dCache.org

dCache, Software for Big Data

Innovation Day 2013, Berlin

Patrick Fuhrmann

Fermilab NDGF Nordic DataGrid Facility EMI European Middleware Initiative DESY Helmholtz Association Federal Ministry of Education and Research LSDMA

About …

- Technology and further roadmap
- Collaboration and partners
- Project status, system deployment and world map.
- Customer base
- What remains to be mentioned.
- Aim for today.

dCache, the technology

dCache.org

- **dCache manages Big Data !**
- It stores and retrieves huge amounts of data, distributed among a large number of heterogeneous disk server nodes, under a single virtual file-system.
- Storing, retrieving and managing data is supported through a variety of standard protocols with full ‘access control, ACL’ support
  - NFS 4.1/pNFS
  - https/WebDAV
  - GridFTP
  - Storage Resource Manager Protocol (SRM)
  - In preparation : The Cloud Data Management Interface, CDMI (SNIA)
- dCache accepts standard credential types and can be customized to accept others. (Pluggable Authentication System)
  - Kerberos
  - X509/Certificates/Proxies
  - User/Password

dCache, the technology

dCache.org

- If configured, dCache transparently moved data from disk to tertiary storage systems, if disk space is running short and retrieves data back to disk if requested. This can be configured to happen automatically or on request, e.g. “Bring Online”
- Data detects disk usage hotspots and automatically migrates data to less used disk nodes.
- Depending on the access profile (random versus streaming) dCache can move data to the most appropriate storage media (SSD versus Spinning Disk)
- dCache can make sure that a minimum number of copies of a file is present in the system (with a single entry in the file system). That allows to shutdown a certain number of storage nodes without affecting the overall availability of the data.
- Tools are build into dCache, allowing to shuffle data around between storage nodes, to
  - add storage nodes and pre-fill them with data to avoid high attraction of new nodes
  - decommission (empty) storage nodes without system service disruption
  - adjust load between storage nodes if access profile is known.

