DØ Computing

Tibor Kurča
IPN Lyon

- **Introduction**

  - Fermilab & Tevatron & DØ Experiment

- **DØ Computing Model**

  1. data grid : **SAM**

  2. computing grid : **SAMGrid**
  MC-production, reprocessing, fixing

  3. grids interoperability : **SAMGrid/LCG (OSG)**
  MC-production, fixing, reprocessing

- **Summary**

DØ – Tevatron - FNAL

- DØ, CDF - 2 experiments
- Fermi National Laboratory

  30 miles west from Chicago
- Tevatron is the world’s highest energy accelerator

  → the most likely place to directly discover a new particles or forces

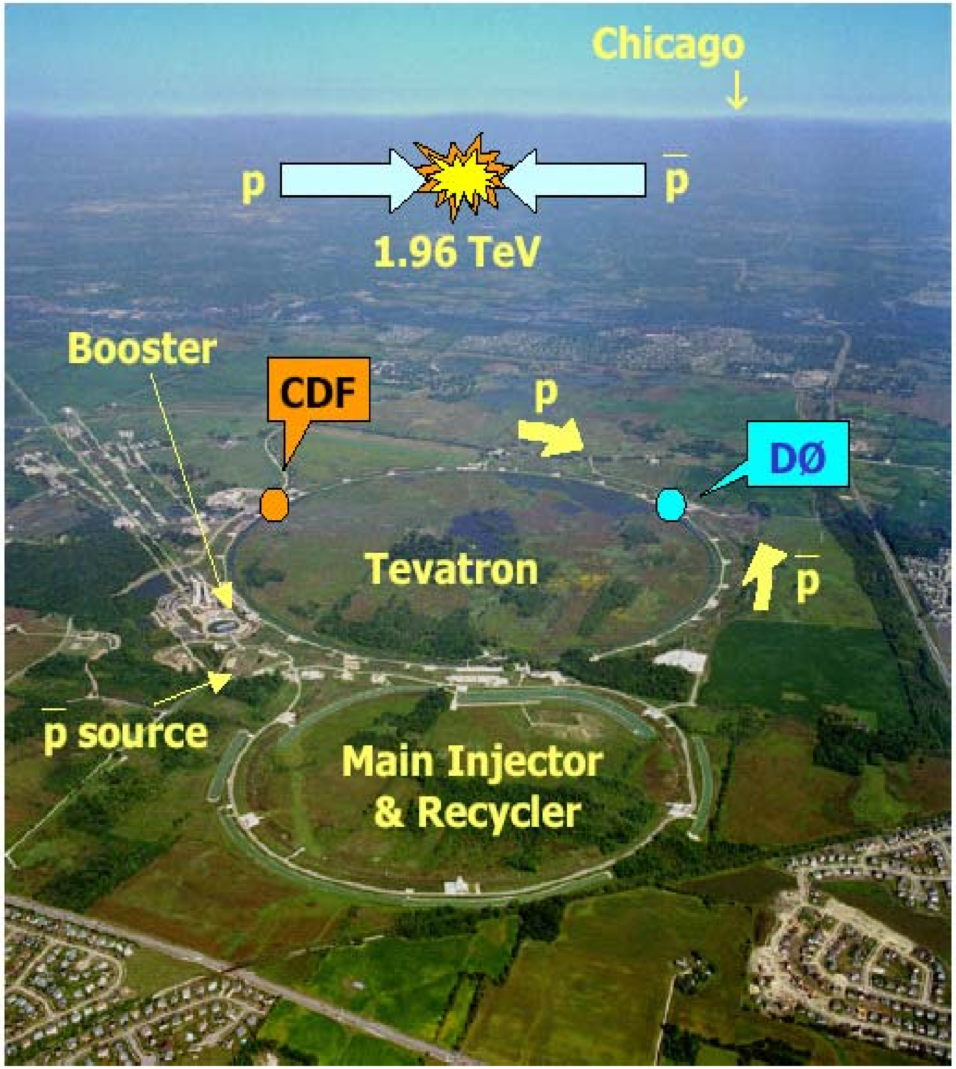
  searches for Higgs, SUSY

  → more general theories predictions can be tested

  →precise measurements of the Standard Model

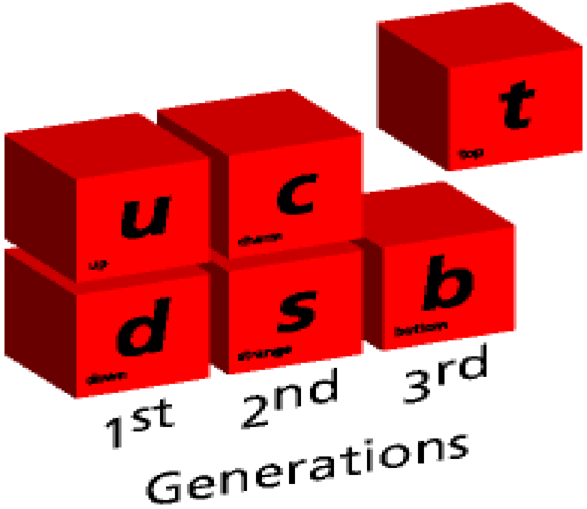
  ……

  →Surprises?



