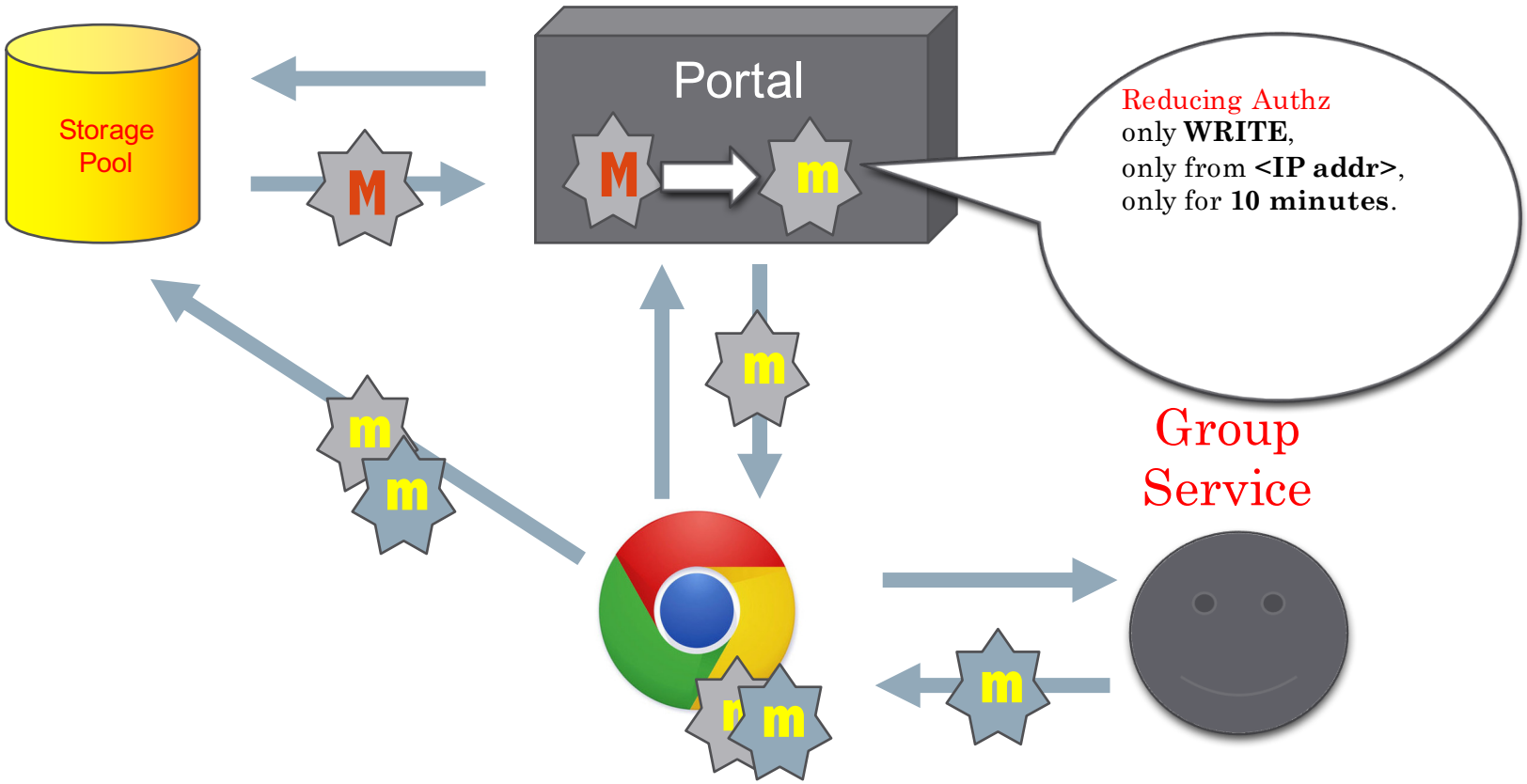
Jülich Aachen Research Alliance, JARA 3D cave

The GPU has to get data from disk fast enough to keep objects moving smoothly (Like here the details of the Human Brain)

