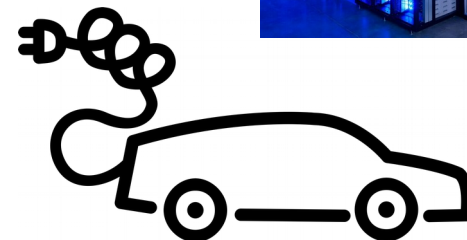


The DataBridge

Laurence Field

Volunteer Computing

- Computer owners donate computing capacity
 - To a cause or project
- Spare cycles on computers
 - Away from office/desk, unused cores
- Other opportunistic resources
 - Idle machines in data centers
 - IoT?
 - Opportunistic Energy?
 - Cars e.g. V2G?

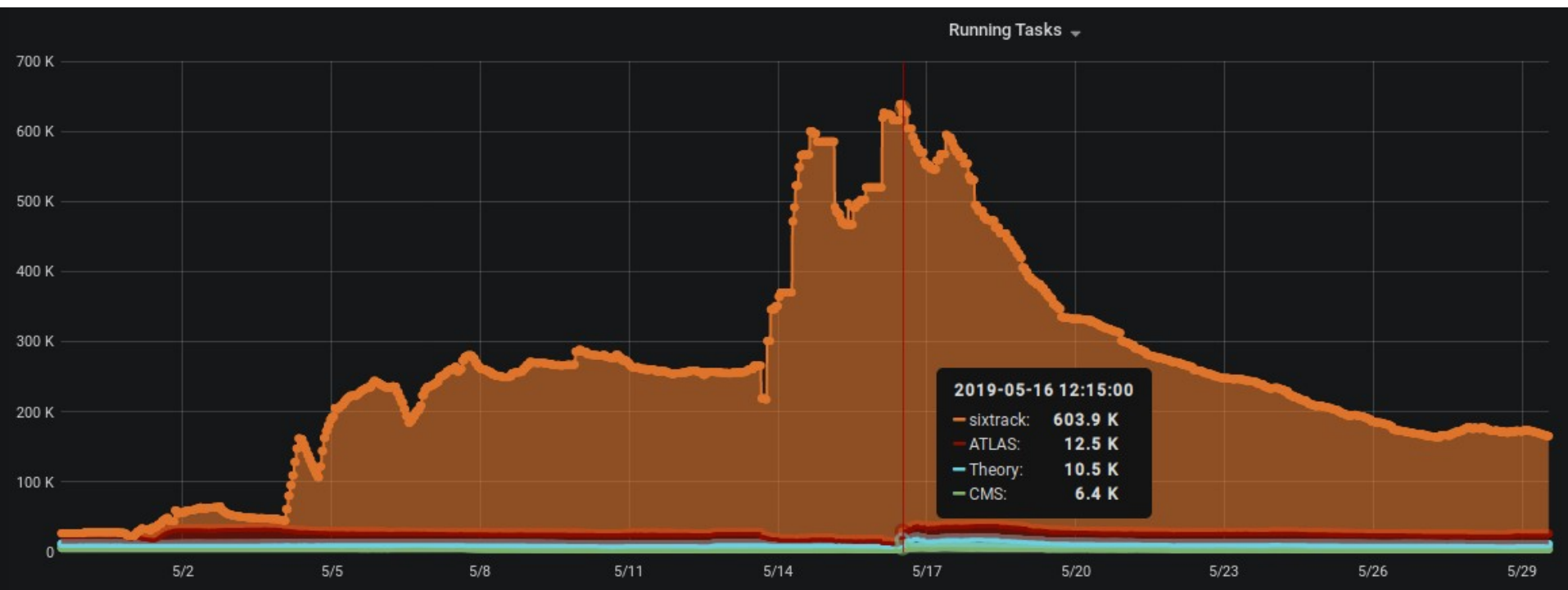


Motivation

- Free* resources
 - 100K hosts achievable for large projects
 - Actual job slot count (number of cores) is higher
- Community engagement
 - Outreach channel
 - Explaining the purpose and value of the science
 - Participation
 - Offering people a chance to contribute
 - Engagement forms a strong bond
- Community support