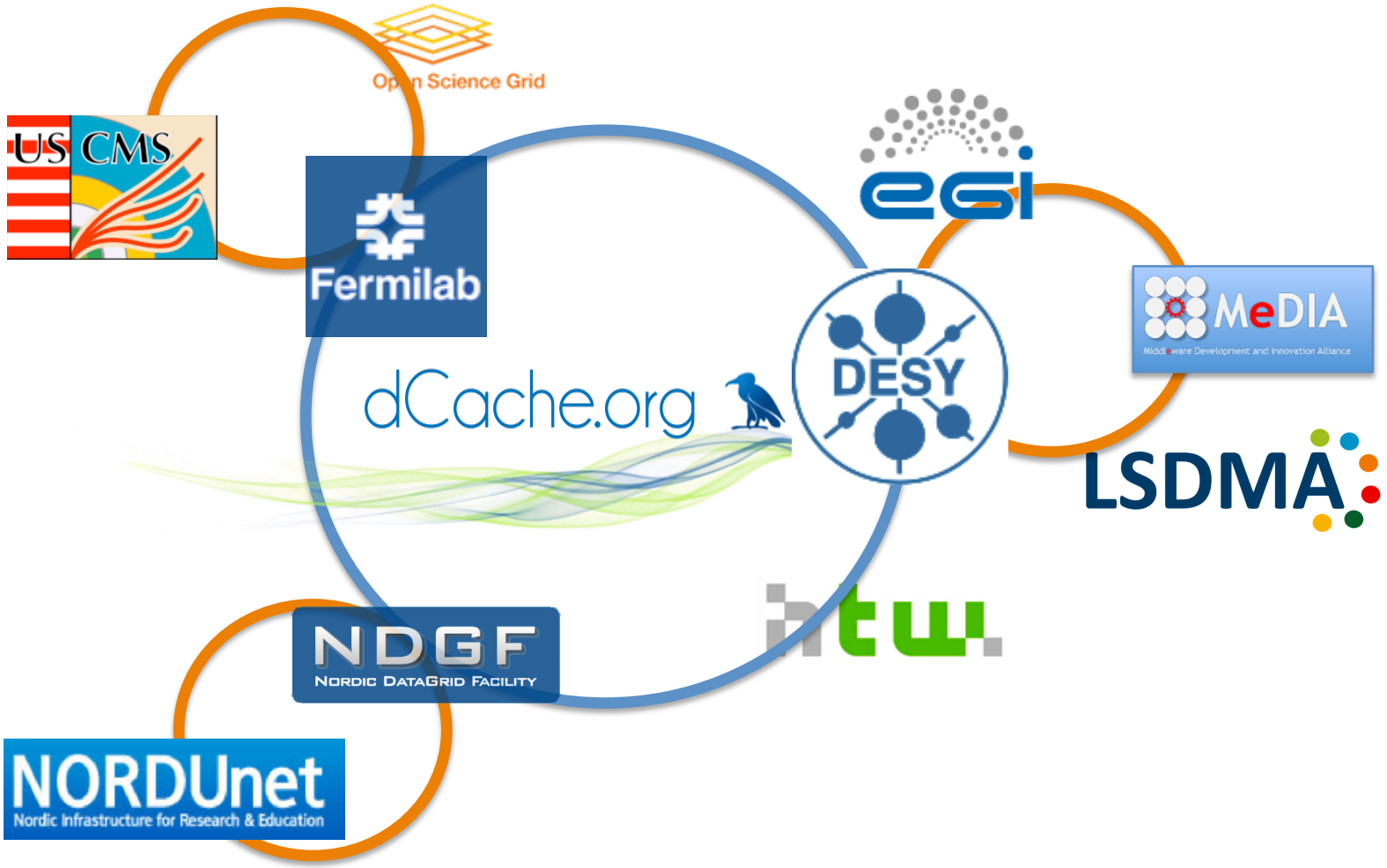
dCache, roadmap

dCache.org

- dCache is setting up a cloud storage instance, in collaboration with the HTW Berlin to serve scientific communities with cheap unlimited high available storage. (located in Europe ☺)
- Focusing on modern (Social Network) sharing mechanisms.
- Easy publishing. (Public URLs, OIDs)
- Data Lifecycle, from high available “extreme low latency” to archiving (tape or sleeping disks)
- Full support of standard Cloud Storage Interfaces (e.g. CDMI from SNIA), including metadata support.

The dCache collaboration and its partners.

