Fermilab

Managed by Fermi Research Alliance, LLC for the U.S. Department of Energy Office of Science

Computing Facilities and Services

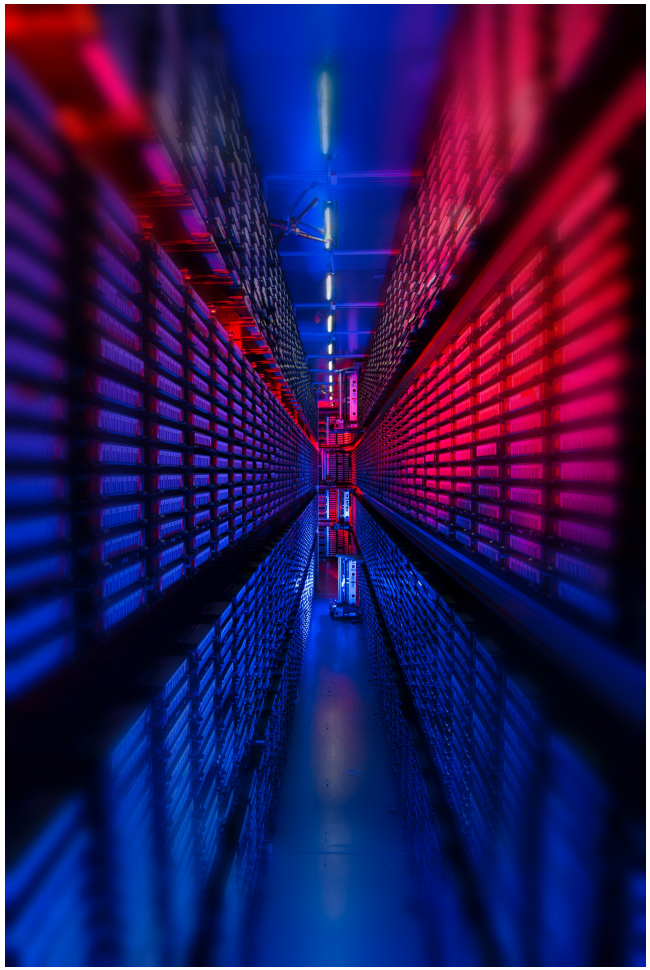
Oliver Gutsche

Fermilab 2015 Institutional Review

10-13 February 2015

Overview

- Scientific Computing
- Facilities
- Services
- Future
- Conclusions



- Apologies if too many acronyms are present in the talk
  - There is a glossary in the backup slides

Scientific Computing

Scientific Computing

- Deliver world-class scientific computing services, operations and software engineering

- Support:
  - NoVA, MicroBooNE, MINOS, Mu2e, Muon g-2, DES, and other Fermilab experiments and projects
  - CMS
  - High-energy physics community at large

- Continue to build out activities for supporting software and computing needs
  - Includes R&D activities to maintain or advance capabilities

Three crosscut areas in Scientific Computing

- We have to leverage effort and resources wisely
  - Develop and maintain core expertise, toolkits and services
  - Build and provide operational support to applications

- We have recently reorganized scientific computing in the Scientific Computing Division (SCD) to help maximize leveraging of expertise and best-of-class tools from partnerships with individual projects

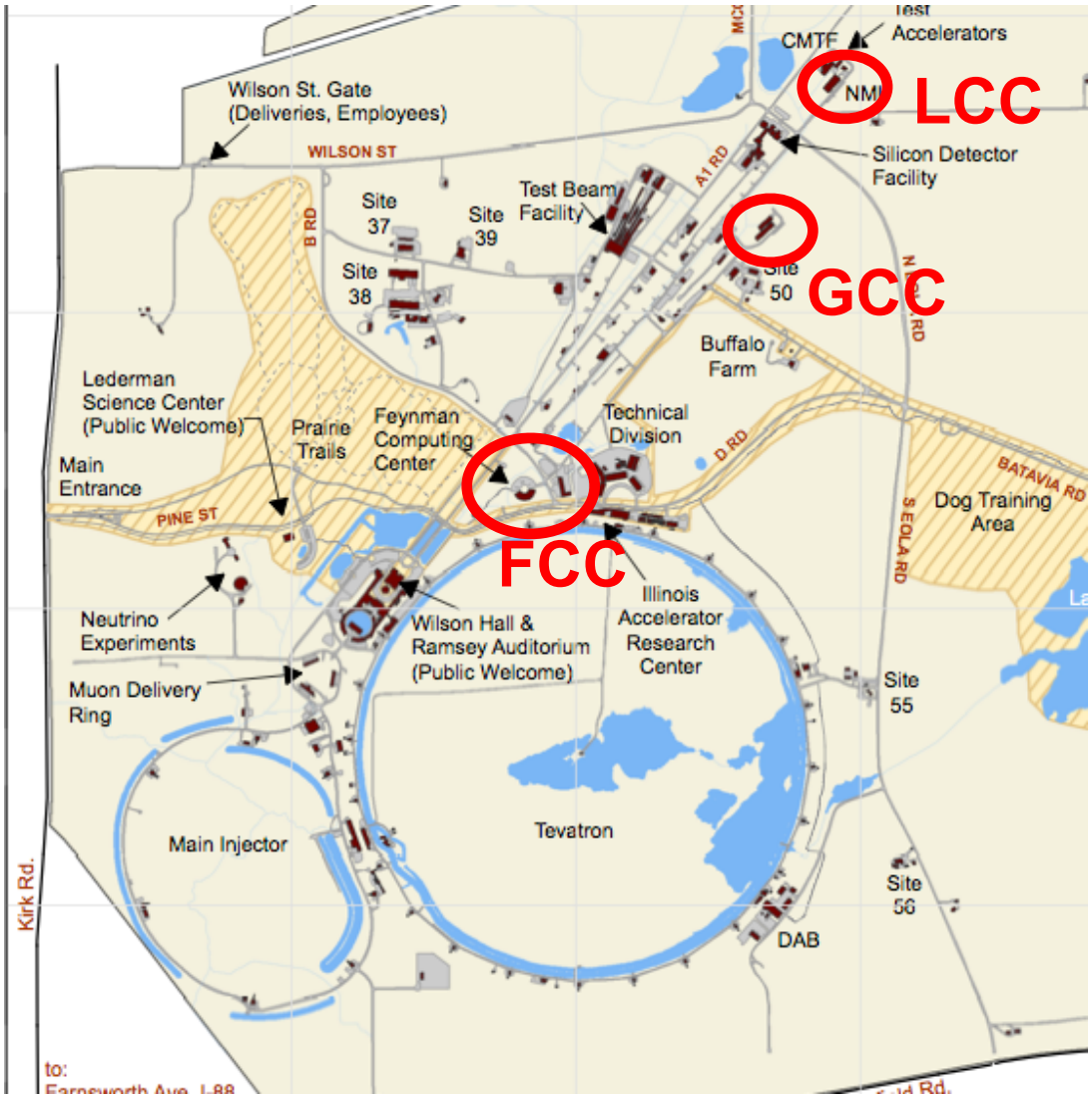
  - We aligned our activities across three crosscut areas.
    - Development, Integration and Research
    - Facilities
    - Science Operations and Workflows

