dCache.org

dCache, a distributed high performance storage system for HPC

ISC 2013

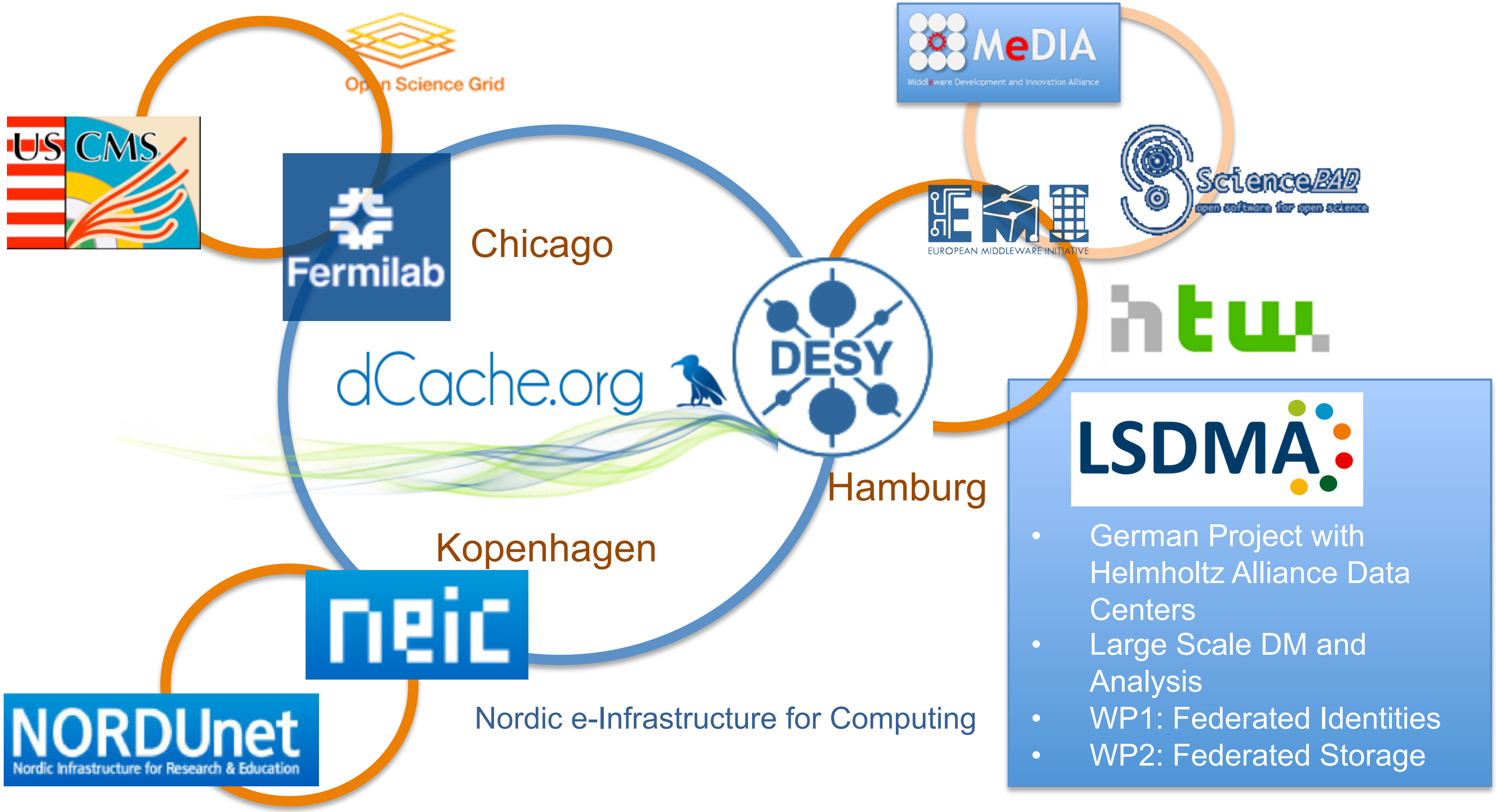
Patrick Fuhrmann

Content

- Who are the dCache people ?
- Where are we coming from ?
  - Some words about WLCG, and others
  - Some dCache deployments
- Where do we want to go and why ?
  - HTC to HPC
  - HPC for our current customers
- But there is more than just performance.
  - Multi Tier storage model
  - Multi Protocol Support
  - Consistent Authentication and Authorization

Who are we ?

dCache is an international collaboration.


Open Science Grid
US CMS
Fermilab
MeDIA
Middleware Development and Innovation Alliance
EMI
EUROPEAN MIDDLEWARE INITIATIVE
ScienceSoft
open software for open science
htw
dCache.org
DESY
LSDMA
neic
NORDUnet
Nordic Infrastructure for Research & Education

Chicago

Hamburg

Kopenhagen

Nordic e-Infrastructure for Computing

- German Project with Helmholtz Alliance Data Centers
- Large Scale DM and Analysis
- WP1: Federated Identities
- WP2: Federated Storage

What do we do ?