Tevatron Collider History

- Oct 13, 1985 - 1st collisions in CDF @ 1.6 TeV (1.6x10^{12} eV)
- Oct 21, 1986 - 1st 900 GeV beam → cms energy 1.8 TeV
- Feb 14, 1992 - DØ detector commissioning
- Mar 3, 1995 - Top quark discovery by CDF&DØ
- Sept, 1997 - end of Run I



- May, 2001 - start of Run II (980 GeV beams → Ecms 1.96 TeV)
- June 2006 - Run IIb

Fermilab ...Then

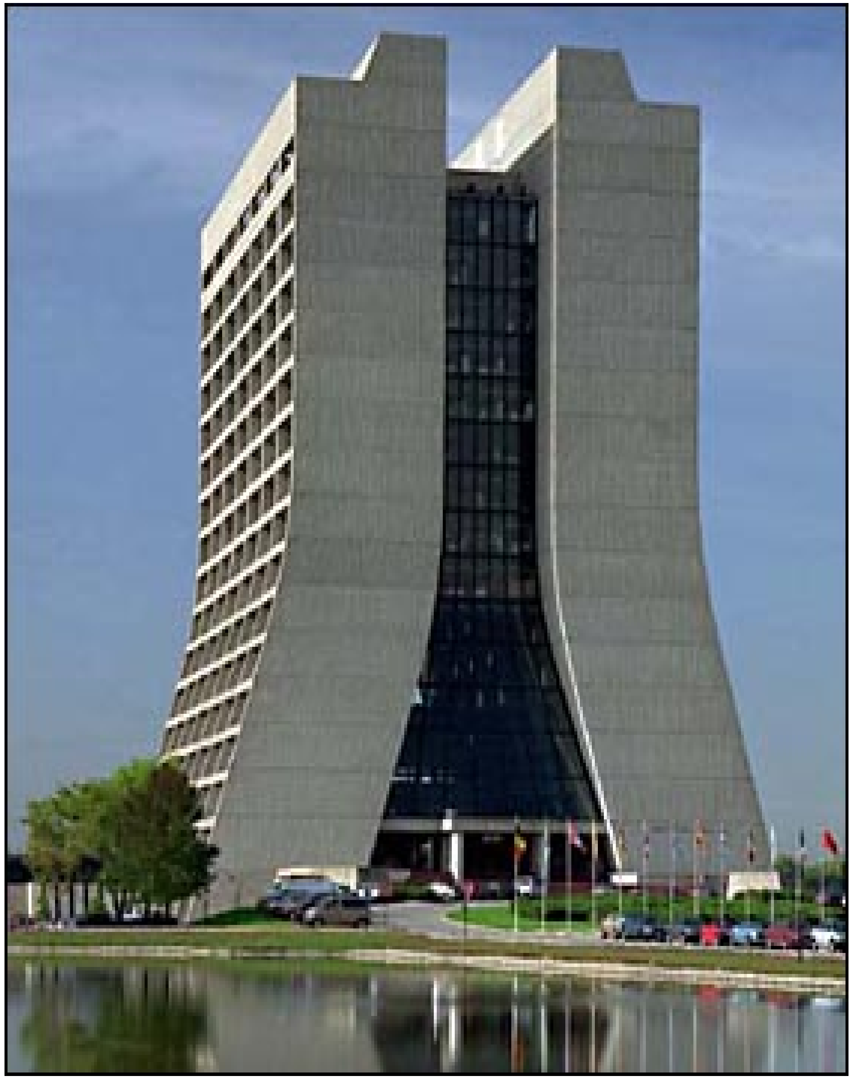
Welcome to
WESTON
FUTURE ATOMIC RESEARCH CAPITOL
Arthur Theriault – Village President
COURTESY OF
HAROLD M. CONN ASSOCIATES
REAL ESTATE BROKERS
161 E ERIE ST. CHICAGO 60611 TEL SU 7-8543

- **1966** **Weston, Illinois** (30 miles west of Chicago) selected as the site for the new **National AcceleratorLaboratory**
- **1 Dec 1968** – **groundbreaking for the 1st linear accelerator**
- **1974** renamed to **Fermi NAL - FNAL** in honour of Enrico Fermi (1938 Nobel Prize)

Fermilab

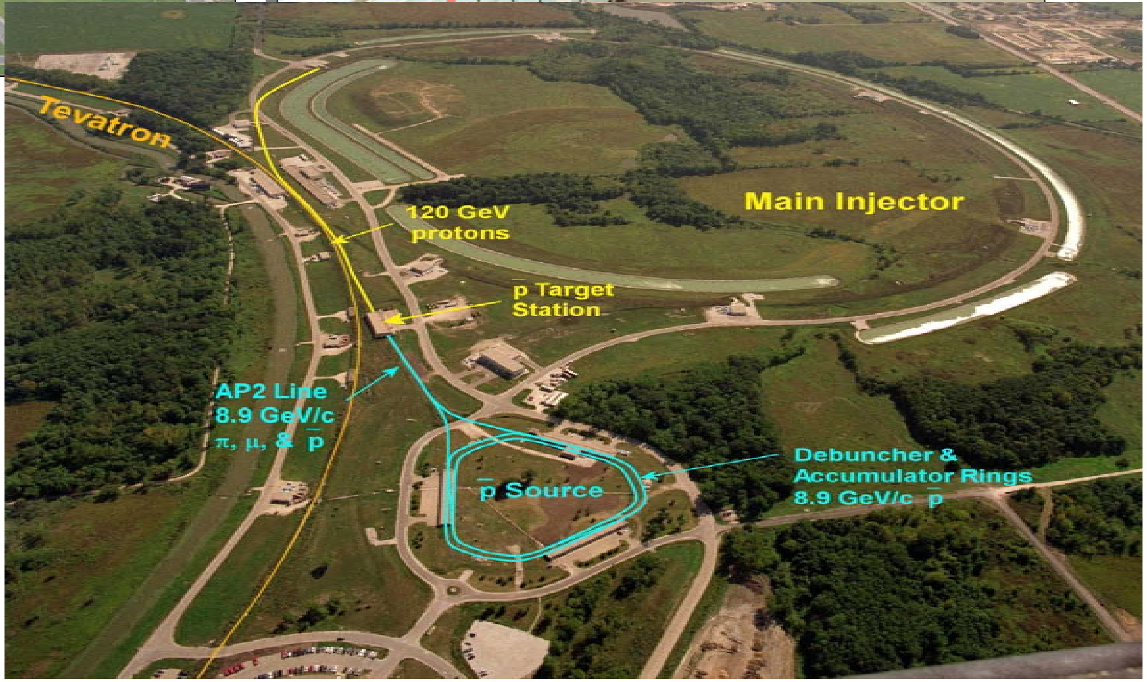
Today …

- 2200 employees;
- Funded by DOE
- Operated by consortium of ~90 Universities (mostly US)
- 6800-acre site (>10 sqmiles)

