A National Scale Federated Caching Infrastructure

Frank Würthwein
OSG Executive Director
UCSD/SDSC
Introducing myself

• Professor of Physics
  – My main physics interests is the search for Dark Matter using data from the CMS experiment at the Large Hadron Collider

• Executive Director of the Open Science Grid
  – The needs of my physics research prompted the creation of a global scale data & computing infrastructure that my group helped develop, and now helps operate.
  – I am responsible for the US contingent of that infrastructure.
The Large Hadron Collider Experiments ATLAS & CMS dominate resources available on and use of OSG
1.6 Billion hours a year

140 Million Core hours in the past 30 days

Over the last 12 months
134 Million jobs consumed
1.6 Billion hours of computing involving 2.3 Billion data transfers
to move 230 Petabytes

This aggregate was accomplished by
federating 60 clusters
that contributed 1h to 100M hours each

http://display.grid.iu.edu

<table>
<thead>
<tr>
<th>In the last 24 Hours</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Jobs</td>
<td>273,000</td>
</tr>
<tr>
<td>CPU Hours</td>
<td>4,354,000</td>
</tr>
<tr>
<td>Transfers</td>
<td>9,048,000</td>
</tr>
<tr>
<td>TB Transfers</td>
<td>1,064</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>In the last 30 Days</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jobs</td>
</tr>
<tr>
<td>CPU Hours</td>
</tr>
<tr>
<td>Transfers</td>
</tr>
<tr>
<td>TB Transfers</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>In the last 12 Months</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jobs</td>
</tr>
<tr>
<td>CPU Hours</td>
</tr>
<tr>
<td>Transfers</td>
</tr>
<tr>
<td>TB Transfers</td>
</tr>
</tbody>
</table>
Each red dot represents a cluster registered in OSG.

In aggregate ~ 200,000 Intel x86 cores used by ~400 projects across 36 fields of science.
Vision
Over hundreds of years, the defining common research service at Universities was the Library.
- defining service was the curation of information to support the creation of knowledge

Modern Science needs so much more …
- compute, storage, networking, …

$=>$ Cyberinfrastructure
The OSG Vision in a nutshell

- Across the nation, institutions invest into research computing to remain competitive
- Science is a team sport, and institutions with an island mentality will underperform.
- Integration is key to success

OSG empowers researchers to use compute & data resources across institutional boundaries.
Science is a Team Sport

The IceCube Collaboration

~100 members, 20 institutions
24 non-affiliated members
+35 associate members

Smithsonian Astrophysical Observatory
Adler Planetarium
Argonne National Lab
Barnard College / Columbia University
Bartol Research Institute / University of Delaware
Georgia Institute of Technology
Iowa State University
Purdue University
University of California, Los Angeles
University of California, Santa Cruz
University of Chicago
University of Iowa
University of Minnesota
University of Utah
Washington University in St. Louis
McGill University, Montreal
University College Dublin
Cork Institute of Technology
Galway-Mayo Institute of Technology
National University of Ireland, Galway

Xenon1T

VERITAS

SPT3G
Large Instruments require large international teams that span many institutions.

They rely on HTC for processing, calibrating, and analyzing their data, and high capacity networks to share their data.
dHTC

• Any scientific problem that can be decomposed such as to benefit from automation of a large number of individually schedulable jobs will benefit from dHTC.
  – CPU, GPU, node level multi-core, data production, data analysis, …

• Things we don’t do: large scale MPI

• Things that require special care: large IO jobs
Transparent Computing across different resource types

OSG integrates computing across different resource types and business models.
Discoveries that relied on HTC

The Nobel Prize in Physics 2017

Rainer Weiss
Prize share: 1/2

Barry C. Barish
Prize share: 1/4

Kip S. Thorne
Prize share: 1/4

The Nobel Prize in Physics 2017 was divided, one half awarded to Rainer Weiss, the other half jointly to Barry C. Barish and Kip S. Thorne “for decisive contributions to the LIGO detector and the observation of gravitational waves”.