We

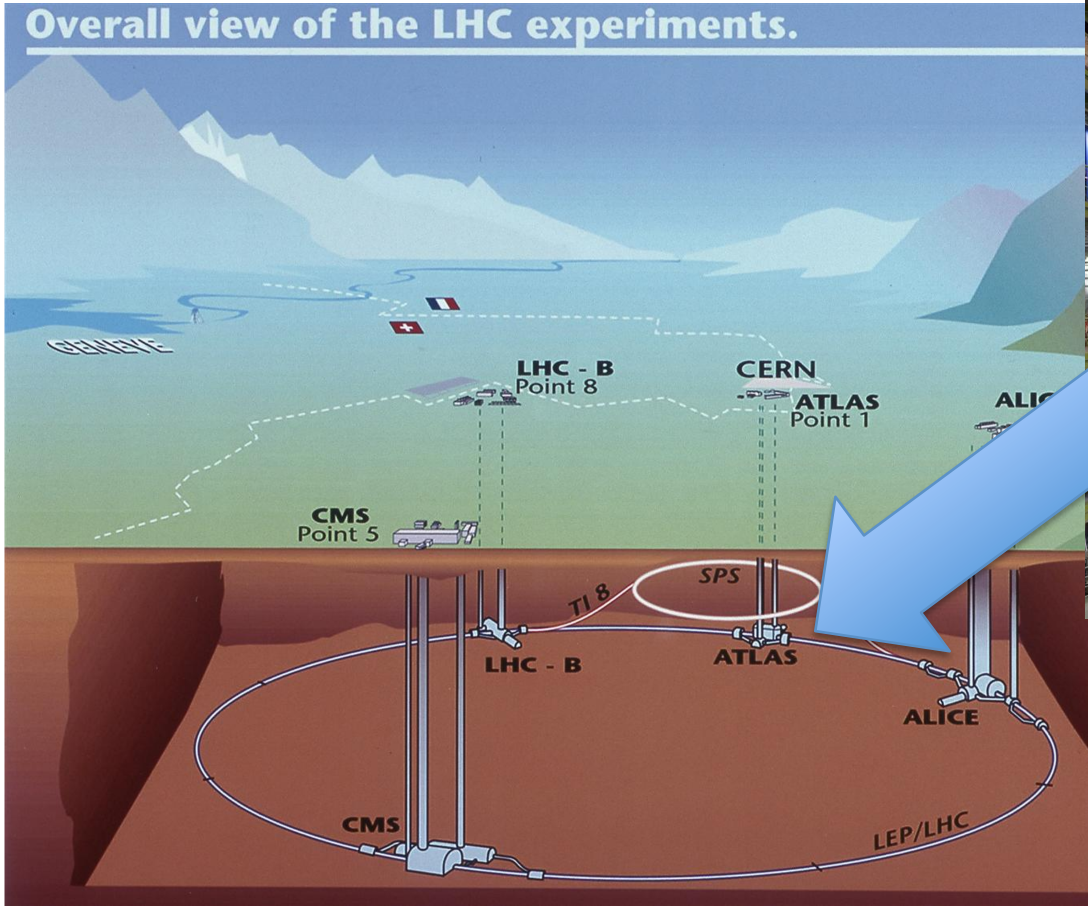
- design
- implement
- and deploy

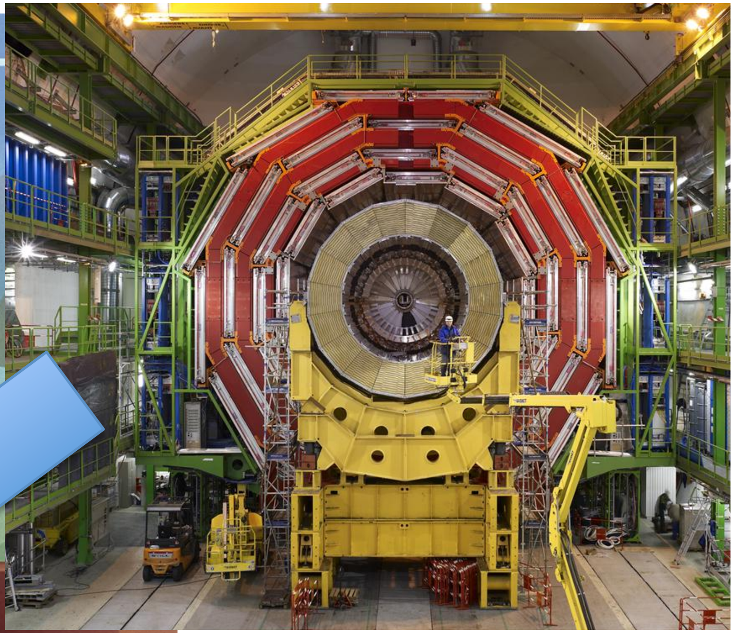
Data Storage and Management software for data intensive communities.

Where are we coming from ?

From
High Energy Physics
HERA and the Tevatron in the past
and now
The Large Hadron Collider in Geneva

High Energy Physics (the sensors)





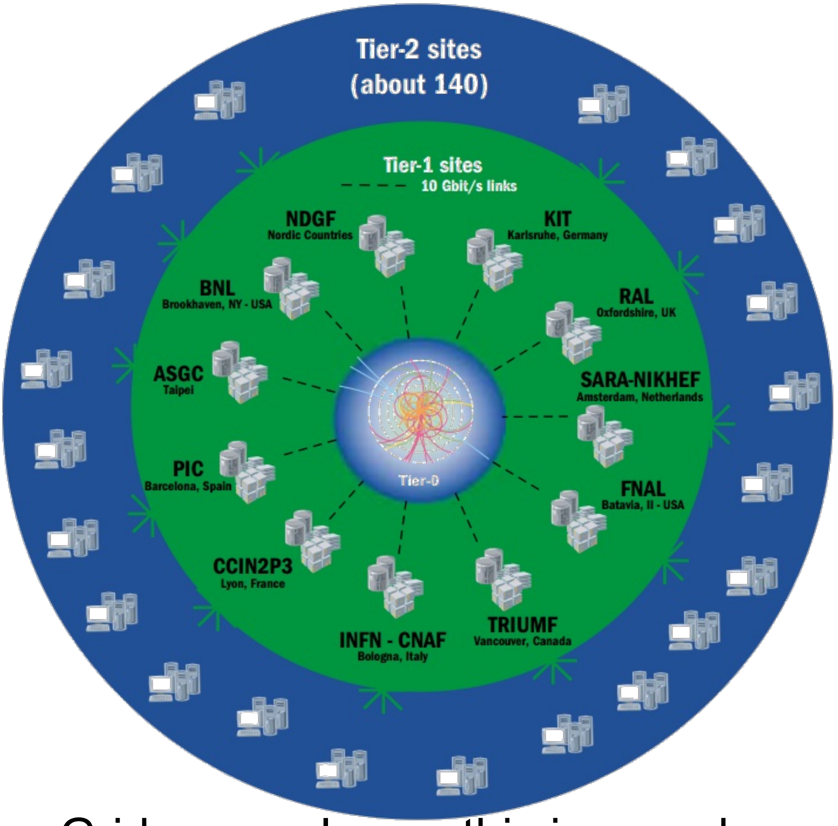
The Atlas detector,
12 m long, 6 in diameter
and 12.000 tonnes

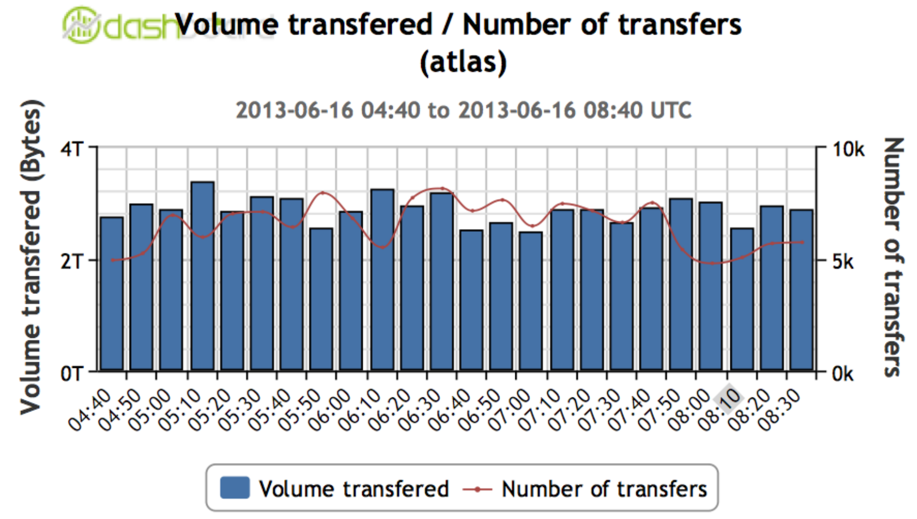
The ring: 27 Km long, -271 degrees cold, some billion Euros and looking for the Higgs and for Dark Matter. Collisions every 25 nsec, filled with 13.000 bunches running with nearly speed of light. The ring needs 120 MW and 50 MW for cooling.

And its computer:

The LHC Computing Grid

High Energy Physics (the computer, GRID)





The Grid never sleeps: this image shows the activity on 1 January 2013, just after midnight, with almost 250,000 jobs running and 14 GiB/sec data transfers.

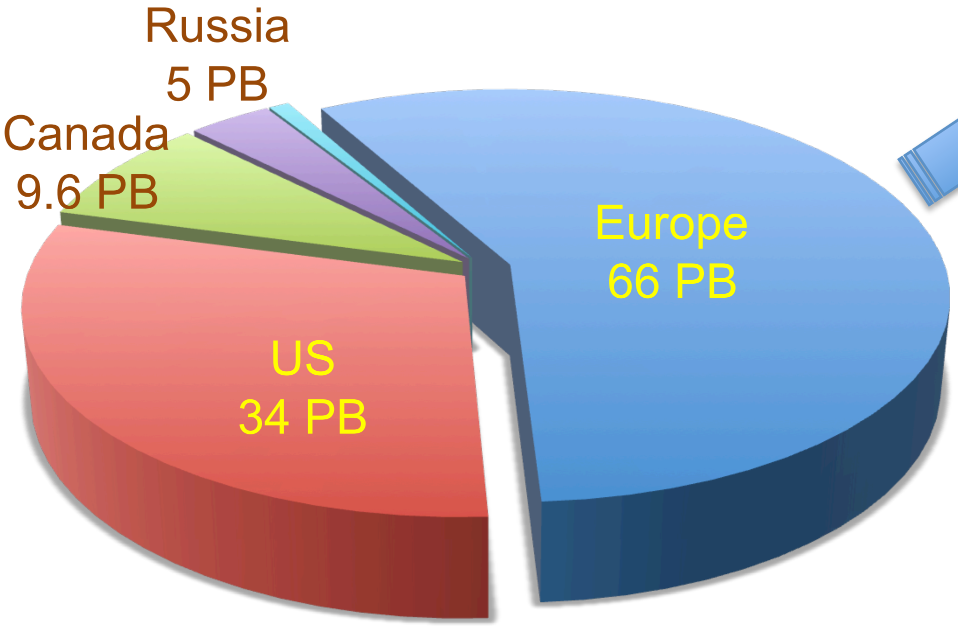
Image courtesy Data SIO, NOAA, US Navy, NGA, GEBCO, Google, U Dept. of State Geographer, GeoBasis, DE/BKG and *April 2013 issue of the CERN Courier*

