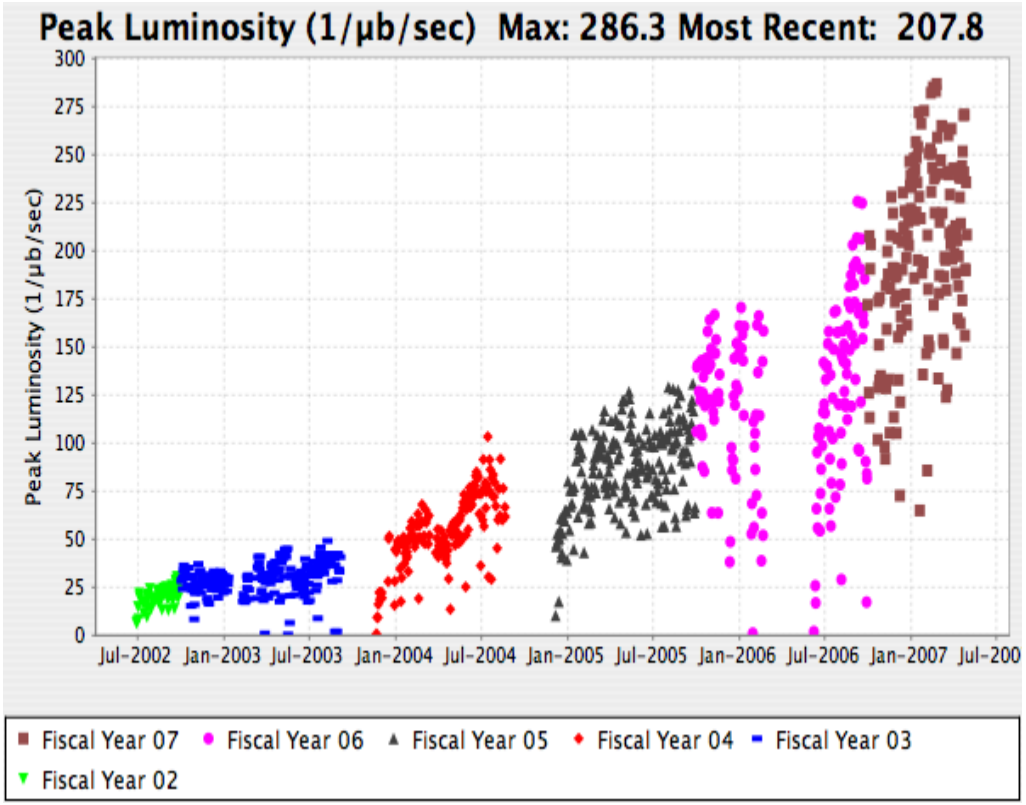
# Computing Tools And Analysis Architecture: The Tevatron Experience

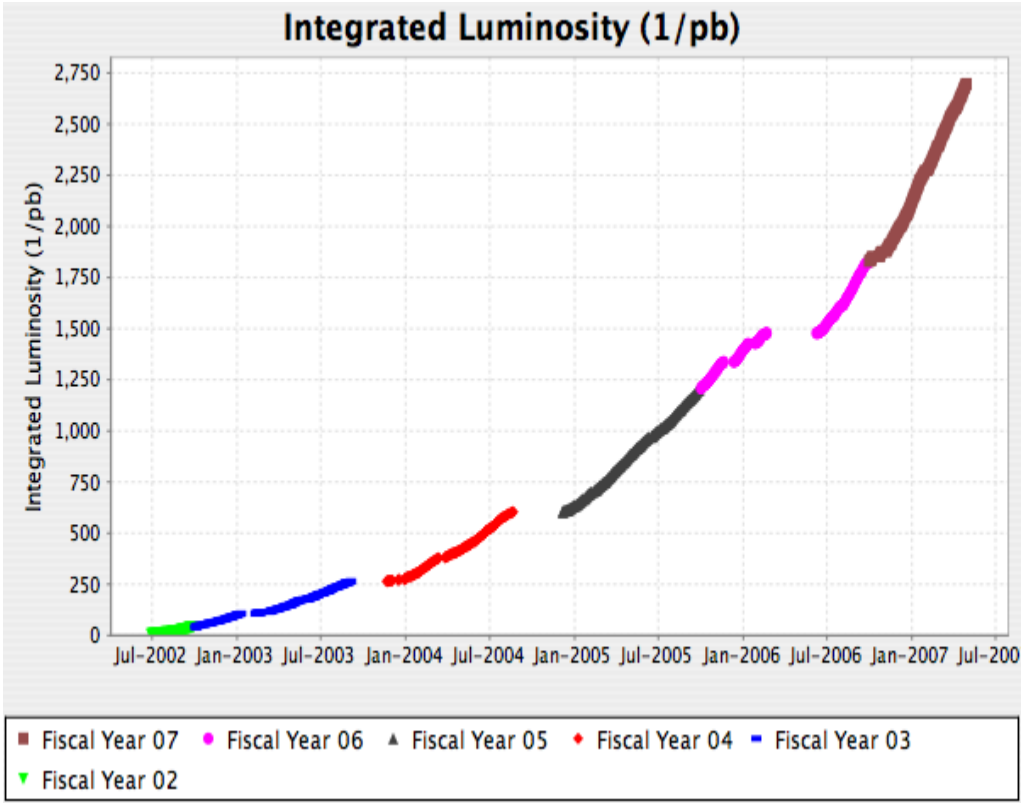
*Viatcheslav Sharyy*

*for CDF and D0 collaborations*

dapnia
cea
saclay

# Tevatron Status



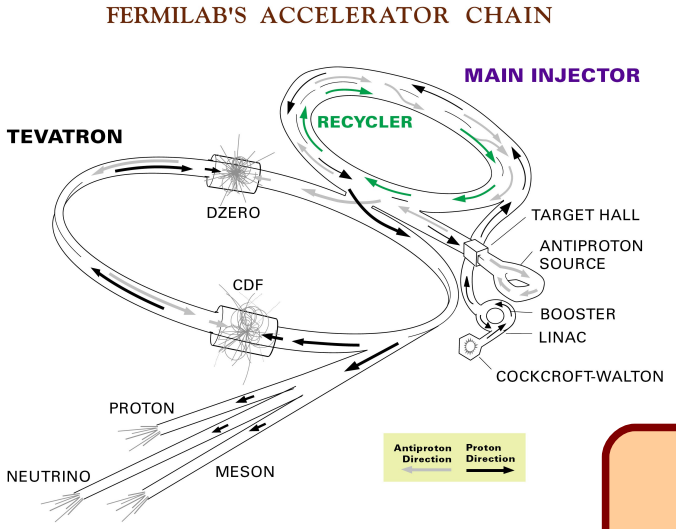

| | |
|---|---|
| Integrated luminosity recorded per experiment | 2.45 $fb^{-1}$ |
| Number of events per experiment | ~$3 \cdot 10^9$ |
| Raw data size per experiment | ~600 TBytes |

# Computing infrastructure

Hardware configuration is typical setup for a large experiment:

- Central analysis farms, mainly dual processor, dual core linux nodes operating under Scientific Linux
- Central production farms of dual processor linux nodes
- Linux disk cache servers
- High speed tape robot
- Large, central database (Oracle, Sun)
- Linux desktops
- Gb Ethernet interconnections on central systems, 100Mb to desktop
- Similar, smaller scale, installations at collaborating institutions including dedicated linux farms (or fraction of farms) worldwide (Gridka, Lyon, INFN, Manchester...)

