Fermilab

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

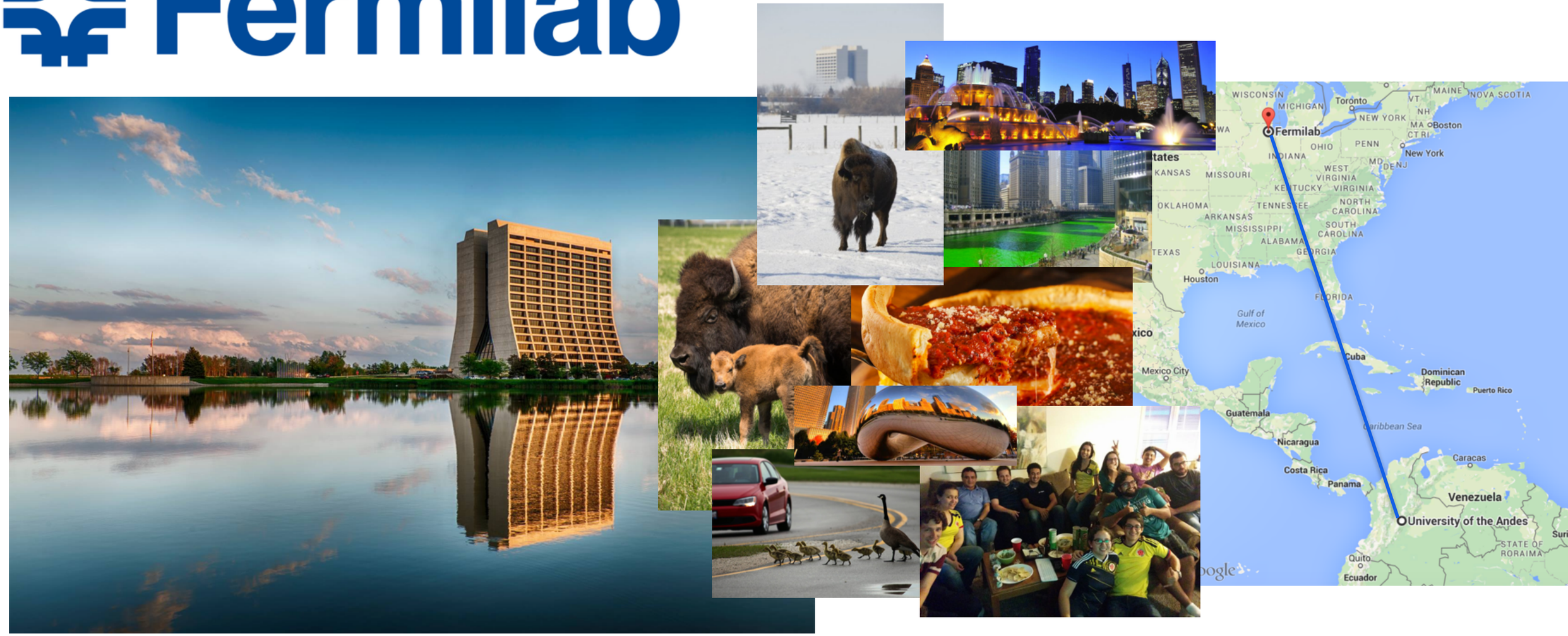
OPOS:
Towards a Common Production Service for Neutrino Experiments at Fermilab

Paola Buitrago

HEP Seminar at Universidad de los Andes - Bogota, Colombia

09 September 2015

Fermilab

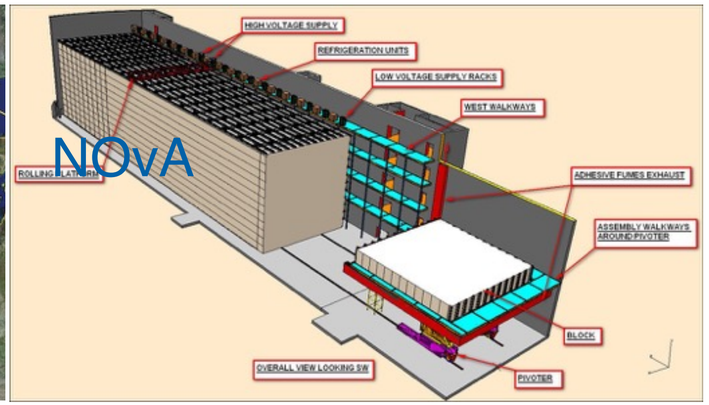


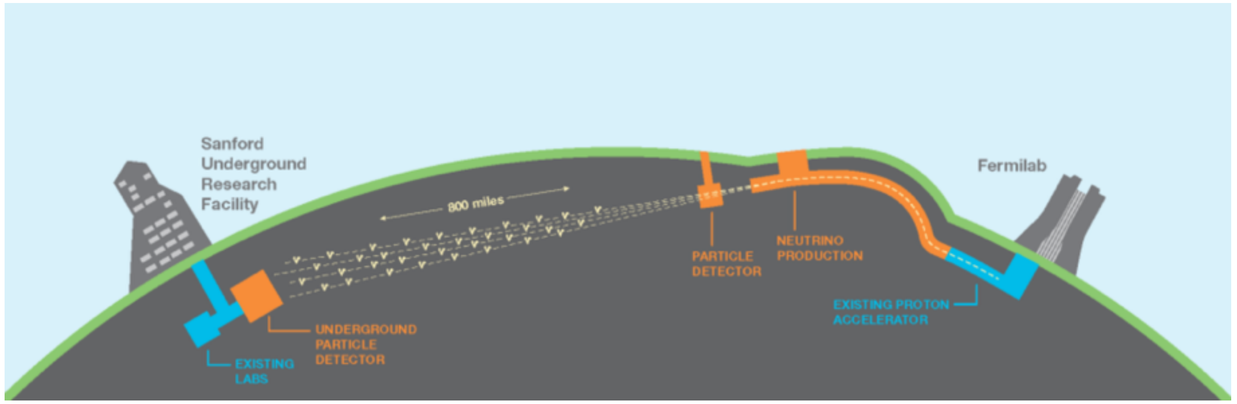
Fermilab

Neutrino Experiments

- Detect neutrinos and measure their masses
- Intense beams of particles.
- Produce neutrino beam and direct it to far detector, compare near and far detector measurements

