(Possible) HEP Use Case for NDN

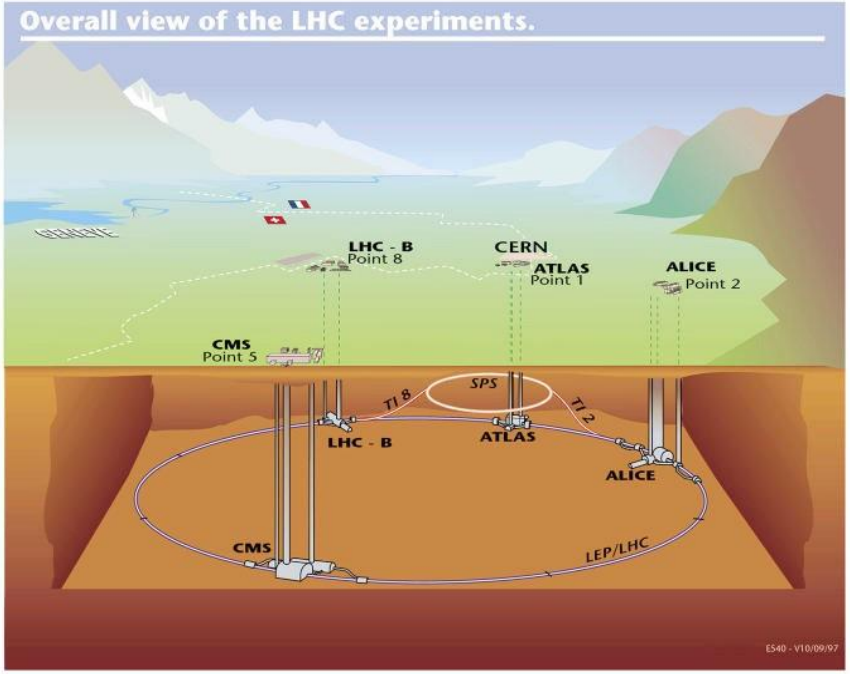
Phil DeMar; Wenji Wu
NDNComm (UCLA)
Sept. 28, 2015

Outline

- LHC Experiments
- LHC Computing Models
- CMS Data Federation & AAA
- Evolving Computing Models & NDN
- Summary

Large Hadron Collider (LHC) 101

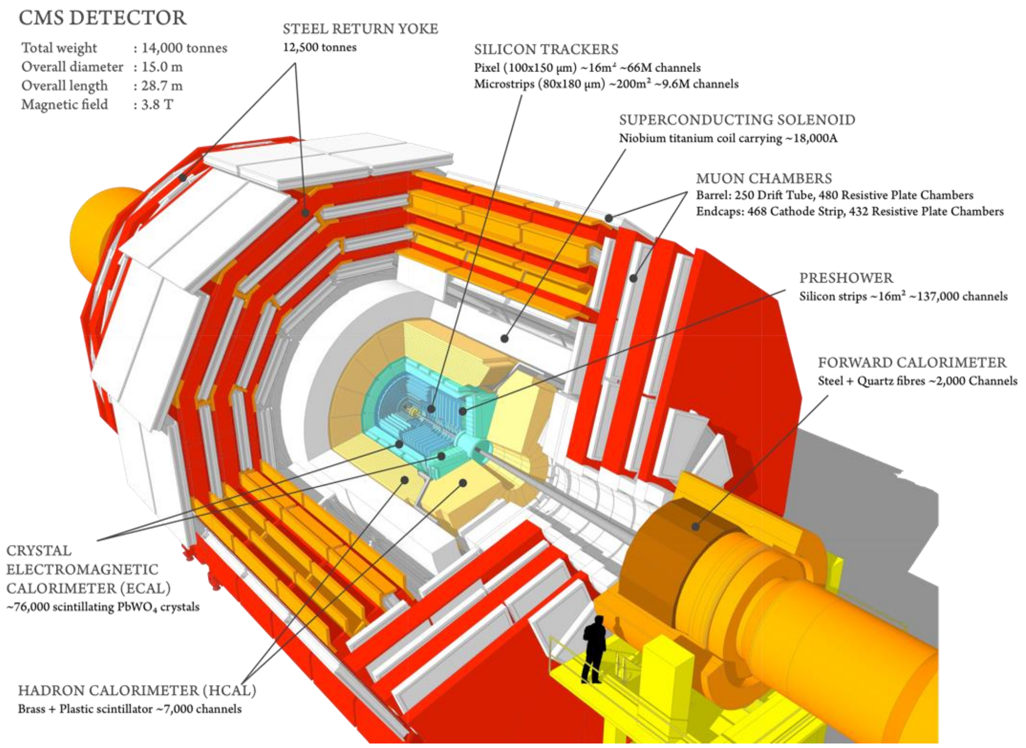
- Circumference: ~ 17 Miles
- 2 proton beams circulating at 99.9999991% speed of light:
- Beams cross and are brought to collision at 4 points:
- Experiments built at those points
  - ATLAS
  - CMS
  - ALICE
  - LHCb