# COMPUTING MODEL




Linux farm + data storage systems worldwide

Data Handling System

Robotic tape storage + disk cache

Reconstruction & analysis farms

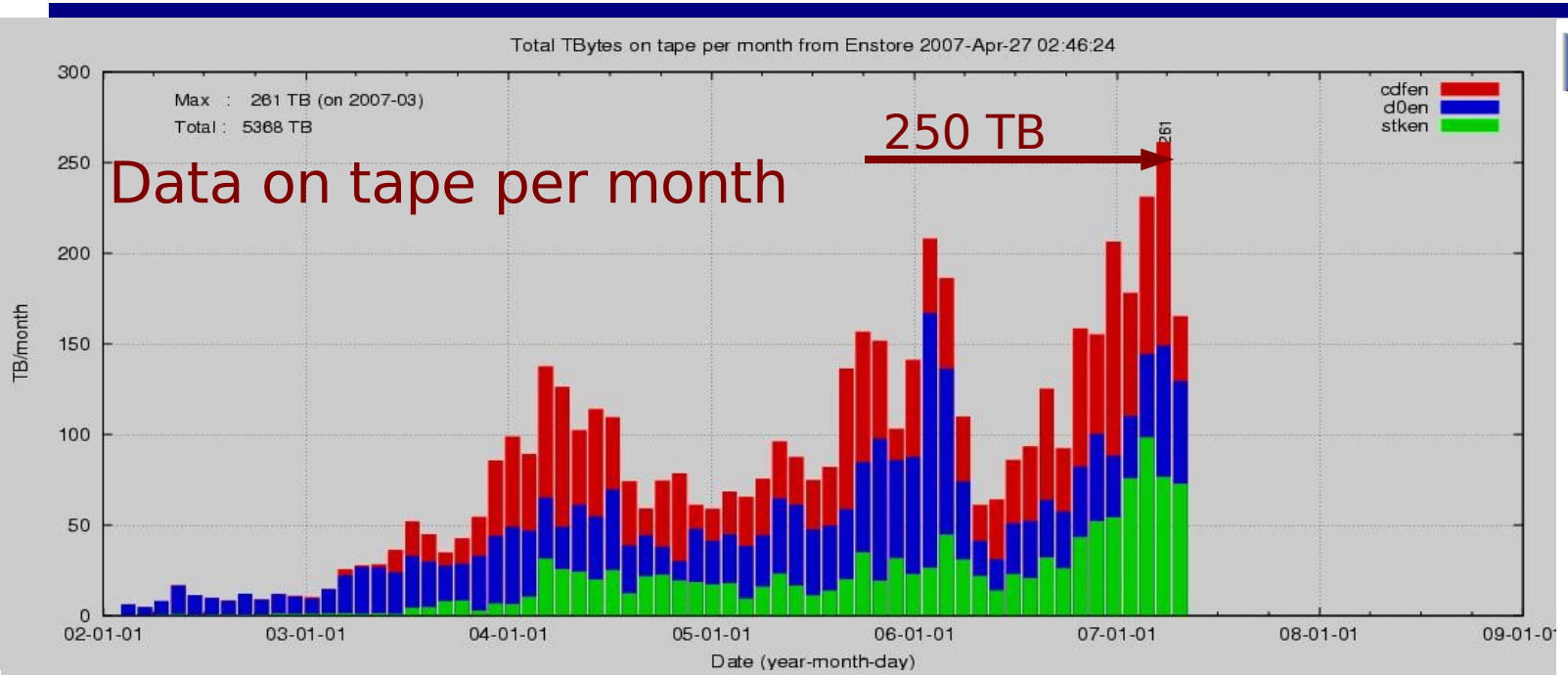
Users desktops worldwide

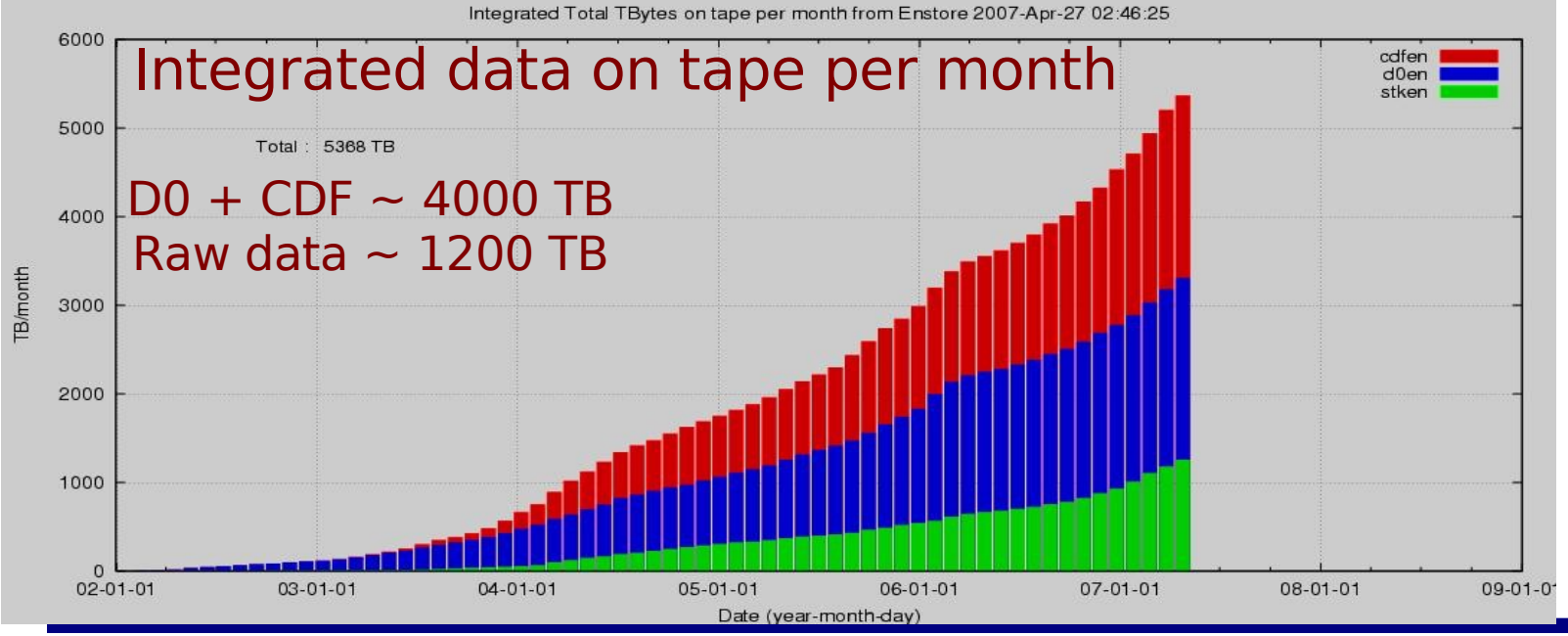
# Computing activities

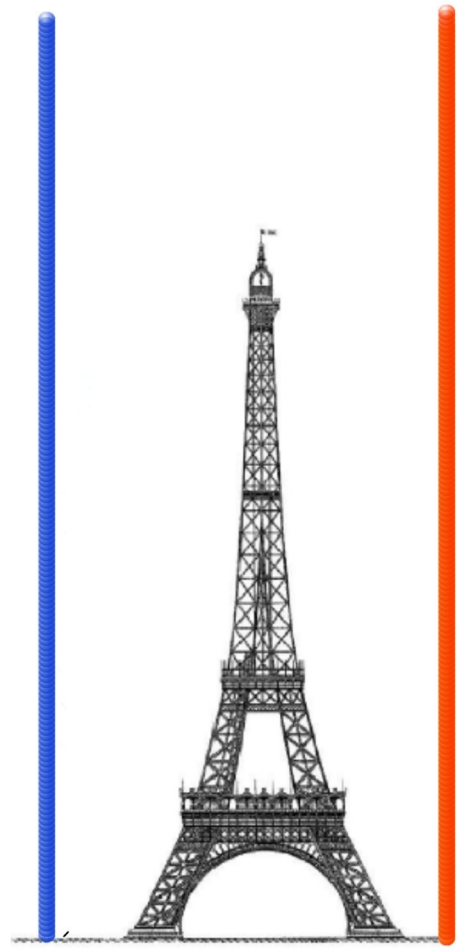
- Raw data reconstruction
- Reconstructed data treatment
  - Additional reprocessing
  - Skimming (secondary dataset production)
  - Saving data in analysis format
- MC production
- Physics analysis

| | CDF | D0 |
|---|---|---|
| Analysis + reconstruction farms at Fermilab | ~6.5 THz | ~8 THz |
| Remote farms | ~2.5 THz | ~2.5 THz |

# Fermilab Mass Storage System



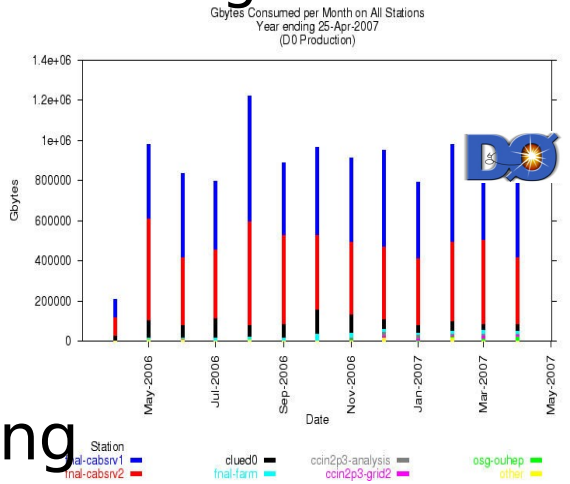



If one writes all D0 and CDF data on DVD (4 GB, 1.2 mm)

# Data Handling

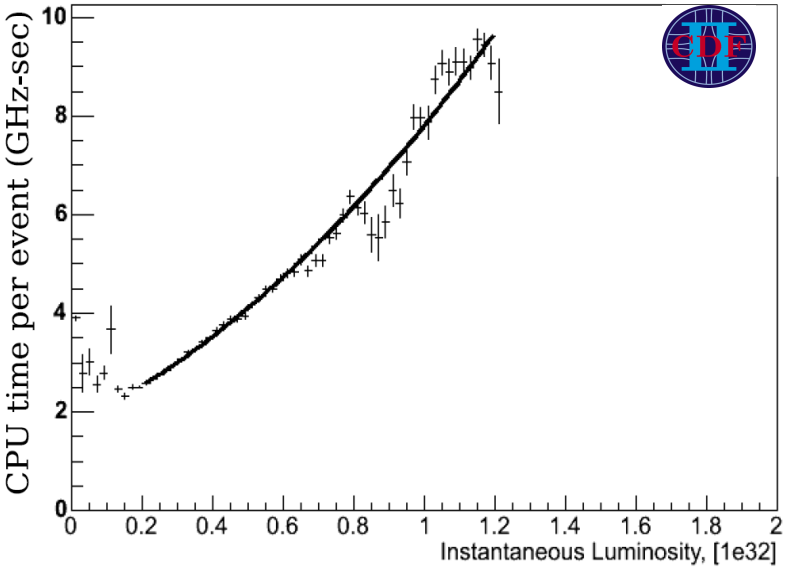
- D0: SAM (Serial Acess via Metadata)
  - File based. Each file characterized by metadata (including the parentage files). Simple tools to define dataset (collection of files) based on the metadata.
  - File tracking information (location, delivery, consumption). Manages multiple copies of the same data units (files).
  - Optimized tape reading (file order is optimize according to the file distribution on tapes + disk cache)
- CDF:
  - SAM for the metadata handling
  - on-site file delivery mechanism: dCache (network-based DCAP protocol, file streaming over network optimize for using with root tree)
  - for off-site data delivery: SAM data delivery utilities
  - dCache-based analysis disk storage model

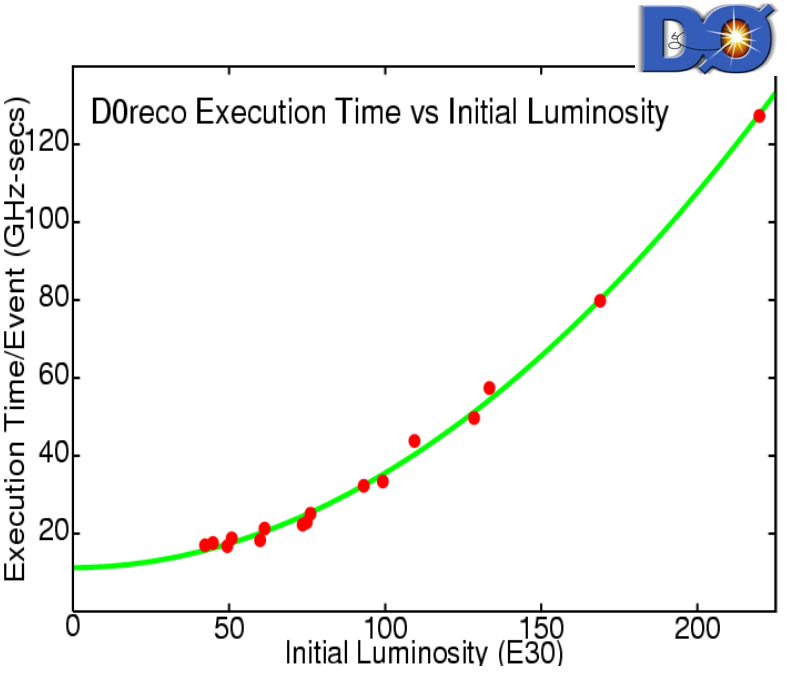


# Raw Data Reconstruction

| | | |
|---|---|---|
| D0 output rate | 50 – 150 Hz | 10 – 30 MB/sec |
| | 5 – 6M events/day | <1 TB/day |
| CDF output rate | 100 – 150 Hz | 20 – 30 MB/sec |
| | 5 – 10 M events/day | < 2 TB/day |

- Done locally on Fermilabs farms. Currently running efficiently and stable with 4M/day (D0) and 20M/day (CDF)
  - D0 expect to run smoothly with 8M/days or higher after summer shutdown.
  - CDF could run up 30M/day
- Plans are to make farms a part of the FermiGrid.
- Work is ongoing on algorithms improvements at high luminosity.