The Nobel Prize in Physics 2013

François Englert
Prize share: 1/2

Peter W. Higgs
Prize share: 1/2

The Nobel Prize in Physics 2013 was awarded jointly to François Englert and Peter W. Higgs “for the theoretical discovery of a mechanism that contributes to our understanding of the origin of mass of subatomic particles, and which recently was confirmed through the discovery of the predicted fundamental particle, by the ATLAS and CMS experiments at CERN’s Large Hadron Collider”.

June 14th 2018
Nobel Prices are just the tip of the iceberg
Engineering a paradigm shift in Networking

• Academia today depends on a mix of regional network (e.g. CENIC), Internet2, and ESNet to support the internet for Science.

• We are working to expand what internet companies have seen as their business by adding storage into the network in order to better support science.
Internet2 deploys storage, OS, and K8s into their network backbone. Internet2 joins the Pacific Research Platform K8s cluster. My group integrates that storage hardware into our Data Federation.
A Data Federation for Science
Science is a Team Sport

Big Science requires large global teams

IceCube compute hours per cluster

CMS disk space per cluster

Federation of resources is essential to do the science.
OSG Data Federation (StashCache)

- Data is placed on an origin.
- Data appears in distributed namespace.
  - Access restricted for privacy
- Namespace is cached locally as part of OSG environment (CVMFS).
- Scientists open files at runtime.
- OSG environment serves files from closest cache.

For more info see DOI: 10.1088/1742-6596/898/6/062044
CVMFS, a virtual filesystem

- fuse mounted on worker nodes on OSG
- Cached on each cluster via squid
- Provides runtime environment & data namespace.

```
/cvmfs/stash.osgstorage.org/user/dweitzel/public/blast/data/yeast.aa
```

hCache
StashCache Use

Served ~12 PB last year at Cache Hit rates of ~88%
Science use of StashCache

- Individual PIs & students
  - Two Bioinformatics groups (UNL & Clemson)
    - Metagenomics and genomics data.

- Instruments
  - DES, Minerva, LIGO, Xenon1T, …
  - Details on LIGO example:
    DOI 10.1145/3093338.3093363
Dive into some technical Details
Data Federation for WAN Reads

Applications connect to local/regional redirector. Redirect upwards only if file does not exist in tree below. Minimizing WAN read access latency this way.

~150 Petabytes of disk accessible this way today.

Many Clusters in US

Many Clusters in EU

Simplified topology as deployed by the CMS Experiment at the LHC

June 14th 2018
• servers can be connected into arbitrary tree structure.
• application can connect at any arbitrary node in the tree.
• application read pattern is vector of byte ranges, chosen by the application IO layer for optimized read performance.
  – Efficient support of partial file reads instead.
Caches added in the Data Federation Tree

• Application client requests file open
  – Cache client requests file open from higher level redirector if file not in cache

• Application client requests vector of byte ranges to read
  – Cache provides subset of bytes that exist in cache, and fetches the rest from remote.
  – if simultaneous writes below configured threshold then write the fetched data to cache.
    ▪ else fetched data stays in RAM, flows through to application, and gets discarded.
  – Cache client fills in missing pieces of file while application processes vector of bytes requested, as long as simultaneous writes below configured threshold.

• Cache can be configured for different behavior on different subsets of the namespace.
Implications

• An organization can join the federation with their own “data origin” and their own partition of the global namespace.
• A cache owner can decide on policies for different parts of namespace.
• This allows the owner to selectively serve only a subset of the community that use the federation.

We can build community specific “storage overlays” with this technology.
The Data Federation does not enforce anything globally.

Everything in the way data is found in the federation works on the premise that any server may disappear at any moment.

- Redirectors lazily forget what they know.
- If a file open fails at a server, the redirection algorithm automatically starts over.

The Data Federation is dynamic by design with this technology.
The PRP/NRP Vision

• Cheap storage is deployed all over the network
  – At end-points inside Science DMZs
  – At various peering points in the network

• Services can be dynamically deployed on top of the storage, e.g. via K8s.
  – A specific cache can grow by adding storage.
  – Additional caches can be added to the tree.