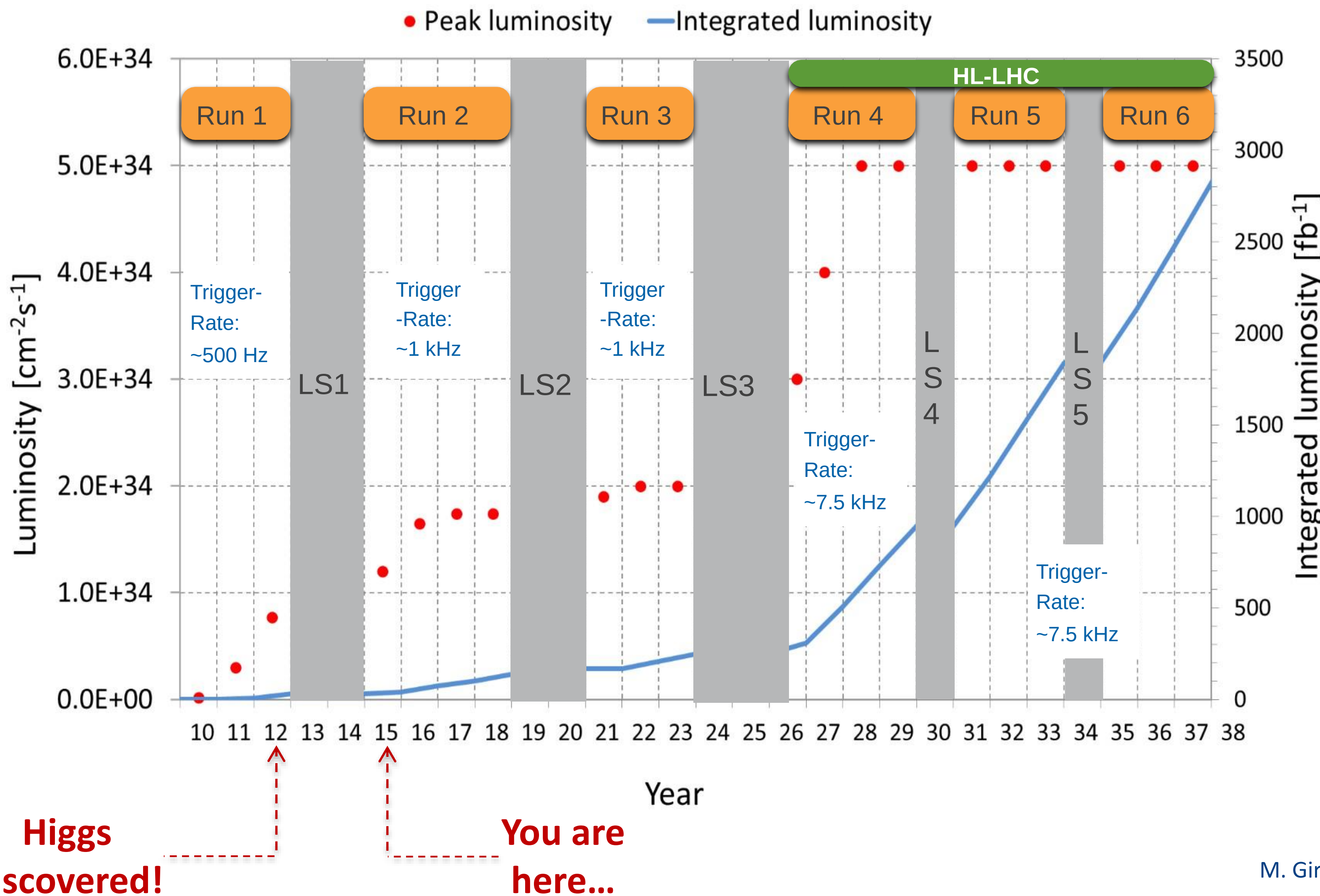
Compact Muon Solenoid (CMS) Experiment

- Detector built around collision point
- Records flight path and energy of all particles produced in a collision
- 100 Million individual measurements (channels)
- All measurements of a collision together are called: **event**