*There are costs associated with their use

2019 Pentathlon



Challenges

- The cost of using the free resources
 - Initial integration requires investment
- Operations and Maintenance
 - Public facing support on all levels
 - Lowered by community supports
- Attracting and retention of volunteers
 - Advertisement and engagement
 - Communications cost for capacity building
- Low level of trust
 - Anyone can register as a volunteer
 - Not the same as internally owned systems
- Running scientific software on Windows
 - ~80% of the resources

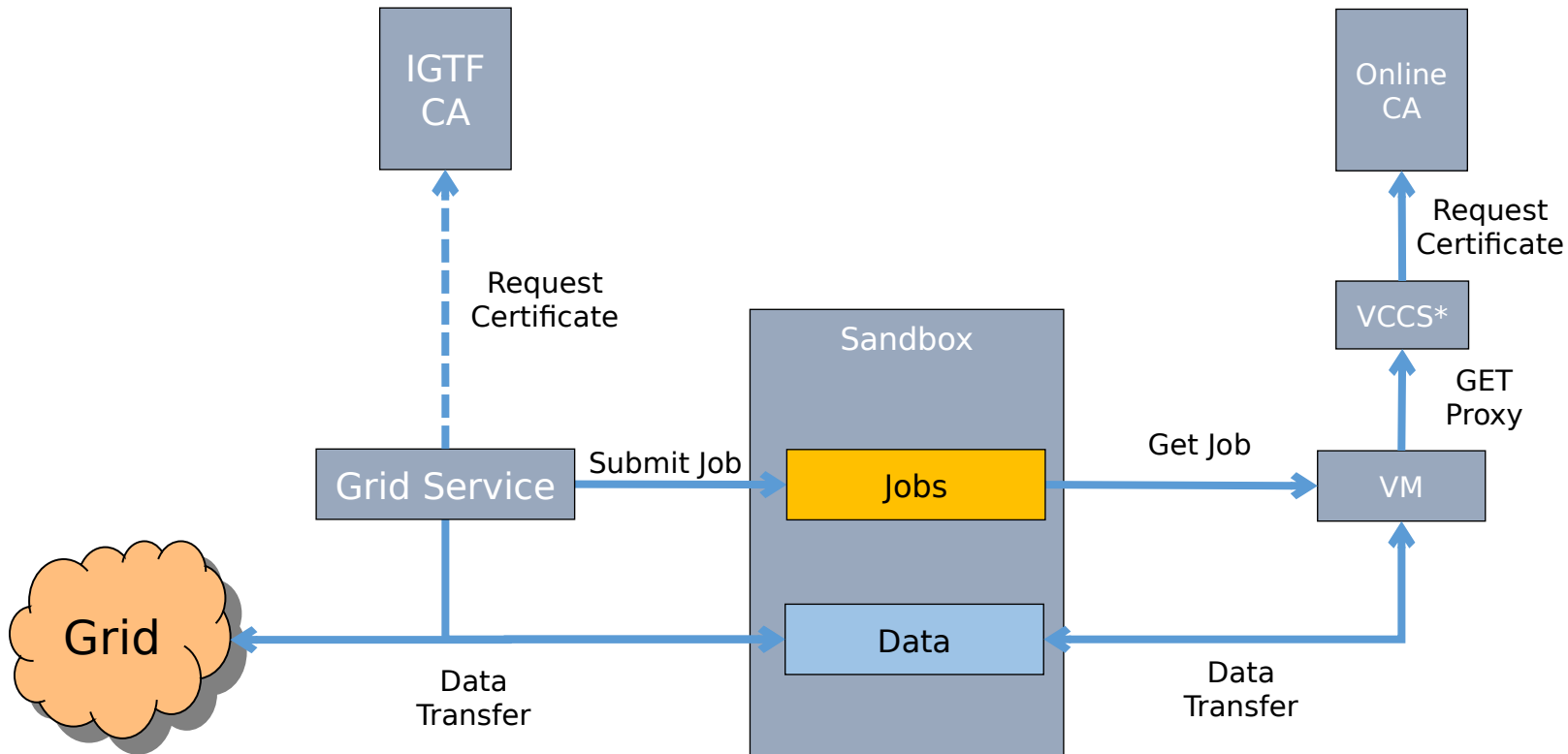
Untrusted Resources

- Volunteer resources are untrusted
 - Not authenticated with an IGTF accredited credential
- Volunteers register with a username and password
 - Low level of assurance
- Need to bridge the trusted and untrusted domains
 - Volunteers need to:
 - Read input files from the Grid
 - Write output files to storage for further validation and processing
 - Pull the job description and execute the job
 - Volunteers must not
 - Delete data
- Authentication and Authorization