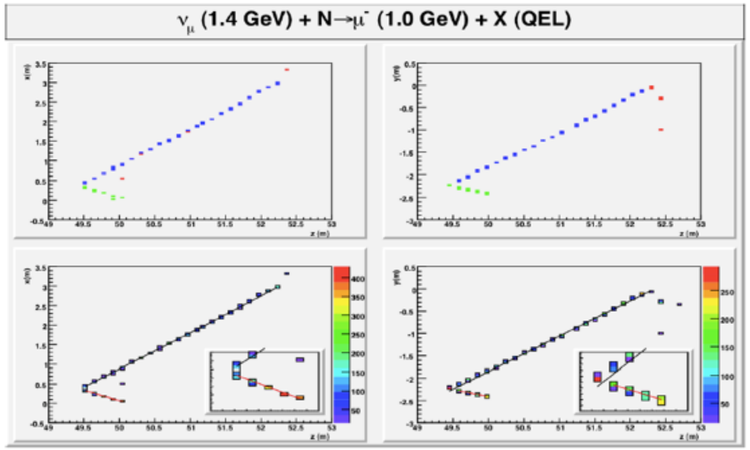
- Current experiments:
  - NOvA
    - 500 miles to Ash River, Minnesota
    - https://goo.gl/sMc9O4
  - MINERvA
  - MINOS+
- Future experiment: DUNE
  - 800 miles to Sanford Underground Research Facility, in Lead, South Dakota



