Challenges in Storage

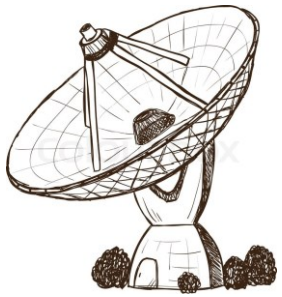
or : A random walk

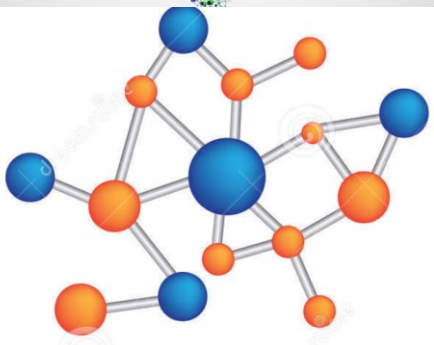
Presented by Patrick Fuhrmann with contributions from many experts.

Contributions and thoughts provided by

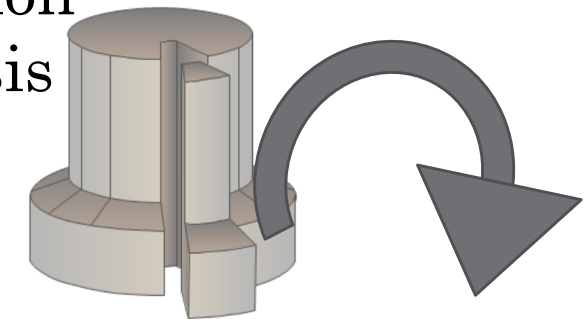
- Oxana Smirnova, NeIC, Lund
- Markus Schulz, CERN IT
- Steven Newhouse, EBI (EMBL)
- Eygene Ryabinkin, KI
- Material from:
  - Martin Gasthuber at al., DESY for XFEL
  - Paul Alexander, SKA
  - Daniele Cecini, CNAF, INFN
  - Hermann Hessling, HTW Berlin
  - Ian Bird, WLCG

My simplified idea of data life cycle

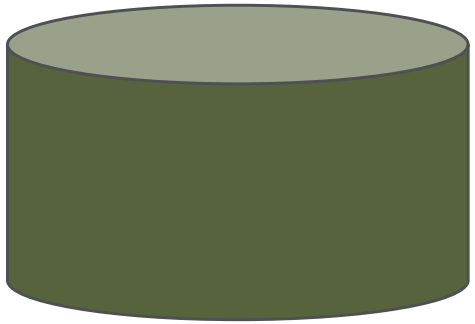




Reconstruction and Analysis



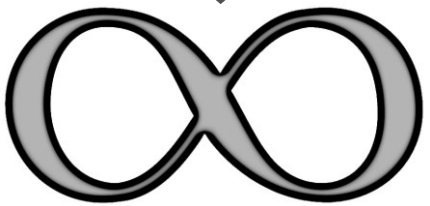
Reduction Algorithms



Long Term Archive

<Meta data>

Visualization Manipulation



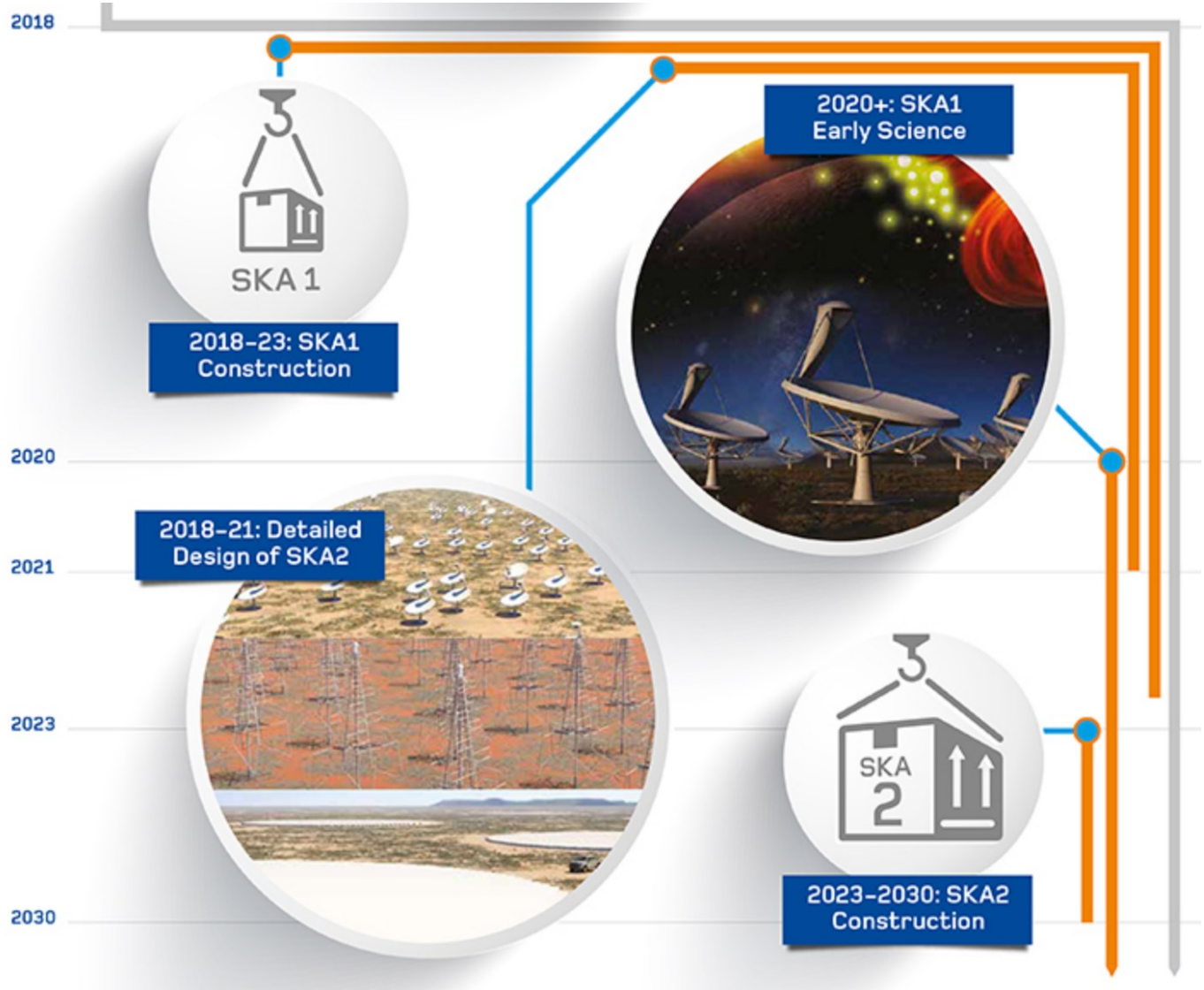
e.g. the Location of a nuclear waste facility.

