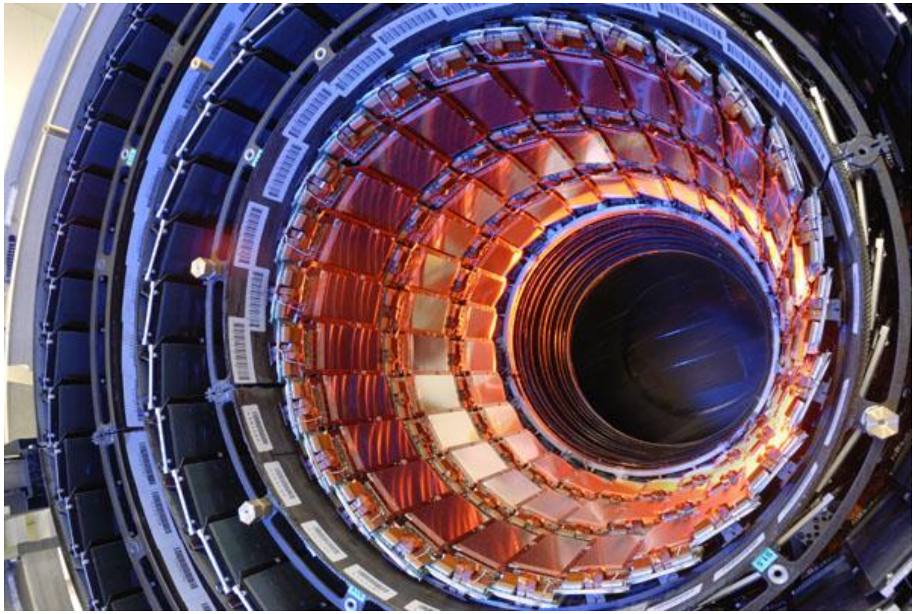
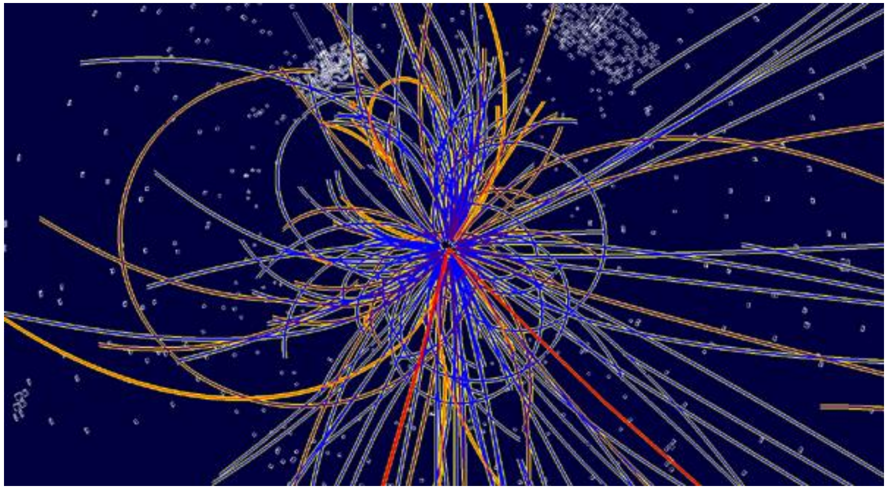
Opportunistic Computing Only Knocks Once: Processing at SDSC