Facilities

Computing Facilities on the Fermilab site

3 main computer room locations

- **Feynman Computing Center (FCC)**
  - 2 rooms with 0.75 MW nominal cooling and electrical power each
    - UPS with independent generator backup
  - Hosts power critical services
    - central services (Mail, web servers, etc.) and disk servers
- **Grid Computing Center (GCC)**
  - 3 rooms with 0.90 MW nominal power each
    - UPS with taps for external generators (no permanent generator)
  - Hosts CPUs and Tape libraries, UPS sustains power during power outages till systems can be powered down in a controlled way
- **Lattice Computing Center (LCC)**
  - 1 room with 0.47 MW nominal power
    - No UPS and no generator
  - Houses CPU and GPU clusters running less power-critical applications



Fermilab

Facility Capacity

CPUs

- **~75k cores**
  - ⅓ Lattice QCD – ⅔ Experiments
    - Difference: Lattice QCD clusters are interconnected with specialized low latency network links, experiment clusters are interconnected with standard Ethernet
- Sizable GPGPU infrastructure and Intel Phi test cluster
  - GPGPU clusters with 76x 8-core NVidia Tesla M2050 GPU nodes and 32 nVidia K40 GPU nodes
  - Intel Phi (Specialized mathematical Co-Processor) test clusters
- Locations: GCC, LCC

Tapes

- **7 robots** with **total capacity of 70k tape cartridges**
- **Total data on tape is 53 PB**
  - dominated by energy frontier experiments,
  - increasing at ~20 PB/yr
  - Additional 33 PB on old tape media, has already been migrated, retained as second copy
  - Average data transfer to/from tape over 1.5 petabyte per month (50 terabytes per day), reached 6 petabytes per month during previous LHC running period including migration
- Locations: GCC, FCC

Disks

- Disk-servers/disk-array managed by Mass Storage Systems (like dCache and EOS);
  Total capacity: 26 PB
- BlueArc Network Attached Storage system;
  Total capacity: 2 PB
- Lattice QCD Lustre installation;
  Total capacity: 1 PB
- Location: FCC

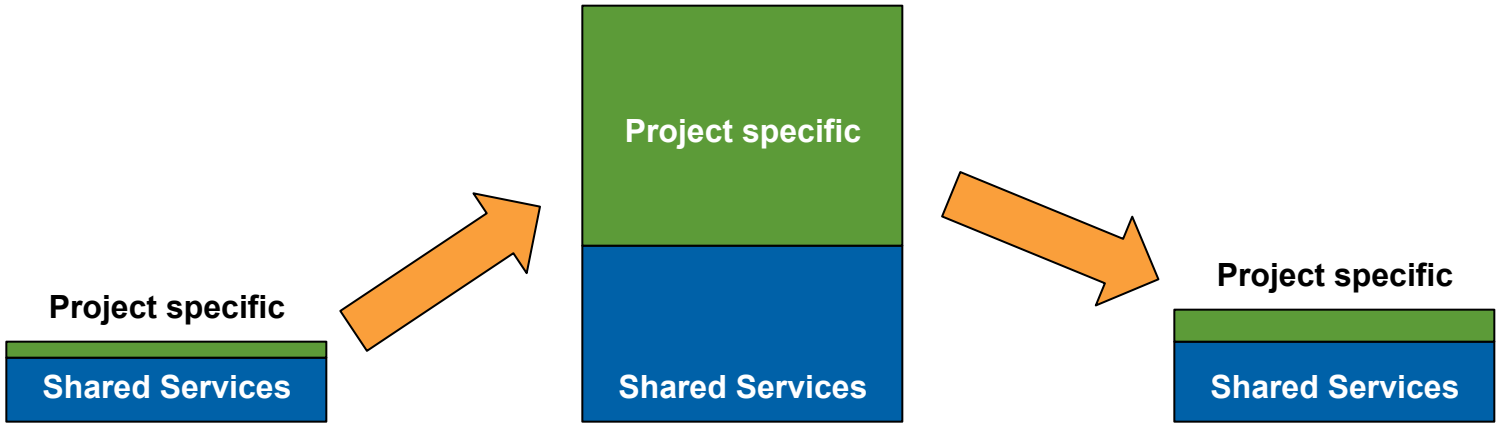
Network

- Internal connections:
  - **~32,000 network ports** connected through network fabric from copper and fiber links
  - **intra-datacenter bandwidth of 1.2 Tb/s** by end of Q1 CY15
- External connection:
  - **130 Gb/s** from from 1x 100 Gb/s and 3x 10 Gb/s fiber links
  - 2nd 100 Gb/s connection for network research will be added in 2015

Service Operations and Workflows

Experiment Support – Fermilab approach

- Fermilab's computing facilities use a shared services model to ensure that all experiments, small and large, are able to make use of the facilities from **"cradle to grave"**