What is the order of magnitude we are talking about ?

Let’s pick two arbitrary ‘big data’ experiments.

The Square Kilometer Array

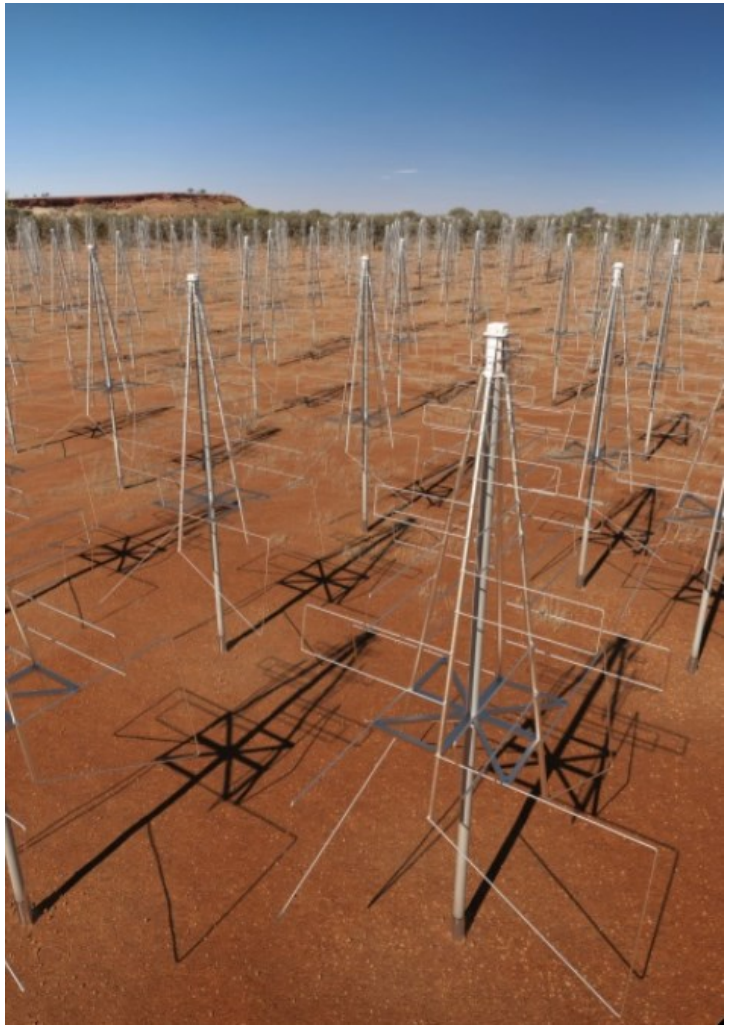
SKA I, 2018

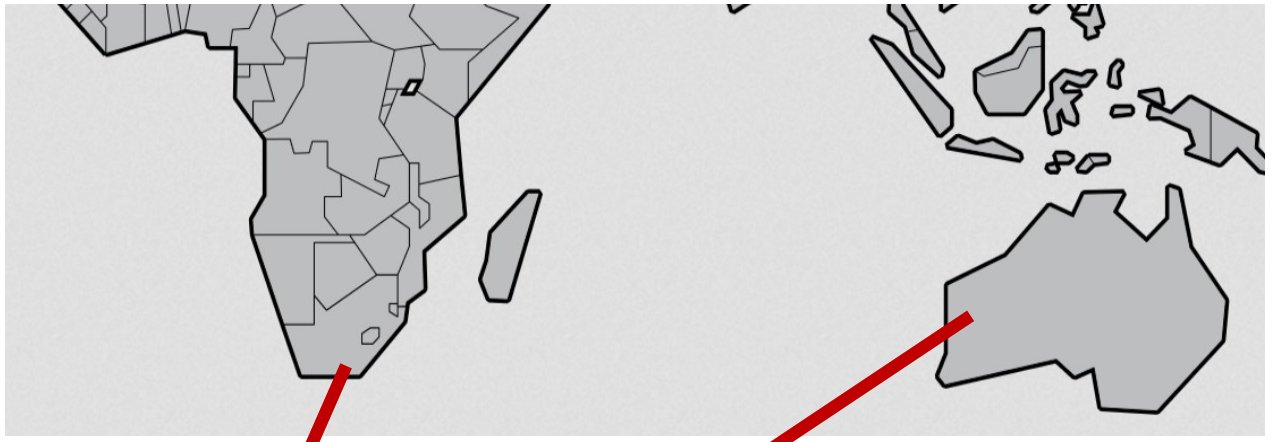


SKA II 2030

Stolen from the SKA Homepage