Sandboxing and Authentication

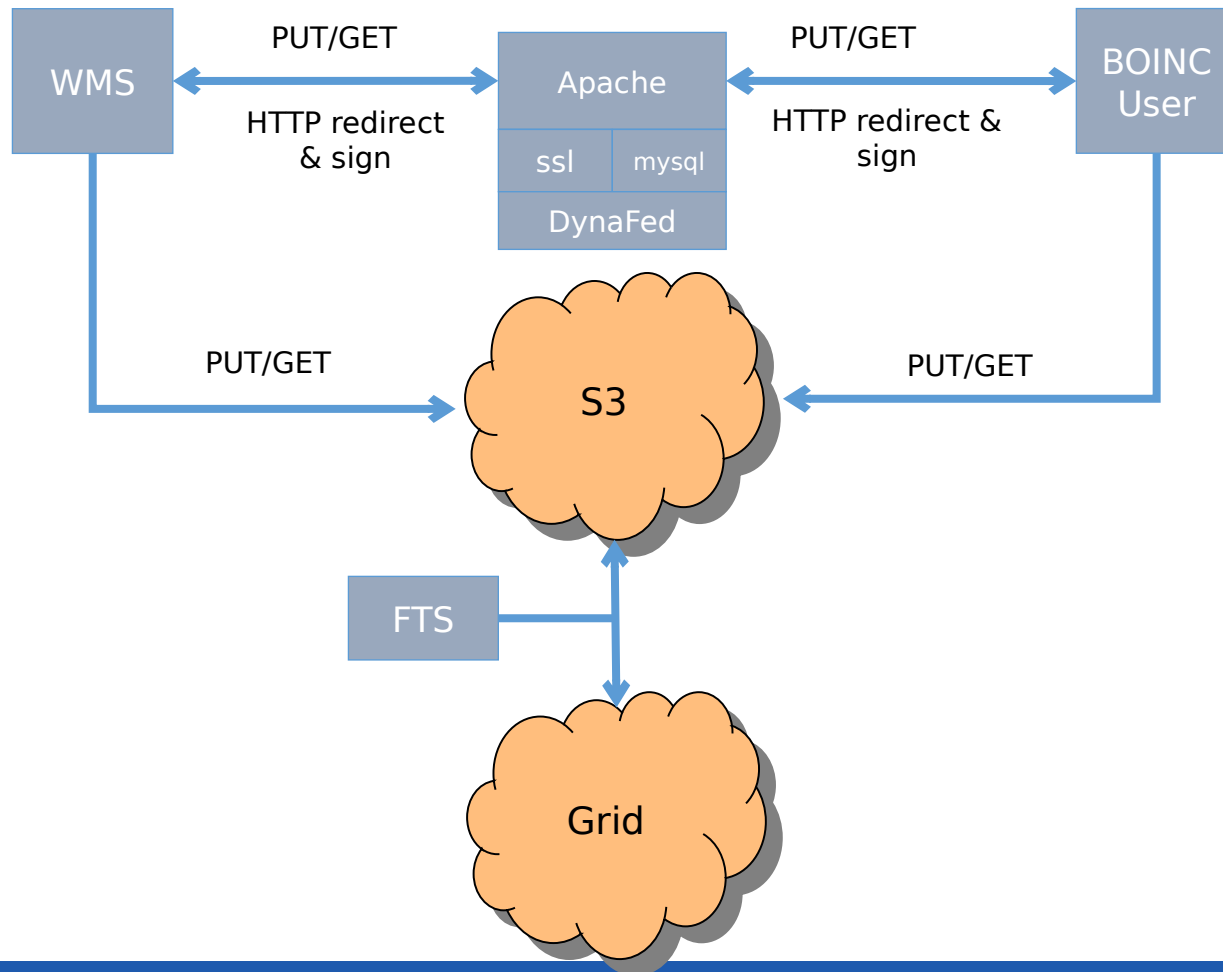
Trusted Domain

Untrusted Domain



*Volunteer Computing Credential Service

DataBridge