Early Period

R&D, Ideas, Simulations
LOI, TDR, Proposals

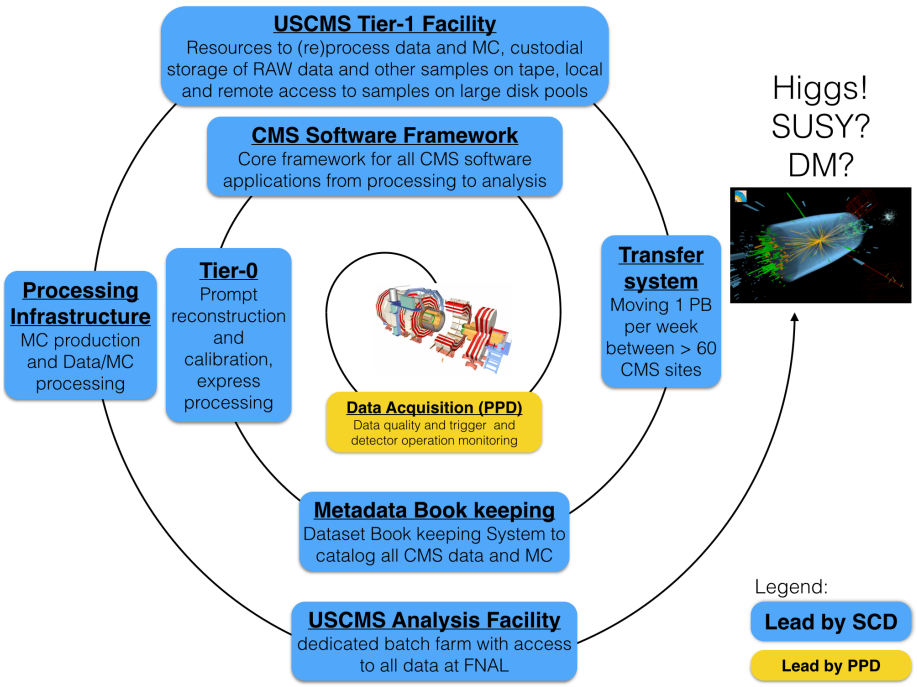
Mature phase

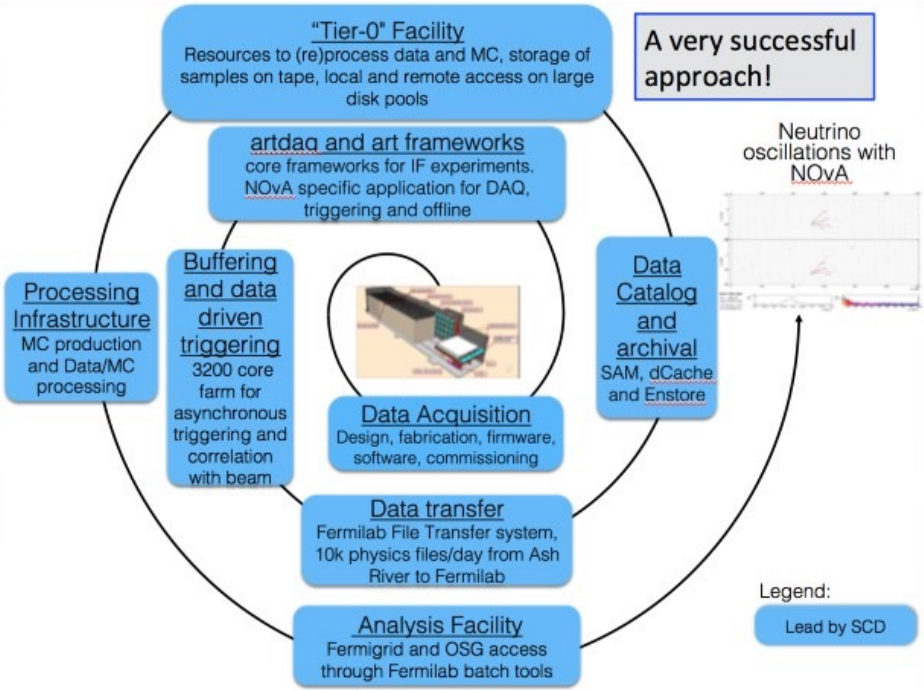
Construction,
operations, analysis

Final data-taking and beyond

Final analysis,
data preservation
and access

Experiments





- Software and computing are needed at all steps from data collection to physics results of any experiment
- Performant and scalable solutions are required to extract physics results quickly and efficiently
- Fermilab is a world leader providing solutions for experiments, we concentrate on:
  - CMS, NoVA, MicroBooNE, MINOS, Mu2e, Muon g-2, DES, and other Fermilab experiments and projects and HEP community at large

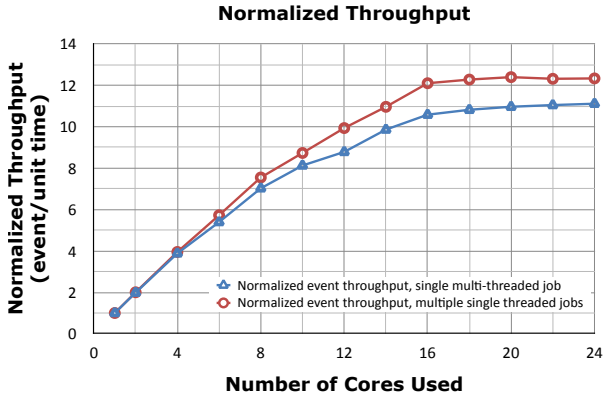
CMS: Software development

- Fermilab: **maintaining and evolving the CMS software and computing infrastructure for LHC Run 2 and beyond**