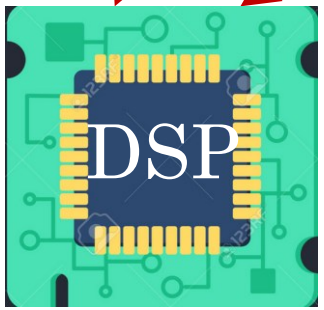
The Square Kilometer Array

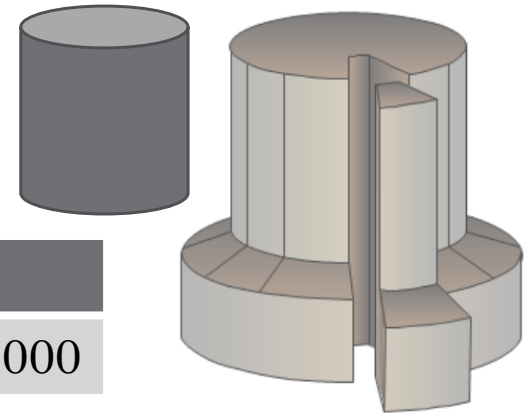




2020	2030
5.000	100.000

Number in PBytes/Day

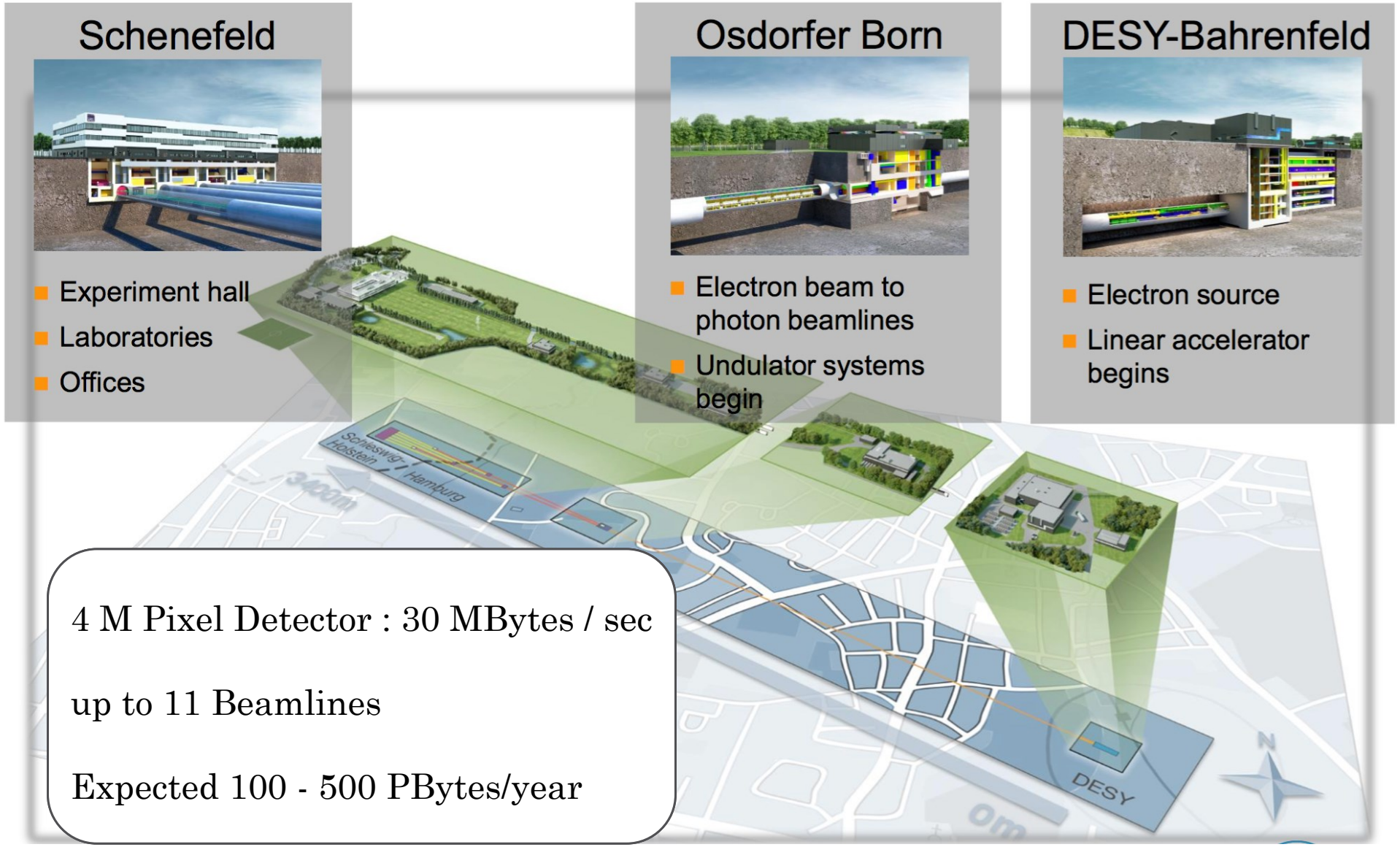




2020	2030
1 - 10	3.000 - 10.000

Stolen from the SKA P. Alexander

The European XFEL