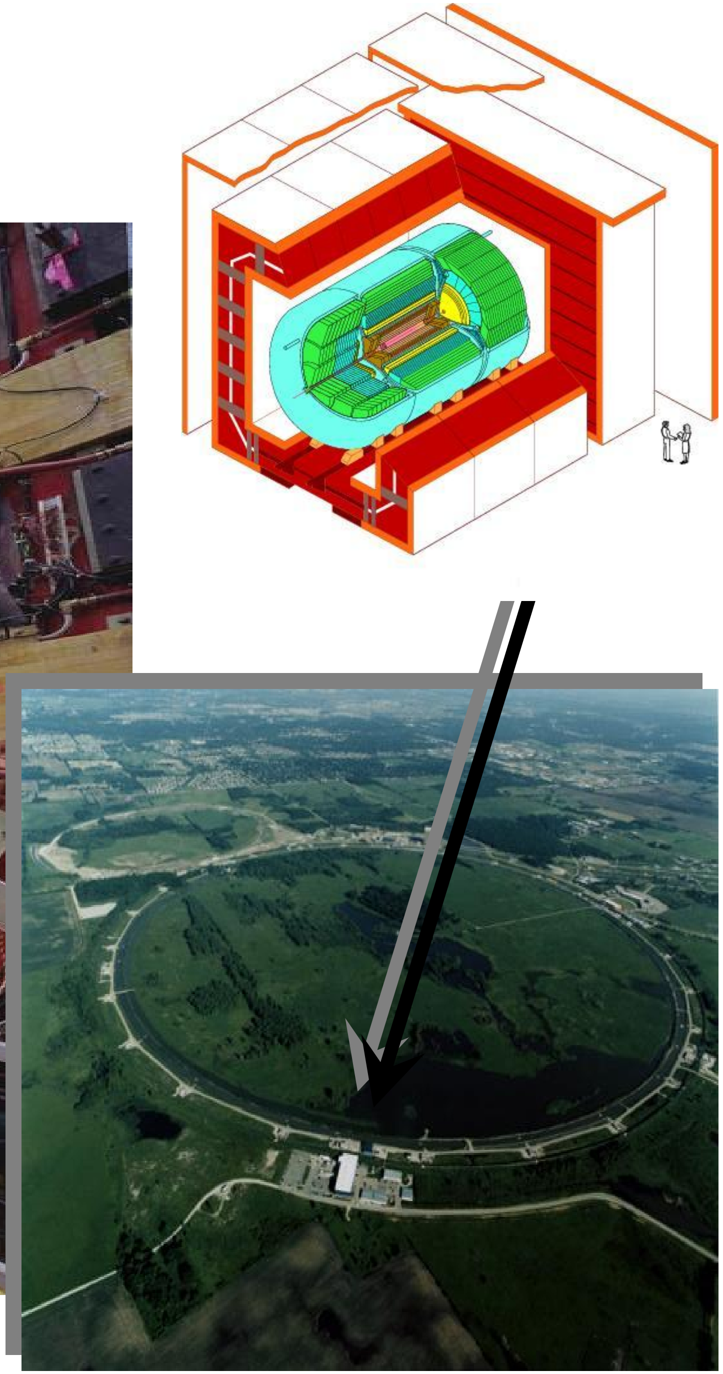


Fermilab

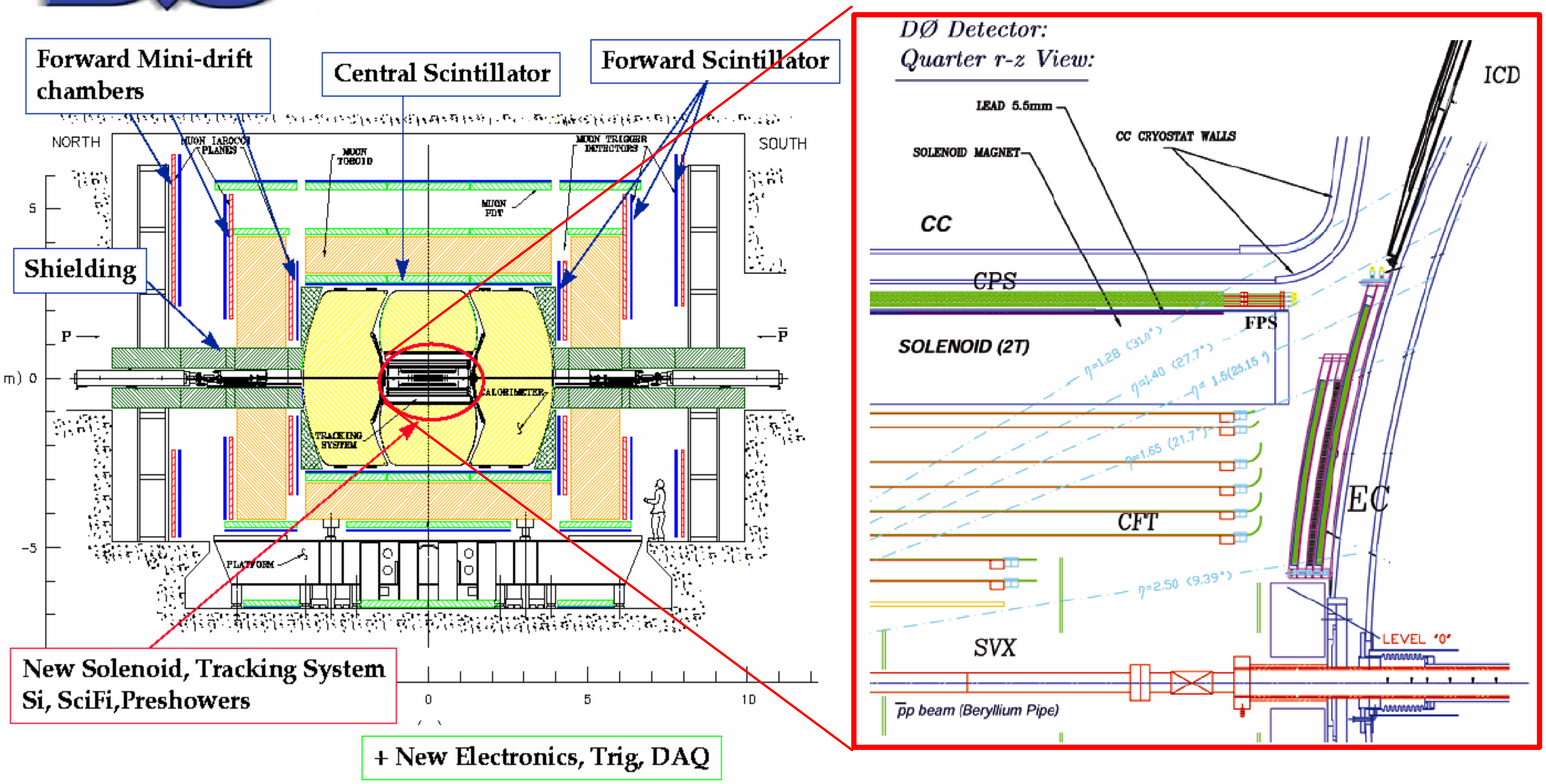
A HEP Laboratory


Tevatron
Main Injector
120 GeV protons
p Target Station
AP2 Line 8.9 GeV/c π, μ, & p̄
p̄ Source
Debuncher & Accumulator Rings 8.9 GeV/c p̄

Detector



Detector – Run II



Forward Mini-drift chambers

Central Scintillator

Forward Scintillator

Shielding

New Solenoid, Tracking System Si, SciFi,Preshowers

+ New Electronics, Trig, DAQ

- ✓ Solenoid (2T)
- ✓ Central tracker (SciFi) |η| ≤ 1.7 → L1
- ✓ Silicon vertex detector |η| ≤ 3 → L1
- ✓ Preshower - Central & Forward

- ✓ Muon forward chamber → |η| ≤ 2
- ✓ Calorimeter electronics
- ✓ Trigger system
- ✓ DAQ system