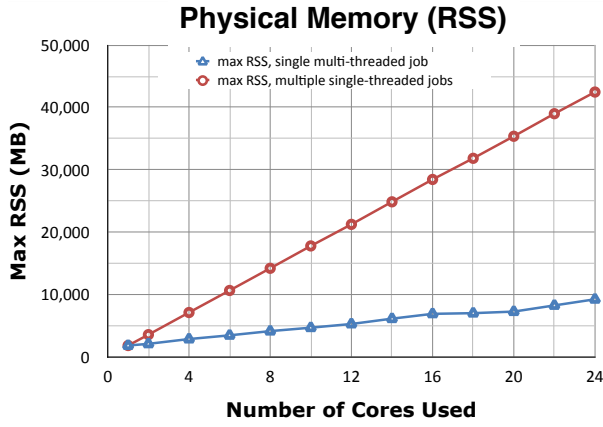
- **CMS software framework**: basis of all data and Monte Carlo production and processing both online and offline, and analysis
  - Fermilab software experts lead and are at the heart of the development team
  - In 2014, significant milestone was reached: multithreaded framework mode to process multiple events simultaneously on several cores

- **Simulation**
  - CMS benefits from the large Fermilab Geant4 development and support team
  - **Cross cutting activity: CMS HCAL group in PPD works with Geant4 group in SCD**
  - As of 2015, Fermilab leads phase 2 upgrade simulation efforts

- Fermilab experts are leading the **development of the most critical computing tools** for the CMS collaboration:
  - Tier-0 and central processing infrastructure and the metadata catalog

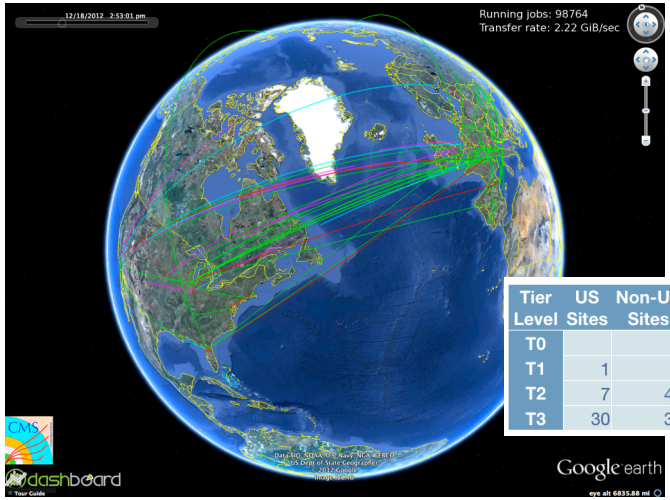


95% efficiency compared to single-threaded jobs at significant memory savings

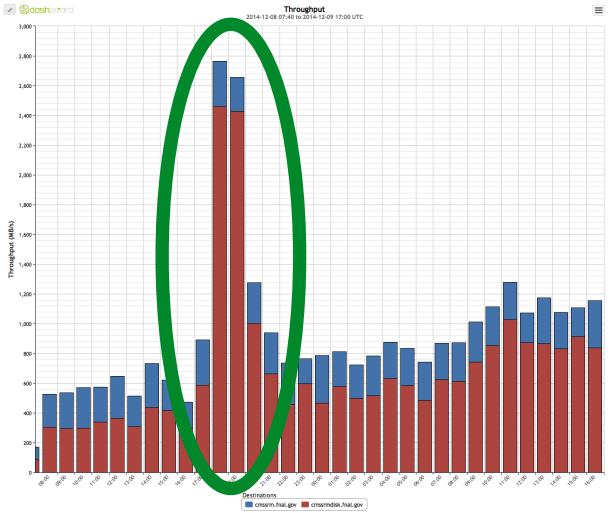


CMS: Facility

- **Largest Tier-1 site for CMS**
  - 40% of the resources of the “global” Tier-1 level
    - FNAL: ~11,000 cores
  - Archives 40% of CMS data and simulation files on tape
    - FNAL: 22 PB
  - Large disk cache for efficient access to files
    - FNAL: 11 PB
  - Strong network connection to all of more than 60 CMS sites worldwide
    - FNAL: 80 Gbps
- Additional resources are available for US CMS use, including the **Analysis Facility for the LPC**
  - Interactive login cluster for all 700 US CMS physicists and also international colleagues
  - Access to all files at the FNAL Tier-1 as well as own disk space
    - FNAL User Facility: 5 PB
  - Large analysis cluster:
    - FNAL User Facility: ~5,000 cores



Tier Level	US Sites	Non-US Sites	Total Sites
T0		1	1
T1	1	6	7
T2	7	45	52
T3	30	33	63

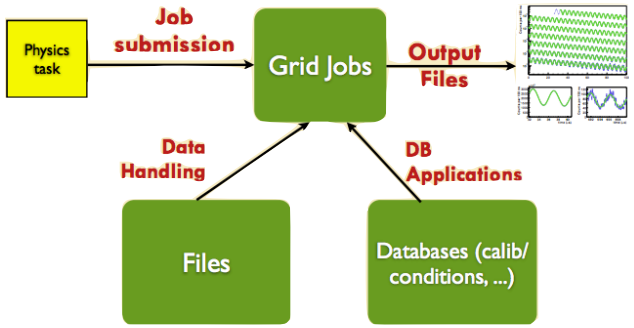


Transfer capabilities from CERN to FNAL: reached 3 GB/s for couple of hours

Fermilab

NoVA, MicroBooNE, MINOS, Mu2e, Muon g-2, DES, …

- **Approach:**
  - Support offline & computing needs of experiments
    - We provide standard interfaces into Fermilab computing resources (CPU, disk, tape)
    - We enable science through a modular toolkit of services and applications