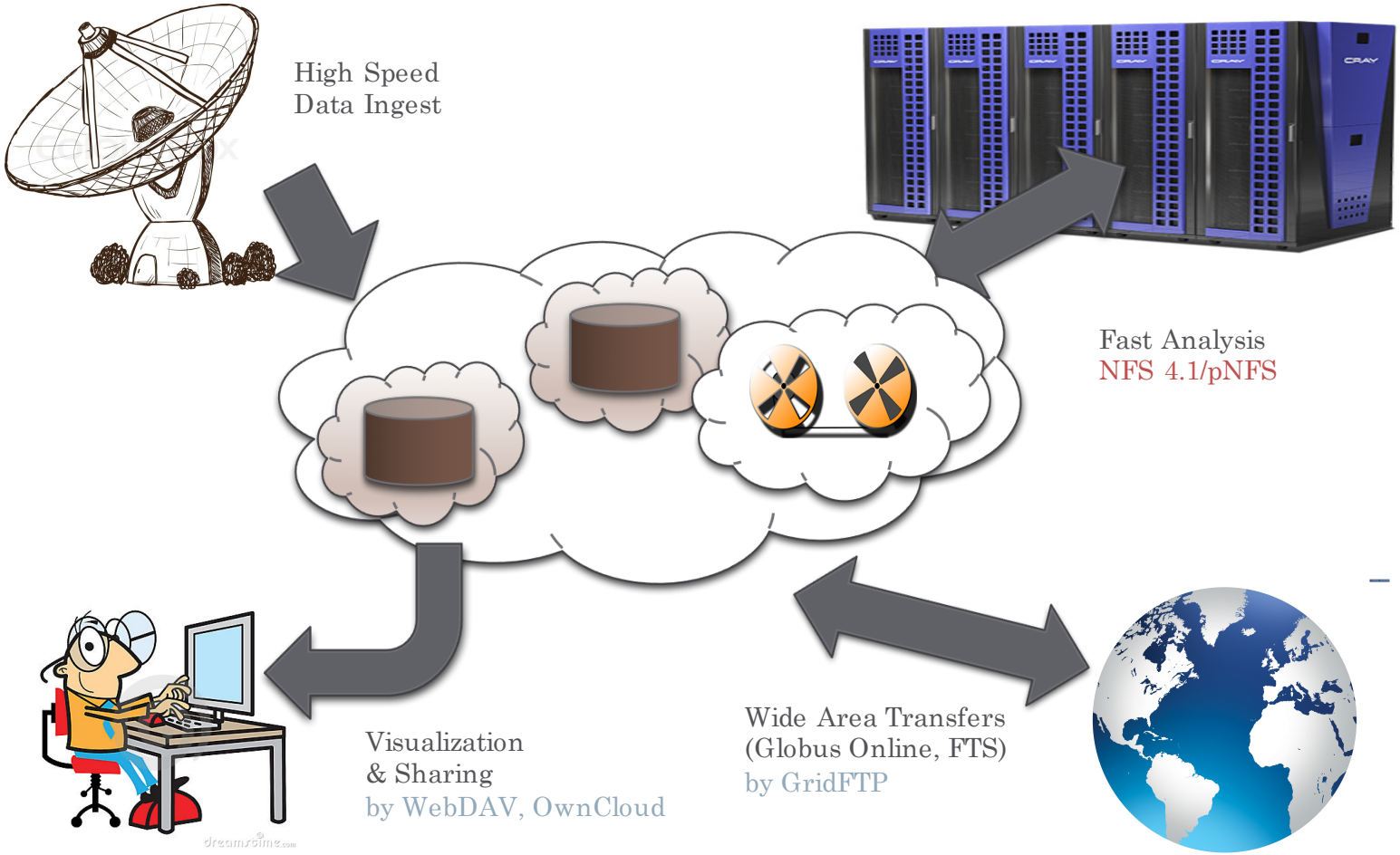
Accessing data by the platform layer



The delegation “Macaroon”



Orchestrating Storage Resources or “The art of Quality of Service” in storage.



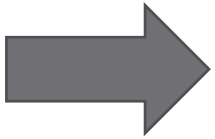
Unfortunately there is no standard protocol resp. API (any more) defining changes in Quality of Service in storage.

It would be needed.

Coming back to your little friend ….

Scientists need to share data .....

chown
chmod
acl set
/afs/cern.ch/home



- Solutions available
  - OwnCloud, NextCloud, Seafile, PowerFolder, ***
- Agreeing on cross platform sharing
  - In progress
- Difficult for Scientific Systems
  - First attempts by EOS and dCache.

Long Term Data Preservation

- Bit file preservation
  - Easiest Case
  - Still has issues : Cold data gets silently corrupted.
- Encrypting your data
  - General encryption mechanisms only last for some year.
  - Companies are offering long term security with distributed systems to reduce the possibility of data getting stolen.
  - Who should hold the master key.
  - Best would be one time pad ☺. Difficult to handle the keys.
- Content preservation.
  - No general solution found yet.
  - DPHEP initiative tries to coordinate some efforts in HEP

Long Term Data Preservation (cont.)

- Legal issue (boring)
  - Ownership question after PI leaves.
  - When should data be made public ?
  - How verifies agreed rules ?

Ending with a sentence Markus Schulz sent as a first reaction on my request for input to this presentation :

There is hardy any field in ”Storage” and “Data Management” which isn’t a challenge.

Enjoy coffee ….