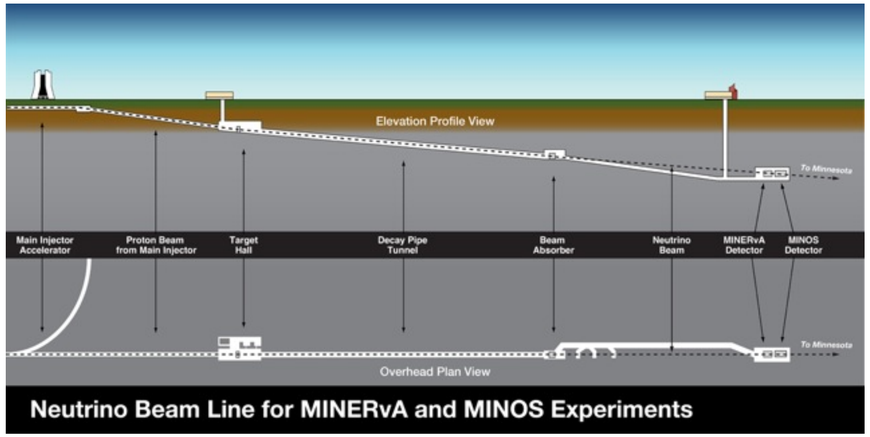


The Scientific Process

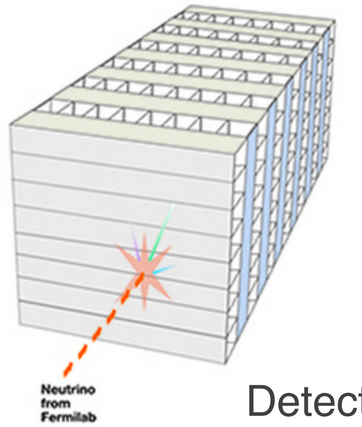
1. Theory



MonteCarlo simulations



Fermilab Numi Beam

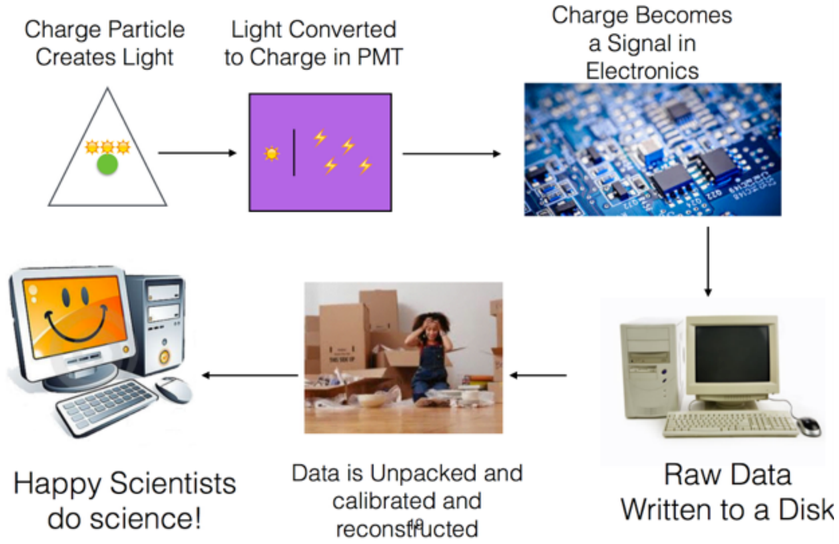


Detector

2. Experiment

3. Compare

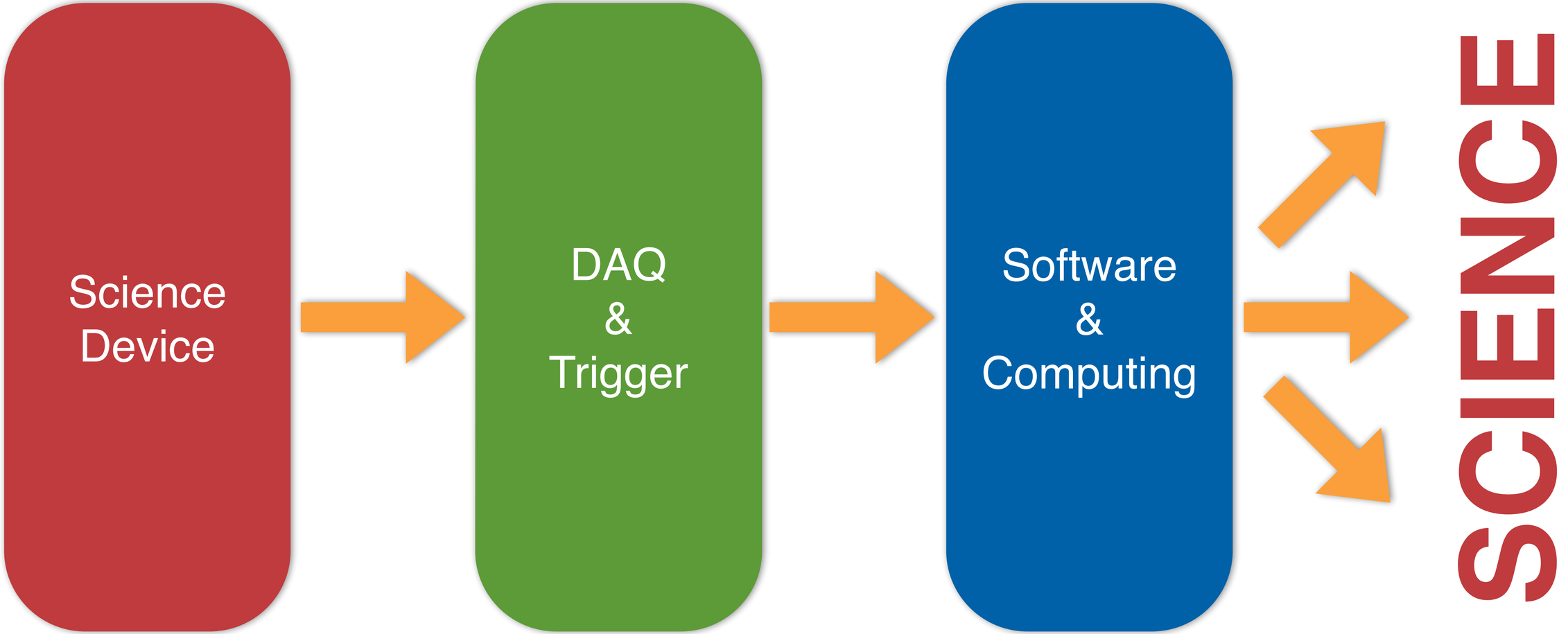
4. Conclude



Data acquisition and processing lifecycle

Fermilab

The Scientific Process

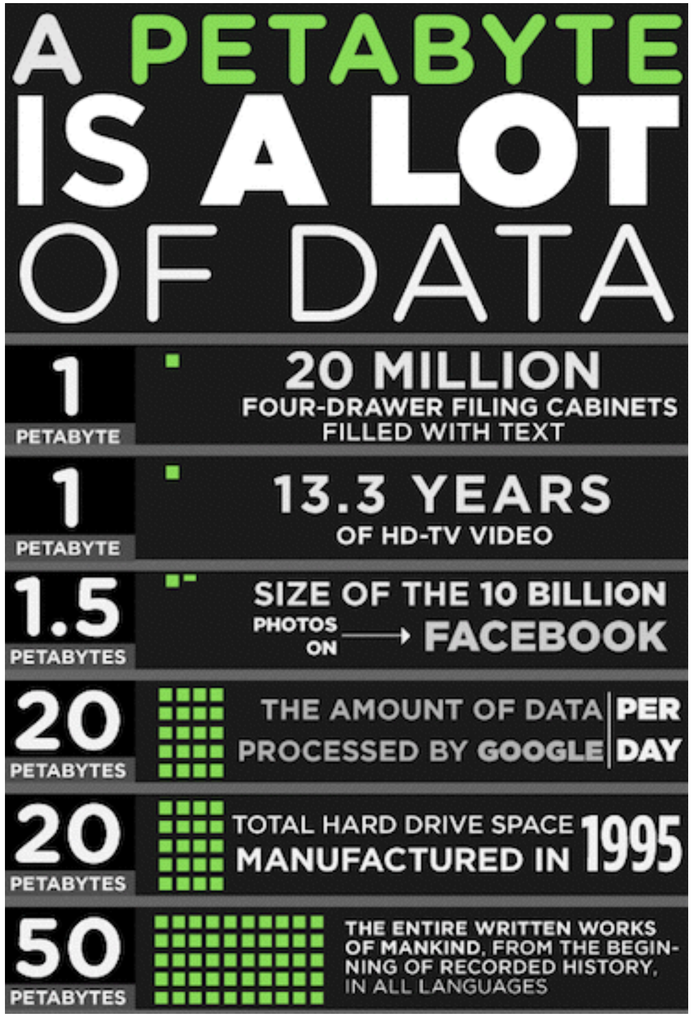


- Software & Computing is an integral part of the scientific process

Computing Challenges

- Requirements are extreme:
  - Huge volumes of data
  - Huge needs of processing

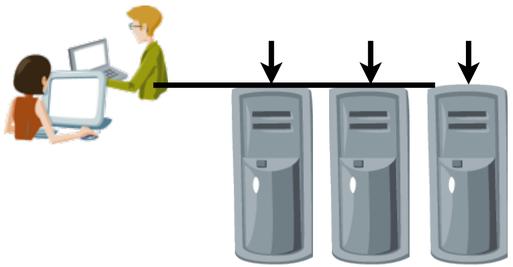
- NOvA needs for FY16
  - 13M CPU hour allocation
  - ~2.5 PB in tape storage
  - 1 PB in disk storage - recently accused datasets
- MINERvA needs for FY15
  - 2000 slots in the local grid
  - ~1.5 PB in tape storage
    - 0.5 PB in disk storage
- This is only two…