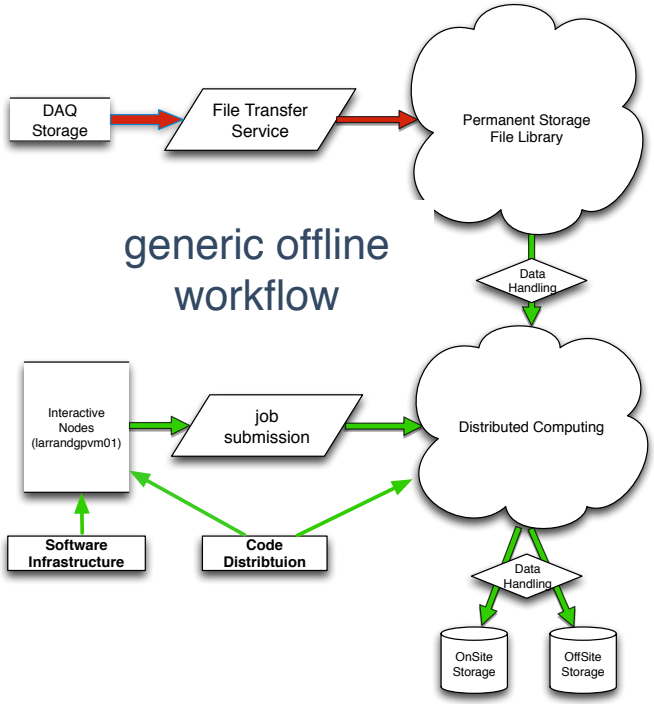
- **Goal**:
  - Help experiments focus on their science
    - Provide infrastructure to experiments that get them to computing resources
    - Provide software framework solution and support community-based reconstruction software
  - Utilize previous solutions and integrate everything into seamless model

- **FIFE project - Strategy**:
  - Address all of the computing needs for experiments
    - Modular enough so that experiments can take what they need
    - Spans all cross cuts, in particular the Neutrino, Muon and Cosmic cross cut
  - Provide mechanism for feedback from experiments to incorporate their tools and solutions
  - Help experiments utilize computing beyond the Fermilab campus
  - Integrate new tools and resources from outside Fermilab and other communities as they develop
    - **Synergy with CMS especially important**

Common Services and Projects toolkit

- FIFE provides access and support for common tools in:

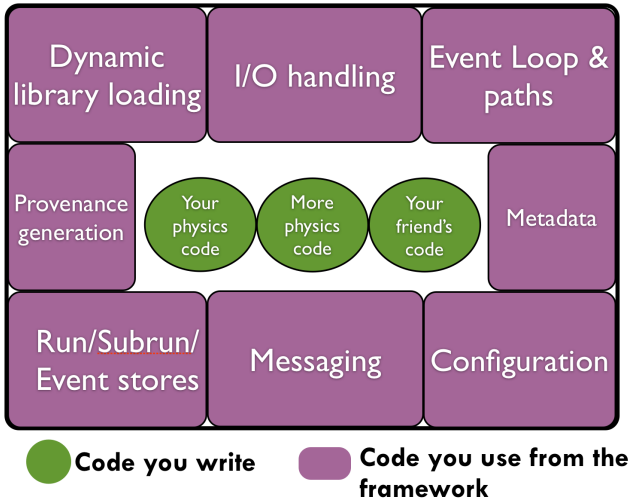
  - **DAQ and Controls**
    - DAQ Status, DAQ Framework
  - **Grid and Cloud**
    - Fermigrid, Fermicloud, Gratia, Job Sub, FIFEMON, OSG, Amazon Web Services (AWS)
  - **Scientific Data Storage and Access**
    - dcache/ enstore, Gridftp
  - **Scientific Data Management**
    - IFDH, SAM Web, F-FTS
  - **Scientific Frameworks and Software**
    - Art, Larsoft
  - **Physics and detector simulation**
    - Genie, Geant4
  - **Databases**
    - IF Beam, Custom Databases, mysql, postgres
  - **Scientific Computing Systems**
    - CVMFS, Interactive machine in GPCF, Experiment Control Rooms
  - **Scientific Collaboration Tools**
    - Redmine, CVS/ Subversion/ Git, ECL



Fermilab

art & Larsoft

- ***art* is a modular event-processing framework for experiments**
  - Fork of the CMS framework: smaller experiments can base their software on a framework from one of the big LHC experiments
- **Many experiments are using *art* as their software framework**
  - This allows shared development and support among the experiments and the developers
    - Weekly stakeholder meetings, mailing list and issue trackers are used to coordinate development
  - Integration into Fermilab's data-handling system, I/O, etc. is an important part of the *art* project
- ***art-daq* is variant of the framework used as part of the daq system**



- **Larsoft**
  - **Technical goal:** Provide an integrated, *art*-based, experiment-agnostic set of software tools to be used by multiple LArTPC neutrino experiments to perform:
    - Simulation, Data reconstruction and Analysis
  - **Broader goal:** By developing common algorithms, services, data structures, architecture, dramatically reduce the cost of developing, maintaining and supporting the reconstruction and simulation software for collaboration members
  - **Community is collectively contributing to the larsoft software**
    - Resolve conflicts through process to integrate development from different stakeholders

Security

- Fermilab is invested to **provide a secure environment for science and continuously strives to remove roadblocks for scientists**, recent achievements:

  - **Easily accessing Fermilab terminals from mobile devices without Kerberos credentials**
    - Instead of Kerberos, Fermilab now collaborates with RSA and distributes an app that works on mobile phones and tablets. With the app, a user can obtain a one time password and login to Fermilab servers.
    - **How it helps users: They can log into Fermilab servers from their phones without setting up Kerberos**
  - **Online, automated certificates.**
    - Fermilab now generates online, automated certificates for their users. Everyone with access to a browser and Fermilab services username and password can get certificates. Entire process takes a minute or so and resulting certificate can be imported into the browser easily. Fermilab websites, such as leave request system, is in works to allow access with these certificates
    - **How it helps users: All steps taken in a browser and whole process takes 1-2 minutes**
  - **Running grid jobs without end user certificates.**
    - Users submitting jobs through GlideinWMS systems no longer needs to have an end user certificate to run jobs. The GlideinWMS will trace the job and obtain necessary permissions to run the jobs. The user only needs to ssh into a submit node to send their jobs
    - **How it helps users: Users will not need certificates to submit jobs.** They only need to ssh into the GlideinWMS system to submit their jobs. The system will take care of the access control management.
  - **Debugging Failing jobs on the worker nodes by using condor-ssh**
    - CMS and Minos requested to access failing grid jobs on the worker nodes via condor-ssh. Normally Fermilab only allows kerberized ssh access to its resources. However, after reviewing condor-ssh, this request was approved with some restrictions.
    - **How it help users: Experts and operations teams can diagnose problems quicker and fewer jobs fail.**