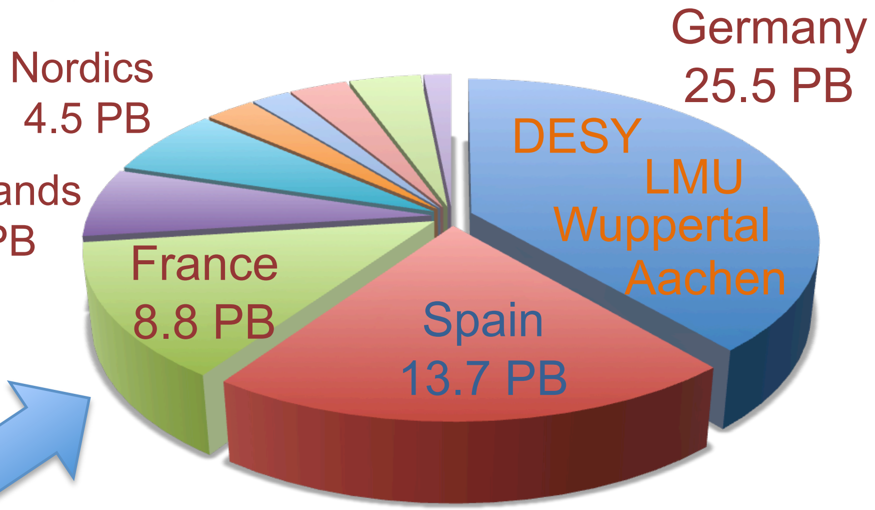
Now, where is dCache ?

We do ½ of their storage
So we have 50% of their famous Higgs ☺

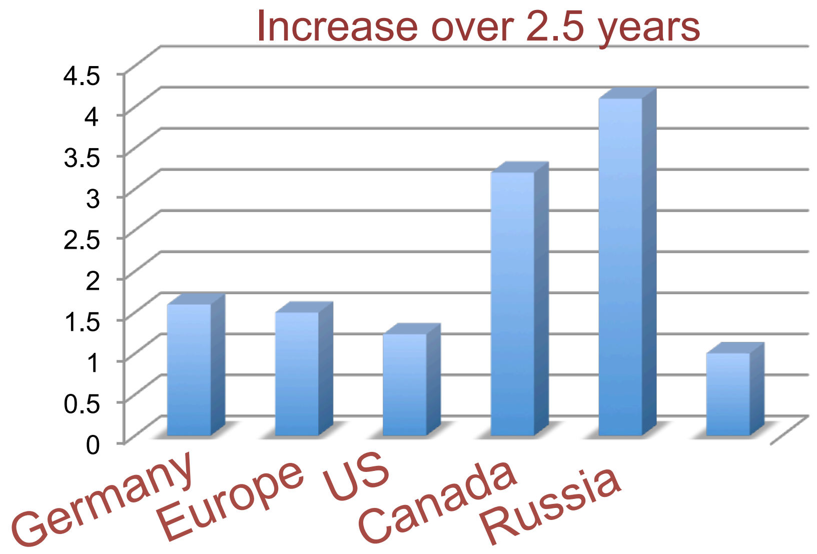
dCache storage for the Large Hadron Collider around the world

About 115 PBytes only for WLCG in 8(+2) out of 11(+3) Tier 1 centers and about 60 Tier 2's, which is about ½ of the entire WLCG data.





Netherlands
4.5 PB



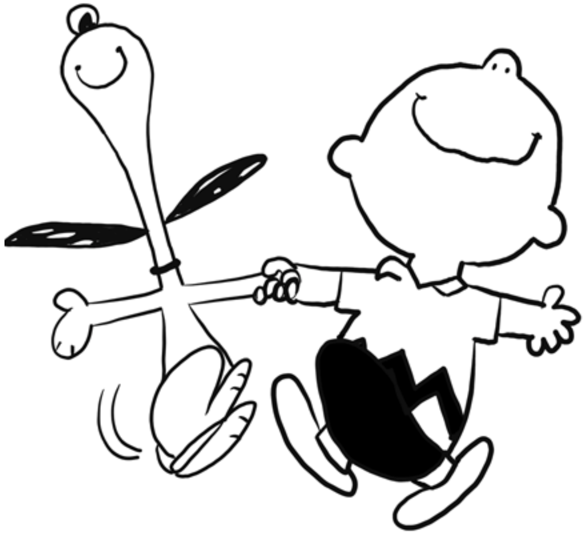
But there are more …

Other customers

European XFEL
ilc
USQCD
EMBL
LOFAR
SNIC
HASYLAB
Mu2e
SLOAN DIGITAL SKY SURVEY
SciBooNE
ICECUBE
SOUTH POLE NEUTRINO DETECTOR
INRA
SCIENCE & IMPACT

THE FRENCH NATIONAL INSTITUTE FOR AGRICULTURAL RESEARCH

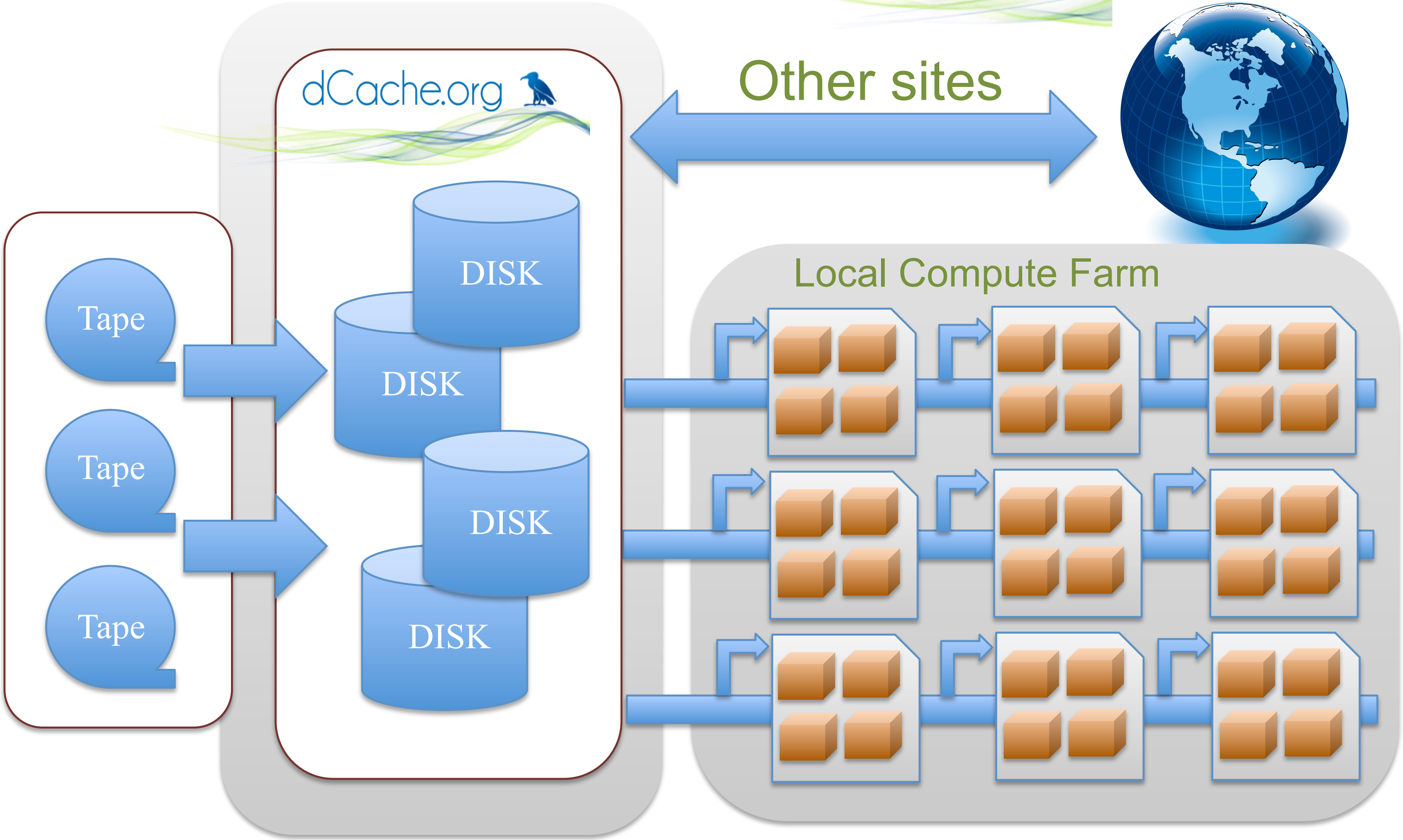
And how do we do this ?