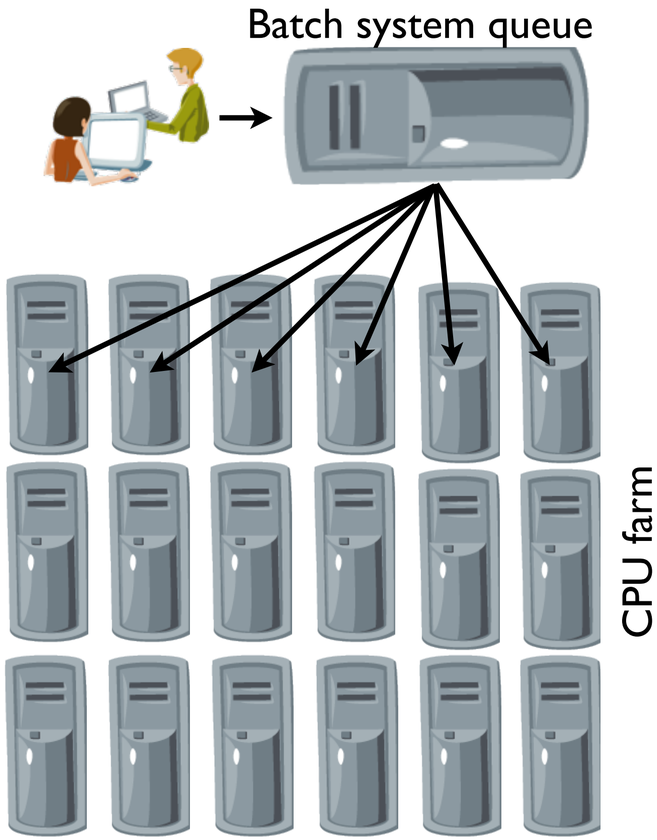


Fermilab

From Single Computer to the GRID

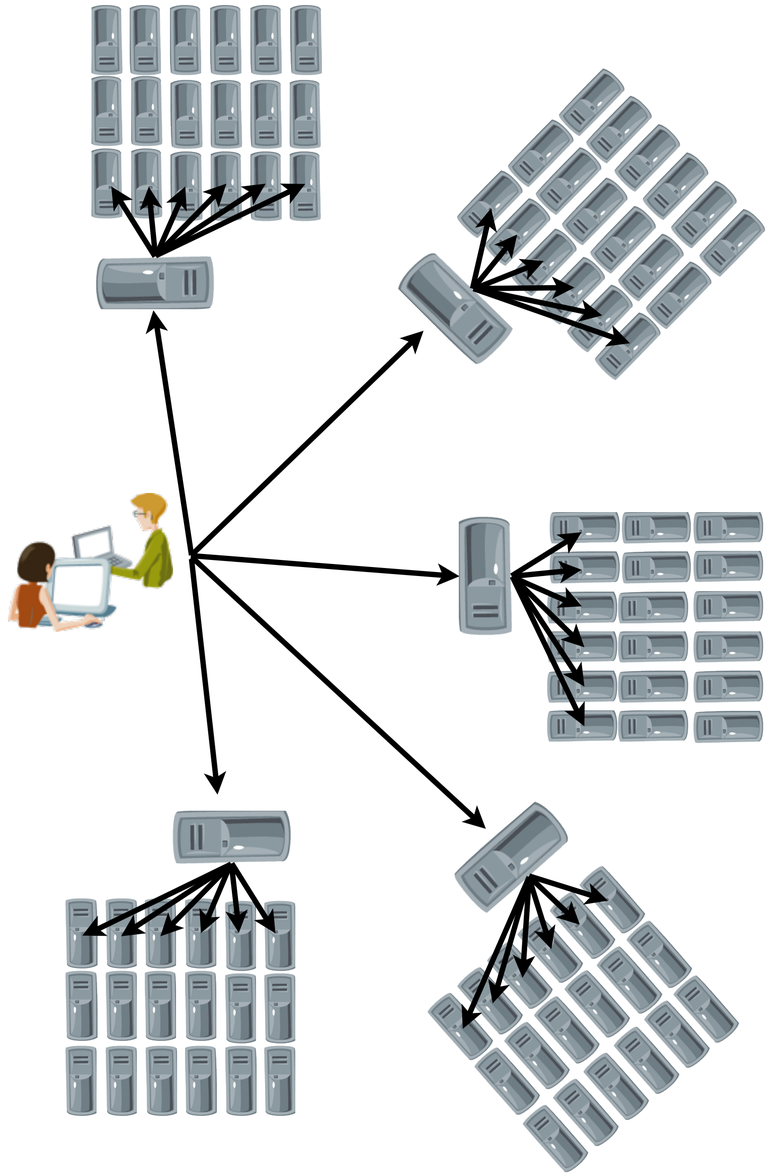
- Start with one PC to reconstruct events
  - Modern PCs have a CPU with multiple cores: in particle physics, we currently can process one event in parallel per core
  - Caveat: you need to manually start the reconstruction program for each core → reconstructing Billions of events takes a long time and is work intensive
- Buy 2 more PCs to be faster
  - You have to start programs on each of the PCs manually
  - Even more work intensive
- Only MINERvA needs a total of 2000 cores allocation.
- We use CPU farms and Batch systems:
  - Individual programs are put in a queue
  - The batch system has access to a farm of (many thousands of) cores
  - The batch system takes the first “job” in the queue and executes it on a free core → fills the farm with jobs





From Single Computer to the GRID

- The GRID, an interconnected network of batch farms
  - Why not a single huge batch farm for CMS:
    - Running 100,000 cores in one installation is very difficult
      - You need a lot of cooling and large amounts of electricity
    - Neutrino experiments are international collaborations, funding agencies (like the Department of Energy in the US) prefer to spend research money in their home countries
  - The GRID enables HPE experiments to have access to enough PCs despite being distributed over the world
  - The GRID software or middleware lets the individual computing centers or farm look like one big farm or center
    - Important: you need strong networks between the centers



Fermilab

OSG Sites: Distributed Over Worldwide

Worldwide OSG sites

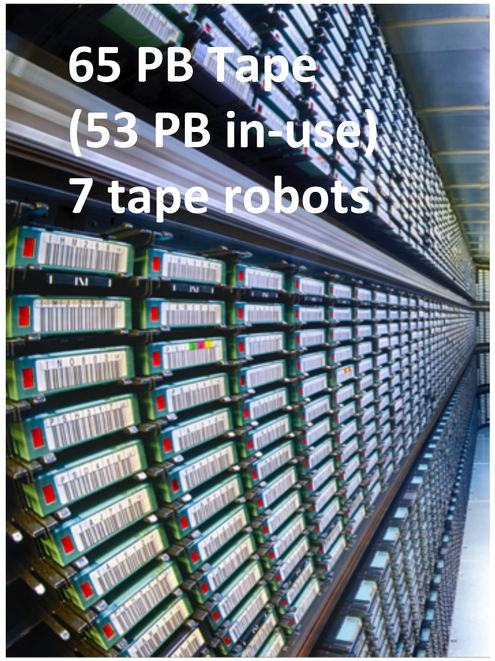
Colombia OSG sites

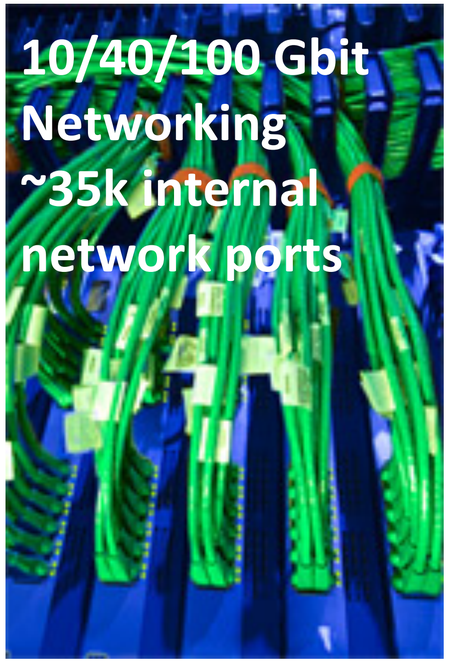
127 OSG sites worldwide

Fermilab

Fermigrid: Fermilab OSG Site

- Provide and manage computing services and resources
- Data recording, storage, access
- Bulk processing, analysis
- Functionality analogous to LHC Tier-0 and Tier-1
- CPU Cores, Online (Disk) and Offline (Tape) Storage, Networking