Setting the Scale I

- **Detector - Raw Data**

  ~1,000,000 Channels

  ~ 250kB Event size

  ~ 50+ Hz Event rate

  ~125 – 250 TB/year

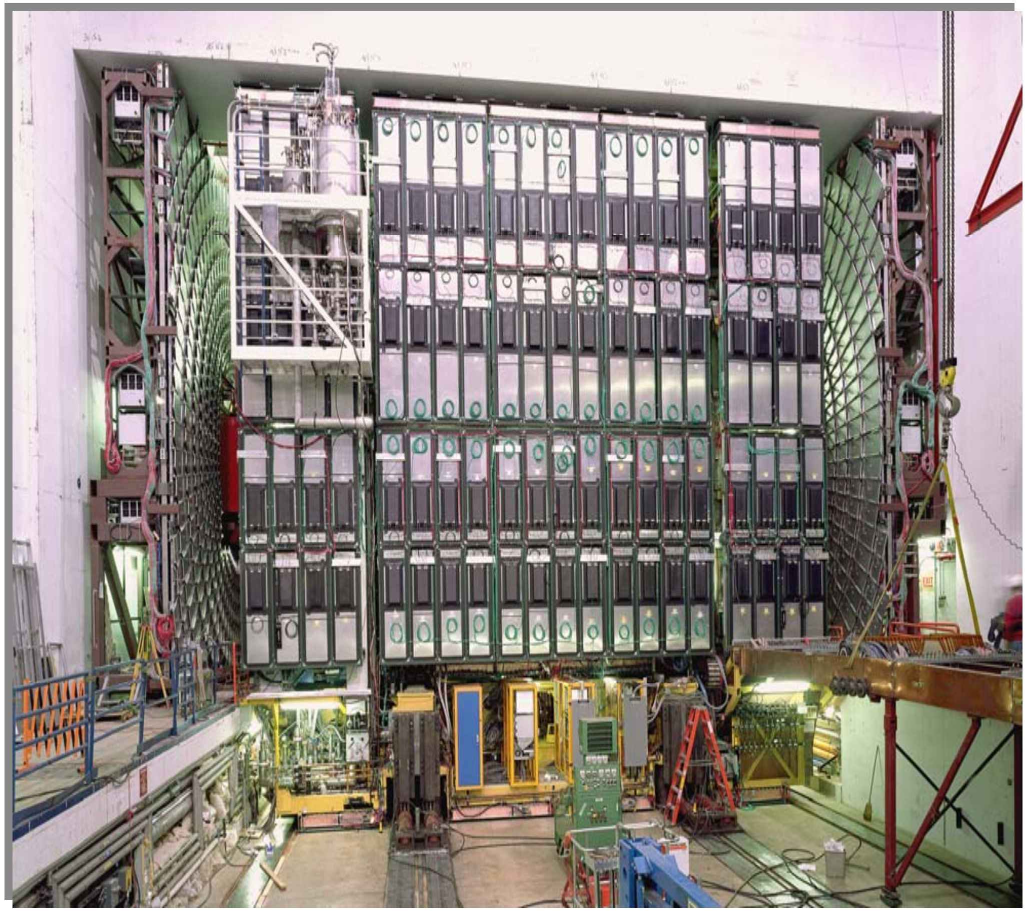
  Now: >1.5 B events

- **Total data**

  - raw, reconstructred, simulated

  Now: > 1.5 PB

  By 2008: 3.5 PB



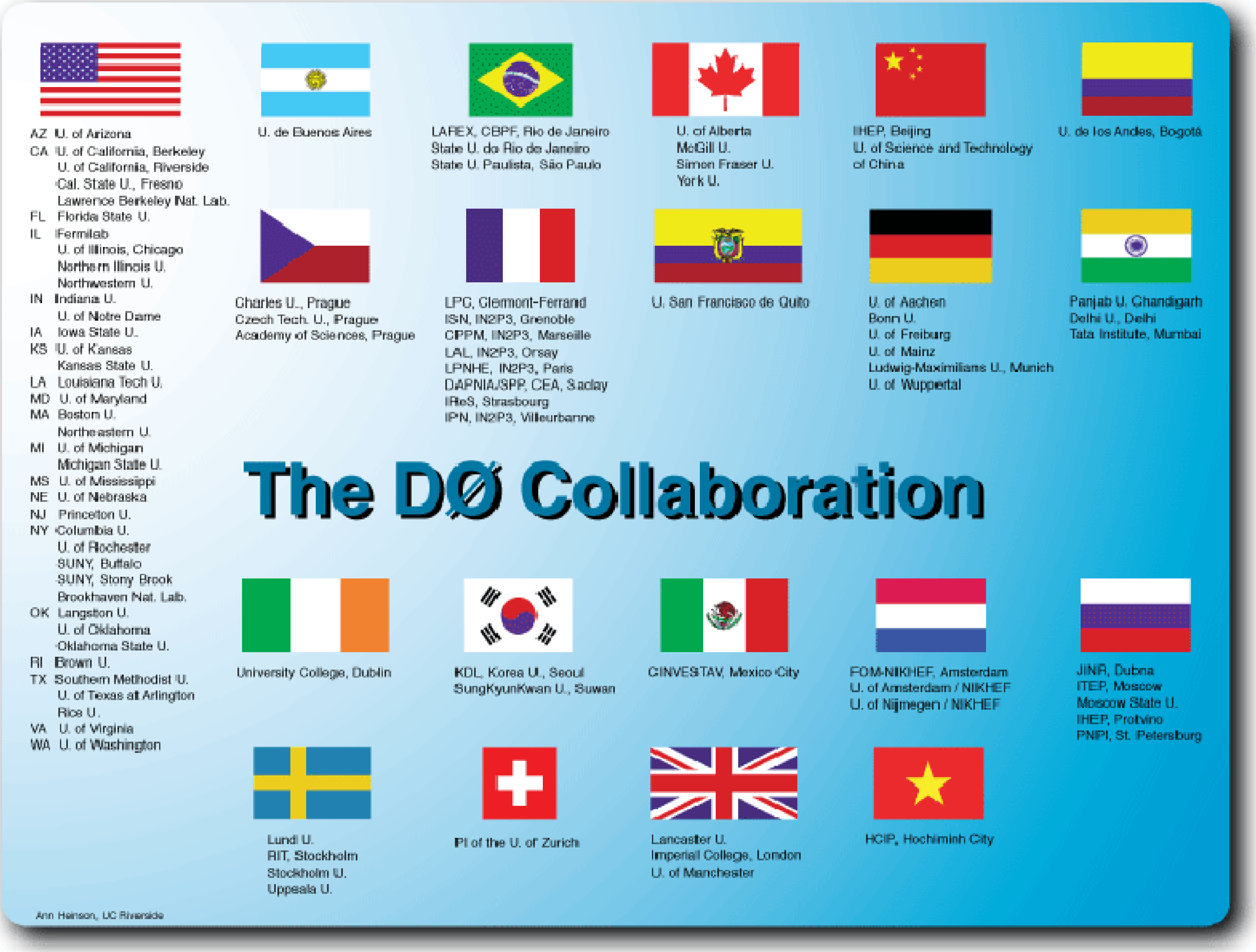
Setting the Scale II

~700 Physicists

~80 Institutions

20 Countries

DØ-France:
8 groups
~80 people



The DØ Collaboration

AZ	U. of Arizona
CA	U. of California, Berkeley
	U. of California, Riverside
	Cal. State U., Fresno
	Lawrence Berkeley Nat. Lab.
FL	Florida State U.
IL	Fermilab
	U. of Illinois, Chicago
	Northern Illinois U.
	Northwestern U.
IN	Indiana U.
	U. of Notre Dame
IA	Iowa State U.
KS	U. of Kansas
	Kansas State U.
LA	Louisiana Tech U.
MD	U. of Maryland
MA	Boston U.
	Northeastern U.
MI	U. of Michigan
	Michigan State U.
MS	U. of Mississippi
NE	U. of Nebraska
NJ	Princeton U.
NY	Columbia U.
	U. of Rochester
	SUNY, Buffalo
	SUNY, Stony Brook
	Brookhaven Nat. Lab.
OK	Langston U.
	U. of Oklahoma
	Oklahoma State U.