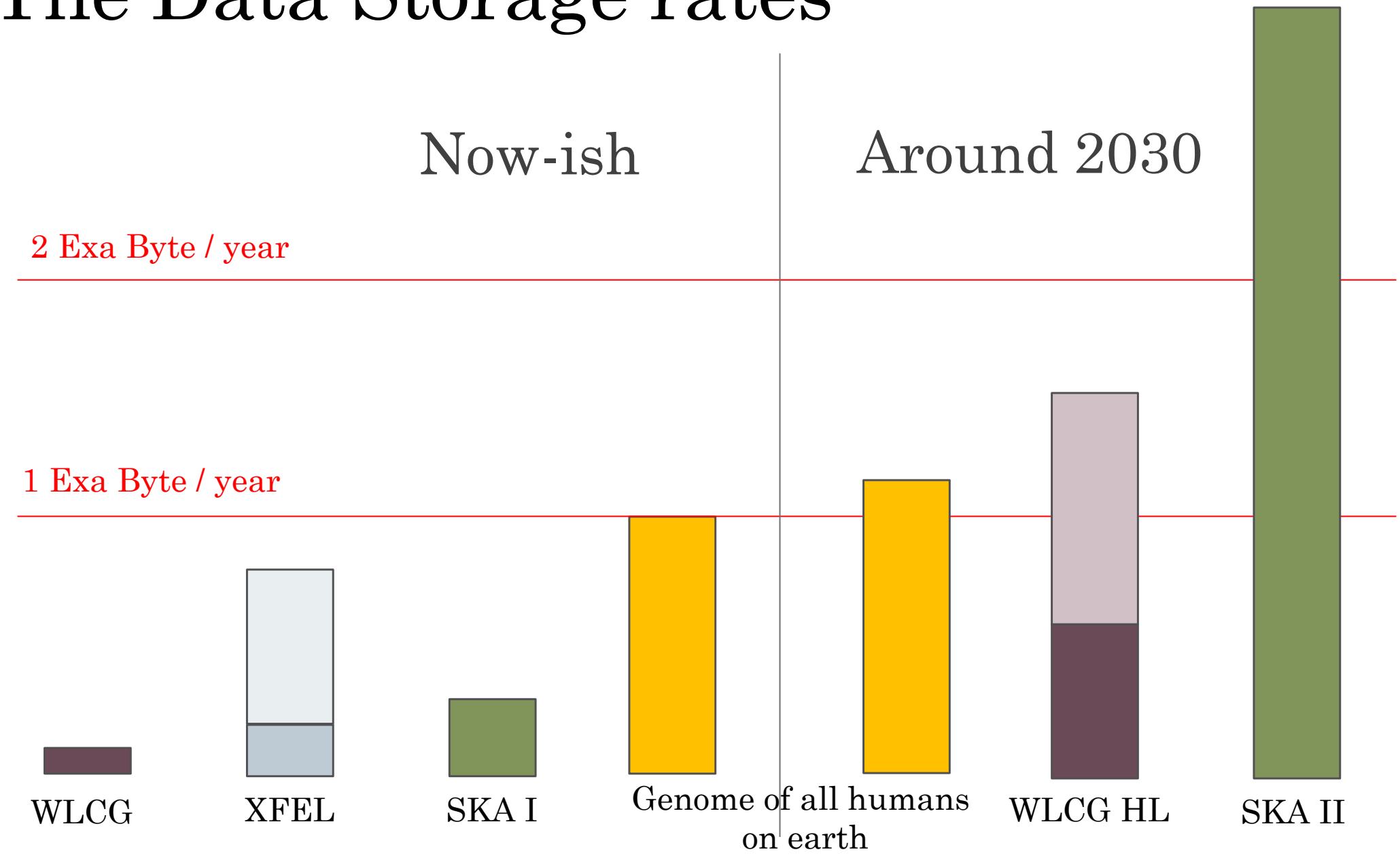


4 M Pixel Detector : 30 MBytes / sec

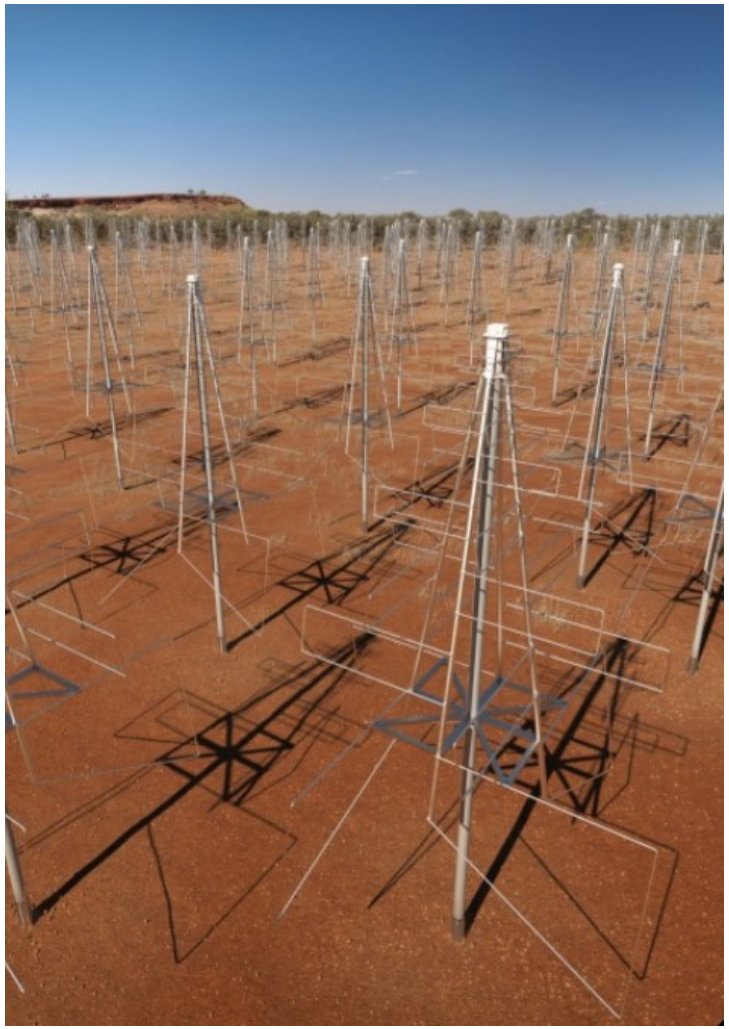
up to 11 Beamlines

Expected 100 - 500 PBytes/year

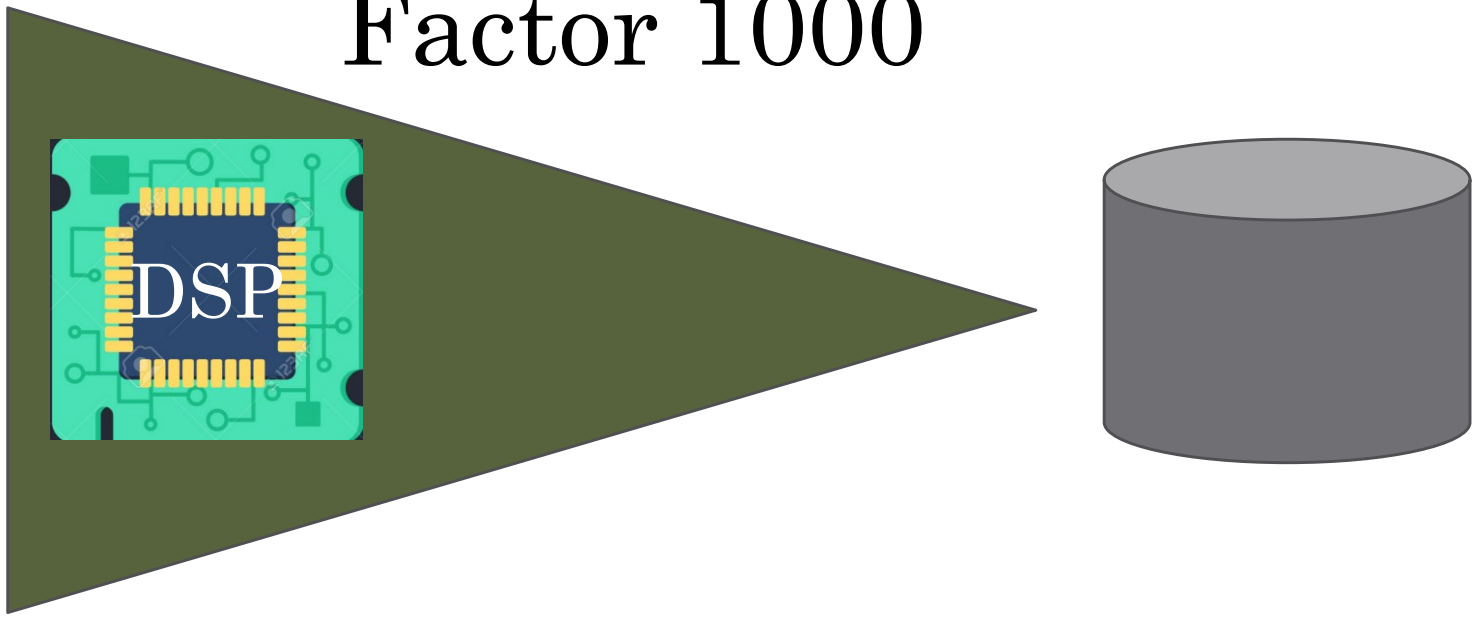
The Data Storage rates



Data Injection



Factor 1000