With joy … and

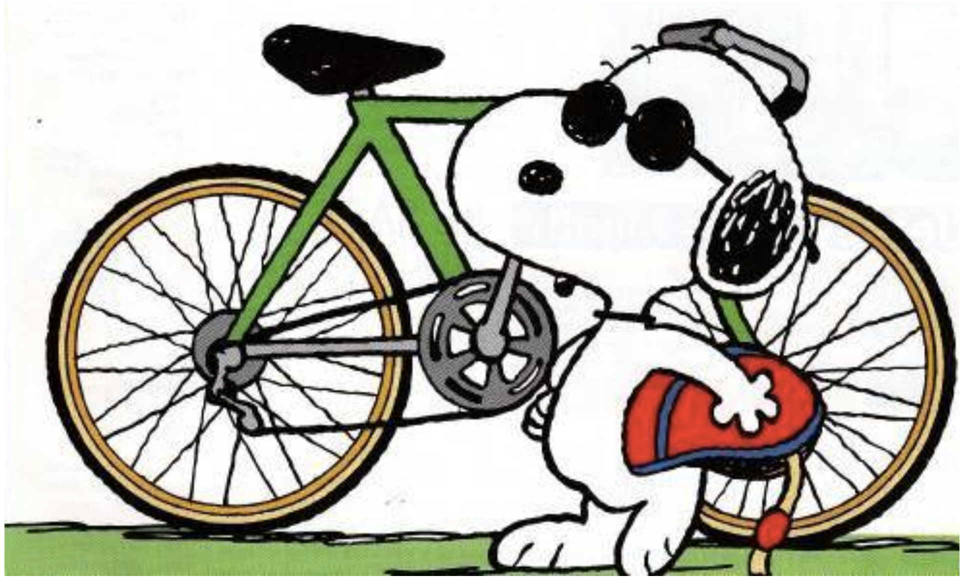
LHC Computing Storage Element dCache.org

medium single stream performance



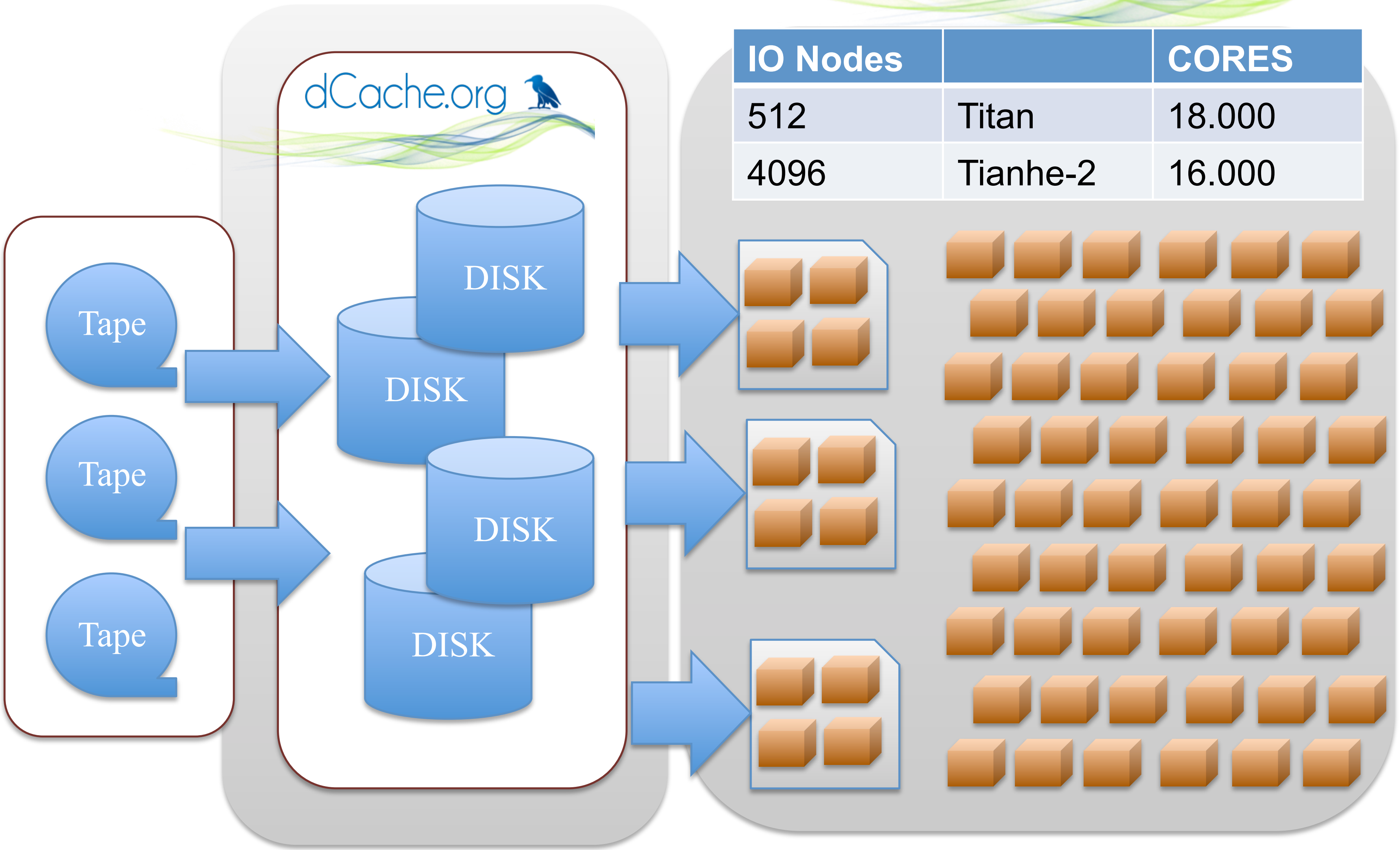
This is quite nice but getting a bit boring ….

So, where do we want to go ?