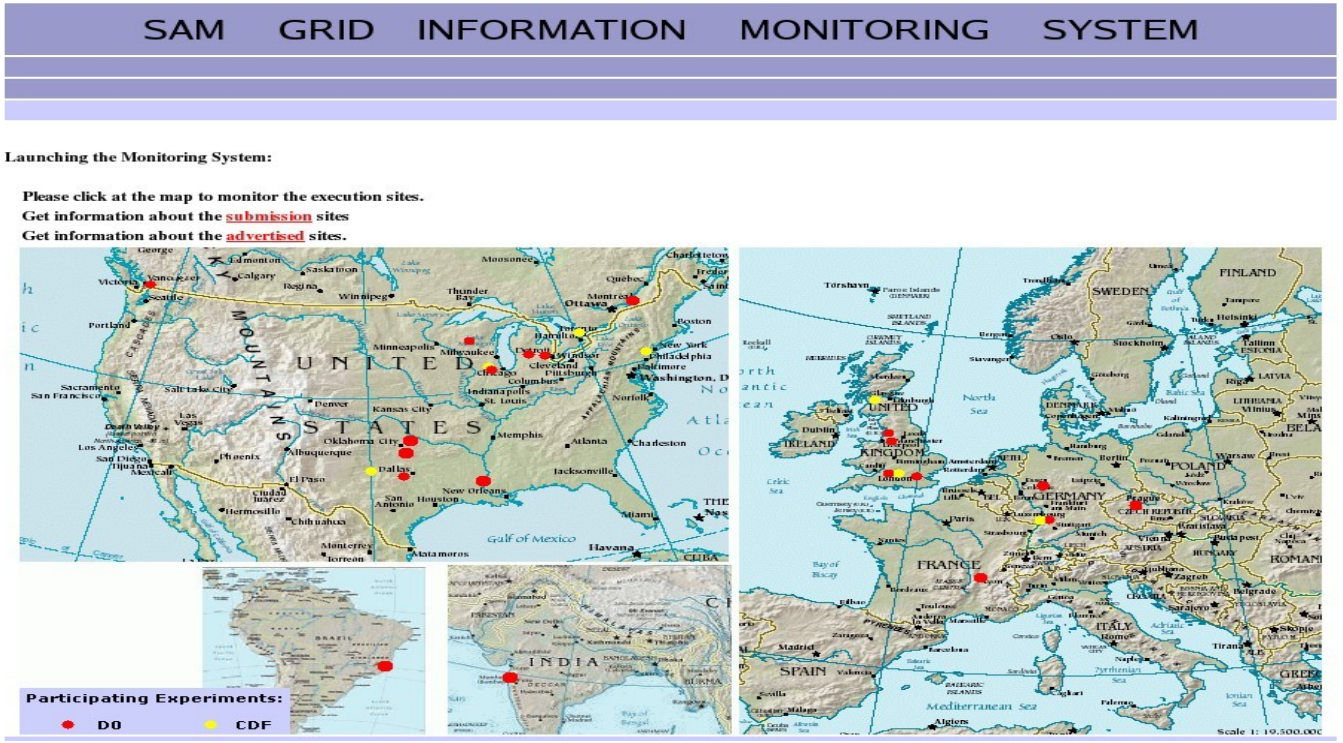



# Distributed computing: CDF

- One GRID “entry point” per continent (using Condor glide-ins)
  - NAmCAF (North American CAF) for the OSG
  - LCGCAF for LCG : providing acces to the European sites running IGEE middleware
  - Asian entry point (PACCAF - 'Pacific CAF') is ready and being commissioned, the logistics works, to use the Asian CPU's efficiently need to improve bandwidth of the WAN to Taiwan/China/Japan - right now have more CPU's than the WAN can support

# Distributed computing: D0

- SAM + JIM (job submission and monitoring) -> SAMGrid
- SAMGrid is used by D0 (~20 sites) for all off-site production,
- Automated access to LCG and OSG.
- http://projects.fnal.gov/samgrid/



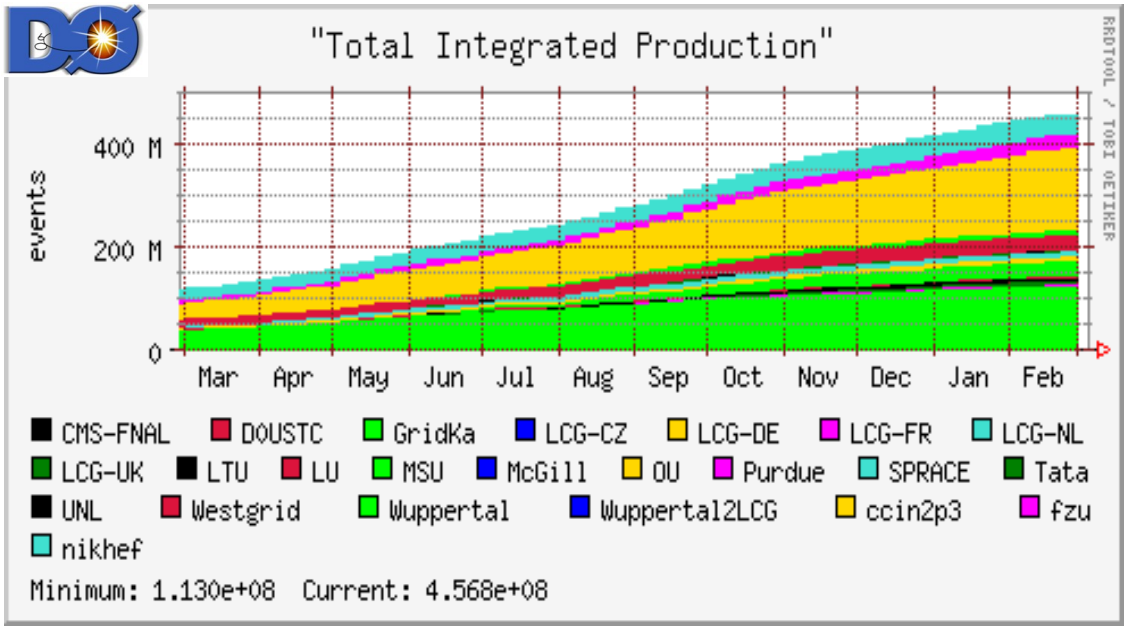
# DO/ CDF distributed computing

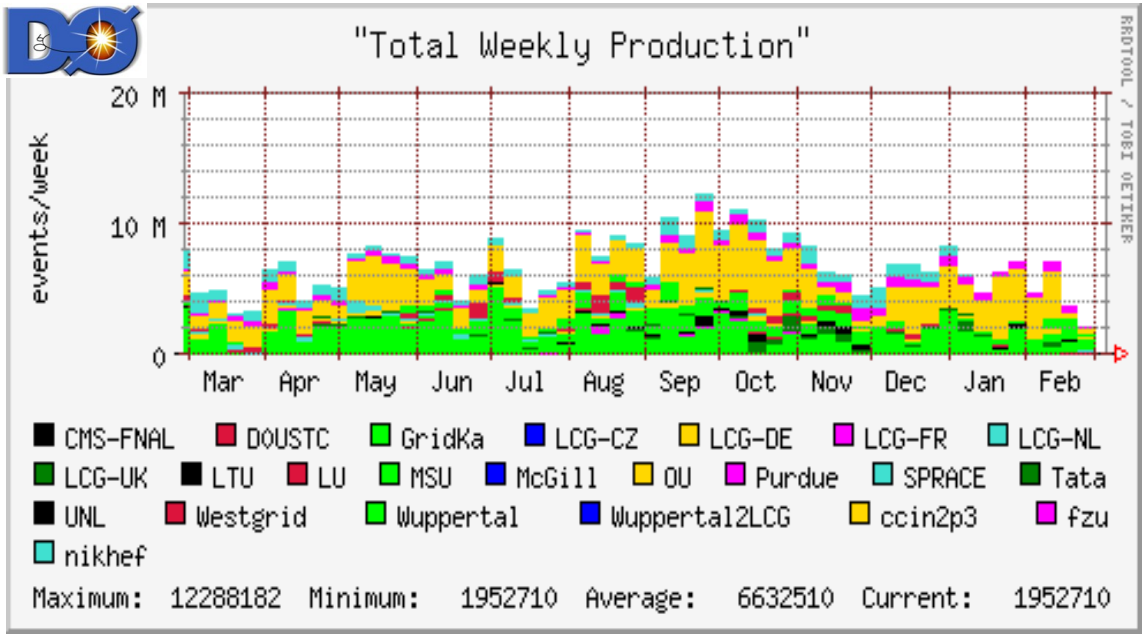
- D0 is doing reprocessing off-site while CDF is doing it on-site with the plan being to do it this way till the end of run II
- GRID interfaces: D0 is using SamGRID, CDF is using Condor-based "pull" model (so-called “glide-ins”)
- CDF is using remote CPU facilities for certain classes of analysis jobs like stripping via grid interface. D0 is using dedicated off-line centers for analysis and just started to explore analysis  on grid.
- Both experiment produce MC mainly outside Fermilab

# MC production

- CDF run-dependent simulation: generate samples that reflect data taking conditions (beam, detector conditions, trigger emulation).
- D0 overlay real “zero bias” events reflecting real luminosity profile in data.
- CDF: Geant3 + calorimeter simulation based on parameterized calorimeter package GFLASH. ~ *20 GHz·sec* for complicated event (e.g. top). CDF produced > $10^9$ events last year.
- D0: Geant 3 based simulation. *~240 GHz·sec* for the top events,  up to 10M/week. The most time consuming part in the Geant simulation.

# MC production: D0 example




- Mainly off-site
  - 80.1% Europe
  - 17.1% N. America
  - 2.2% S. America
  - 0.6% Asia
- Up to 10M events per week

# Example: D0 Reprocessing 2007

- Reprocessing to apply the updated calibration constants for RunIIb and algorithms adapted to the higher luminosity
  - ~90% off-site
  - Initial plans: 400M events / 4 months off-site → 3M/day
- First large experience on OSG: very challenging to get a full production system running for many months at high efficiency on the grid. Numerous problems in the beginning
  - No resource selector available
  - Monitoring issues
  - Finding bugs can be difficult
  - Turn around time to fix problems can be long
  - Hardware failures
  - Remote sites changing configurations
- OSG set-up “trouble shooting team”. See results on the next slides