RI	Brown U.
TX	Southern Methodist U.
	U. of Texas at Arlington
	Rice U.
VA	U. of Virginia
WA	U. of Washington

U. de Buenos Aires

LAFEX, CBPF, Rio de Janeiro
State U. do Rio de Janeiro
State U. Paulista, São Paulo

U. of Alberta
McGill U.
Simon Fraser U.
York U.

IHEP, Beijing
U. of Science and Technology
of China

U. de los Andes, Bogotá

Charles U., Prague
Czech Tech. U., Prague
Academy of Sciences, Prague

LPC, Clermont-Ferrand
ISN, IN2P3, Grenoble
CPPM, IN2P3, Marseille
LAL, IN2P3, Orsay
LPNHE, IN2P3, Paris
DAPNIA/SPP, CEA, Saclay
IReS, Strasbourg
IPN, IN2P3, Villeurbanne

U. San Francisco de Quito

U. of Aachen
Bonn U.
U. of Freiburg
U. of Mainz
Ludwig-Maximilians U., Munich
U. of Wuppertal

Panjab U. Chandigarh
Delhi U., Delhi
Tata Institute, Mumbai

University College, Dublin

KDL, Korea U., Seoul
SungKyunKwan U., Suwan

CINVESTAV, Mexico City

FOM-NIKHEF, Amsterdam
U. of Amsterdam / NIKHEF
U. of Nijmegen / NIKHEF

JINR, Dubna
ITEP, Moscow
Moscow State U.
IHEP, Protvino
PNPI, St. Petersburg

Lund U.
RIT, Stockholm
Stockholm U.
Uppsala U.