HPC Computing

Possibly high single stream performance

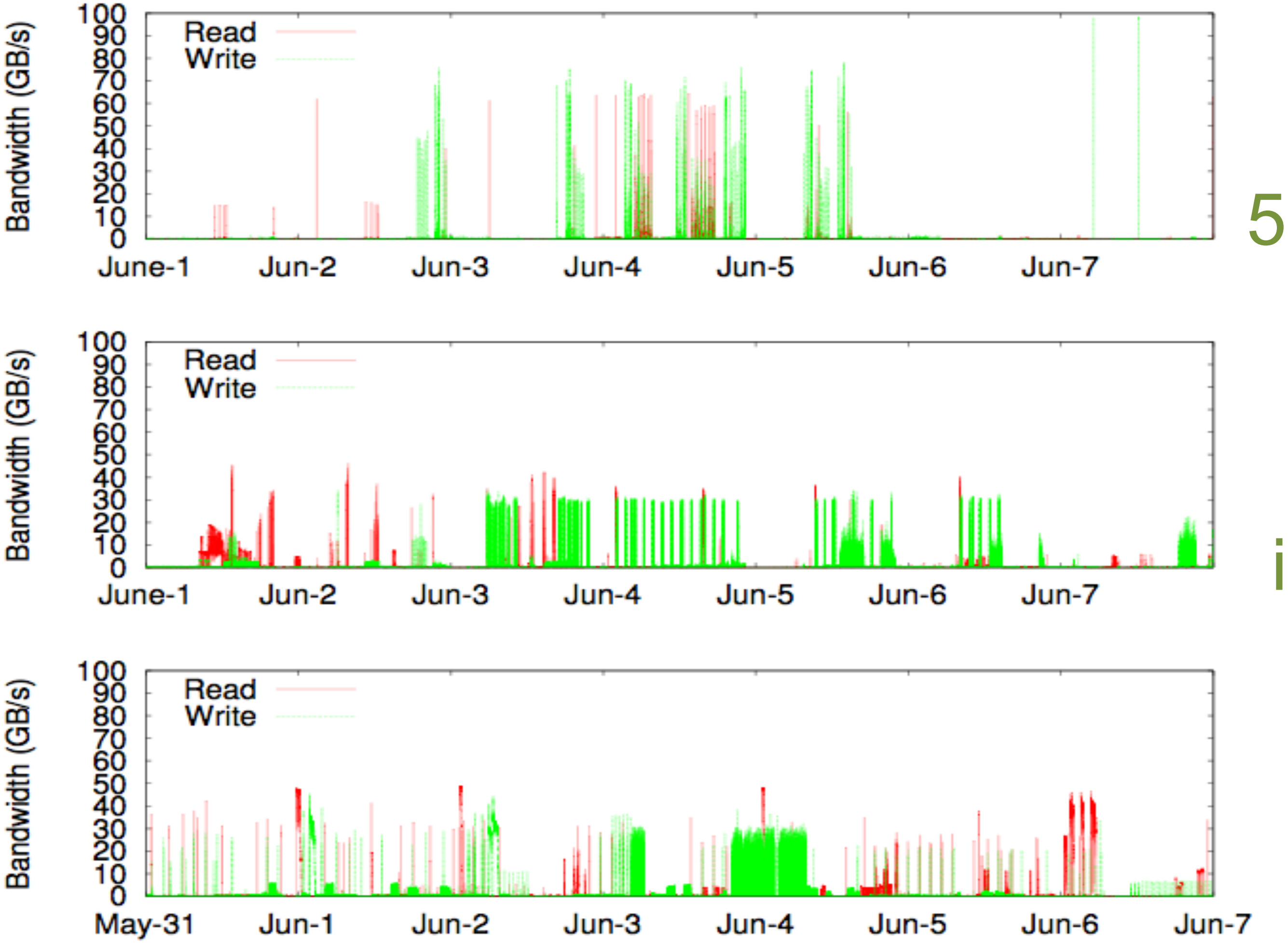


IO Nodes		CORES
512	Titan	18.000
4096	Tianhe-2	16.000

Having a look into real HPC performance numbers.

Real Data Rates for HPC

Three file system partitions in front of the Titan



Equivalent to 5000 and 10000 Cores per File system in GRID Terms

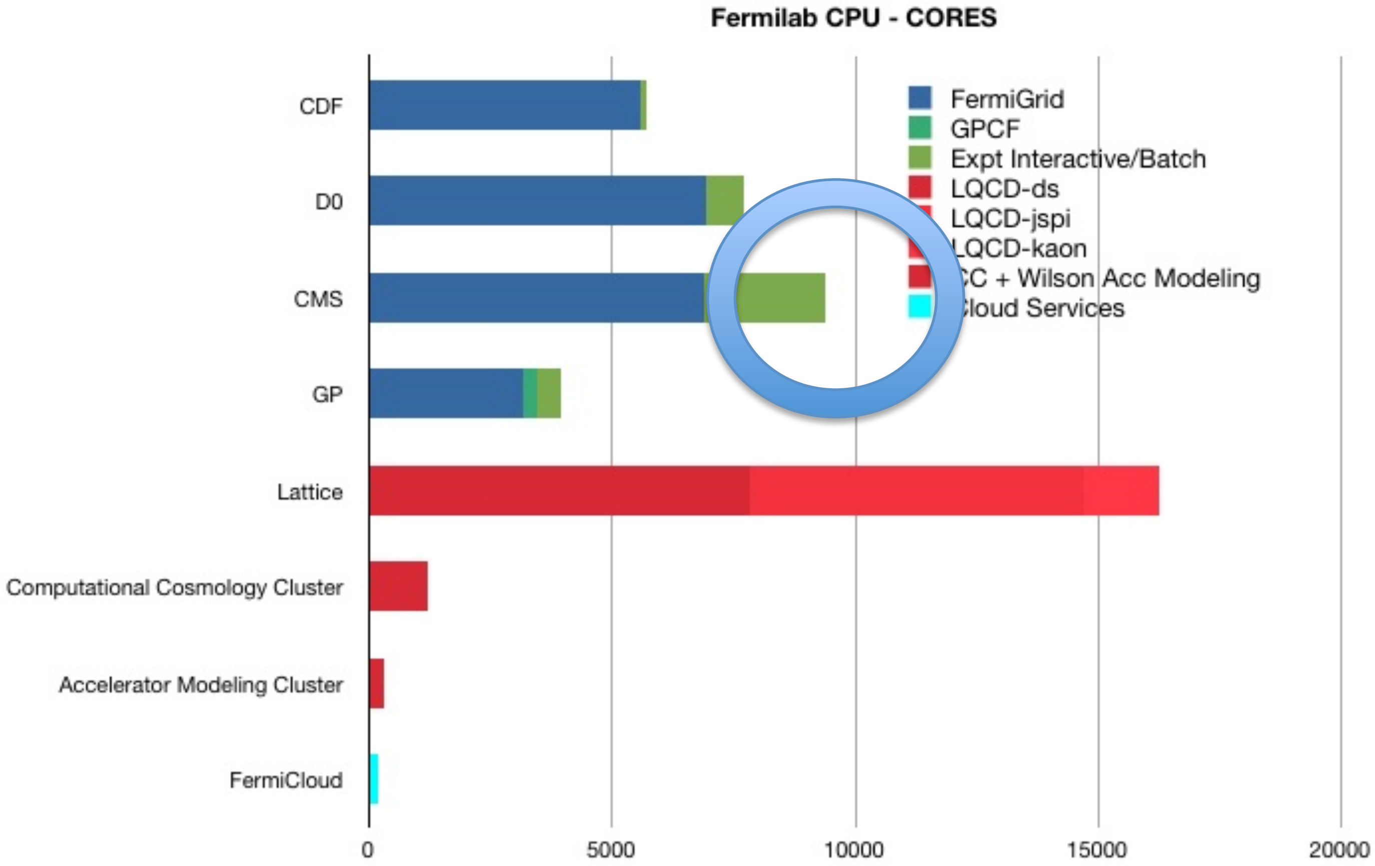
Each GRID core consumes about 6Mbytes/s

Courtesy: Oak Ridge National Lab, Spider FS Project

So the question arises …

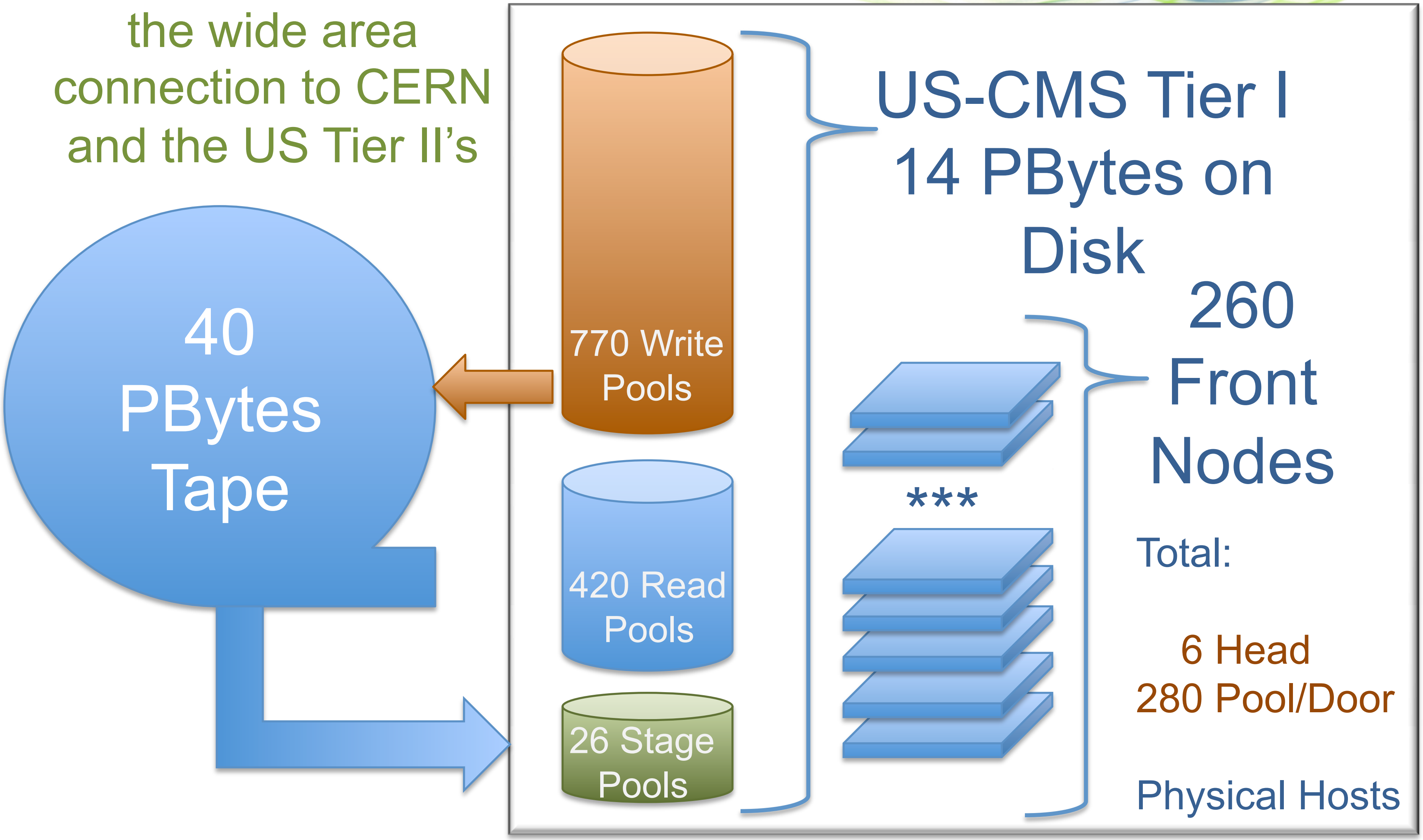
Can dCache do this in a single instance ?