# Grid efficiency: D0 example

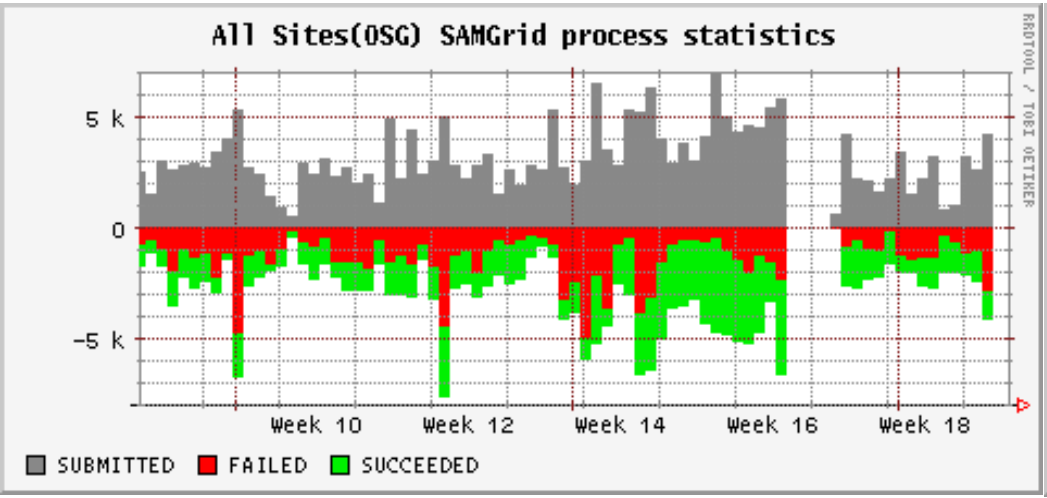
Number of jobs per day

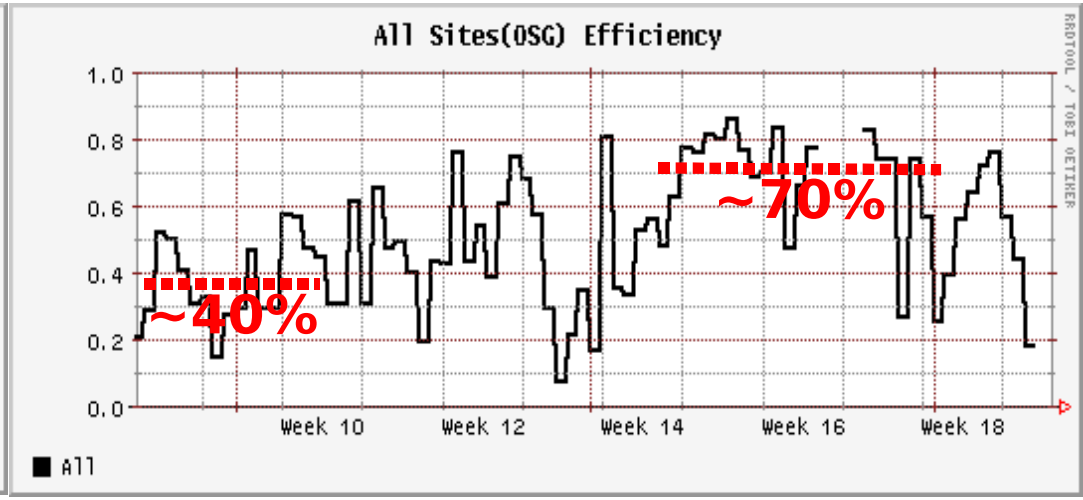
**Input Jobs**

**Job Results:** **Success** **Failed**

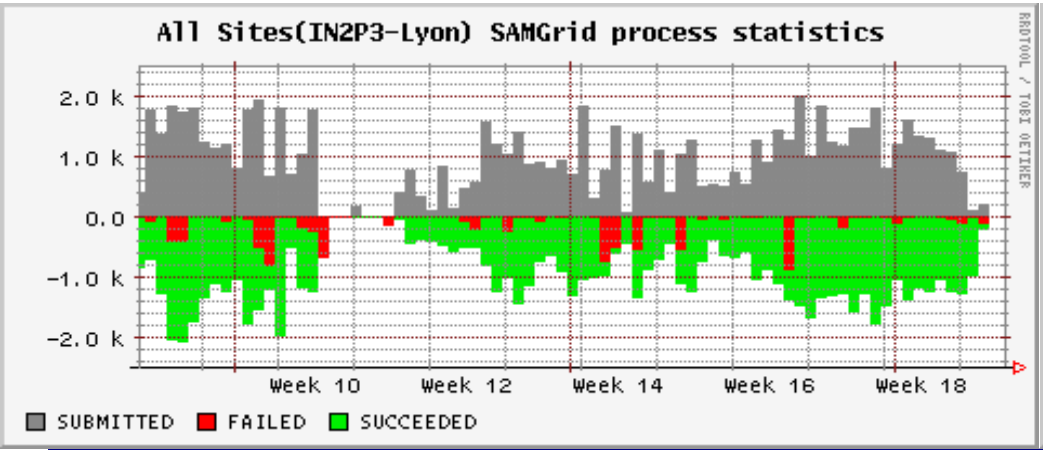
Efficiency
= N jobs success / N jobs in

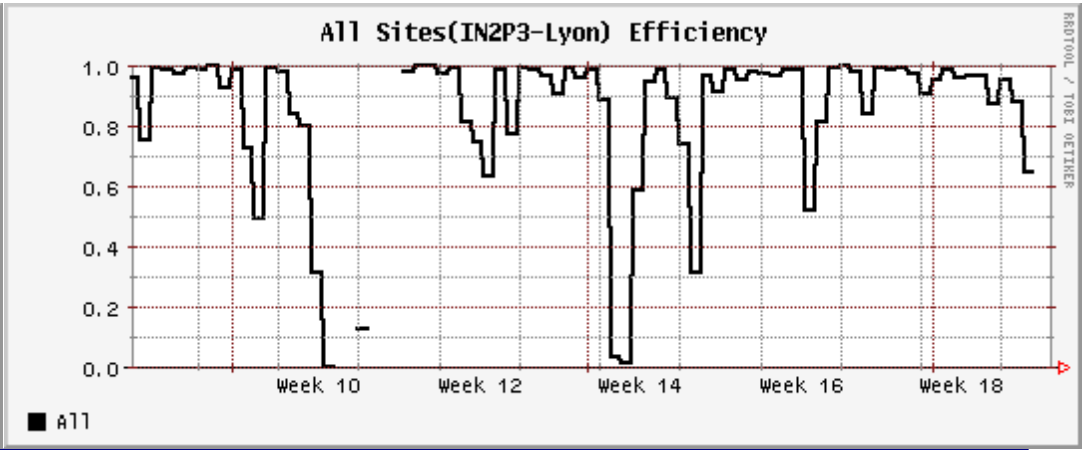
**OSG sites**



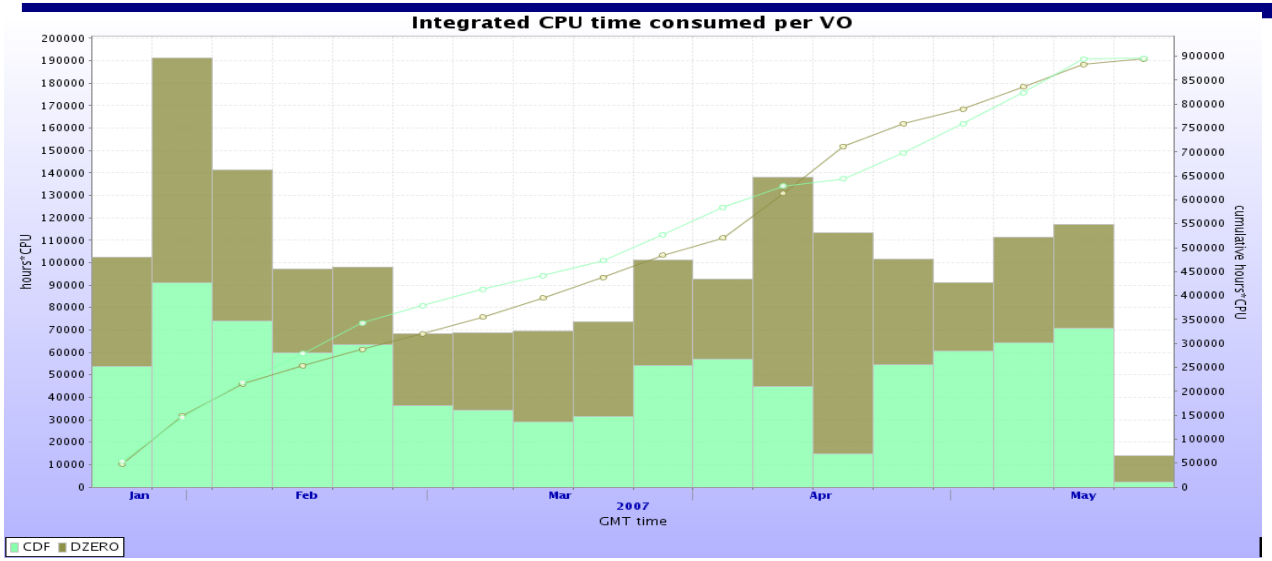




**CCIN2P3 (native SAMGrid)**



# OSG CPU Usage




Last week in average:

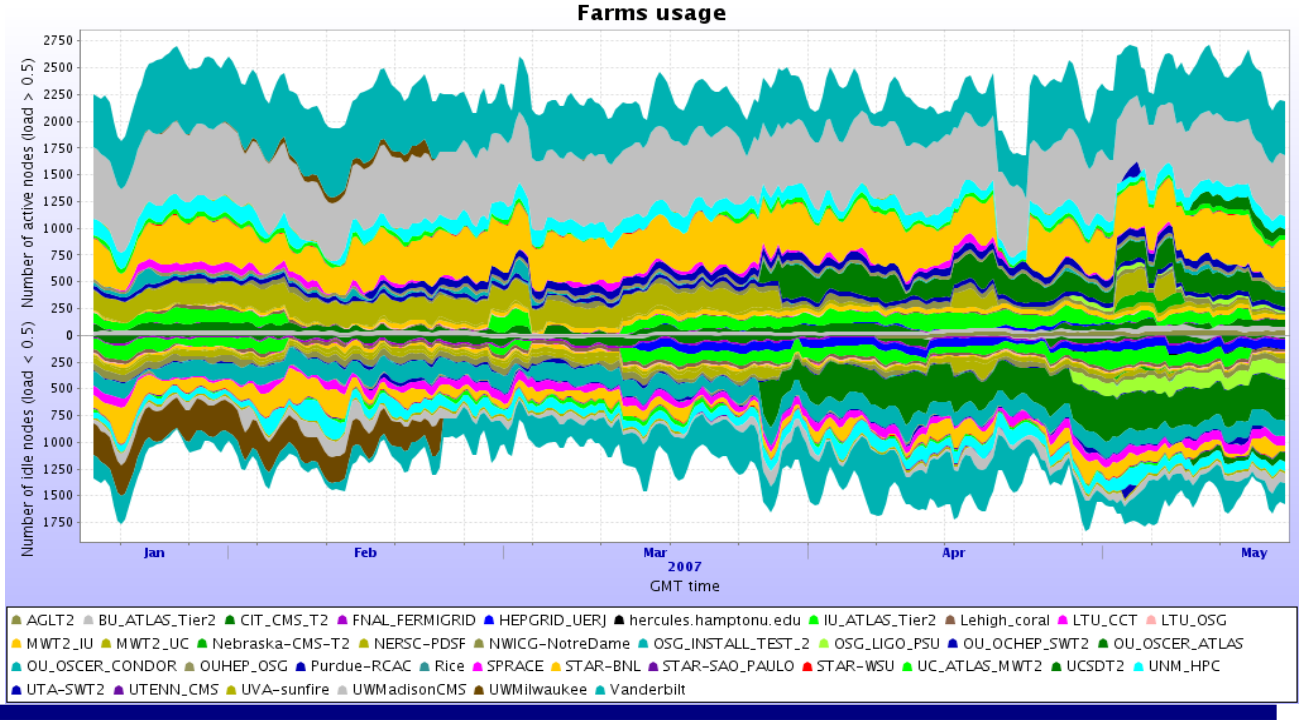
D0 ~270 CPU

CDF ~420 CPU

Last 4 months in average per experiment: ~310 CPU

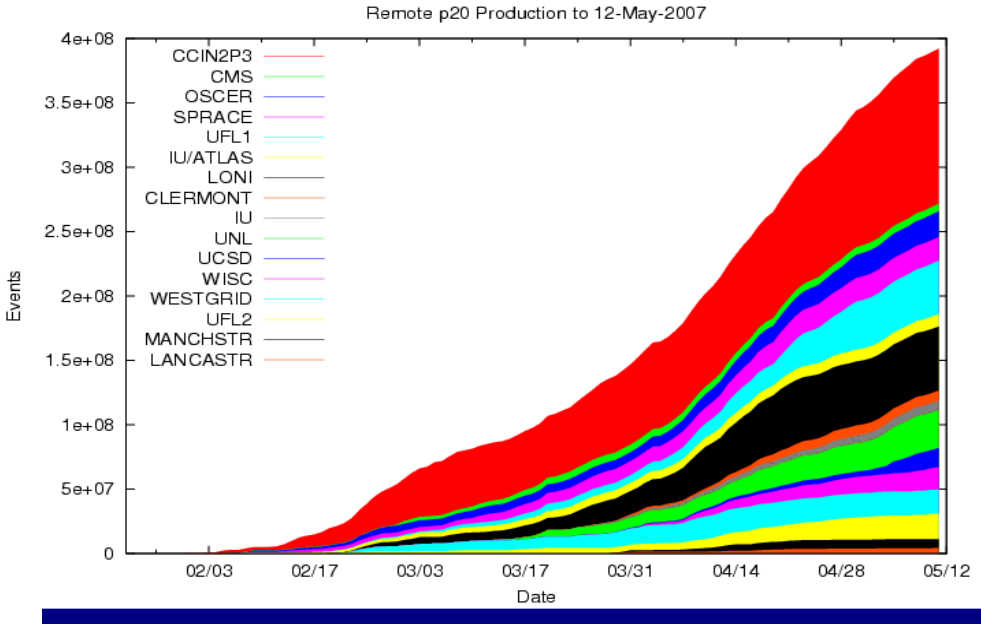
http://grid02.uits.indiana.edu:8080

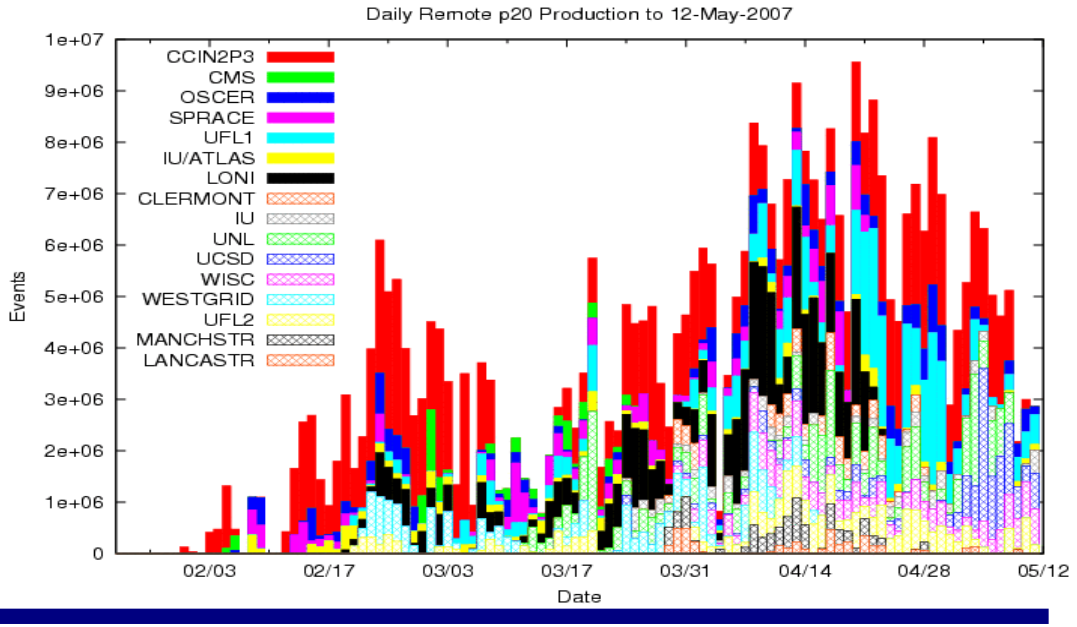
4 months usage ~65% (all users)



# Example: D0 Reprocessing 2007

- OSG Trouble shooting team setup to solve problems. → Much higher efficiencies seen at OSG sites
- SAM Grid team also solved many causes of inefficiencies
- Successful reprocessing effort: D0 able to process up to 8M/day when all sites available and 4M in average per day off-site: higher rate as originally anticipated



# User Analysis

- CDF: Fermilab analysis farm + external resources
  - Analysis running on-site and on dedicated and grid resources
  - 3 standard ntuples flavors covered 95% of users
- D0: Fermilab analysis farm + external resources
  - Analyzes are running on-site and on dedicated resources
  - Grid is used for production, started to for analysis. For example, top mass matrix element analysis requested 300,000 CPU hours and was able to obtain these amount within one month on grid
  - One common ROOT Tree format + B-physics one

| Event size | RAW | Reconstructed | Analysis |
|---|---|---|---|
| D0 | 190 – 260 kB | ~145 kB | 35 – 60 kB |
| CDF | 150 – 200 kB | 150 – 220 kB | ~ 50 kB |

| | D0 | CDF |
|---|---|---|
| Production time per event | ~1 sec/evt | ~0.05 sec / evt |
| Average time in user analysis | 0.1 – 0.005 sec/evt | 0.01 – 0.001 sec/evt |

# Common Analysis Format

## Does a X experiment needs the common format?

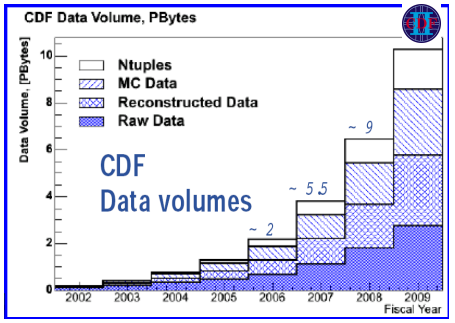
- **Disadvantages**: usually have more information than needed individual analysis ⇒ larger size of the data sets, lower processing speed
- **BUT**: D0 had a very painful experience before introducing the common analysis format because:
  - Tools developed by different physics groups are not compatible
  - No way to ensure the “certified” object selection (electron, jets, muon, ...)
  - Maintaining different formats are difficult (efforts duplication). Limited documentation, usually obsolete
  - Computing resources used inefficiently: the same dataset could be translated several times to the different formats
- One format is an ideal, which is difficult to reach. Practically experiments converged on 2 – 3 formats.

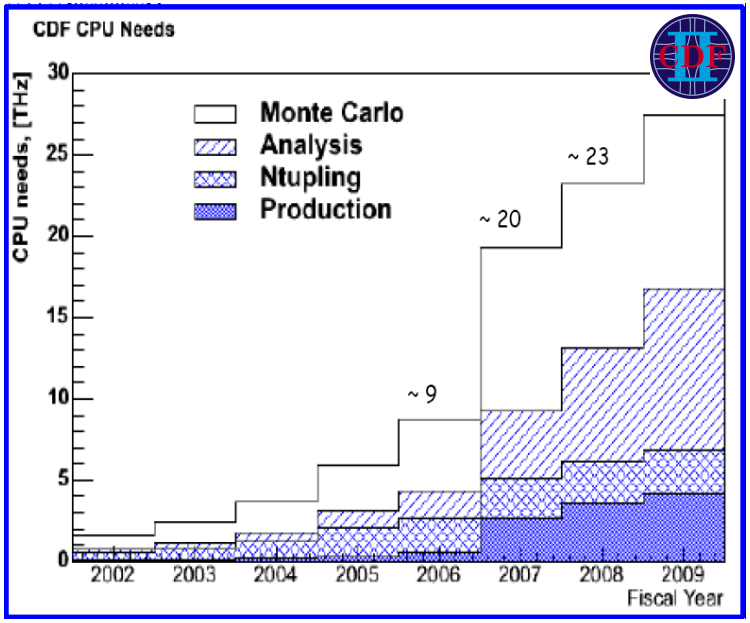
# Common Analysis Format: D0 example

- ROOT Tree based format use C++ classes for object description and cross-references between different objects
- Allow direct skimming of trees (centrally coordinate production)
- Configurable: the branches with non used information could be dropped out to reduced analysis data set size
- Common analysis format is decoupled from the D0 software and can be used in stand alone mode (suitable for making analysis at any linux node)
- Contains set of the standard tools
  - Standard object selection (several certified qualities per object)
  - Standard MC corrections (no, we can't reproduce all effects in native MC and need to apply ad-hoc corrections)
  - Statistics tools (efficiencies calculation)
  - Tools for systematics studies

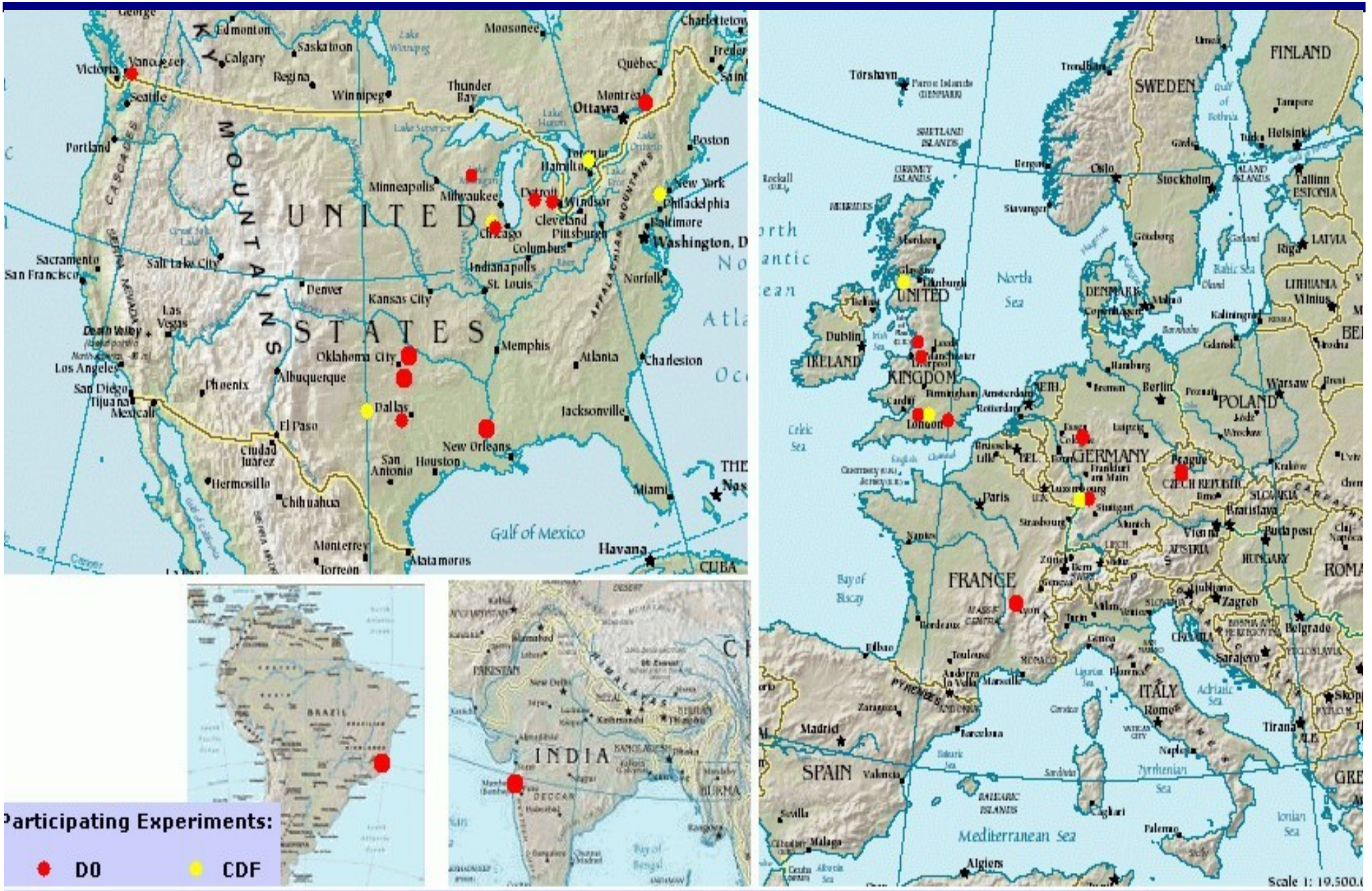
# Conclusion

- 3x increasing in integrated luminosity could not be manged with Fermilab computing resources only. Distributed computing for most of the tasks is a solution
- Both D0 and CDF use actively dedicated and opportunistic off-site resources.
  - Most of the MC production done off-site
  - Will move raw reconstruction to FermiGrid (part of OSG)
  - Expanding all other computing activities off-site.
  - Use grid for the analysis purposes.
- Use of the common format is important to provide the solid physics results, but practically 2 – 3 formats seems achievable in the experiment.