Ian Fisk
FNAL
On behalf of the CMS Collaboration

Overview

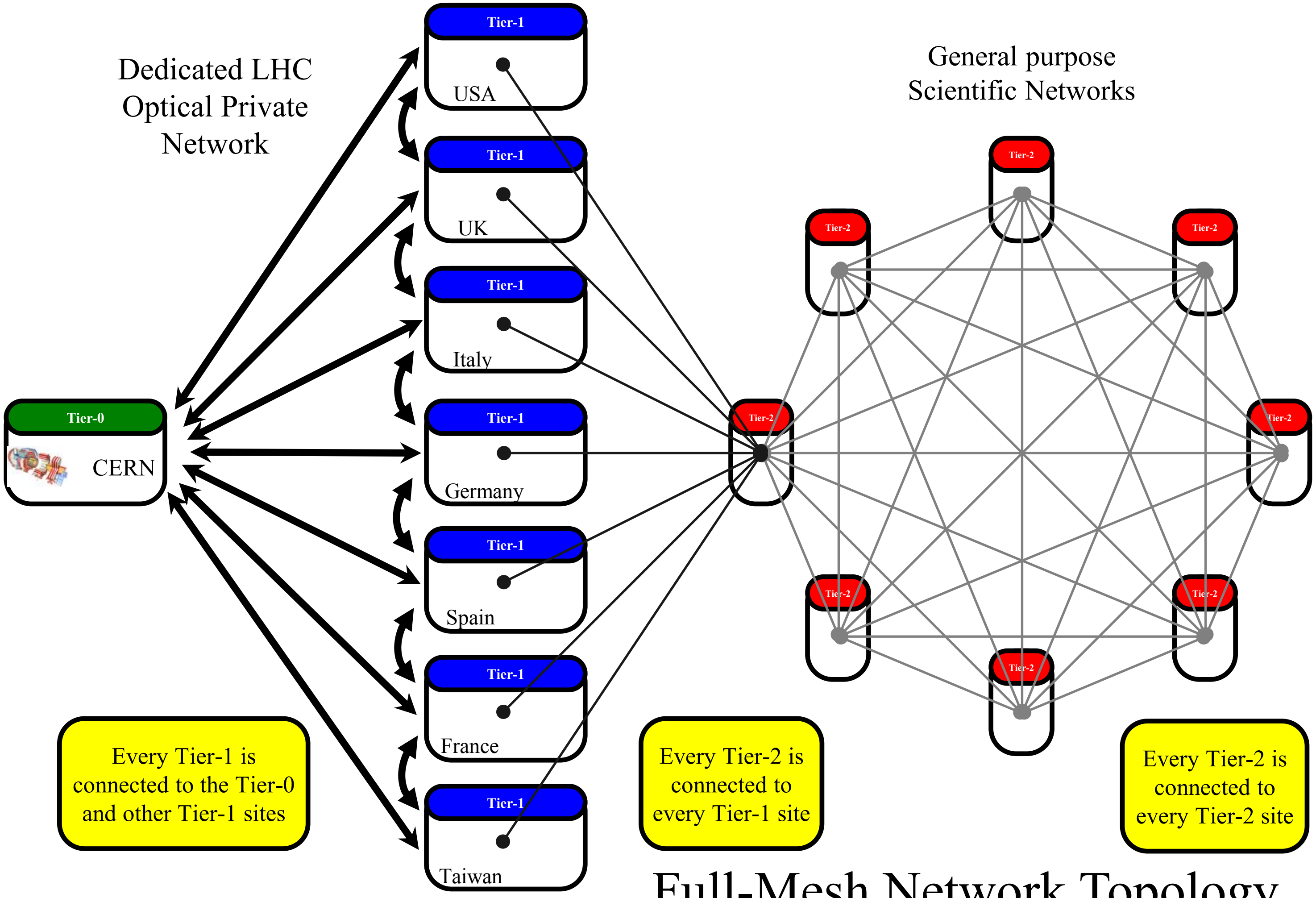
- In 2012 CMS took more data than could be immediately processed using the available resources
- The extra was referred to as “Parked Data”
  - Samples expected to be looked at during the long shutdown



CMS Computing Infrastructure

7 Tier-1 sites

52 Tier-2 sites



Every Tier-1 is connected to the Tier-0 and other Tier-1 sites

Every Tier-2 is connected to every Tier-1 site

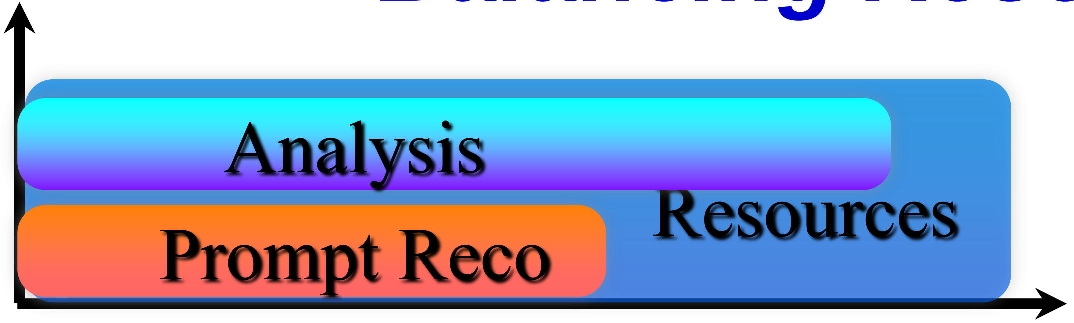
Every Tier-2 is connected to every Tier-2 site