Core Count of FERMIlab



Courtesy: Vicky White, Institutional Review 2011

US CMS dCache Setup to serve the farm and the wide area connection to CERN and the US Tier II's



Information provided by Catalin Dumitrescu and Dmitry Litvintsev

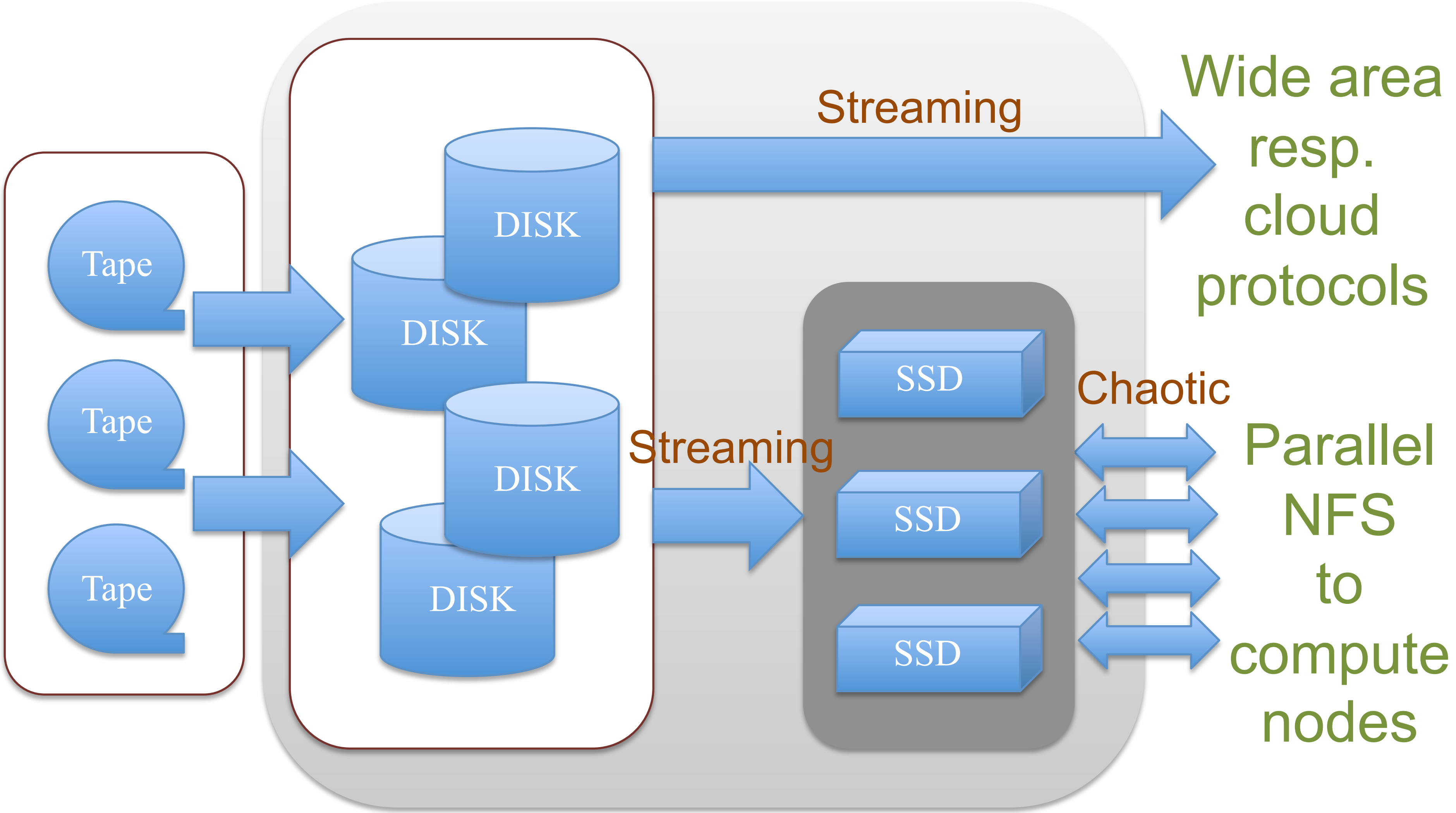
As network and spinning disks are becoming the bottleneck, we can even do better …

Or

Using Multi-Tier Storage

Multi Tier Storage

dCache.org

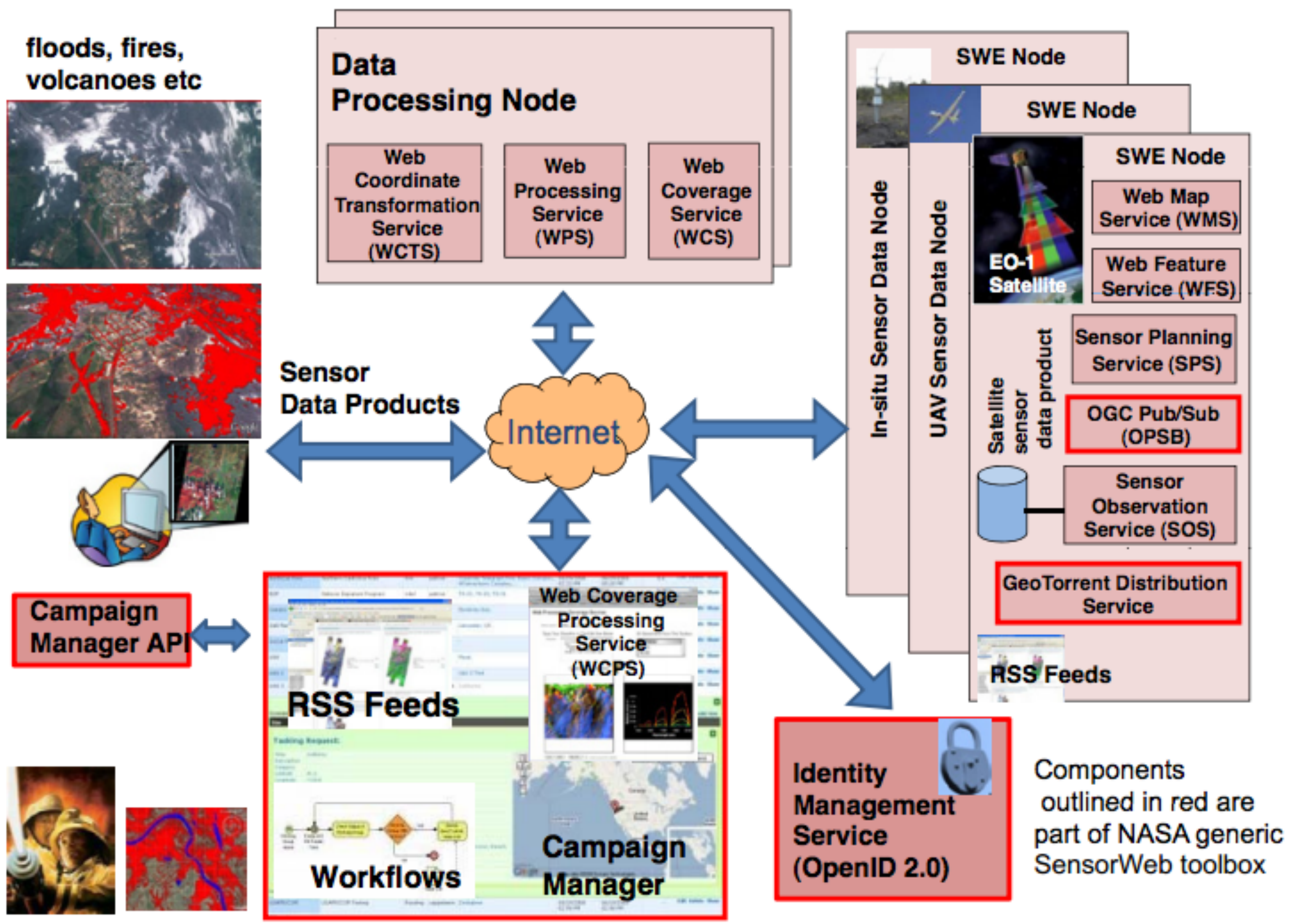


Why do we want to go HPC ?