# BACK-UP

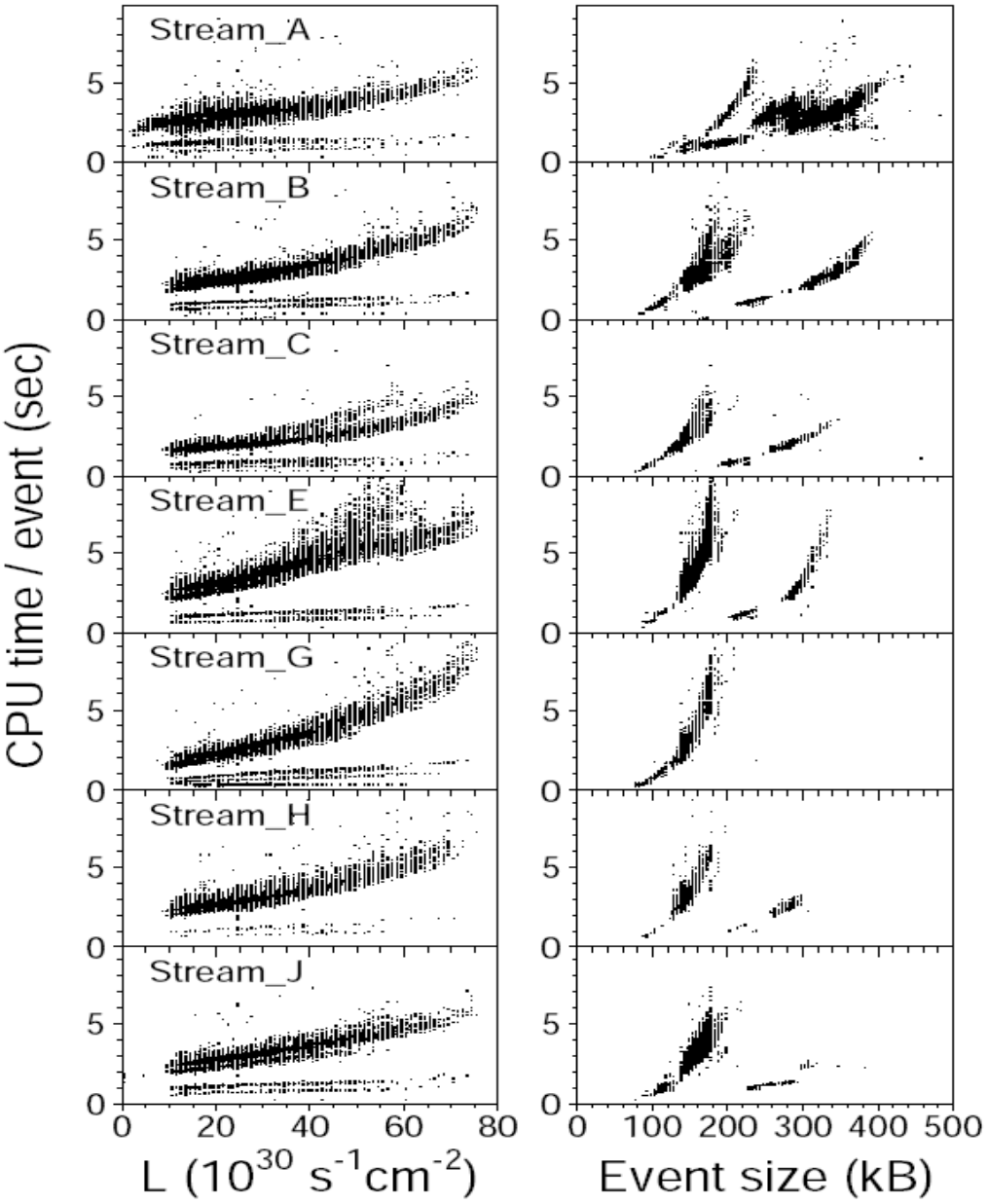

UNITED STATES
MOUNTAINS
Gulf of Mexico
Havana
CUBA
Mexico City
Monterrey
Matamoros
Torreon
Hermosillo
Chihuahua
Ciudad Juarez
El Paso
Albuquerque
Phoenix
Tijuana
Mexicali
Los Angeles
San Diego
Las Vegas
Death Valley
Sacramento
San Francisco
Salt Lake City
Denver
Kansas City
Oklahoma City
Dallas
Houston
San Antonio
New Orleans
Memphis
Atlanta
Jacksonville
Charleston
Miami
Chicago
Minneapolis
Milwaukee
Detroit
Cleveland
Pittsburgh
Columbus
Indianapolis
St. Louis
Norfolk
Baltimore
Washington, D
Philadelphia
New York
Boston
Montreal
Ottawa
Quebec
Toronto
Hamilton
Windsor
Thunder Bay
Winnipeg
Regina
Saskatoon
Calgary
Edmonton
Victoria
Vancouver
Seattle
Portland
Moosonee
Charlottetown
Fredericton
Saint John
Lake Superior
Lake Huron
Lake Michigan
Lake Erie
Lake Ontario
Lake Winnipeg
Great Salt Lake
Great Bear Lake
CASCADES
APPALACHIAN MOUNTAINS
SIERRA MADRE
BRAZIL
INDIA
PAKISTAN
HIMALAYAS
DECCAN
BURMA
Bay of Bengal
FINLAND
SWEDEN
Torshavn
Faroe Islands
SHETLAND ISLANDS
ORKNEY ISLANDS
Stavanger
Bergen
Oslo
Stockholm
Helsinki
Tampere
Gulf of Bothnia
ALAND ISLANDS
Tallinn
ESTONIA
LATVIA
Riga
Baltic Sea
LITHUANIA
Vilnius
Kaliningrad
Gdansk
Warsaw
POLAND
Berlin
Bremen
Hamburg
Leipzig
Dresden
Wroclaw
Poznan
Lodz
Krakow
Prague
CZECH REPUBLIC
SLOVAKIA
Bratislava
Vienna
AUSTRIA
HUNGARY
Budapest
Zagreb
SLOVENIA
Belgrade
Sarajevo
BOSNIA AND HERZEGOVINA
YUGOSLAVIA
Tirana
ALB.
ROMA
GERMANY
Frankfurt am Main
Stuttgart
Munich
Strasbourg
Luxembourg
Paris
Lille
Brussels
Amsterdam
Rotterdam
NETH.
Cologne
Essen
Leipzig
Dresden
Wroclaw
Poznan
Lodz
Krakow
Prague
CZECH REPUBLIC
SLOVAKIA
Bratislava
Vienna
AUSTRIA
HUNGARY
Budapest
Zagreb
SLOVENIA
Belgrade
Sarajevo
BOSNIA AND HERZEGOVINA
YUGOSLAVIA
Tirana
ALB.
ROMA
UNITED KINGDOM
IRELAND
Dublin
Cardiff
London
Birmingham
Liverpool
Manchester
Edinburgh
Glasgow
Belfast
North Sea
Celtic Sea
English Channel
Irish Sea
DENMARK
Copenhagen
Malmo
Gothenburg
Aarhus
Esbjerg
Skagerrak
Kattegat
FRANCE
Bay of Biscay
Bordeaux
Toulouse
Lyon
Nantes
Marseille
Nice
Monaco
Andorra la Vella
Pyrenees Mountains
SPAIN
Madrid
Barcelona
Valencia
Zaragoza
Bilbao
Seville
Malaga
Balearic Islands
Mediterranean Sea
ITALY
Rome
Naples
Palermo
Cagliari
Sardinia
Corsica
Tyrrhenian Sea
Adriatic Sea
Ionian Sea
Algiers
Scale 1: 19,500,0
Participating Experiments:
D0
CDF