CMS detector



LHC schedule



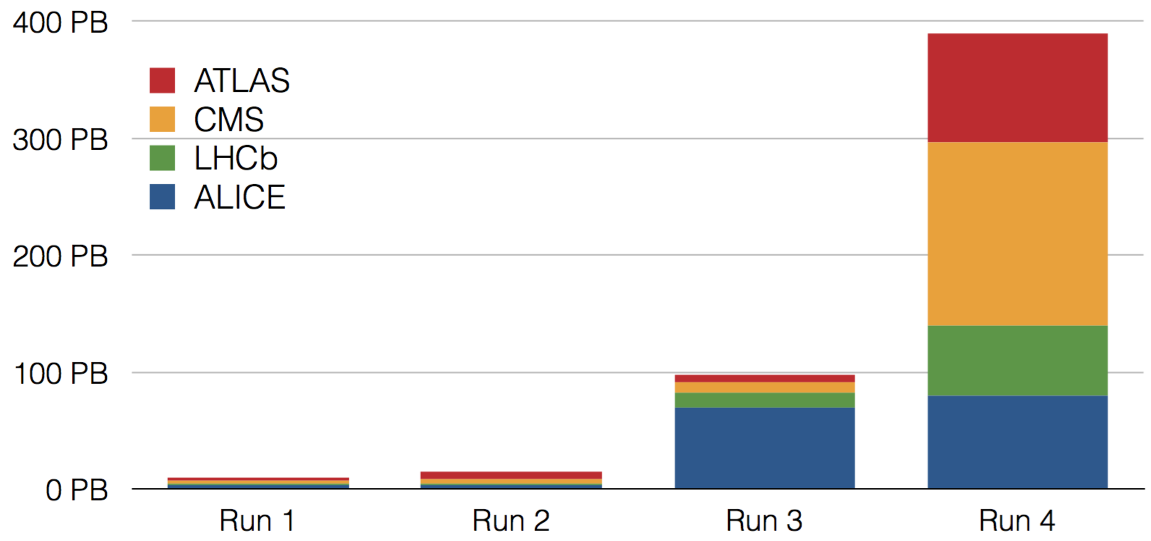
M. Girone (CERN)