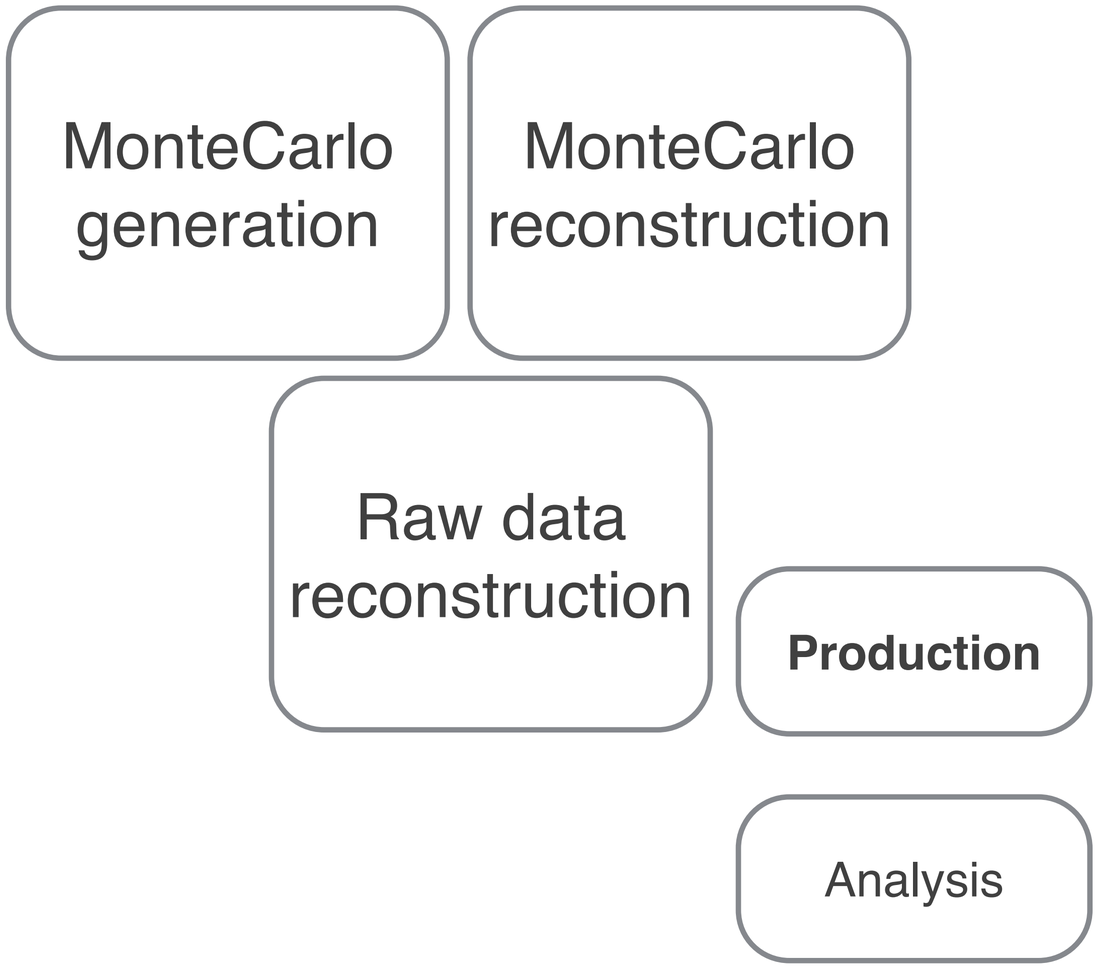




Fermilab

Computing in the Neutrino Experiments

Processing Stages



Analysis Processing

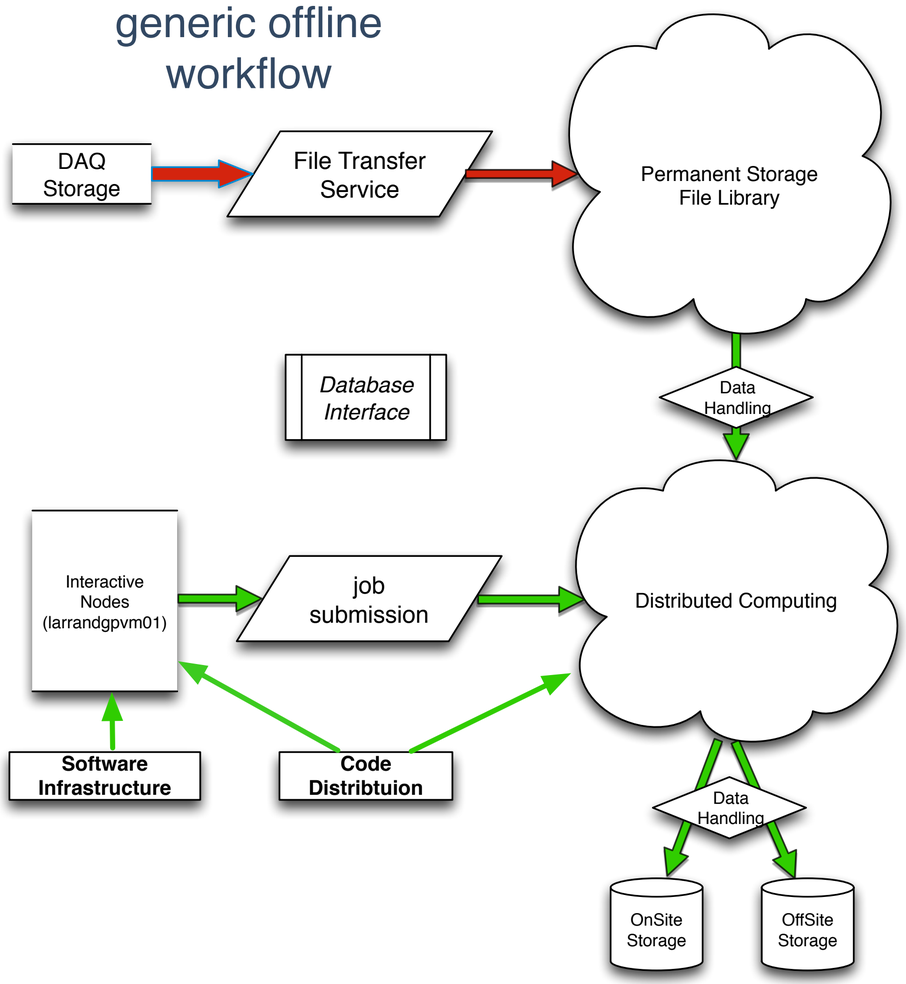
- Non stable/standard processing
- Perform by any member of the collaboration
- Uses official reconstructed files

Production Processing

- Reconstruct data
- Produce official reconstructed files for the collaboration
- Stable/standard processing
- Must be done on a daily basis
- Performed by the production group inside each collaboration

Neutrino Experiments Computing Needs

- Data transfer and storage
  - Store files
  - Store files metadata
  - File transfer to grid node
  - File transfer from grid node
- Parallel Computing
  - Submit jobs
  - Orchestrate jobs
  - Distribute software
  - Networking
- Bookkeeping
  - Which raw data has been recorded?
  - Which files have been generated? How were they generated?