# CDF output data streams

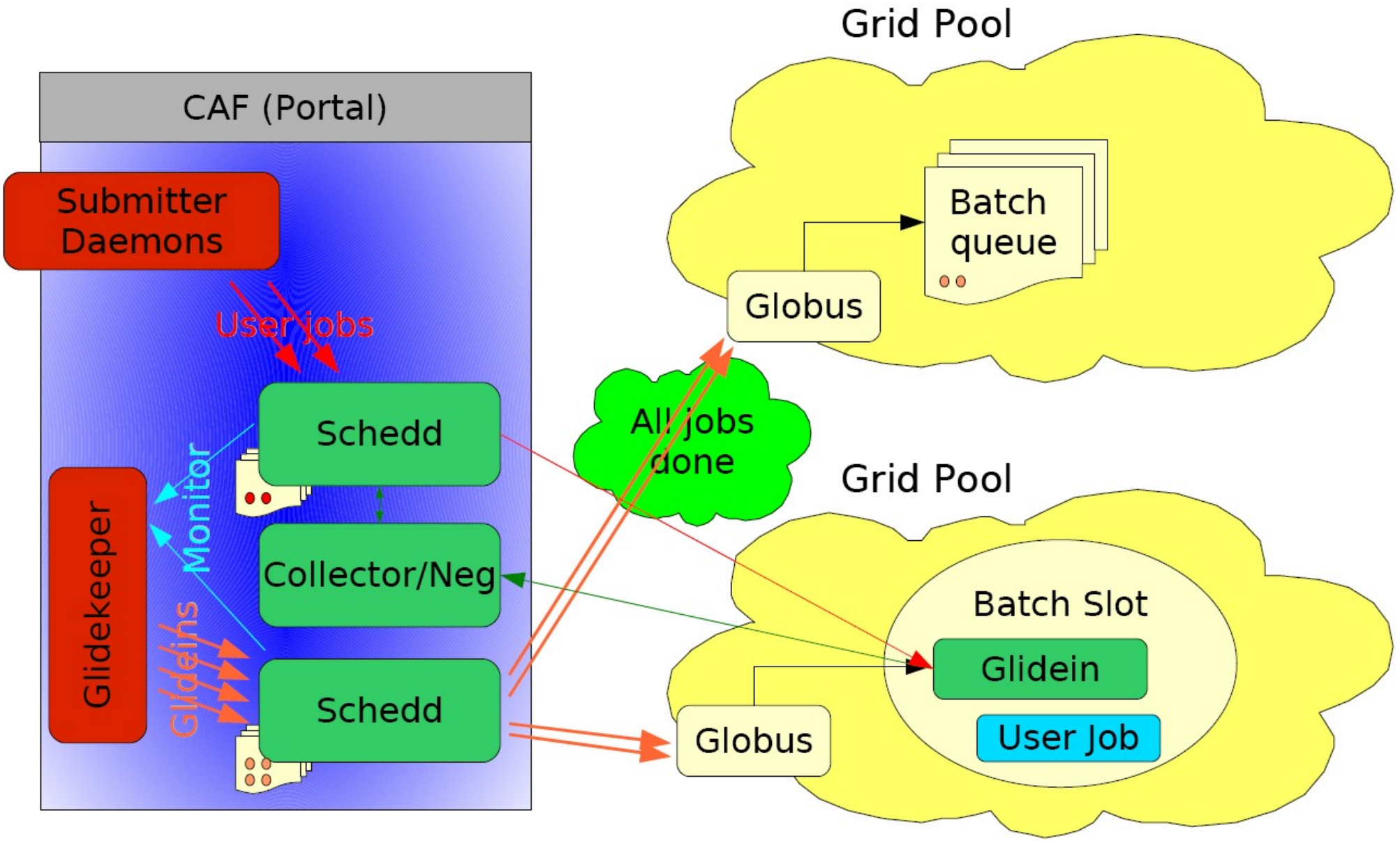
| Stream | trigger contents | |
|---|---|---|
| | event size (kByte) | ratio to total size (%) |
| A | monitoring | |
| | 195 | 2.5 |
| B | high $E_T$ leptons, missing $E_T$ | |
| | 140 | 11.5 |
| C | high $E_T$ photons, di-photons | |
| | 122 | 13.0 |
| D | study stream | |
| | 344 | 11.2 |
| E | tau, lepton pair, Z to bb, higgs | |
| | 146 | 16.0 |
| G | QCD jet, Di-jet, Miss $E_t$ plus jets, high $P_T$ b-jet | |
| | 139 | 11.2 |
| H | $B$ to $\pi\pi$, $B_s$ to $D_s\pi$ | |
| | 136 | 24.0 |
| J | $J/\psi$ to leptons, $\Upsilon$ to leptons | |
| | 146 | 10.6 |

TABLE I

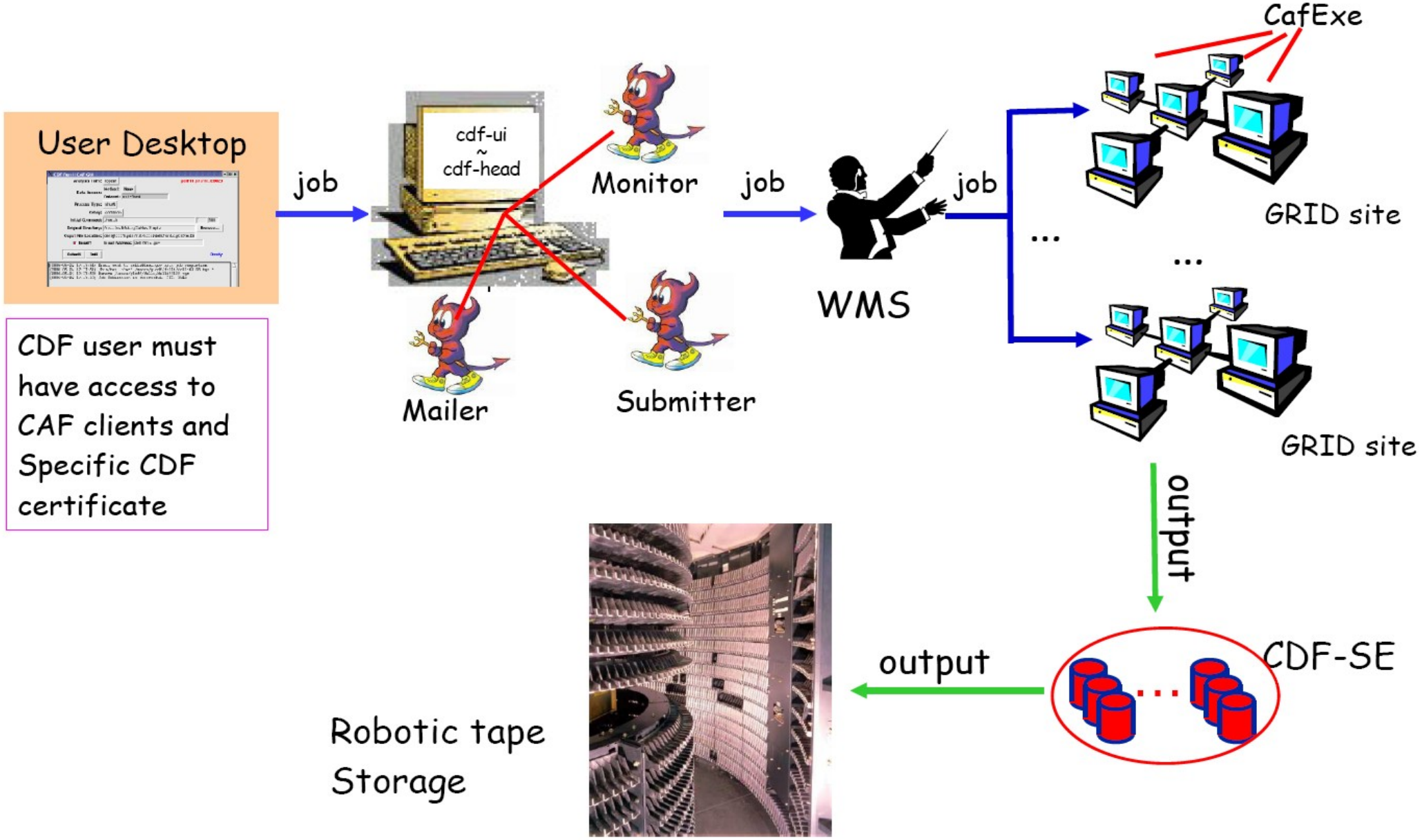
STATISTICS OF DATA STREAMS IN 2005. LISTED ARE THE TRIGGER CONTENTS, AVERAGE EVENT SIZE AND THE RATIO TO TOTAL SIZE.



# CDF NAmCAF


Grid Pool
CAF (Portal)
Submitter Daemons
User jobs
Schedd
Collector/Neg
Schedd
Glidekeeper
Monitor
Glideins
All jobs done
Grid Pool
Batch queue
Globus
Batch Slot
Glidein
User Job
Globus

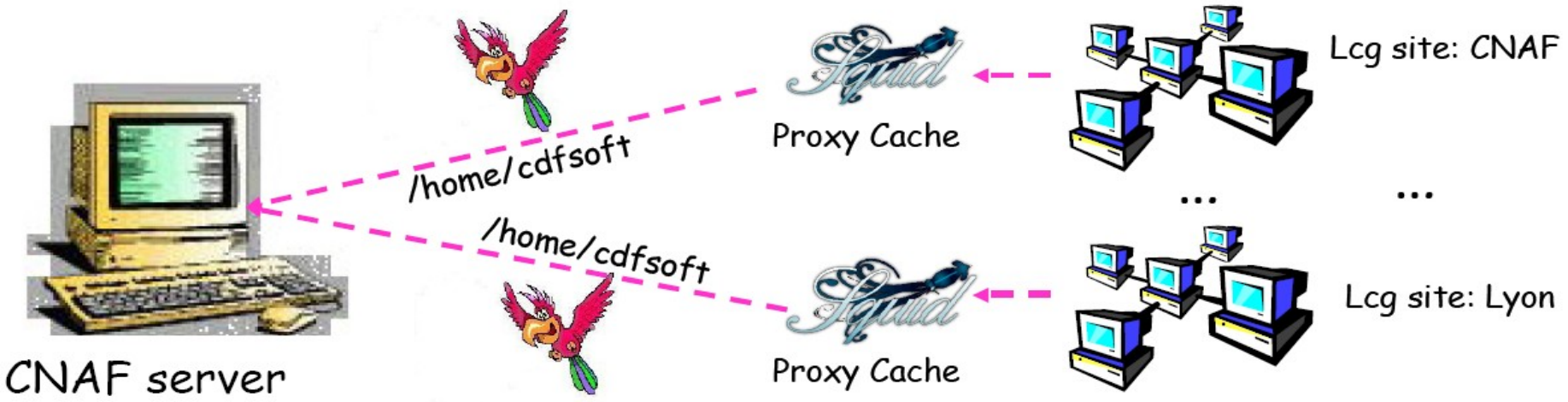
# CDF: LcgCAF




CDF user must have access to CAF clients and Specific CDF certificate

Developed by: F.Delli Paoli, D.Jeans, D.Lucchesi, S.Sarkar, I.Sfiligoi

# CDF code distribution




In order to run, a CDF MC job needs available to the WN:

- CDF code distribution:

  Parrot is used as virtual file system for CDF code

  To have good performances with Parrot, a Squid web proxy cache is also needed.