Fermilab

Projected LHC data volumes

RAW

M. Girone (CERN)

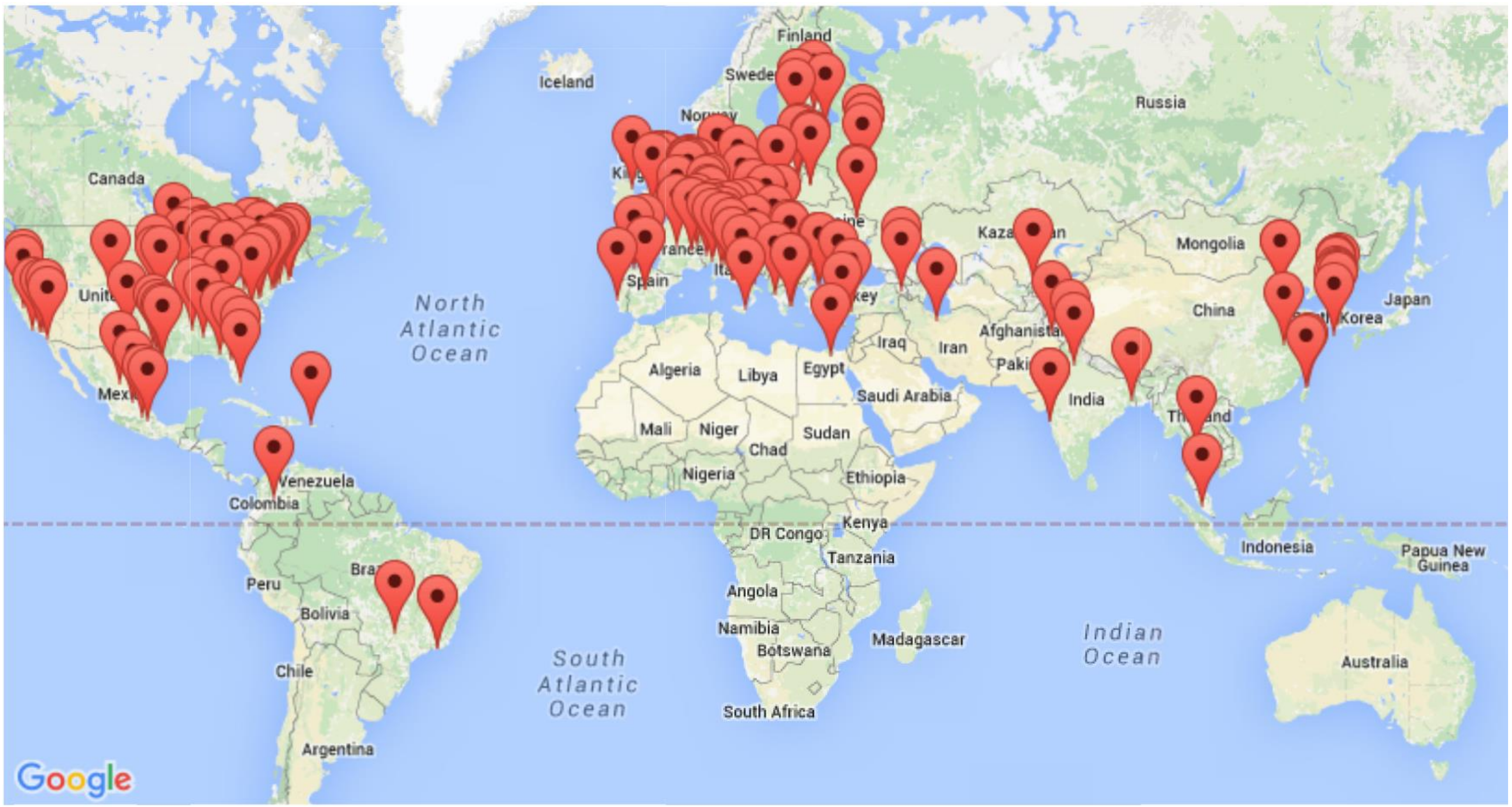


Exabyte era…

- Raw data = generated by detector(s)
- Derived data = reconstructed data, simulation data, summary data sets, etc…)
  - (derived data volumes) ~= (raw data volumes) x 8

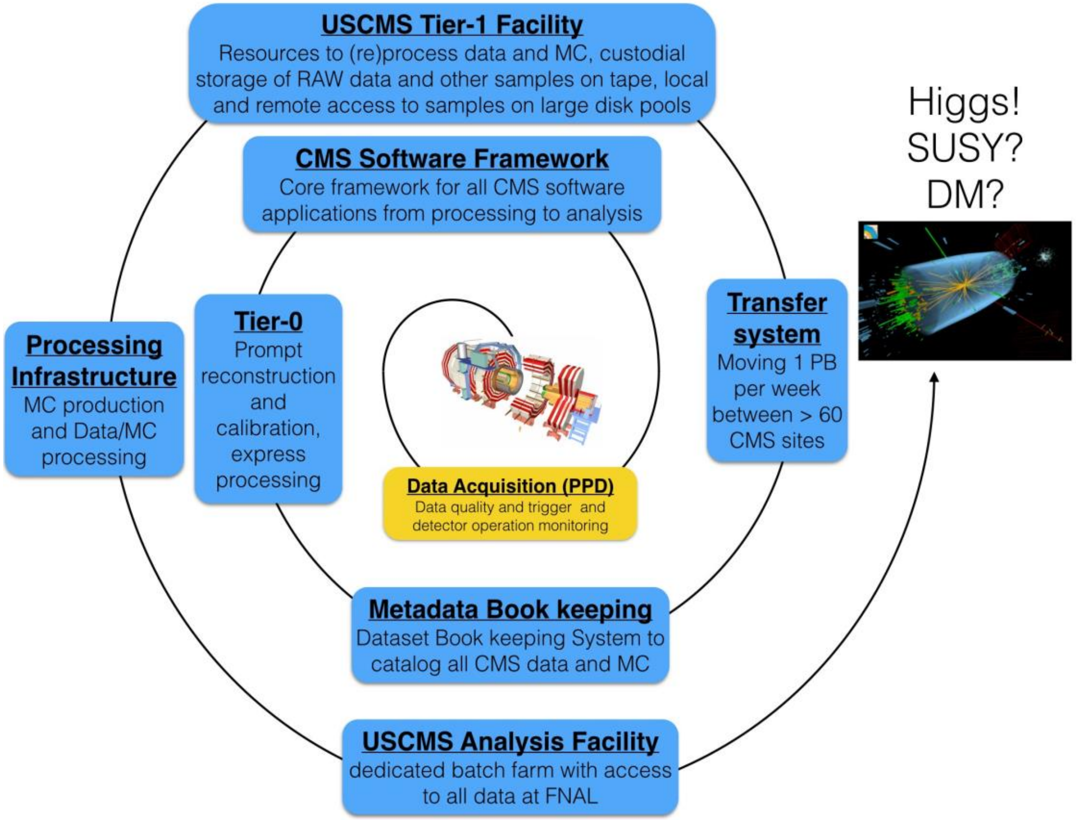
CMS Collaboration

- 186 institutions (globally distributed)
  - High b/w R&E networks support experiment data movement