Neutrino Experiments - Processing Workflow

Submission time

1
Define input data to be processed

2
Update bookkeeping system

3
Configure and submit jobs to the grid

Local machine

Worker node time

Queueing system

A
Identify file(s) to process

B
Get file (s) to process

C
Process file(s)

D
Copy out output files, update bookkeeping

Optional

Batch system

Post processing activities

Optional I
Postprocess output files

II
Output files to permanent storage

FNAL machines

Fermilab

Neutrino Experiments - Processing Workflow

Submission time

1
Define input data to be processed

2
Update bookkeeping system

3
Configure and submit jobs to the grid

Local machine

Worker node time

A
Identify file(s) to process

MONITORING

C
Process file(s)

D
Copy out output files, update bookkeeping

Queueing system

Optional

Batch system

Post processing activities

Optional

ERROR TRIAGING

permanent storage

FNAL machines

Fermilab

FIFE

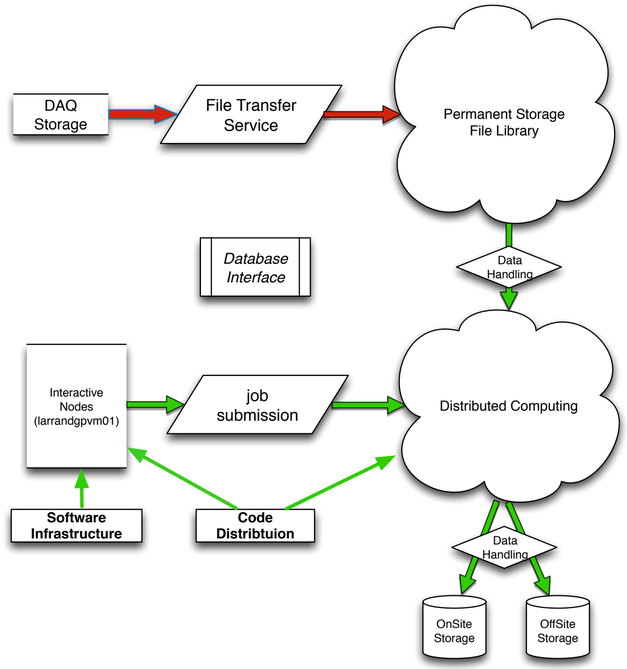
Fabric for Frontier Experiments

Common Services and Projects toolkit

FIFE provides access and support for common tools in:

- DAQ and Controls
- Grid and Cloud
- Scientific Data Storage and Access
- Scientific Data Management
- Scientific Frameworks and Software
- Physics and detector simulation
- Databases
- Scientific Computing Systems
- Scientific Collaboration Tools

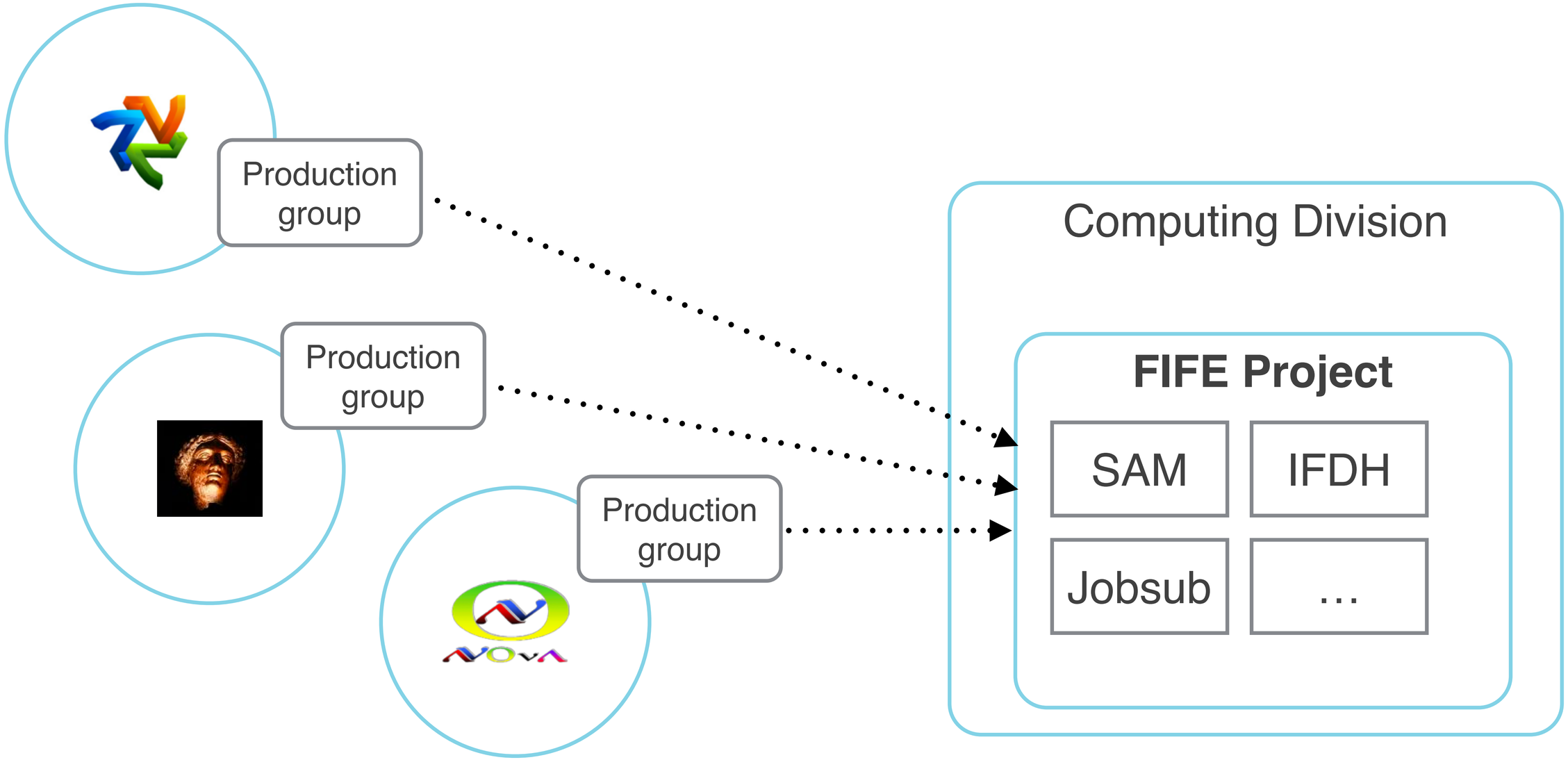
generic offline workflow



Fermilab

Computing and Neutrino Experiments Overview

Neutrino Experiments and Computing Division



OPOS

Offline Production Operations Service

Neutrino Experiments - Processing Workflow

Submission time

1
Define input data to be processed

2
Update bookkeeping system

3
Configure and submit jobs to the grid

Local machine

Worker node time

Queueing system

A
Identify file(s) to process

B
Get file (s) to process

C
Process file(s)

D
Copy out output files, update bookkeeping

Optional

Batch system

Post processing activities

Optional
I
Postprocess output files

II
Output files to permanent storage

FNAL machines

Fermilab

OPOS - Offline Production Operation Service

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

- Summer 2014 : group founded to support NOvA, MINOS and Minerva (experiments that are currently taking data), helping in running production workflows.