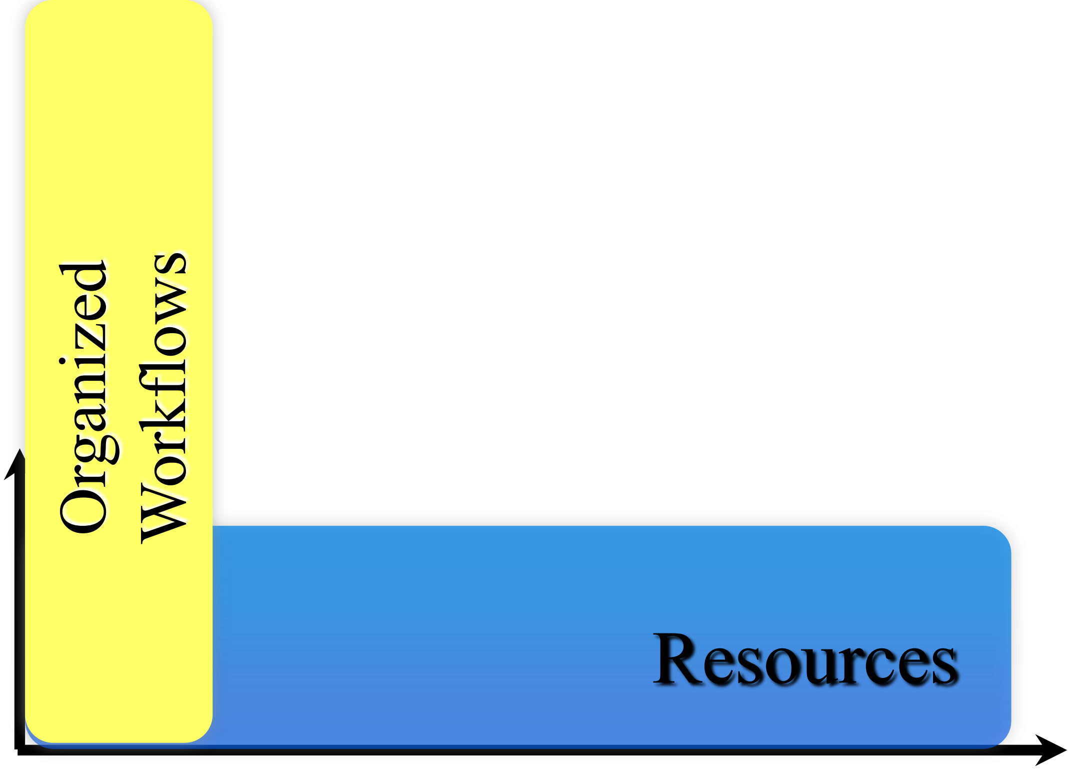
Full-Mesh Network Topology

Resources

- **CMS has roughly 25k processor cores dedicated to reconstruction tasks at Tier-0 and Tier-1 computing centers**
  - These resources are negotiated years in advance and scheduled for full utilization
    - Processing data and reconstructing simulation events occupy the bulk of the time
  - If we want to do something outside the schedule or faster
    - We need to kick out something else, or we need to find resources
      - – One option would be commercial clouds, but then we are looking for money
      - – Opportunistic computing is the only way to get significant increases in a short term