LHC Computing Models

Computing Lifecycle: CMS

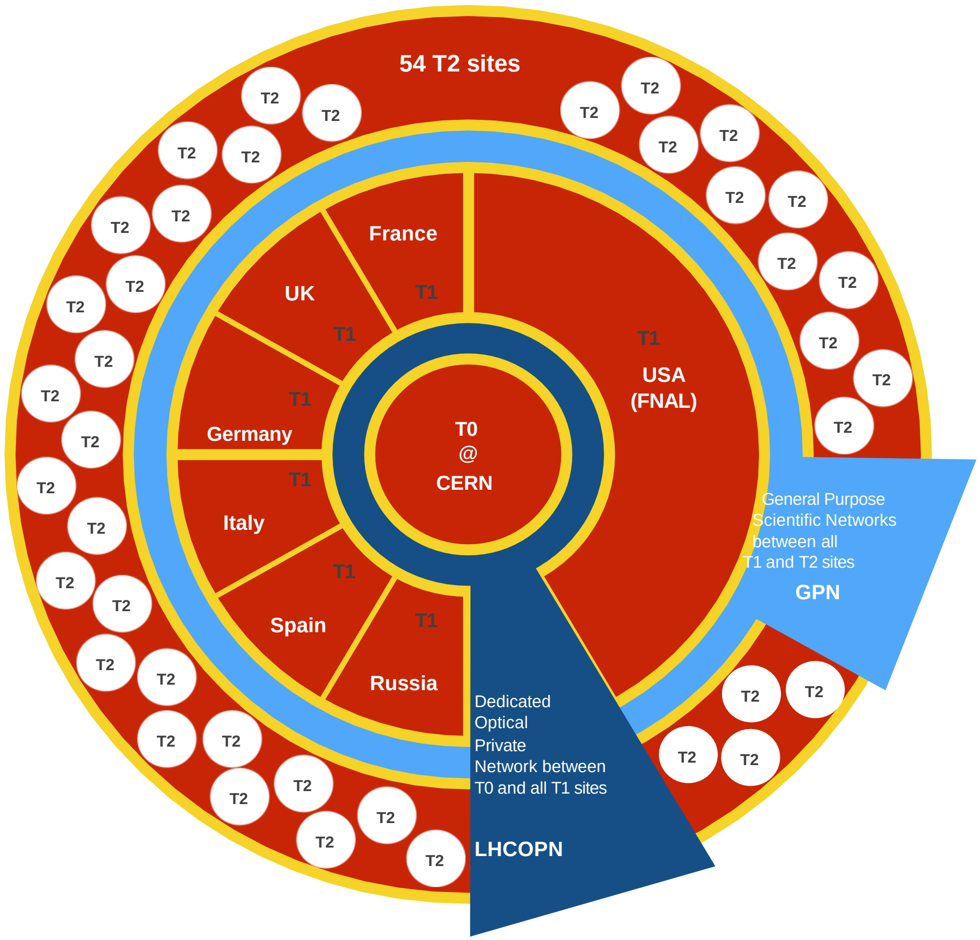


O. Gutsche (FNAL)

- Tier structure for computing (MONARC):
  - Tier 0 = CERN
  - Tier 1 = National data centers for event reconstruction & archiving
  - Tier 2 = Computing facilities for Monte Carlo production & event analysis
  - Tier 3 = Collaboration sites
  - Tier 4 = Physicist desktops

CMS Computing GRID infrastructure

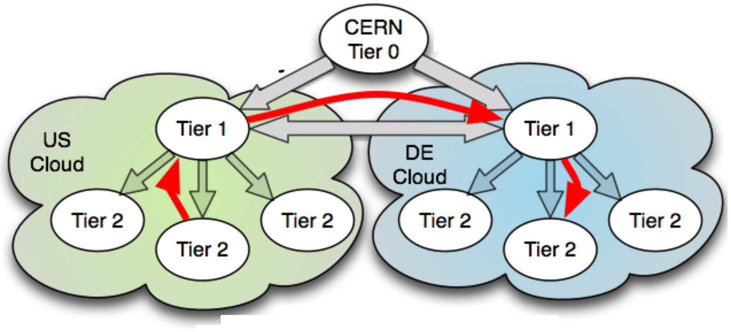
- CERN (T0) at the center
- 7 Tier-1 centers:
  - Connected to T0 by a "dedicated" network
- 54 Tier-2 facilities
  - Connected to T1s by R&E networks
- ~120,000 cores
- ~75PB disk
- ~100PB tape