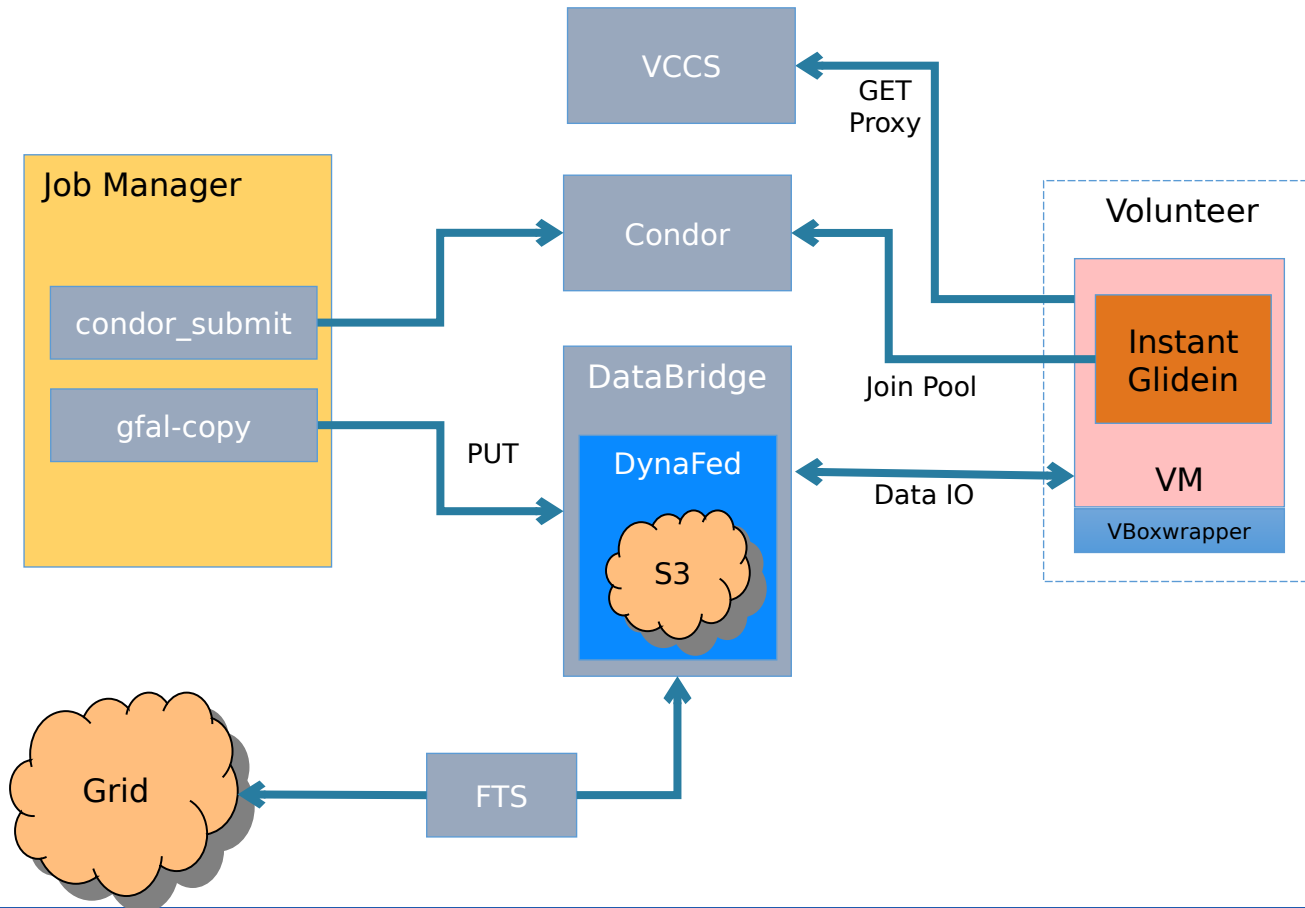


The Volunteer Cloud

Application Server

Common Infrastructure

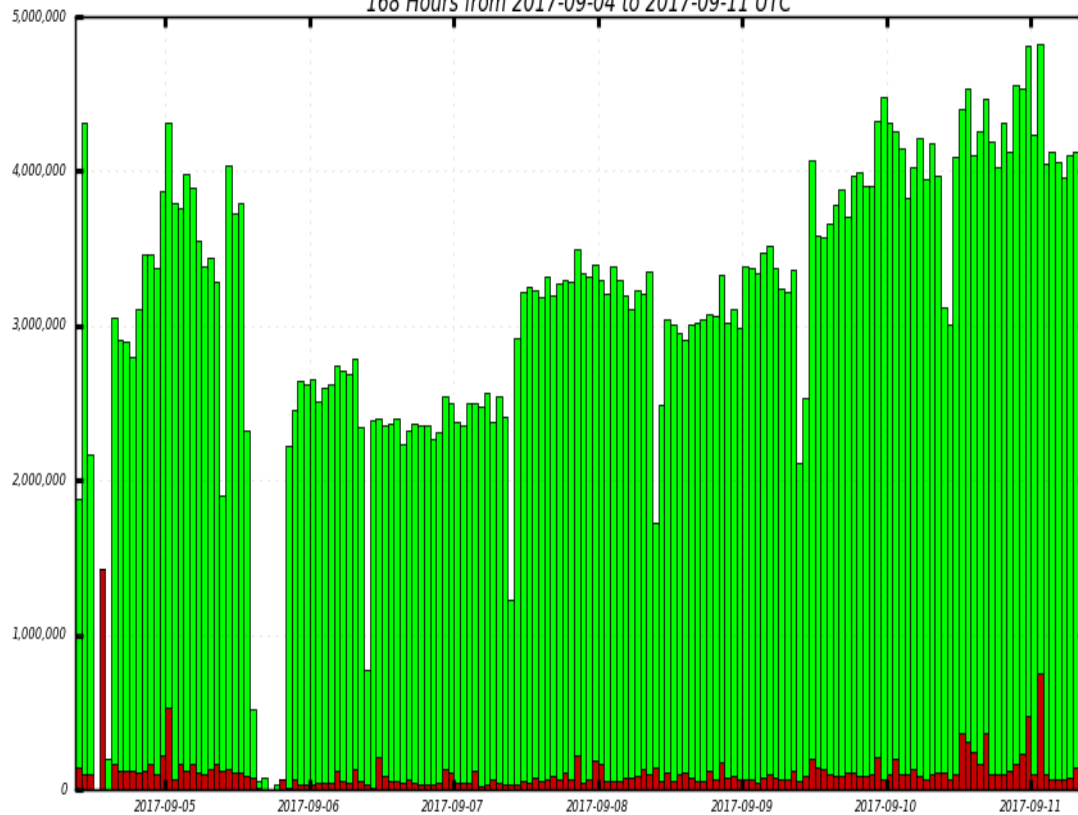
Volunteer's machine



Efficiency of CMS@home



WallClock Consumption for Successful, Failed and/or Aborted Jobs
168 Hours from 2017-09-04 to 2017-09-11 UTC