User support

- About **4,000 scientists worldwide use Fermilab** and its particle accelerators, detectors and computers for their research. About **2,000 researchers from 34 countries collaborate on Fermilab experiments**.
  - On-site users (users who at some point were on the Fermilab site and still have a valid badge): 1,800
  - Additional 2,200 users interact with Fermilab from off-site and have registered Fermilab computing accounts

- **Interactive access**
  - Fermilab provides interactive access to the LPC Analysis Facility for US CMS and collaborating scientists
    - These facilities are accessible from inside and outside of Fermilab and serve the US community at large.
  - Collaborators of NoVA, MicroBooNE, MINOS, Mu2e, Muon g-2, DES, and other experiments and frontiers are provided with accounts and can use Fermilab's interactive facilities

- **Fermilab's resources are on the OpenScience GRID**
  - CPUs are accessible through the GRID interfaces
  - OSG Virtual Organizations of non-FNAL experiments can run on our resources opportunistically

Training

- **User support and training for members of the community are very important to Fermilab**

  - CMS is conducting regular Data Analysis Schools for young scientists to not only learn physics tools but also get to know the computing tools necessary to succeed in analysis
  - *Art* has an excellent workbook that is praised by the community for its usefulness and conducts regular trainings
  - Recently we conducted day-long trainings on how to use the FIFE GRID job submission tool
  - We conduct a yearly FIFE workshop to inform the community and foster information exchange
  - Re-started offering a C++ course with emphasis on common analysis tools
  - Provide training on accelerator modeling tools in the context of USPAS
  - Specialized trainings help users and experts to stay informed about latest technologies and techniques
    - example: Parallel Programming and Optimization with Intel Xeon Phi Coprocessors

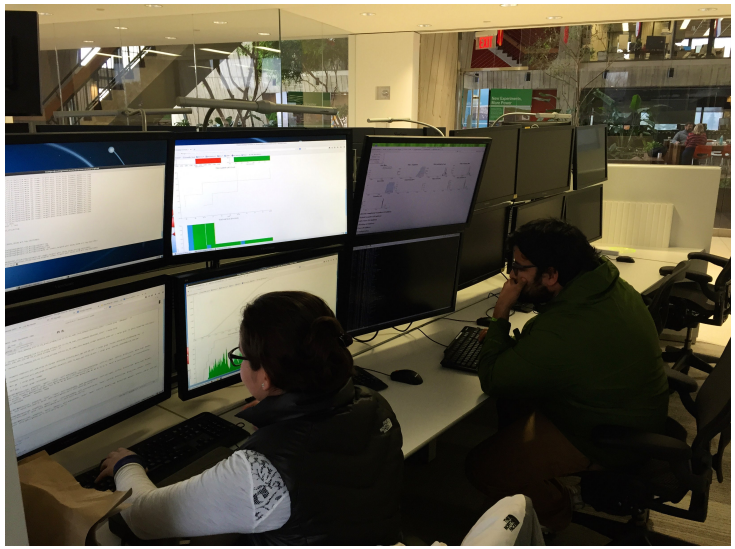
Operations groups

- **Operating the complex computing services requires significant knowledge and experience**
  - Especially valid for job execution at high scales on various resources

- **Fermilab provides operation groups to aide in operating central workflows**

  - CMS: Fermilab experts are involved in the complex operation of the different and diverse components and workflows, from processing/production to maintaining the petabyte-scale transfer system to keeping the over 60 GRID sites operational

  - **New**: Fermilab founded the operations group to support NOvA, MINOS and Minerva (experiments that are currently taking data), helping in running crucial workflows like data keep-up processing and MC production