• The people who own the storage hardware shall not have to know anything beyond the container orchestration, e.g. K8s.

• A given server may generally run one container to measure network performance, and a second container to provide the cache service.
  – The two containers may be managed by different organizations.
Aside on Intelligent Storage
R&D on Caching and Intelligent Storage

• We have a joint project with UCSC CS on understanding the case for partial file reads and partial file caching.

• A CMS user typically analyzed entire datasets of $O(1)$-$O(100)$ TB
  – The same identical application processes all of the dataset via a collection of independent jobs.
  – Each job typically only needs a subset of the data.

If the storage understood the data content structure then it could prefetch based on the first job out of the collection of independent jobs.
Optimize Data Structure for Partial Reads

- 1, CE$9, /@ DG$C- ) ' 4: //; $/:' , 4, D 56$ef $' ( ) *:+; $D $ g C* . DB$4, , ; H$
  - ', 4( , $ J: . D$ # #$. 56$4 C@1$
  - ', 4( , $ 9B CD$ # #$. 56$15 CD$ . SL1W$D $: C* : $5DE: $J: . D$/ $: D$. 4 : ( . 1 $ D$ 5$5$

- 4$7WV$9C$A/ D 4$*: $. $ *@/ $ $ $ b$WZ:
- -= D $95 $*: $4 ) $ bb$WZ$D $ b$ Z$
Experience from Production Deployment in CMS
Production Goal:
Distributed cache that sustains 10k clients reading simultaneously from cache at up to 1MB/s/client without loss of ops robustness.

Provisioned test systems:
UCSD: 9 x 12 SATA disk of 2TB @ 10Gbps for each system.
Caltech: 30 SATA disk of 6TB @ 2x40Gbps per system

Applications can connect at local or top level cache redirector.
⇒ Test the system as individual or joint cache.
Example Cache use today

Serving data unaffected by cache filling

Cache is filling

Data Federation & caches are step towards a Content Delivery Network
The existing production cache caches only small part of total CMS data namespace. 
Namespace cached ~ 700+ TB 
Disk space in cache ~ 200+ TB 

Whether or not there are jobs running that use that part of namespace is essentially random. 

We expanded cached namespace to a couple PB and cache space to ~600TB since February
Exabytes of data in 2026

Disk needs projections for the CMS experiment at the LHC.
Summary & Conclusion

• Science is a Team Sport
• Cyberinfrastructure is the new Library
• Federation is essential for science to maximally benefit from global investment into Cyberinfrastructure
• At SDSC, we are a Cyberinfrastructure Leader
  – We invent, design, construct, deploy, and operate cyberinfrastructure for science.

Let’s work together to support Science!
Some plots on performance testing of Caches
Performance Testing

• Start with an empty cache.
• Submit jobs to cluster where each job reads a random file from a few thousand file dataset.
• Record:
  – running jobs
  – data send and received at NIC
  – data written to and read from disk
Let’s walk through a single 8h test in some details
# of clients hitting the cache

Up to 8980 clients simultaneously hit the cache with read requests
Aggregate view of all 9 NICs

Bits received means IO from WAN to cache

Bits send means IO from cache to jobs

Bits read from disk cache = Bits send – Bits received
Aggregate view of all 9 NICs

Peak of 57.8 Gbps sent
Bits send means IO from cache to jobs

Peak of 45 Gbps received
Bits received means IO from WAN to cache

Bits read from disk cache = Bits send – Bits received
Aggregate view of all 9 NICs

Bits received means IO from WAN to cache

Bits send means IO from cache to jobs

Bits read from disk cache = Bits send – Bits received

20+ Gbps read from disk cache

~10 Gbps read from disk cache for most of the remainder of the 8h
Single Server View

- Write to disk peak at ~3Gbps
- Reads from disk pretty steady at ~2Gbps

Note: both reads and writes are limited by complicated interplay of cache behavior, hardware performance, and requests from jobs.
Single server view

Bits received means IO from WAN to cache

Bits send means IO from cache to jobs

Disk IO