■ WallClock Consumption of Successful Jobs

■ WallClock Consumption of Failed and/or Aborted Jobs

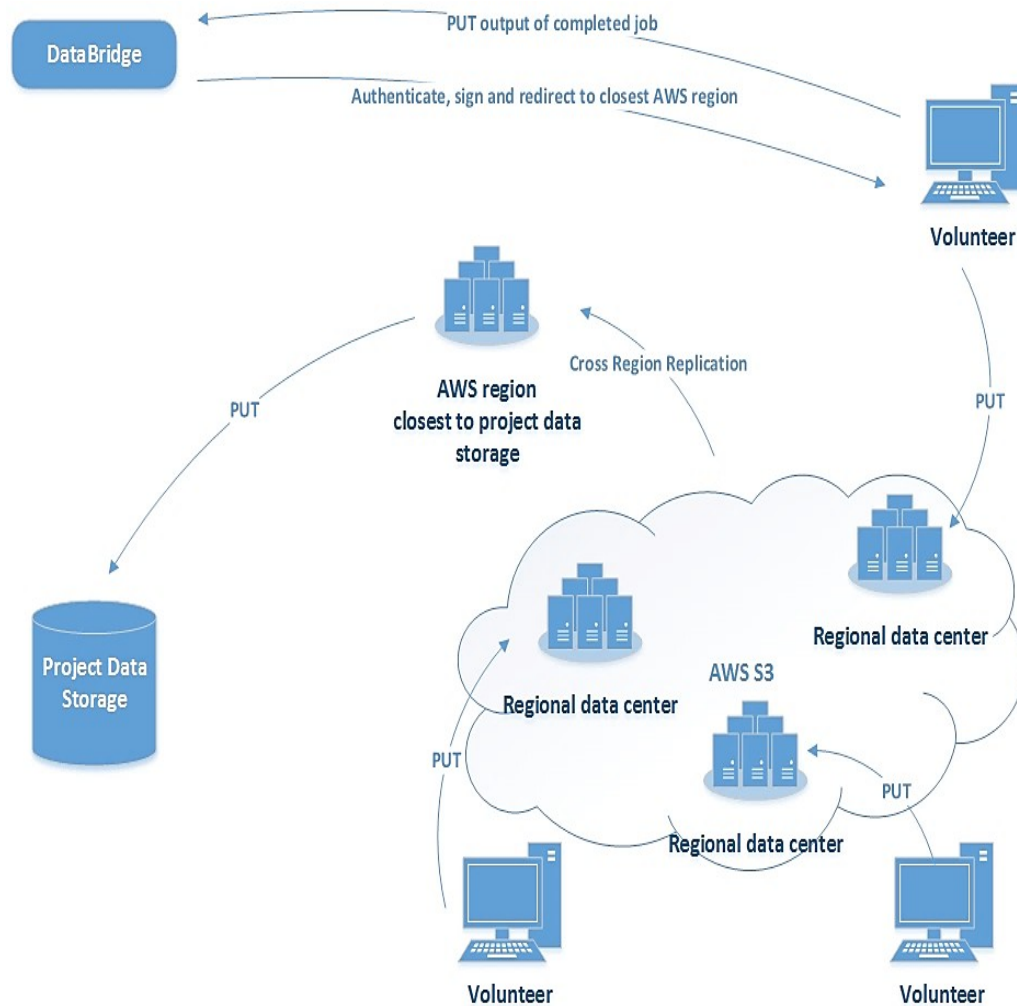
~95%

Maximum: 4,822,344 , Minimum: 0.00 , Average: 3,097,391 , Current: 3,056,933

DataBridge Federation

- How to reliably collect output files from volunteers?
 - Thousands of miles away from the central storage location
- Larger distances results in higher RTTs
 - Increased chance of packet loss during transmission
 - Thus lowering the transfer performance
 - Analysis of CMS@home logs showed that
 - network related error codes occur more frequently
 - as the distance from CERN increases
- Provide a closer upload endpoint to reduce latency
 - Federate the DataBridge upload endpoints
 - Redirect upload requests to a close endpoint