Fermilab

OPOS - Mission

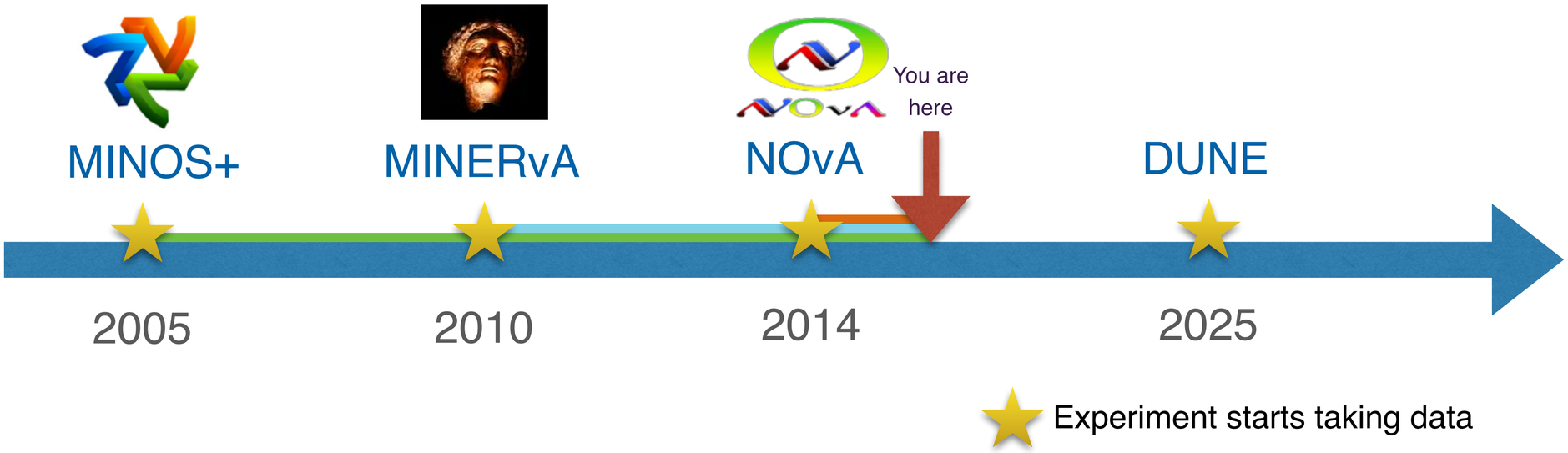
- What we do:
  - Assist experiments with running their large scale production workflows and other large scale offline production activities.
  - Provides the experiments with domain experts familiar with the supported tools for offline processing at Fermilab.
  - Schedule and run the large scale production tasks for the experiments.
  - Monitor the tasks and provide the experiment with status and progress information on requested tasks.
  - Assist the experiment with limited triage and analysis of failed jobs and workflows.
  - Dacilitate communication!!!

OPOS Challenges

- Interface with several collaborations
- Understand each experiment workflow
  - Process stage configuration
  - Error messages
- Interface with computing division support groups
- Talk different languages
- Handle huge processing volume in a timely manner
  - Guarantee efficient processing
  - **Monitor** and **report**
  - Job triaging
  - Limited monitoring tools (change to a positive statement)
  - Monitor different offline processing tools (difficult to link)

OPOS Challenges

- Collaborations at different stage of their life cycle

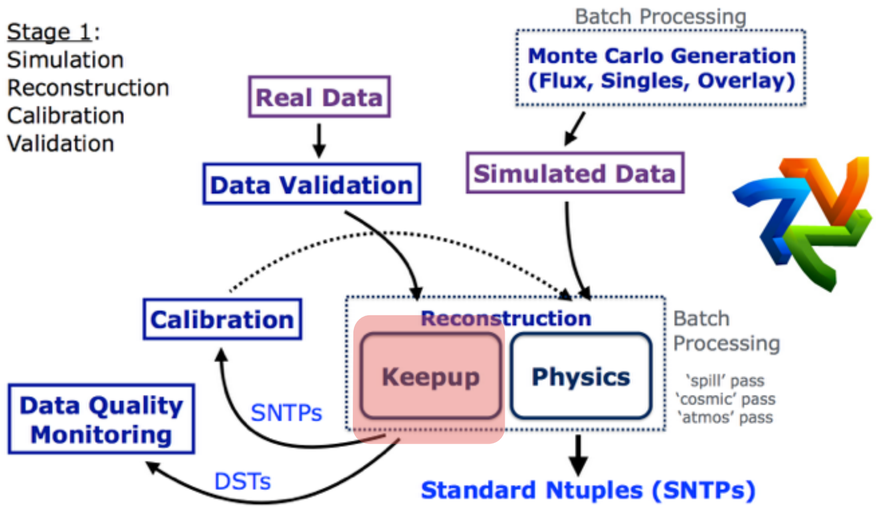


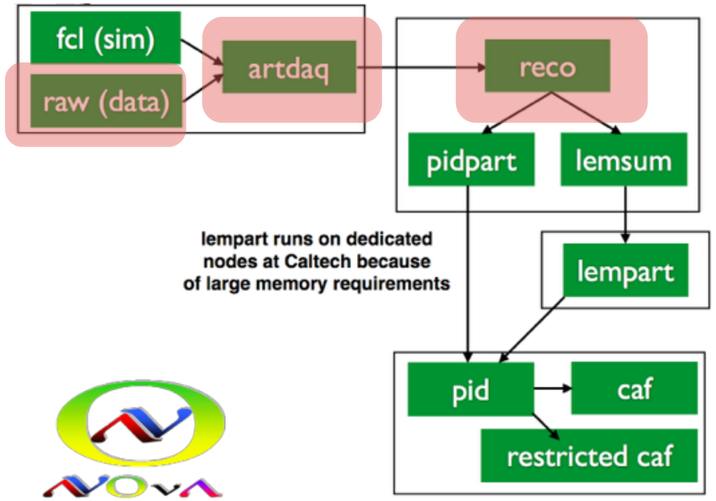
OPOS Challenges

- Keep knowledge/ Disseminate knowledge
- Accountability
  - What have we processed and with what efficiency?
- Scalability
  - More experiments, more stages, more jobs, more logs to process, more reports

OPOS Where we stand

- Running/monitoring production jobs for the three data taking IF experiments: NOvA, MINERvA, MINOS





MC data processing

OPOS Addressing Challenges

- Interface with several collaborations
  - Service Now (Service Management PaaS) Request
  - Colaboration Redmine Issues
- Understand each experiment workflow and keep knowledge
  - OPOS Wiki
- Interface with Computing Division support groups
  - Service Now (Service Moment PaaS) Incident Filling
- Handle huge processing volume in a timely manner and accountability
  - Production DB - proof of concept
  - Need a robust management system
- Scalability
  - Need a robust management system

OPOS - Production Operation Management System

- Production Management System: software developing
  - OPOS proof of concept
  - Connect computing and physics semantics
  - Records submission info
  - System that talks with different monitoring systems and automatically retrieves relevant metrics to monitor and report
  - Relevant metrics: two clicks away
  - In the future
    - Interface with experiments
    - Bookkeeping
    - Intelligent error triaging
    - Knowledge store

What have we accomplished?