- The LHC experiment (e.g. ATLAS) are seriously looking into HPC. They would like to utilize free resources in HPC worldwide. Feasibility evaluations are ongoing. If they decide to go for it, they need Grid Storage Elements to ensure access to their worldwide data federation.
- The HPC community begins to share data. Right now this is still all manual. But they could learn from the LHC Grid. We share and transfer data automatically for about a decade, including proper authentication and authorization at the source and endpoints.

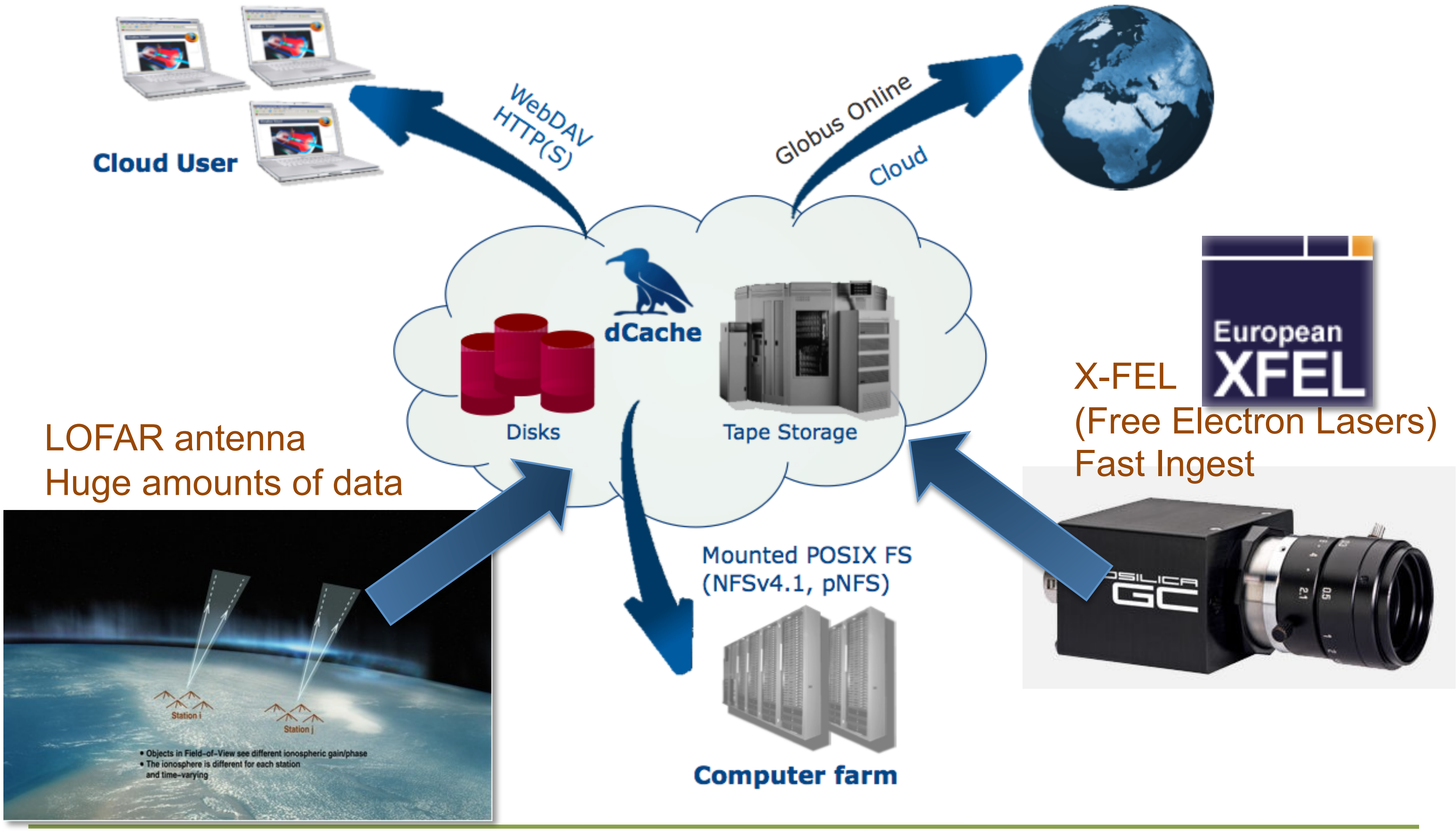
Just performance is not sufficient for
BIG DATA
in the future

NASA Evaluation of scientific data flow


SensorWeb High Level Architecture
floods, fires, volcanoes etc
Data Processing Node
Web Coordinate Transformation Service (WCTS)
Web Processing Service (WPS)
Web Coverage Service (WCS)
Sensor Data Products
Internet
SWE Node
SWE Node
SWE Node
In-situ Sensor Data Node
UAV Sensor Data Node
EO-1 Satellite
Web Map Service (WMS)
Web Feature Service (WFS)
Satellite sensor data product
Sensor Planning Service (SPS)
OGC Pub/Sub (OPSB)
Sensor Observation Service (SOS)
GeoTorrent Distribution Service
RSS Feeds
Campaign Manager API
Web Coverage Processing Service (WCPS)
RSS Feeds
Workflows
Campaign Manager
Identity Management Service (OpenID 2.0)
Components outlined in red are part of NASA generic SensorWeb toolbox

Courtesy: Goddard Tech Transfer News | volume 10, number 3 | summer 2012

Scientific Storage Cloud



Cloud User

WebDAV
HTTP(S)

Globus Online

Cloud

dCache

Disks

Tape Storage

X-FEL
(Free Electron Lasers)
Fast Ingest

LOFAR antenna
Huge amounts of data

Mounted POSIX FS
(NFSv4.1, pNFS)

Computer farm

European XFEL
Station i
Station j
• Objects in Field-of-View see different ionospheric gain/phase
• The ionosphere is different for each station and time-varying

Scientific Storage Cloud

- The same dCache instance can serve
  - Globus-online transfers via gridFTP
  - FTS Transfers for WLCG via gridFTP or WebDAV
  - Private upload and download via WebDAV
  - Public anonymous access via plain http(s)
  - Direct fast access from worker-nodes via NFS4.1/pNFS (just a mount like GPFS or Lustre but with standards)