O. Gutsche (FNAL)

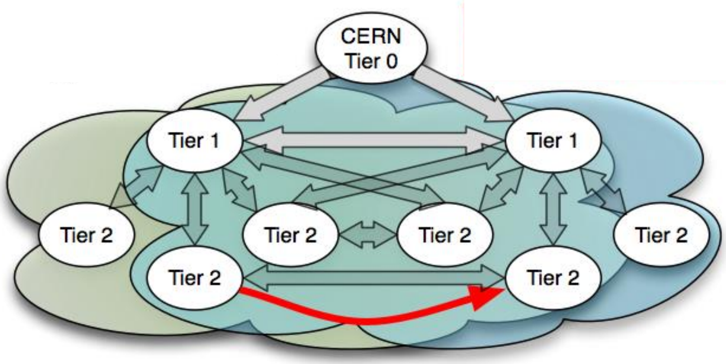
Fermilab

Tier Model for Data Movement Abandoned



- MONARC hierarchical model
- Based on expectation of low b/w & modest storage at T2s
- CMS abandoned MONARC before the LHC even started…
  - ATLAS followed suit during Run I

- Any CMS T1/T2 site could be used as a data source
- Encouraged more flexible data placement & replication
- Enabled more efficient utilization of available resources

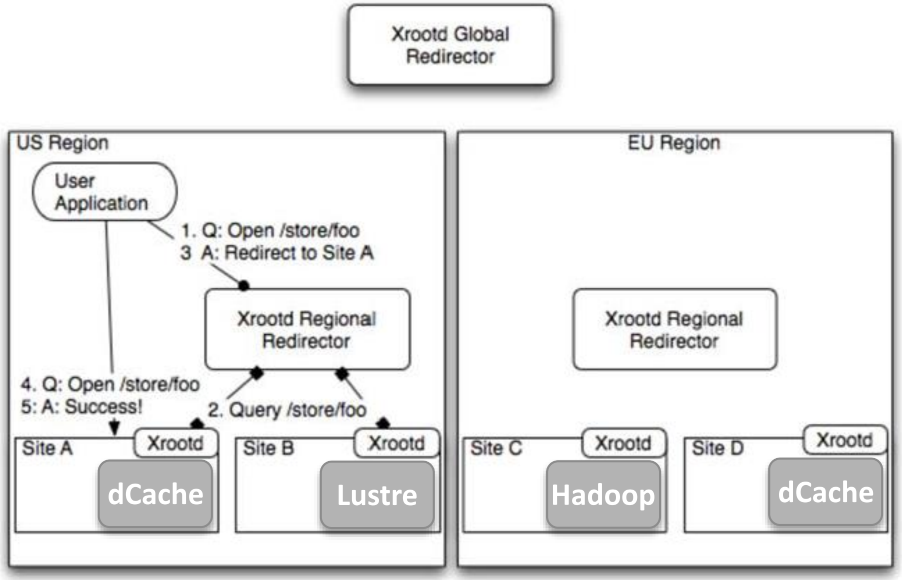


T. Wenaus (BNL)

CMS Data Federation & AAA

Data Federation - XrootD

- LHC experiments have implemented federated data storage, made possible by:
  - High bandwidth WAN connectivity across all tiers
  - Global data namespace(s)
- Based on XrootD:
  - "Hides" local file storage systems
  - Hierarchical, w/ regional, global, & local redirectors
  - Maintains catalog of known file locations
    - Negative cache as well
  - Tree-walk redirects to locate file