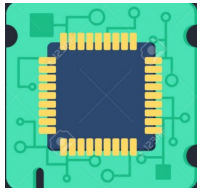
Experiment Specific :

- 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.

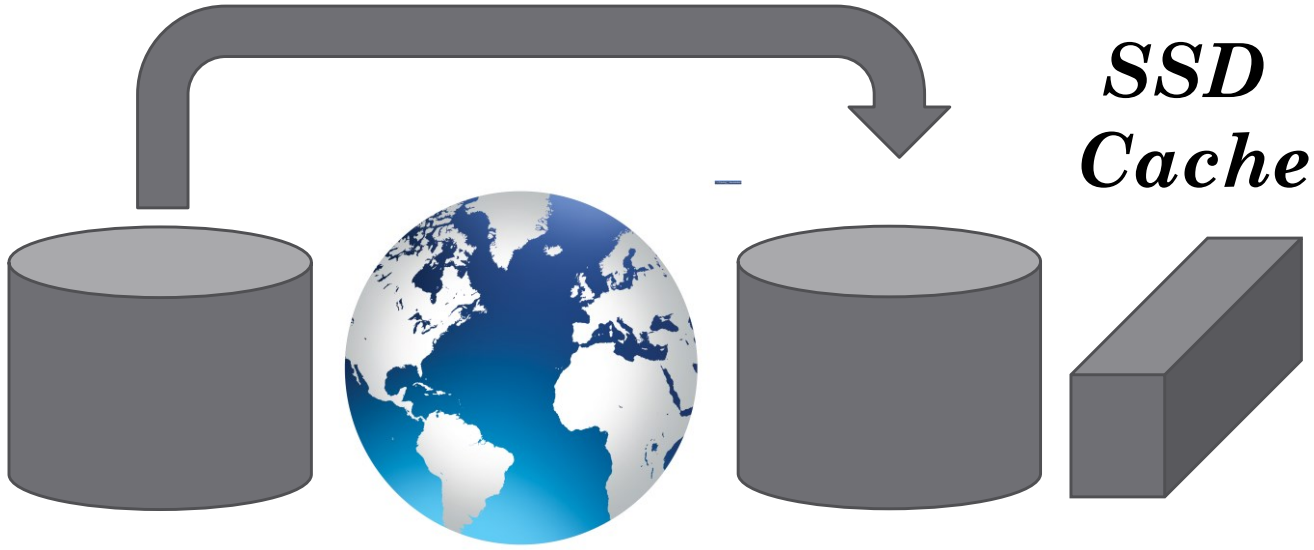
Next up …

Get data back for processing (analysis)

Current Data Access Structure



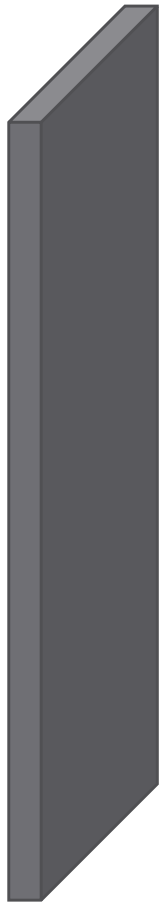
Super Bottleneck: Process power is expensive, don't let your CPU/core wait.



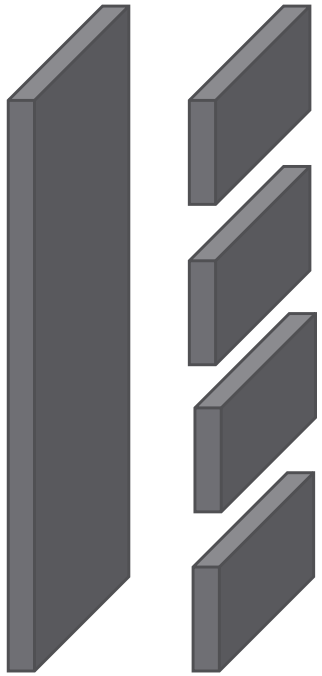
Distributed Storage

SSD Cache

Files System Cache



Core Cache



Core

Core

Core

Core

CPU Level x 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)