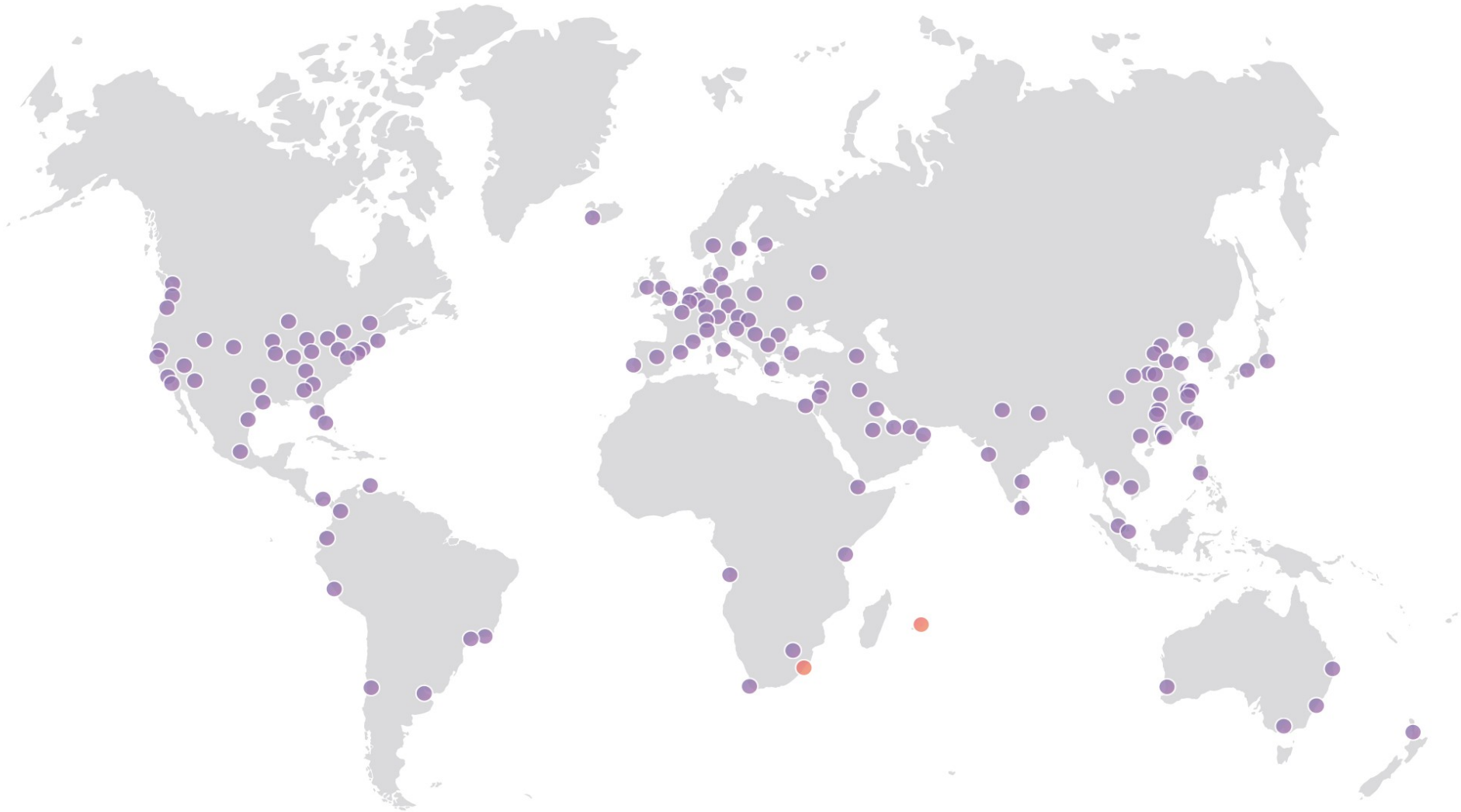
AWS Implementation (CHEP 2016)