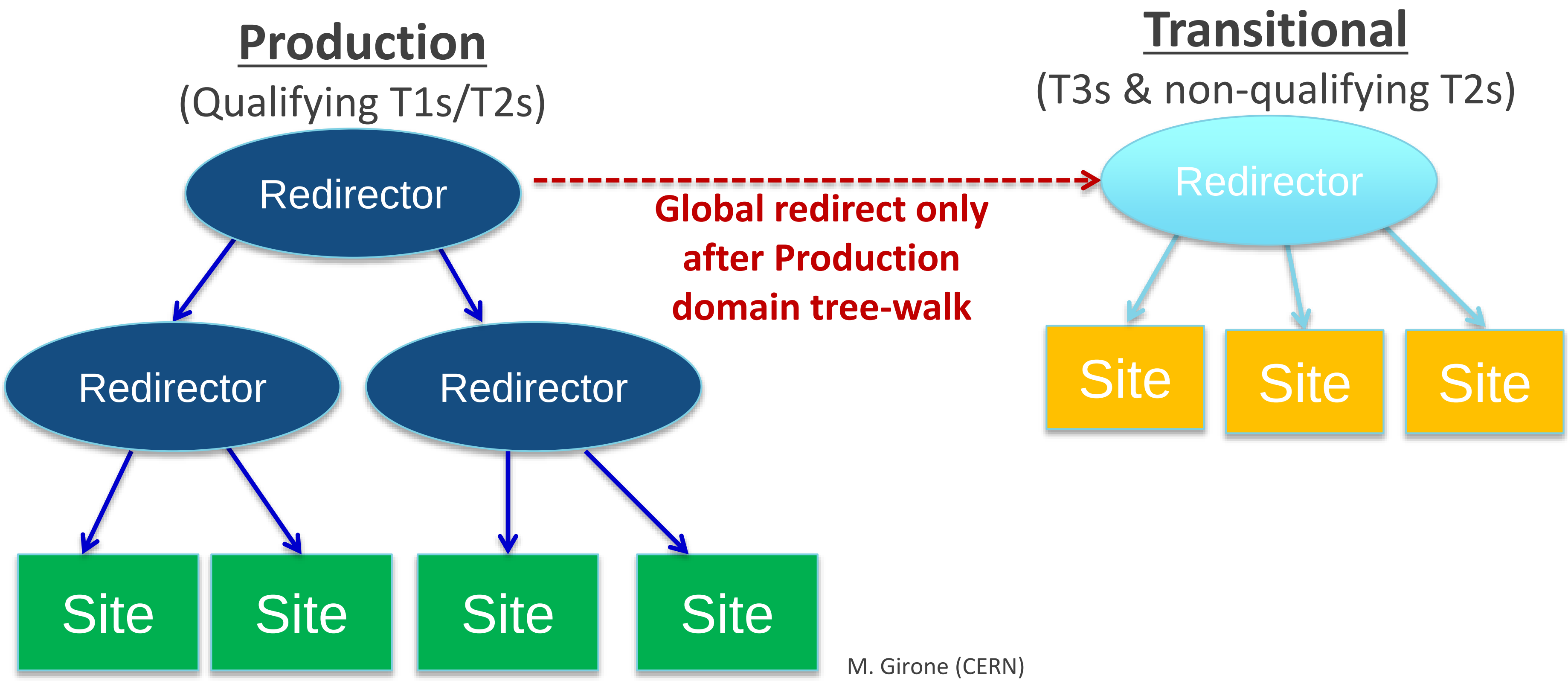


Any Data, Any Time, Anywhere (AAA)

- AAA is CMS's implementation of federated storage:
  - Based on XrootD
  - Finds data based on logical file name
  - Transfers data to application
- High-level philosophy: remote storage ~= local storage:
  - In practice: CPU efficiency slightly lower w/ remote data
- Principally driven by (macro) economics:
  - Maximizes efficiency of collaboration computing resources
  - Fallback data access & overflow job redistribution capabilities
- A few numbers:
  - Nearly all (95%+) CMS data available via AAA
  - Projection is 20%+ of CMS Run II data access through AAA
    - Local storage access is not through AAA…

AAA's Two-domain Federation

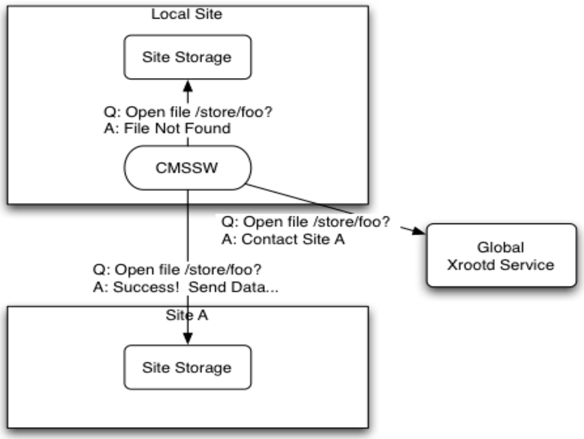
- Production domain for AAA performance-certified sites
  - Transition domain for sites not meeting performance standards
  - All CMS T1s and most T2s are now Production-certified



M. Girone (CERN)

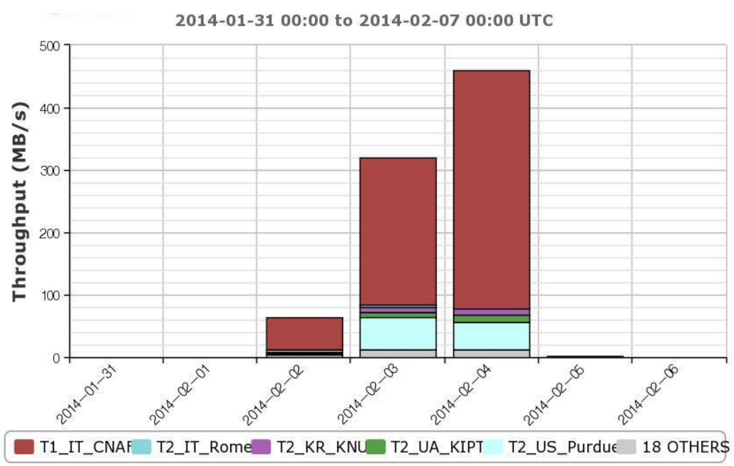
AAA Fallback Mode

- Job unable to access local data:
  - AAA fallback capability locates remote copy of data
  - Job is able to complete…