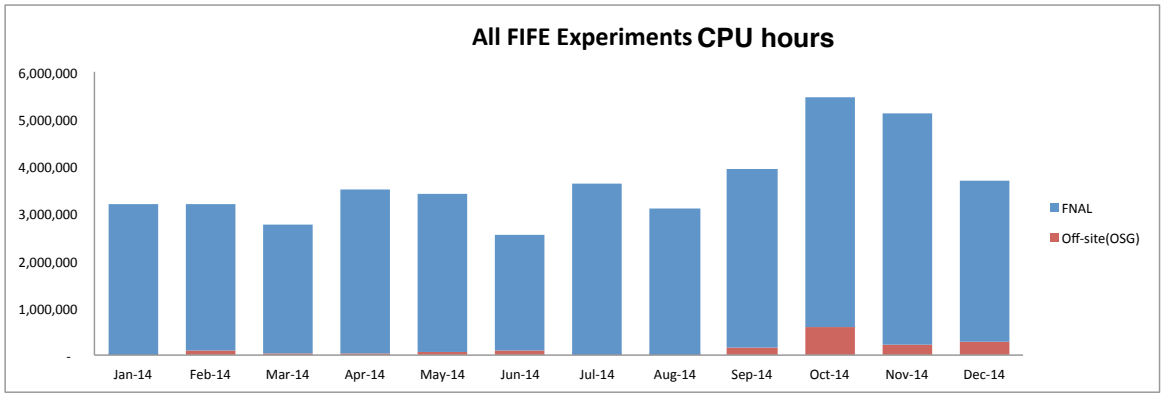


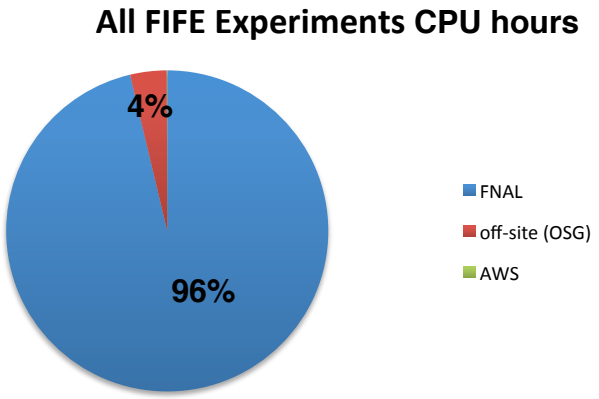
Supported experiments and software

- **Collider experiments**
  - CMS
  - CDF
  - D0
- **Neutrino experiments**
  - ArgoNeut
  - LAr1-ND
  - LArIAT
  - LBNE
  - MicroBooNE
  - MINERvA
  - MiniBoone
  - MINOS
  - NOvA
  - NUMI
  - SNO+
- **Astroparticle/Cosmology Experiments**
  - CDMS
  - COUPP
  - DAMIC
  - Dark Energy Survey
  - DarkSide-50
  - DESI
  - Holometer
  - LSST
  - SDSS
- **Flavor experiments**
  - MIPP
  - Mu2e
  - Muon g-2
  - SeaQuest
- **Software**
  - Geant4
  - GENIE
  - MARS
- **R&D**
  - CHIPS
  - GENDEtRD
  - MCDRD
  - Next
  - SCENE

Full overview of experiments and which services they use available here: https://indico.fnal.gov/getFile.py/access?resId=3&materialId=4&confId=9353

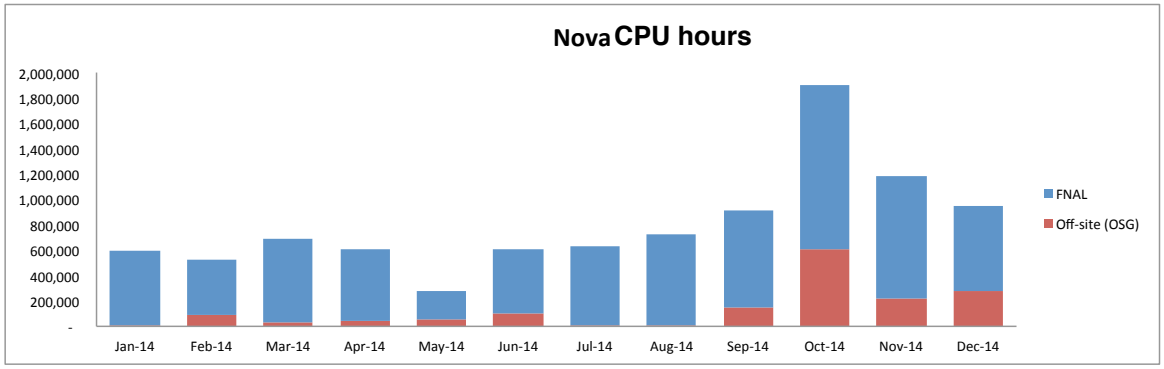
FIFE Experiments - Activity in 2014

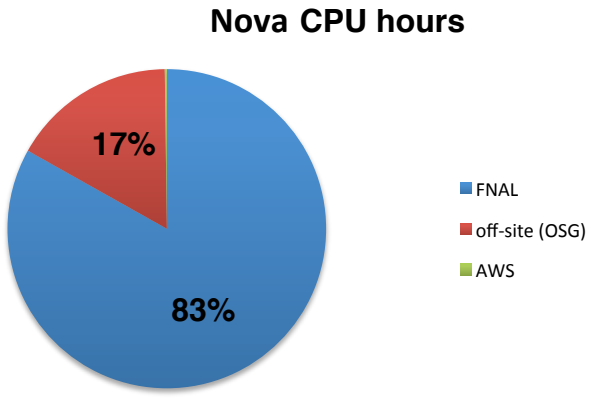




Off-site utilization is ramping up

In 2014 of 45M CPU hours, almost 2M hours were run off-site including first Amazon Web Services (AWS) usage of 20k CPU hours worth $3.3k as a pilot project





Future

Active archival facility

- Fermilab provides **world-class scientific data management capabilities** developed by the High Energy Physics community
  - The **active archival facility** provides these services to other science activities in the US, allowing them to preserve the integrity and availability of important and irreplaceable scientific data.
  - The benefit to Fermilab and DOE is to sustain expertise and experience that is "best of class" in the field of scientific data management, in support of ongoing and future scientific missions of the laboratory
  - The “active archive infrastructure” technologies utilize the unique wide–area transfer protocols and cached storage systems at Fermilab

  - Example done as demonstration of capability:
    - Data management and access for a community of genomic researchers
    - Recently imported ~300TB of data primarily from Amazon S3
      - 2 100TB samples were exported from FNAL using GridFTP to Iceland and Oregon for additional processing
      - The community created and made publicly available 11TB of diversity project data
        - 300 people from all over the planet

Services

- We are continuing to **document and formalize our relationships with the experiments**
  - **Very successful for both experiment and scientific computing at Fermilab**
  - Clarifies expectations on services and facilities on both sides
  - We are working with the liaisons to write SLAs and TSWs
  - On-boarding of services onto ITIL is progressing well

- **Monitoring** on both service provider and user level will be a **focus area**
  - Providing easy to use and easily understandable monitoring for the user community is a continuation of the successful documentation and training program
  - Improving service provider monitoring will help us increase the efficiency of providing resources and services

- Recently, we completed **enabling major experiments to run their central workflows on GRID resources** without the need for specialized resource setups
  - Enables the experiments to benefit from opportunistic resources on the OSG and other facilities without having to change their workflows
  - This will become a necessity when we move to the virtual facility

Virtual Facility

- **Fermilab provides resources via well-defined interfaces to high-energy physics experiments**
  - Provisioning of resources is offered as a transparent service
  - The provided resources are currently Fermilab-owned and physically located on the Fermilab campus
  - We are planning to optimize the FNAL resource provisioning in a transparent way for the resource users

- **Expand the resources transparently** to commercial clouds (e.g. Amazon Web Services (AWS)), opportunistic grid (OSG), owned and shared private clouds (i.e. OpenStack)
  - **Transparent to the users - economic decision is made internal to the facility**
    - Cost models will be continuously evaluated, facilitating low-latency high-demand scenarios
    - Resource provisioning has been exercised — process and procedures are being fleshed out