Balancing Resources





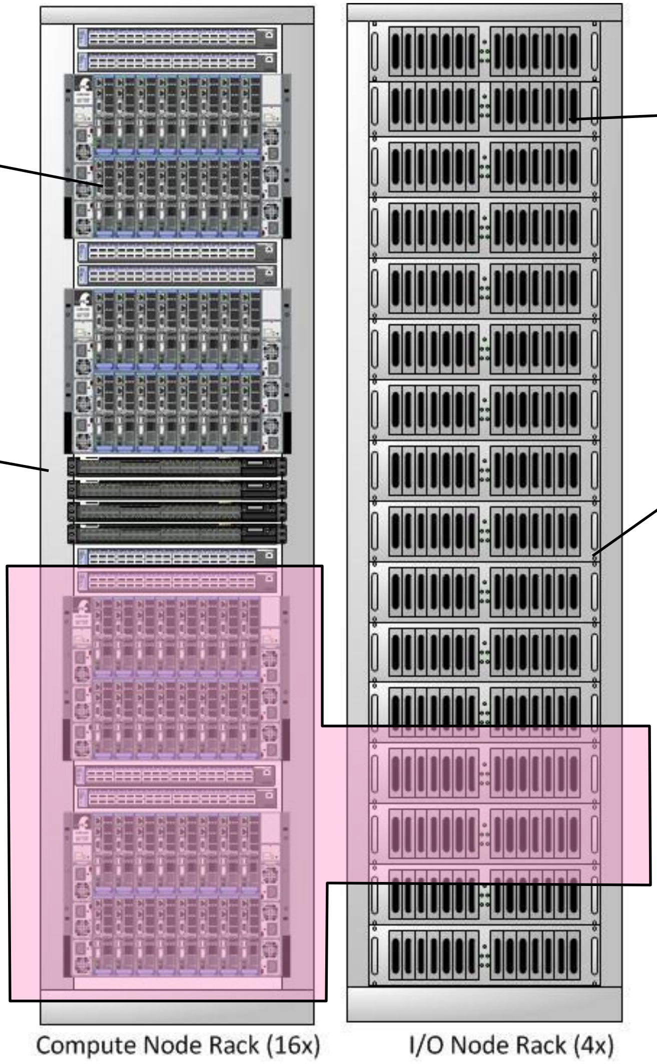
For many parts of the program we do use an average load,

However there are benefits to growing to peaks that are much larger than the average and then have sustained period of lower than average usage

Overview

- Frank Würthwein (UCSD, CMS Tier II lead) approaches Mike Norman (Director of SDSC) regarding analysis delay
- A rough plan emerges:
  - Ship data at the tail of the analysis chain to SDSC
  - Attach Gordon to CMS workflow
  - Ship results back to FNAL
- From CMS perspective, Gordon becomes a compute resources
- From SDSC perspective, CMS jobs run like a gateway

Gordon Overview



- 1,024 2S Xeon E5 (Sandy Bridge) nodes
- 16 cores, 64 GB/node
- Intel Jefferson Pass mobo
- PCI Gen3

- 3D Torus
- Dual rail QDR

- Large Memory vSMP Supernodes
- 2TB DRAM
- 10 TB Flash

- 300 GB Intel 710 eMLC SSDs
- 300 TB aggregate

- 64, 2S Westmere I/O nodes
- 12 core, 48 GB/node
- 4 LSI controllers
- 16 SSDs
- Dual 10GbE
- SuperMicro mobo
- PCI Gen2

“Data Oasis”
Lustre PFS
100 GB/sec, 4 PB

CMS Components

- CMSSW: Base software components, NFS exported from IO node
- OSG worker node client: CA certs, CRLs
- Squid proxy: cache calibration data needed for each job, running on IO node
- glideinWMS: worker node manager pulls down CMS jobs
- BOSCO: GSI-SSH capable batch job submission tool
- PhEDEx: data transfer management

CMS "My Friends" Stack

- **CMSSW release environment**
  - NFS exported from Gordon IO nodes
  - Futu
    -

This is clearly complex !!!