- “Educate” experiments
- Force experiments to document
- Relieve effort from collaborations
- Standardized the software
- Improved job efficiency
- Computing tools improved
- Standardized workflows
- Define procedures and standards

Oldest collaboration

Summary and Outlook

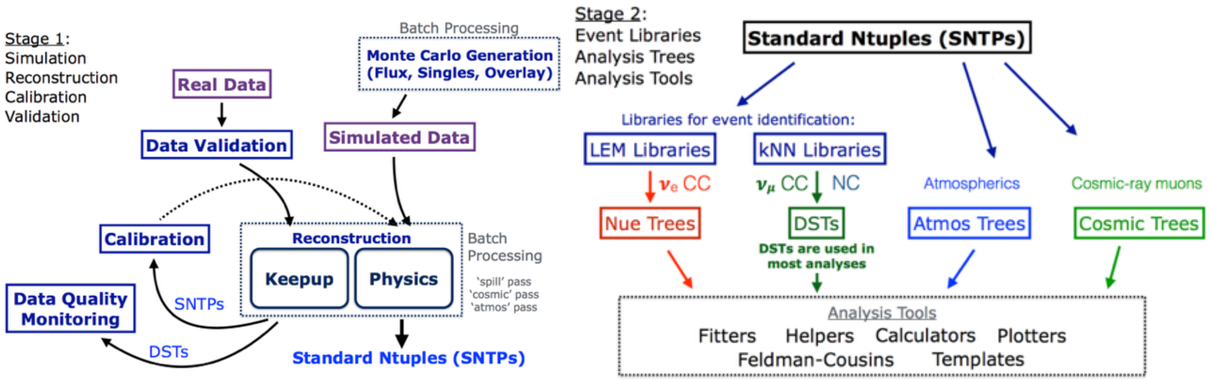
Summary and Outlook

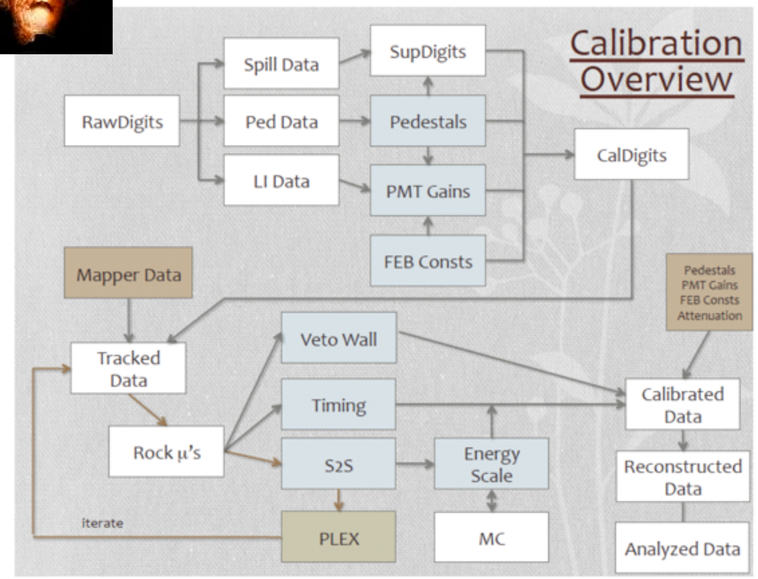
- The “ecosystem” in which OPOS team work is highly complex
- OPOS team has managed to be up to the challenges
- OPOS has defined procedures and standards to properly define it’s scope
- OPOS has helped improved the tools offered by CD and also the collaboration software
- OPOS has helped in transferring the knowledge and good practices between experiments
- OPOS service is highly appreciated by the collaborations
- OPOS will work on on-boarding more experiments and more stages of the production workflows

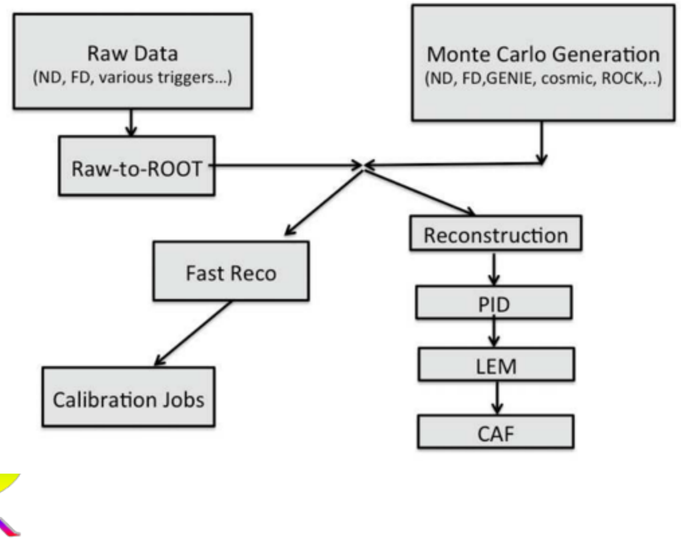
Thank you!

Backup Slides

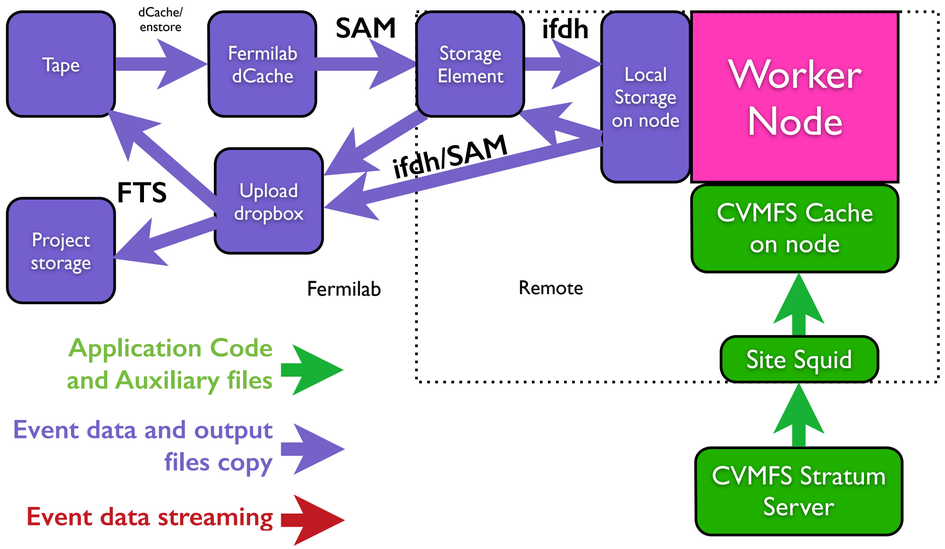
Data Taking Frontier Experiments - Production Processing







Fermilab


Tape
dCache/
enstore
Fermilab
dCache
SAM
Storage
Element
ifdh
Local
Storage
on node
Worker
Node
ifdh/SAM
FTS
Upload
dropbox
Project
storage
CVMFS Cache
on node
Fermilab
Remote
Site Squid
Application Code
and Auxiliary files
Event data and output
files copy
Event data streaming
CVMFS Stratum
Server