LHC@home Squids For CVMFS

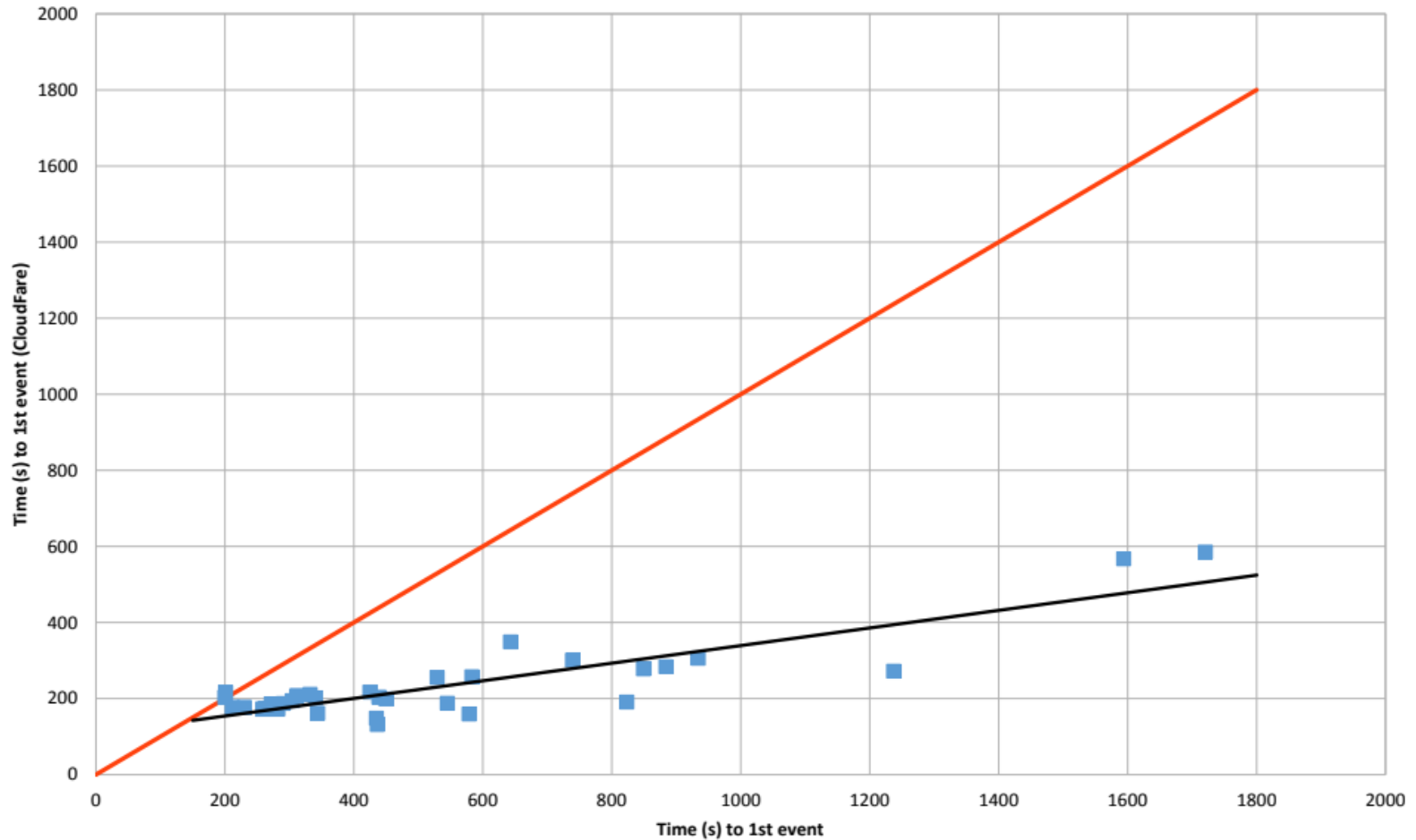
- Used to have two services
 - lhchomeproxy.cern.ch
 - lhchomeproxy.fnal.gov
 - Two machines behind each service
- Used CVMFS_PAC_URLS
 - <http://lhchomeproxy.cern.ch/wpad.dat>
- Geographical load balancing
 - Redirect to the nearest instance

CDN: Cloudflare



CMS@home Job Speedup

CloudFare



Summary

- Links the trusted and untrusted domains
 - Implements different policies per domain
- Provides an AuthN and AuthZ layer for S3
 - Hides admin passwords from the volunteers
- Can be used to build a global data federation
 - To improve the efficiency of uploads from volunteers
- A very reliable service for the volunteer cloud
 - Can be useful in other cloud contexts
- Dynafed is a really useful component



Volunteer
computing
for the LHC

Search

HOME

ABOUT ▾

PROJECTS ▾

JOIN US!

HELP & FAQ

CONTACT