PI of the U. of Zurich

Lancaster U.
Imperial College, London
U. of Manchester

HCIP, Hochiminh City

Ann Heinson, UC Riverside

Computing – Data Analysis

Real Data

Beam collisions

Particles traverse detector

Readout:

Electronic detector signals written to tapes

→ raw data

Monte Carlo Data

Event generation:

software modelling beam particles interactions
→ production of new particles from those collisions

Simulation:

particles transport in the detectors

Digitization:

Transformation of the particle drift times, energy deposits into the signals readout by electronics
→ the same format as real raw data

Reconstruction:

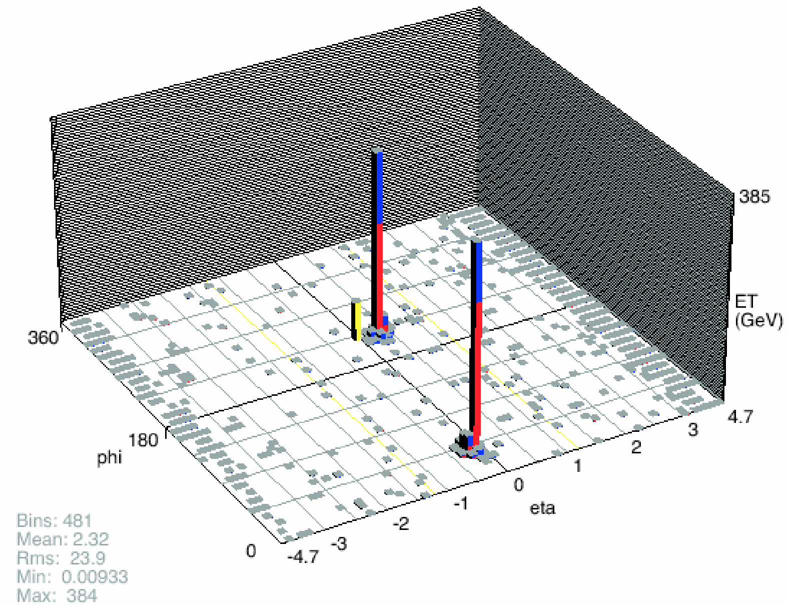
physics objects, i.e. particles produced in the beams collisions -- electrons, muons, jets…

Physics Analysis

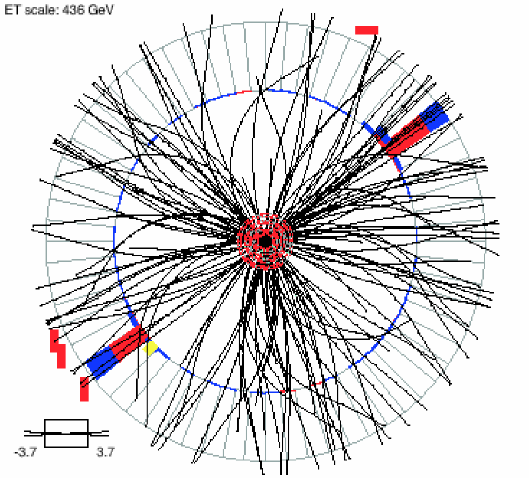
High Pt event

jet 1	jet 2
$p_T = 616\,\mathrm{GeV}$	$p_T = 557\,\mathrm{GeV}$
$y = -0.19$	$y = 0.25$
$\phi = 0.65$	$\phi = 3.78$
$M_{jj} = 1206\,\mathrm{GeV}$	

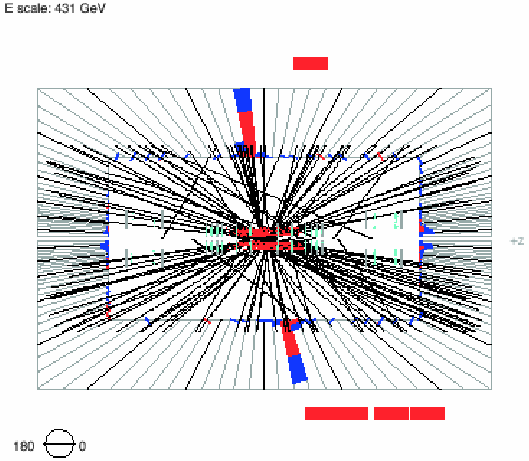
Run 178796 Event 67972991 Fri Feb 27 08:34:03 2004