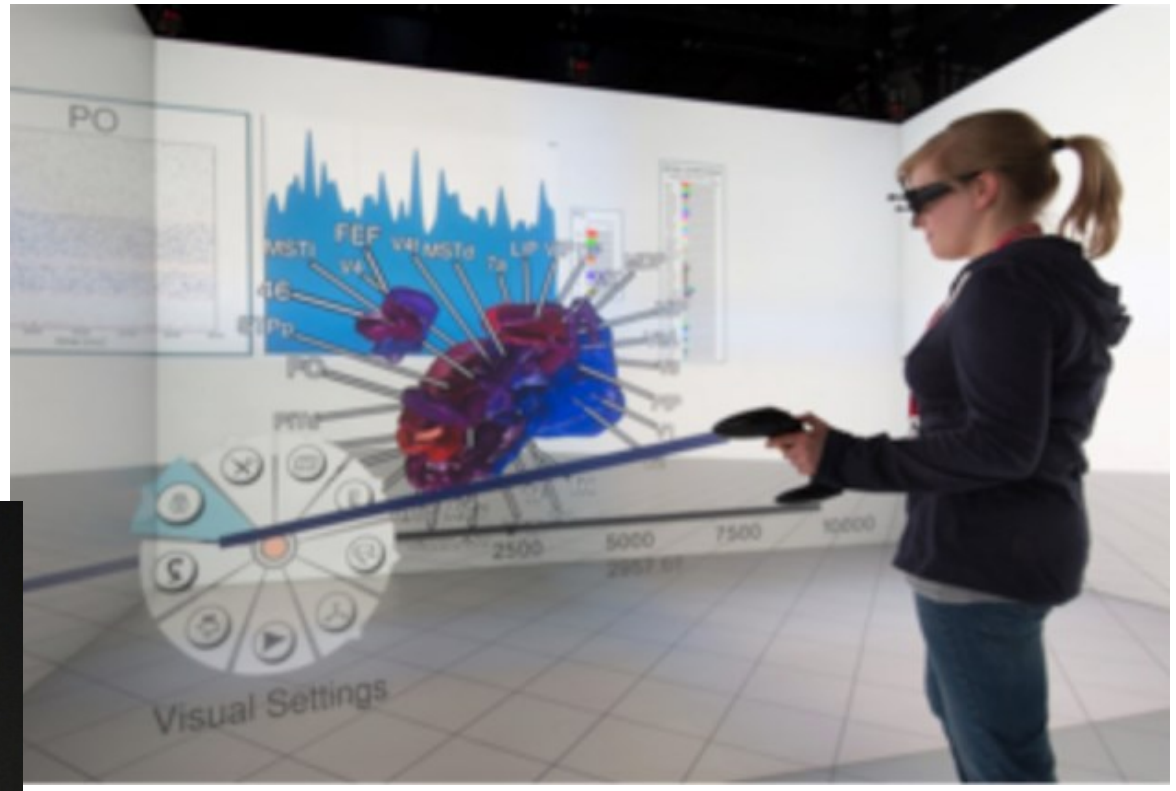
- Simplify access layer (avoid name space lookups)
  - Option : Object stores.
  - Skips name space lookups, client calculates the location of the data itself.
  - Experiment frameworks store IDs anyway.
- Application Software
  - Improve algorithms
  - Teach ’best practice programming’
  - Learn to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