Now, performance seems to be ok…

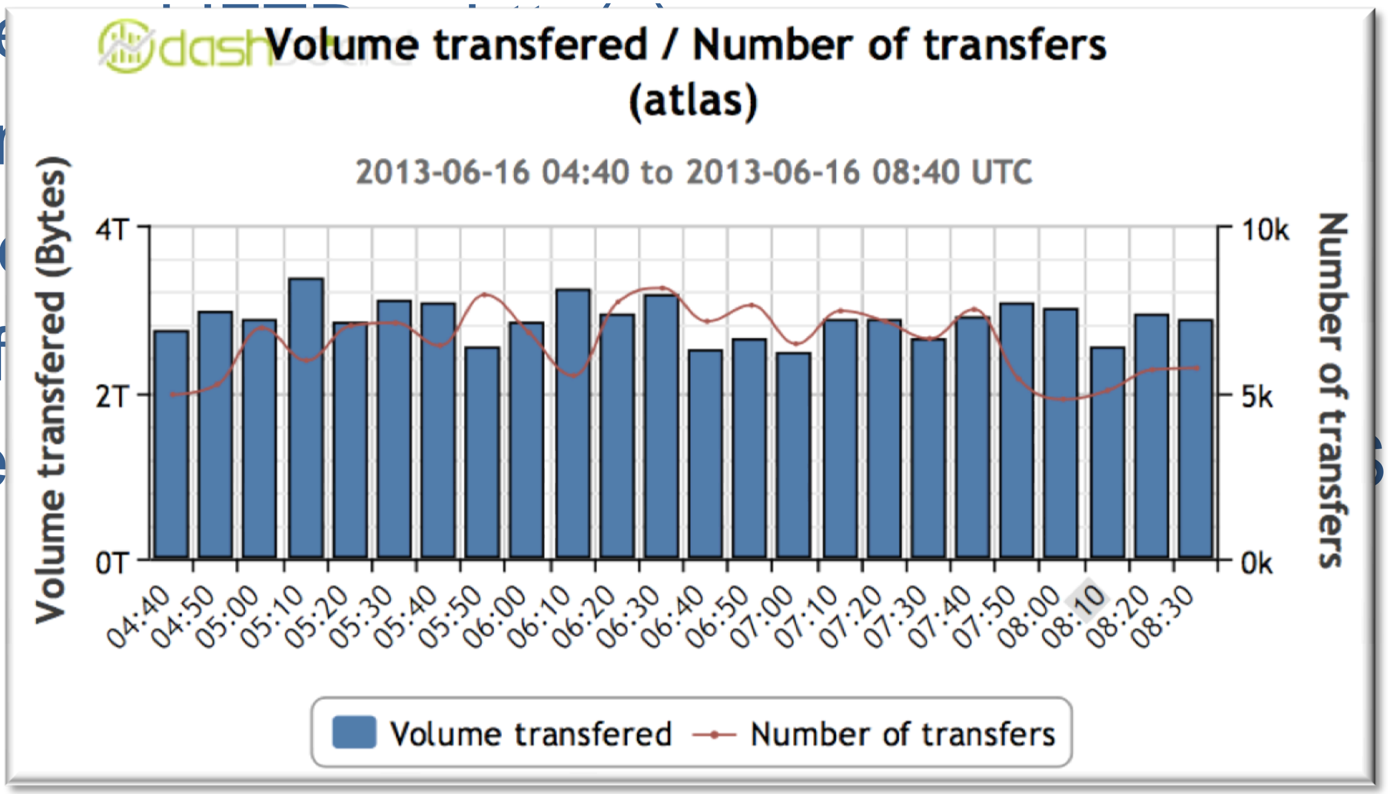
how about automated worldwide data transfers ?

How can you do worldwide automated transfers I

- Use ‘globus online’ a worldwide transfer services.
- dCache provides the necessary interfaces, including authentication.

How can you do worldwide automated transfers II

- Run your own “File Transfer Service, FTS(3)”.
- The Software is provided by EMI/CERN DM.
- FTS use
- FTS ca
- FTS do
- FTS is f
- dCache S, includin

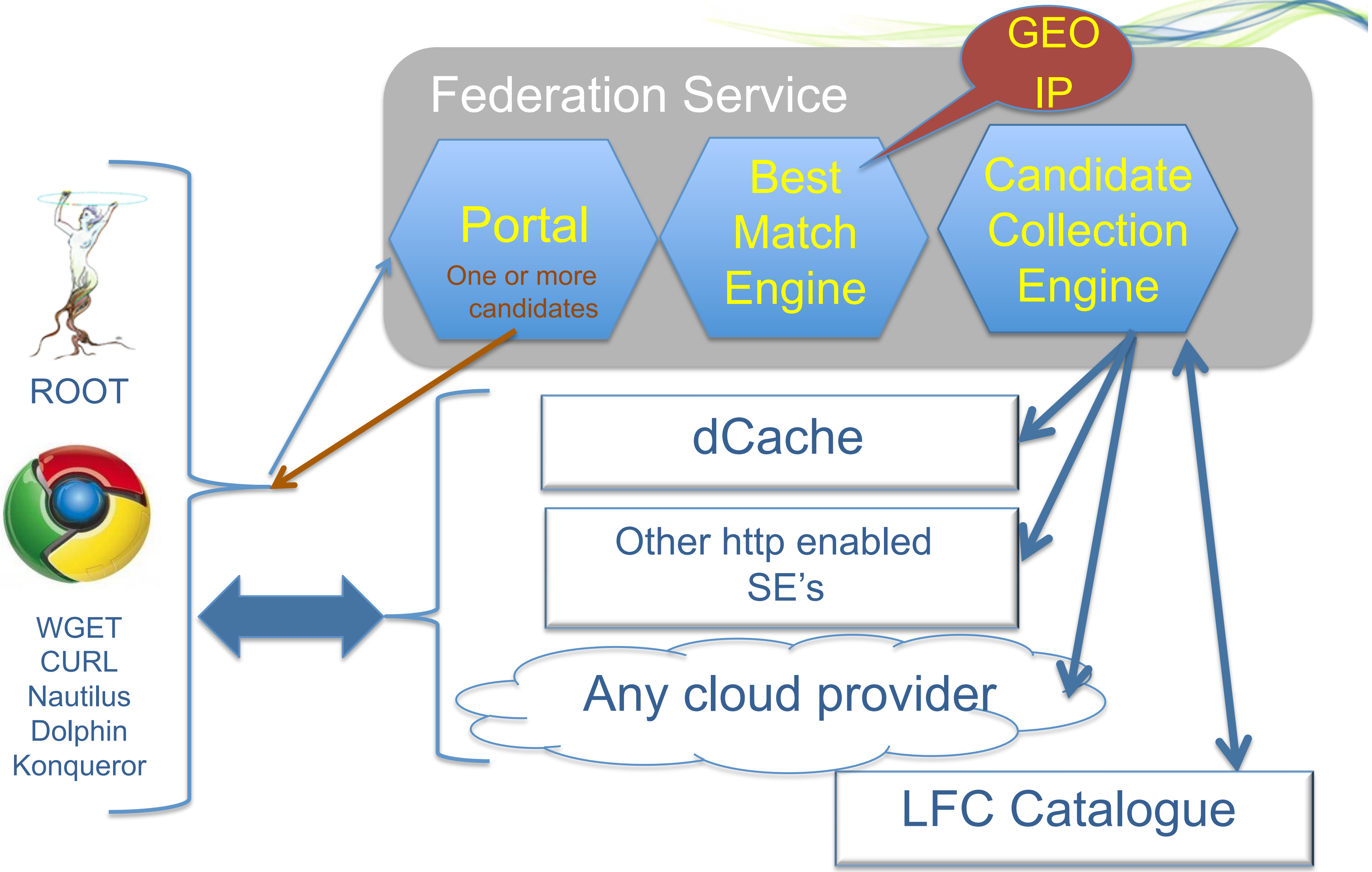


How can you do worldwide automated transfers III

The Dynamic http/WebDAV federation

- Still prototype status
- Collaboration between dCache.org and CERN DM, started with EMI

Dynamic Federation



Some remarks on authentication and authorization

- A user (individual) usually holds a set of credentials to identify him/herself against services.
  - Passport, Driver license, credit card
  - Google account, Twitter,
  - X509 Certificates (GRID, Country Certificate Authority)
- Federated Data Services should
  - Understand as many as possible of those credentials
  - Be able to map different ones to the same individual
- dCache does all this with :
  - User/password
  - Kerberos
  - X509 Certificates and Proxies
  - SAML assertions (in development within LSDMA)

In summary

- dCache has a long history in serving Big Data communities with PetaBytes of local and remote storage and Gbytes/sec of transfer performance.
- dCache is successfully moving into the “Scientific Cloud” direction, incorporating HTC and HPC.
- Focusing on High Individual Throughput as well as scaling out.
- Moreover, making sharing of scientific data easy and secure.
  - Making all data available via a set of industry access protocols:
    - NFS 4.1/pNFS for local high performance access (like local mount)
    - WebDAV and http for Web Sharing
    - CDMI and S3 (in preparation) for cloud acces.
    - GridFTP for fast wide area transfers.
  - Mapping various different credentials to a single ‘user’
    - X509 Certificates
    - User/Password
    - Kerberos
    - SAML/OpenID

The End

further reading

www.dCache.org