- Useful in redirecting jobs to other sites in overflow situations

- Real life example:
  - DB error results in “missing” local data at FNAL
  - AAA failover locates replica at CNAF (Italy)
  - Jobs run for 2 days using CNAF data, without anyone noticing…



Evolving Computing Models & NDN

Additional Trends in CMS Computing Model…

- Dynamic data placement (ALICE/ATLAS):
  - Distributing/redistributing (abbreviated) data sets by popularity
  - Subset of larger trend for dynamic data management in general
- Cloud & High Performance Computing (HPC) cycles:
  - Amazon Web Service spot CPU cycles already highly economic
  - Next gen. super computers will have massive computing power

M. Ernst (BNL)

System attributes	NERSC Now	OLCF Now	ALCF Now	NERSC Upgrade	OLCF Upgrade	ALCF Upgrades	
Name Planned Installation	Edison	TITAN	MIRA	Cori 2016	Summit 2017-2018	Theta 2016	Aurora 2018-2019
System peak (PF)	2.6	27	10	> 30	150	>8.5	180
Peak Power (MW)	2	9	4.8	< 3.7	10	1.7	13
Total system memory	357 TB	710TB	768TB	~1 PB DDR4 + High Bandwidth Memory (HBM) +1.5PB persistent memory	> 1.74 PB DDR4 + HBM + 2.8 PB persistent memory	>480 TB DDR4 + High Bandwidth Memory (HBM)	> 7 PB High Bandwidth On-Package Memory Local Memory and Persistent Memory
Node performance (TF)	0.460	1.452	0.204	> 3	> 40	> 3	> 17 times Mira
Node processors	Intel Ivy Bridge	AMD Opteron Nvidia Kepler	64-bit PowerPC A2	Intel Knights Landing many core CPUs Intel Haswell CPU in data partition	Multiple IBM Power9 CPUs & multiple Nvidia Voltas GPUS	Intel Knights Landing Xeon Phi many core CPUs	Knights Hill Xeon Phi many core CPUs
System size (nodes)	5,600 nodes	18,688 nodes	49,152	9,300 nodes 1,900 nodes in data partition	~3,500 nodes	>2,500 nodes	>50,000 nodes
System Interconnect	Aries	Gemini	5D Torus	Aries	Dual Rail EDR-IB	Aries	2nd Generation Intel Omni-Path Architecture
File System	7.6 PB 168 GB/s, Lustre®	32 PB 1 TB/s, Lustre®	26 PB 300 GB/s GPFS™	28 PB 744 GB/s Lustre®	120 PB 1 TB/s GPFS™	10PB, 210 GB/s Lustre initial	150 PB 1 TB/s Lustre®

CMS Computing (today…) vs NDN

Warning!!! My interpretation only!
Subject to large error bars on both ends…

	CMS (today)	NDN
Namespace	Global logical file names	Hierarchical data name space
Content-based data retrieval	Middleware service	Basic network service
Routing optimization	Some architectural & middleware optimizations	Basic network service
Caching optimizations	Middleware optimizations	Basic network service (?) (not clear how this would work with LHC scale data volumes)
Scalable Repository	Open Science Grid Stashcache (middleware) [?]	Repo-Se (?)

But Don’t Confuse Us with NetFlix…

- NetFlix delivers streaming video content to ~20M users
  - Regarded as largest content provider for internet traffic
- CMS much smaller user base & generates only a fraction of NetFlix’s traffic
  - But CMS aggregate amount of data is 1000X NetFlix
  - NetFlix deals with much lower amount of data, which is much easier to efficiently replicate or cache

	NetFlix	CMS
Users	20M	100K
Total Data	20TB	20PB

O. Gutsche (FNAL)

NDN Activities in High Energy Physics (HEP)…

- Climate Data Sciences NDN test bed (C. Papadopoulos, etc.) has ties with HEP community
  - Caltech Network Research group (H. Newman) is involved
- Imperial College London (D. Rand, etc.) evaluating NDN in a local test bed:
  - Application-level (ROOT)
  - Repository-level
- Caltech & FNAL funded to create small NDN test bed for CMS app evaluations

Summary…

- LHC experiments heading toward exascale data volumes:
  - Terabit networks will be needed to handle that data
- LHC computing models are becoming increasingly distributed in nature:
  - Both data storage & CPU
  - This creates greater demands on network services beyond b/w
- LHC computing is already implementing content-based data services at the middleware level
- There seems to be a natural fit for NDN with LHC computing:
  - Performance optimizations within the exascale data / terabit network environment will be key

Questions?