Authentication

Portals

Delegation

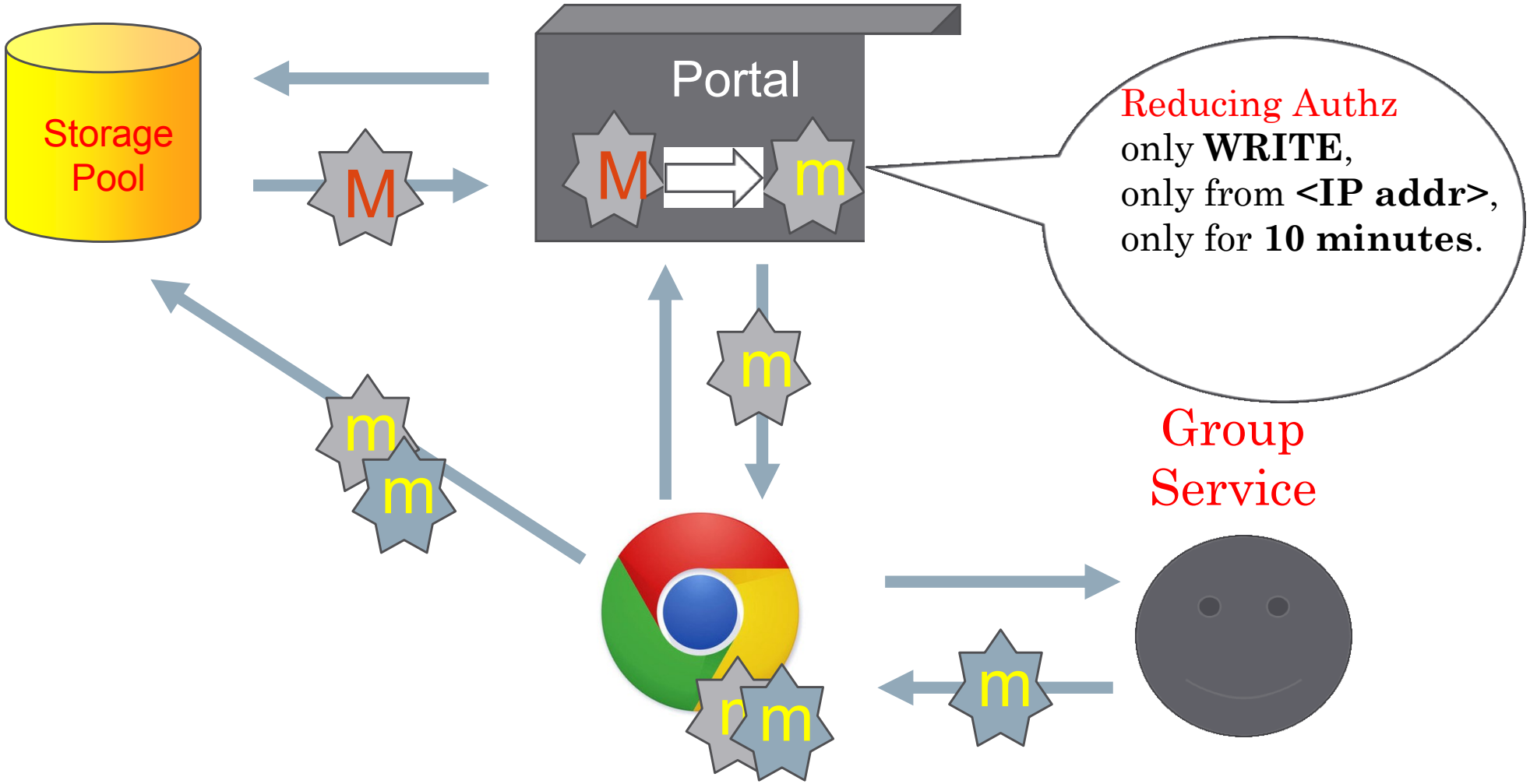
Platform Layer (PaaS)

Tape

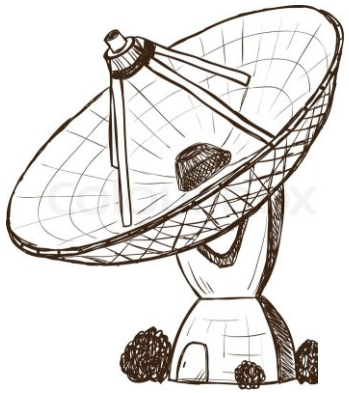
Storage

Open Stack

The delegation “Macaroon”

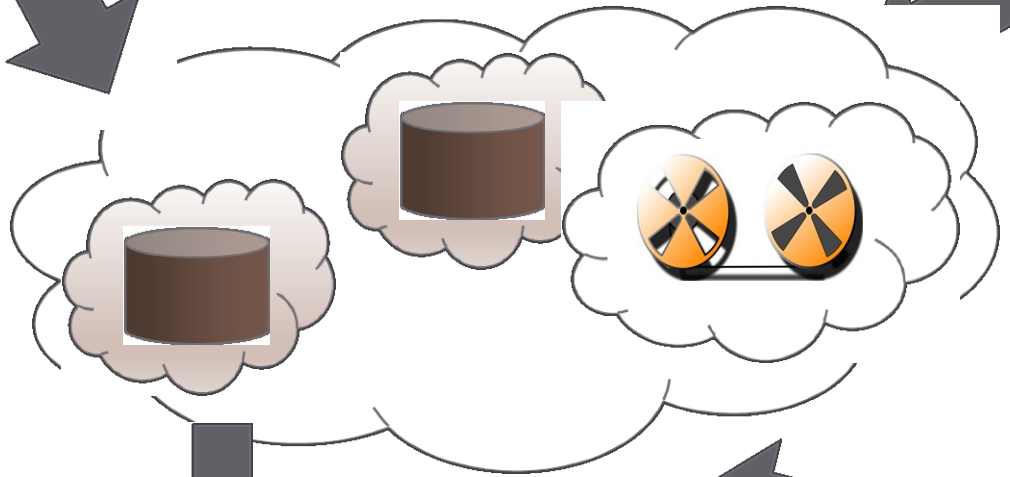


Orchestrating Storage Resources or "The art of Quality of Service" in storage.



High Speed
Data Ingest

Fast Analysis
NFS 4.1/pNFS



Visualization
& Sharing
by WebDAV, OwnCloud

Wide Area Transfers
(Globus Online, FTS)
by GridFTP

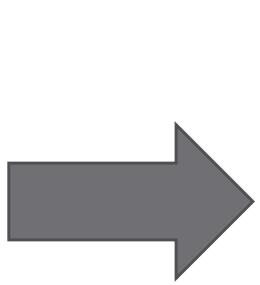
Unfortunately there is no standard protocol resp. API (any more) defining changes in Quality of Service in storage.

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 ….