Run 178796 Event 67972991 Fri Feb 27 08:34:15 2004



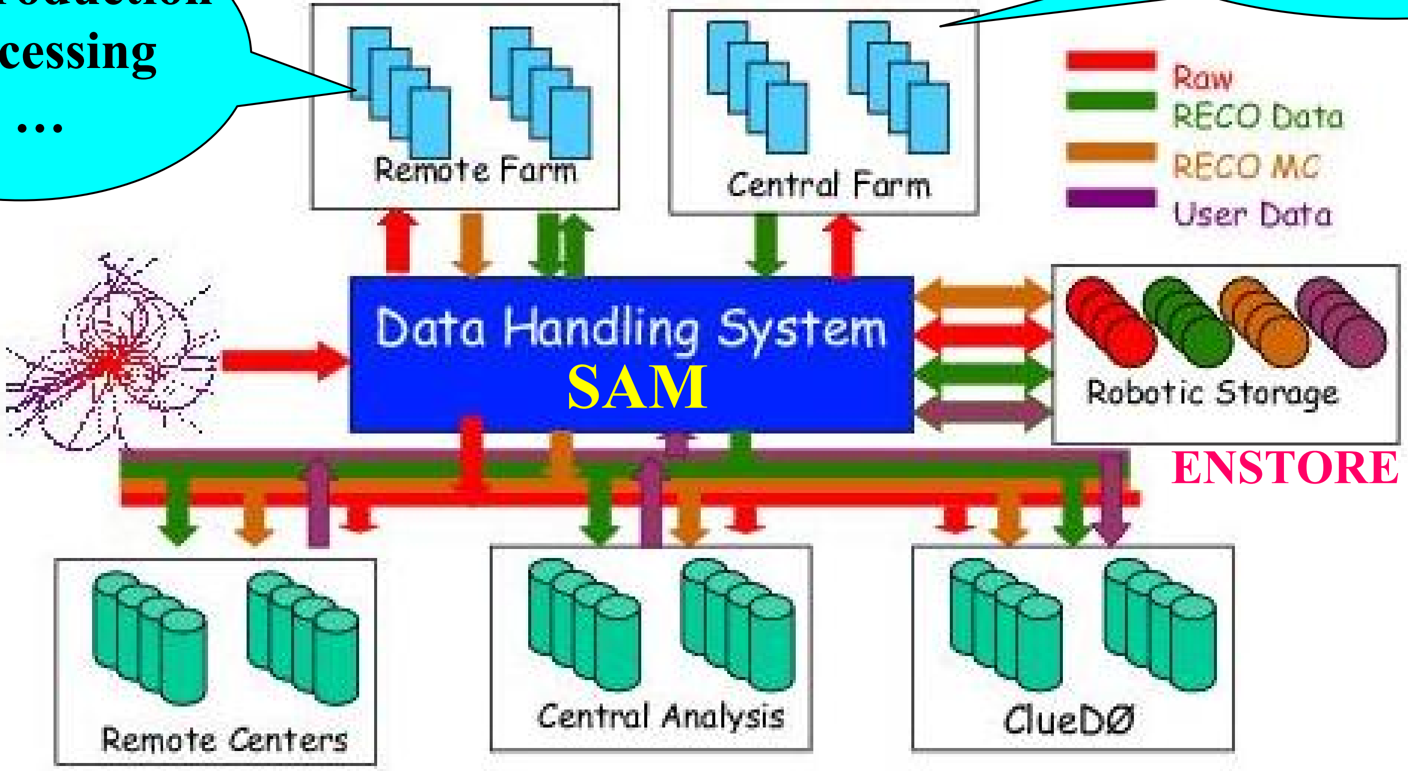
Run 178796 Event 67972991 Fri Feb 27 08:34:09 2004



Computing Model I

MC-production
Reprocessing
Fixing …

1st reconstruction



Analysis , Individual production …

Computing Model II

- DØ – active, data taking experiment !

  - amount of data growing
  - production vs development
    corrections, fixing → rerun only part of the code
  - nevertheless improvements are necessary even vital !

- Many of the tasks, problems already on the LHC scale
- So how do we cope with ever increasing demands ?
- DØ computing model built on SAM

SAM - Data Management System

- SAM (Sequential data Access via Metadata)

  - distributed Data Handling System for Run II DØ, CDF experiments

  - set of servers (stations) communicating via CORBA

  - central DB (ORACLE @ FNAL)

  - project started in 1997 by DØ

  - designed for PETABYTE sized datasets !