- **Fermilab personnel have experience with enabling technologies** (i.e. glideinWMS)

Conclusions

Conclusions

- Scientific Computing is one of the corner-stones of the scientific process for all experiments at Fermilab
- Facilities and services are providing excellent tools for the experiment to carry out their scientific missions
- The community aspect and the transfer of knowledge between all computing users at Fermilab are key to a sustainable and flexible scientific computing landscape
- More details can be found in the auxiliary summary of scientific computing
  - https://indico.fnal.gov/getFile.py/access?resId=2&materialId=4&confId=9353

Backup: Facility Upgrades

- Recent facility upgrades
  - Cooling upgrade and cold aisle containment in GCC, FCC and LCC completed in FY14
  - Fermilab was awarded the energy star award for 4 years in a row, 5th year is pending
  - Plans to interconnect CRAC air conditioning units to regulate room temperature, also switch to variable speed fans for GCC and FCC
- Network upgrades:
  - A 2nd 100 Gbps connection for network research and development will be added in 2015.
  - Internal network fabric between all our data centers with intra-datacenter bandwidth of 1,260 Gb/s (1.2 Tb/s) by end of Q1 CY15. Data centers will use a combination of 100 Gb/s and 10 Gb/s intra-datacenter links.
  - Wireless upgrade to 802.11ac for WH and FCC completed.
  - Village network upgraded to VDSL2+ this past December.

Backup: Glossary

Art	modular event-processing framework for experiments
AWS	Amazon Web Services - Amazon's cloud
BlueArc	network-attached storage (NAS) systems that are sold either as appliances bundled with storage, or as "NAS heads" supporting third-party storage area network connected storage
CVMFS	CernVM File System (CernVM-FS). is a network file system based on HTTP and optimized to deliver experiment software in a fast, scalable, and reliable way
CVS	The Concurrent Versions System (CVS), also known as the Concurrent Versioning System, is a client-server free software revision control system in the field of software development.
daq	Data Acquisition
DAQ	Data Acquisition System
dCache	is a system for storing and retrieving huge amounts of data, distributed among a large number of heterogeneous disk server nodes, under a single virtual file system tree with a variety of standard access methods.
enstore	Enstore provides distributed access to and management of data stored on tape
docDB	is a powerful and flexible collaborative document server.
ECL	Electronic Logbook - log book solution for experiments
EOS	is a Xroot-managed disk pool for analysis-style data access
F-FTS	Fermi-FTS: File transfer system – easy and robust uploading files to SAM
Fermicloud	OPenNebula cloud instance at FNAL
Fermigrid	GRID resources at FNAL
FIFEMON	Fifemon gathers information from many different sources and graphs it on a common timeline to help experiments understand their computing usage and identify problems
Gbps	giga bits per second
GEANT4	(for GEometry ANd Tracking) is a platform for "the simulation of the passage of particles through matter," using Monte Carlo methods.
GENIE	(Generates Events for Neutrino Interaction Experiments) is a universal object-oriented neutrino MC generator supported and developed by an international collaboration of scientists whose expertise covers a very broad range of neutrino physics aspects, both phenomenological and experimental.
Genie	GENIE (Generates Events for Neutrino Interaction Experiments) is a universal object-oriented neutrino MC generator

Backup: Glossary

GIT	is a distributed revision control and source code management (SCM) system with an emphasis on speed.
Git	Git is a free and open source distributed version control system
GPCF	General Physics Computing Facility (GPCF), providing interactive login nodes for experiments
GPGPU	General-Purpose computing on Graphics Processing Unit
Gratia	The Gratia Project designs and deploys robust, scalable, trustable and dependable grid accounting, publishes an interface to the services and provides a reference implementation.
Gridftp	GridFTP is an extension of the standard File Transfer Protocol (FTP) for high-speed, reliable, and secure data transfer. The protocol was defined within the GridFTP working group of the Open Grid Forum.
I/O	input/output or I/O (or informally, io or IO) is the communication between an information processing system (such as a computer) and the outside world, possibly a human or another information processing system.
IFBeam	Beam Conditions Database for Intensity Frontier Experiments
IFDH	(Intensity Frontier Data Handling), is a suite of tools for data movement tasks for Fermilab experiments
IntelPhi	Brand name for Intel's Many Integrated Core Architecture or Intel MIC (pronounced Mike): multiprocessor computer architecture developed by Intel
ISO20k	first international standard for IT service management to reflect best practice guidance contained within the ITIL framework
ITIL	The Information Technology Infrastructure Library (ITIL) is a set of practices for IT service management (ITSM) that focuses on aligning IT services with the needs of business.
JobSub	Batch Submission System For Intensity Frontier Experiments
Larsoft	integrated, art-based, experiment-agnostic set of software tools to be used by multiple LArTPC neutrino experiments
LArTPC	Liquid-Argon Time Projection Chamber
LOI	Letter of Intent
LT04	Linear Tape-Open (or LTO) is a magnetic tape data storage technology, LT04 capacity: 800 GB
Lustre	is a type of parallel distributed file system, generally used for large-scale cluster computing
OpenStack	OpenStack is a free and open-source cloud computing software platform
OSG	OpenscienceGrid

Backup: Glossary

redmine	Redmine is a free and open source, web-based project management and bug-tracking tool.
RSA	RSA is one of the first practicable public-key cryptosystems and is widely used for secure data transmission
SAM	SAM is a data handling system organized as a set of servers which work together to store and retrieve files and associated metadata, including a complete record of the processing which has used the files.
SLA	Service Level Agreement
Subversion	Subversion is an open source version control system
SVN	Apache Subversion (often abbreviated SVN, after the command name svn) is a software versioning and revision control system distributed as free software under the Apache License.
TDR	Technical Design Report
TSW	Technical Scope of Work
UPS	uninterruptible power supply
Xrootd	is a fully generic suite for fast, low latency and scalable data access, which can serve natively any kind of data, organized as a hierarchical file system-like namespace, based on the concept of directory