dCache.org

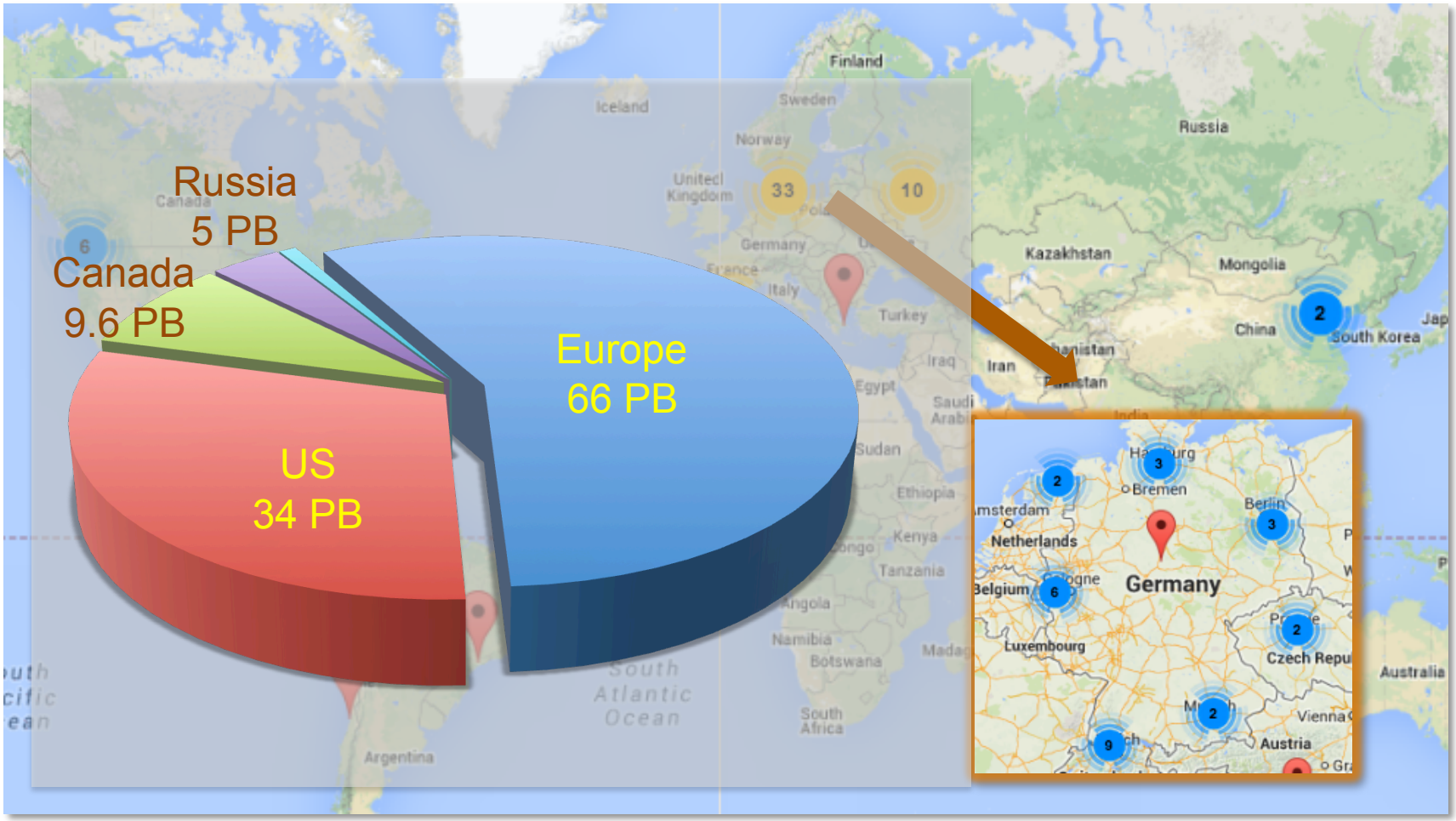


Deployed Production Instances

dCache.org

- dCache is in production for about 10 years.
- At the time being, about 80 sites around the world are using dCache to manage their Big Data repositories.
- The three largest Systems are managing about 14 Petabytes on disk and 40 Petabytes on tape on about 300 physical storage nodes each. (Karlsruhe, NYC, Chicago)
- The geographically biggest single system spans Norway, Finland, Sweden and Denmark.
- About 8 instances are in the 5 – 10 Petabyte range on disk and about the same amount on tape.
- About 70 instances are up to 5 Petabytes w/o tape backend
- The smallest instance is on a Raspberry Pi.

Worldwide Distribution



Customers/Communities

dCache.org

- With very little exceptions, dCache is used by science communities only.
- Largest Community, with about 120 Petabytes of active data in dCache worldwide, is the *Large Hadron Collider, LHC, WLCG* community.
- However, as more and more communities need Big Data solutions, additional costumers are *Photon Science (DESY), Neutrino Physics (FERMIlab)* and the European *LOFAR* antenna (Amsterdam, Jülich).

What remains to be mentioned

- dCache is available under AGPL v3
- dCache.org can offer customized Licenses.
- dCache.org provides professional
  - Release Management
  - Product testing and certification
  - User Ticket (Bug Reporting) mechanisms
  - European and Asian, Workshops and Schools
- Documentation, presentations and papers are available at www.dCache.org

What dCache.org expects from today

- We are looking for partners in industry, interested in taking over support for dCache at big sites.
- Industry partners, integrating dCache into appliances, e.g. disk systems.
- Integrating dCache into their software, for internal use or as part of their software distribution (portfolio). Preferred would be an ‘Open Source’ collaboration.

Thanks

further reading

www.dCache.org