- Run Condition DB:

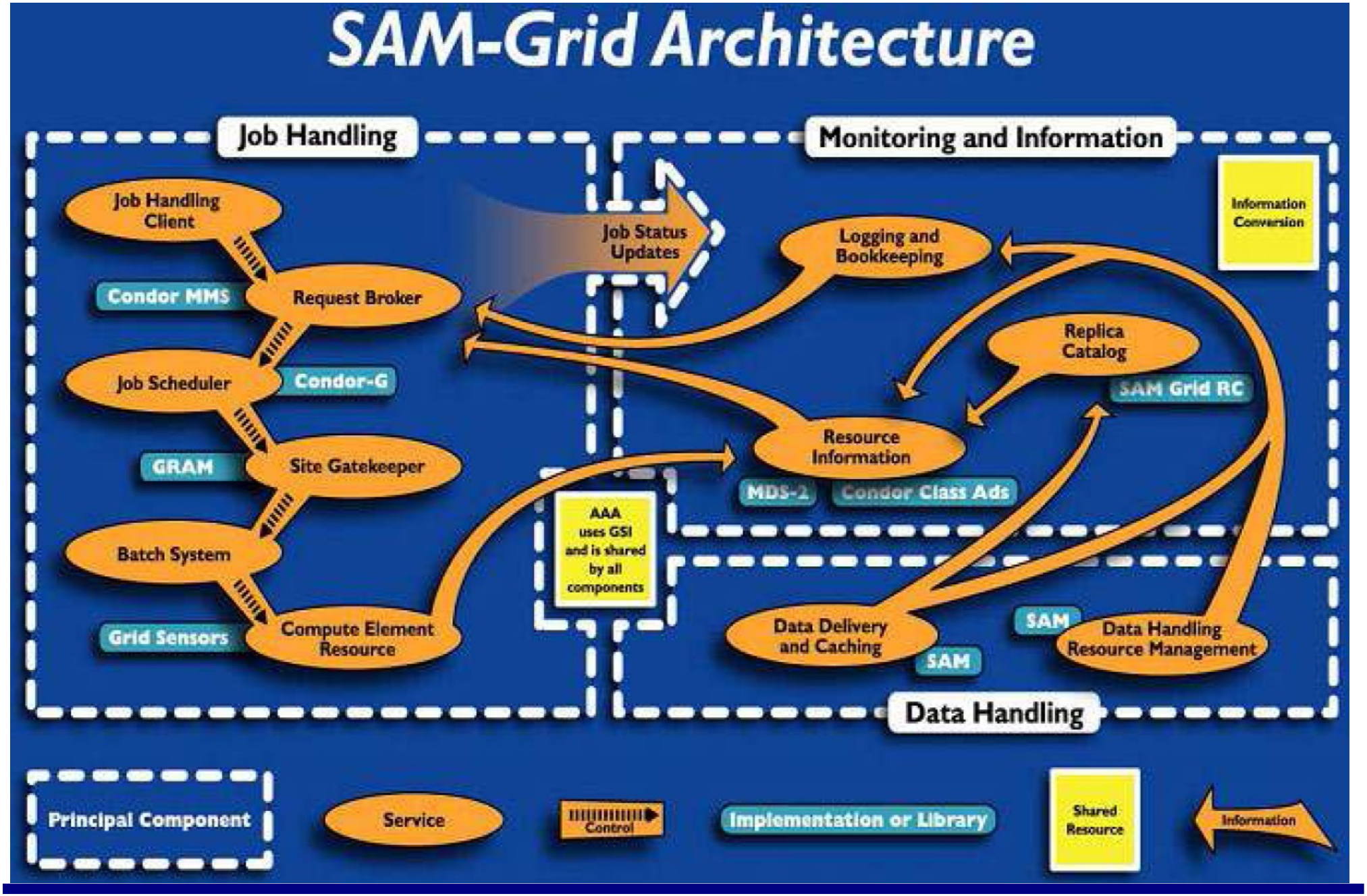
  FroNTier is used for DB access (...in NamCAF too)

- Job monitor (implementation different from CAF, same functionalities)

# SAMGrid

SAMGrid is a general data handling system designed to be performant for experiments with large (petabyte-sized) datasets and widely distributed production and analysis facilities. The components now in production provide a versatile set of services for data transfer, data storage, and process bookkeeping on distributed systems. Components now in testing add the capability of job submission to a Grid, built around standard Grid middleware from Condor and Globus. SAMGrid is used by D0 and is being tested for use by CDF and MINOS.

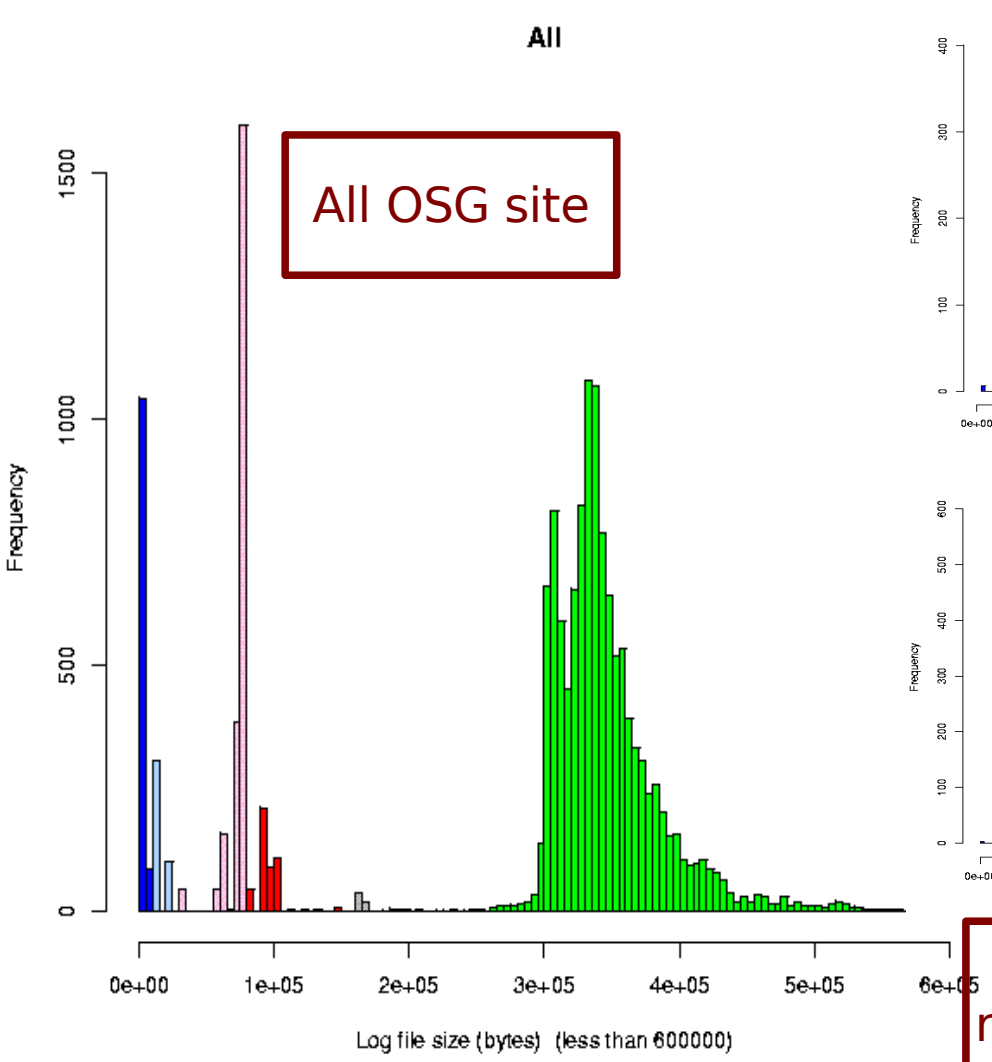
http://projects.fnal.gov/samgrid/

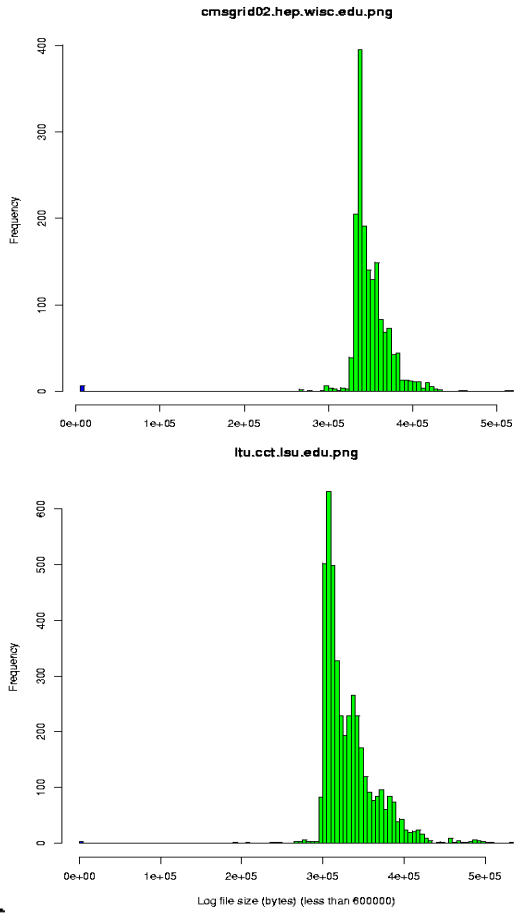

SAM-Grid Architecture
Job Handling
Monitoring and Information
Job Handling Client
Condor MMS
Request Broker
Job Scheduler
Condor-G
GRAM
Site Gatekeeper
Batch System
Grid Sensors
Compute Element Resource
Job Status Updates
AAA uses GSI and is shared by all components
Logging and Bookkeeping
Information Conversion
Replica Catalog
SAM Grid RC
Resource Information
MDS-2
Condor Class Ads
Data Delivery and Caching
SAM
SAM
Data Handling Resource Management
Data Handling
Principal Component
Service
Control
Implementation or Library
Shared Resource
Information

# Failure analysis (log file size)

**Failures:**

- Worker node incomparability. Lost standard output. OSG no assign.
- Forwarding node problems.
- SAM problem.
- SAM problem.
- D0 start-up script crash.
- **Successful**




Some sites are running just perfect

# Open Science Grid (OSG)

The Open Science Grid is a distributed computing infrastructure for scientific research.

The OSG Consortium's unique alliance of U.S. universities, laboratories, scientific collaborations and software developers brings petascale computing and storage resources into a uniform shared cyberinfrastructure. Members of the OSG Consortium contribute effort and resources to the OSG infrastructure and reap the benefits of a shared infrastructure that integrates computing and storage resources from more than 50 sites in the United States, Asia and South America.

The OSG is supported by the National Science Foundation and the U.S. Department of Energy's Office of Science.

http://www.opensciencegrid.org

The mission of the LHC Computing Project (LCG) is to build and maintain a data storage and analysis infrastructure for the entire high energy physics community that will use the LHC.

http://lcg.web.cern.ch

# US CMS Grid

The US CMS Grid is part of the WorldWide LHC Computing Grid for LHC science analysis. The US CMS grid environment is part of the Open Science Grid infrastructure. The national and international grid infrastructures such as OSG and the LHC Computing Grid Project provide the basis for the Worldwide CMS computing system. The USCMS Grid software infrastructure is a collaboration across CMS, with the Fermilab Computing Division, with other physics groups in the US and with computer scientists at various universities.

http://www.uscms.org/SoftwareComputing/Grid

# FermiGrid

In order to better serve the entire program of Fermilab, the Computing Division has undertaken the strategy of placing all of its production resources in a Grid "meta-facility" infrastructure called FermiGrid. This strategy is designed to allow Fermilab:

- to insure that the large experiments ... to have first priority usage of those resources that are purchased on their behalf. To allow opportunistic use of these dedicated resources, as well as other shared Farm and Analysis resources, by various Virtual Organizations (VO's) ...
- to optimise use of resources at Fermilab. To make a coherent way of putting Fermilab on the Open Science Grid. To save some effort and resources by implementing certain shared services and approaches. To work together more coherently to move all of our applications and services to run on the Grid.
- to better handle a transition from Run II to LHC in a time of shrinking budgets and possibly shrinking resources for Run II worldwide. To fully support Open Science Grid and the LHC Computing Grid and gain positive benefit from this emerging infrastructure in the US and Europe.
- http://fermigrid.fnal.gov

# Disk Cache (dCache)

The goal of this project is to provide a system for storing and retrieving huge amounts of data, distributed among a large number of heterogenous server nodes, under a single virtual filesystem tree with a variety of standard access methods.

http://www.dcache.org

# Condor

- The goal of the Condor® Project is to develop, implement, deploy, and evaluate mechanisms and policies that support High Throughput Computing (HTC) on large collections of distributively owned computing resources.
- http://www.cs.wisc.edu/condor