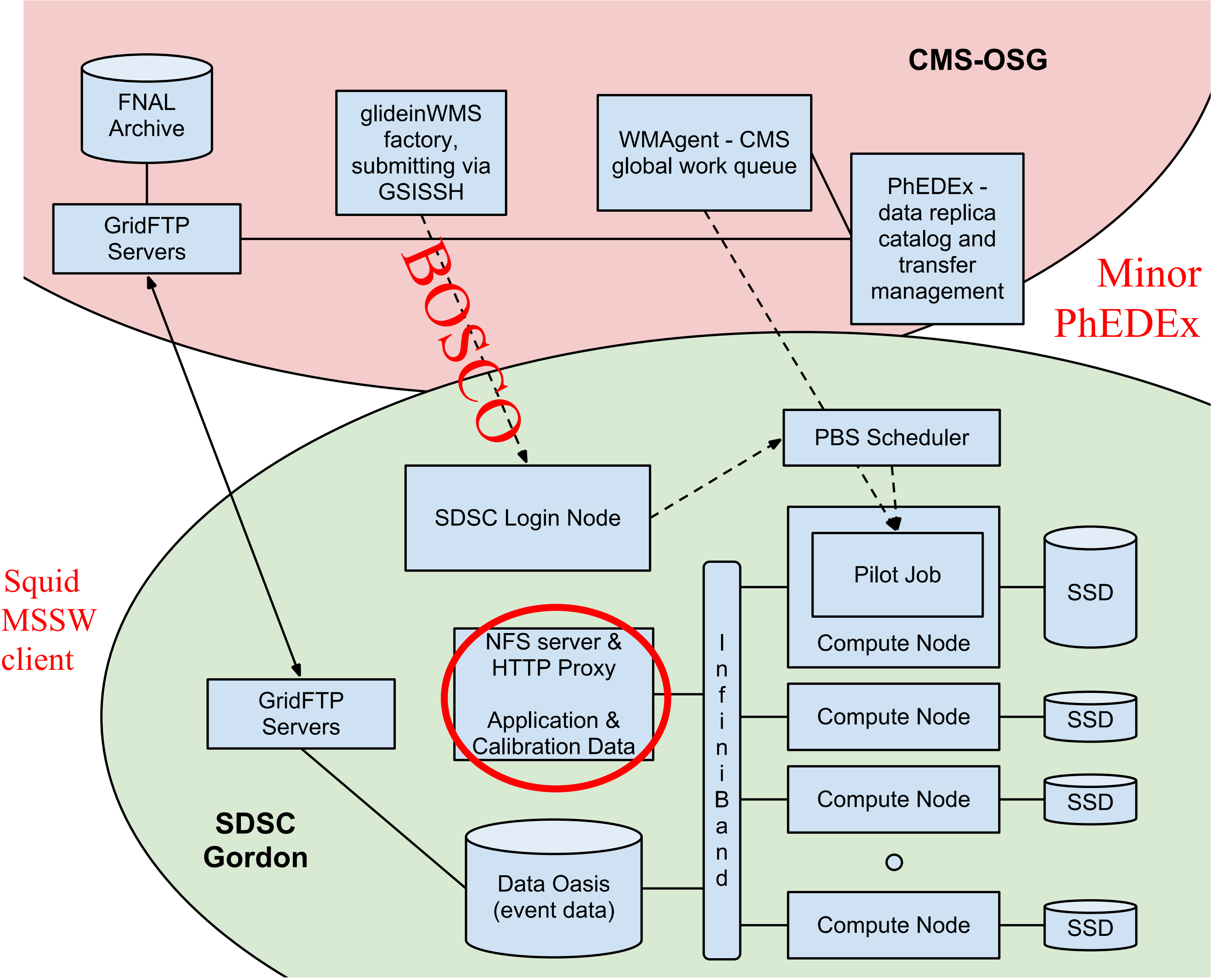
- **Security Context (CA certs, CRLs) via OSG worker node client**
- **CMS calibration data access via FroNTier**
    - B. Blumenfeld *et al*; 2008 *J. Phys.: Conf. Ser.* **119** 072007
  - Squi
- **glidei**
    -
  - Imp g
  - Sul
- **WMA**
- **PhED**
  - Uses SRM and gridftp

So let's focus only on the parts that are specific to incorporating Gordon as a dynamic data processing center.

Job environment

Data and Job handling

Items in red were deployed/modified to incorporate Gordon



CMS-OSG

FNAL Archive

GridFTP Servers

glideinWMS factory, submitting via GSISSH

WMAgent - CMS global work queue

PhEDEx - data replica catalog and transfer management

BOSCO

SDSC Login Node

PBS Scheduler

Pilot Job

Compute Node

SSD

NFS server & HTTP Proxy

Application & Calibration Data

InfiniBand

GridFTP Servers

Data Oasis (event data)

SDSC Gordon

Minor mod of PhEDEx config file

Deploy Squid Export CMSSW & WN client

Results

- Work completed in February to March 2013

- 400 million collision events reconstructed
- 125TB in, ~150 TB out

- Normal Job completion rates

Thoughts & Conclusions

- In a matter of weeks CMS was able to connect to a large opportunistic resource
  - We were able to accelerate the processing of a sample for physics
- A proof of concept moving forward to use diverse resources and augment the capacity at low cost.