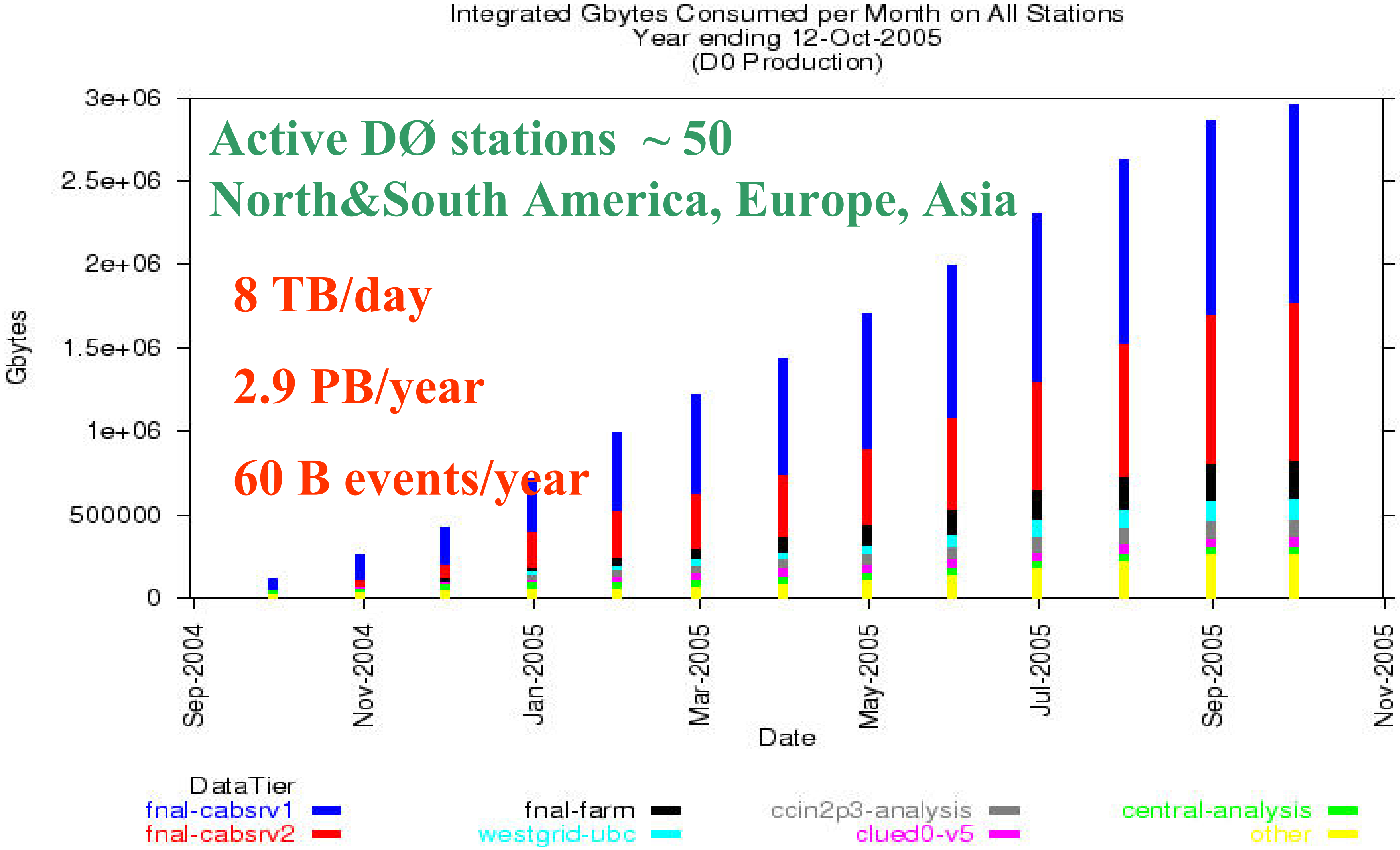
SAM Functionalities

- **file storage** from online and processing systems
  → MSS - FNAL Enstore, CCIN2P3 HPSS…
  disk caches around the world
- **routed file delivery**
  - user doesn't care about file locations
- **file metadata cataloging**
  → datasets creation based on file metadata
- **analysis bookkeeping**
  → which files processed succesfuly by which application
  when and where
- **user authentication** **-** registration as SAM user
- **local and remote monitoring capabilities**
  http://d0db-prd.fnal.gov/sam_local/SamAtAGlance/
  http://www-clued0.fnal.gov/%7Esam/samTV/current/

SAM Terms and Concepts

- A project runs on a station and requests delivery of a dataset to one or more consumers on that station.
- **Station:** Processing power + disk cache + (connection to tape storage) + network access to SAM catalog and other station caches
  Example: ccin2p3-analysis
- **Dataset**: metadata description which is resolved through a catalog query to a list of files. Datasets are named.
  Examples: (syntax not exact)
  - data_type physics and run_number 78904 and data_tier raw
  - request_id 5879 and data_tier thumbnail
- **Consumer**: User application (one or many exe instances)
  Examples: script to copy files; reconstruction job

SAM Data Consumption 2005



Active DØ stations ~ 50

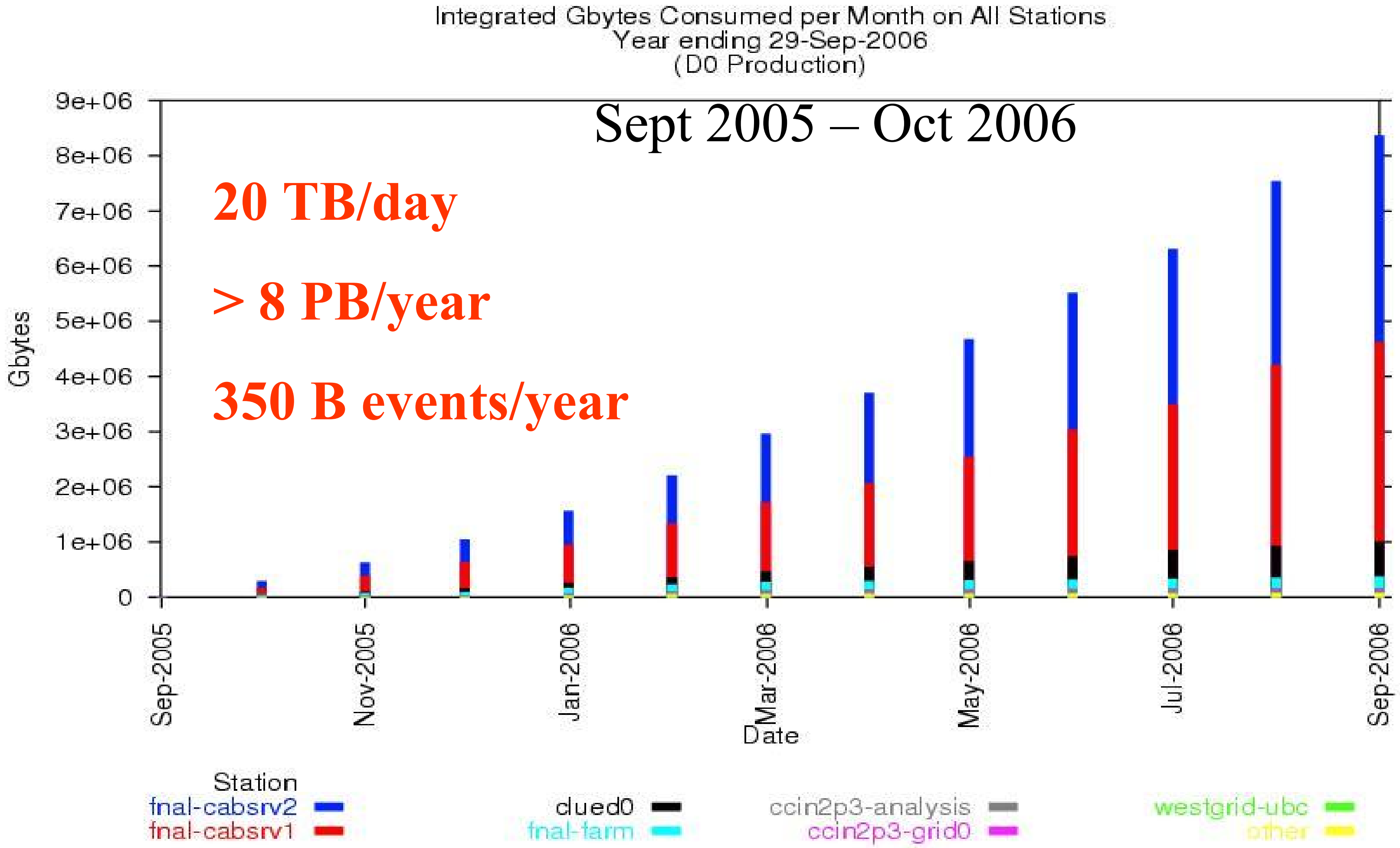
North&South America, Europe, Asia

8 TB/day

2.9 PB/year

60 B events/year

SAM Data Consumption 2006



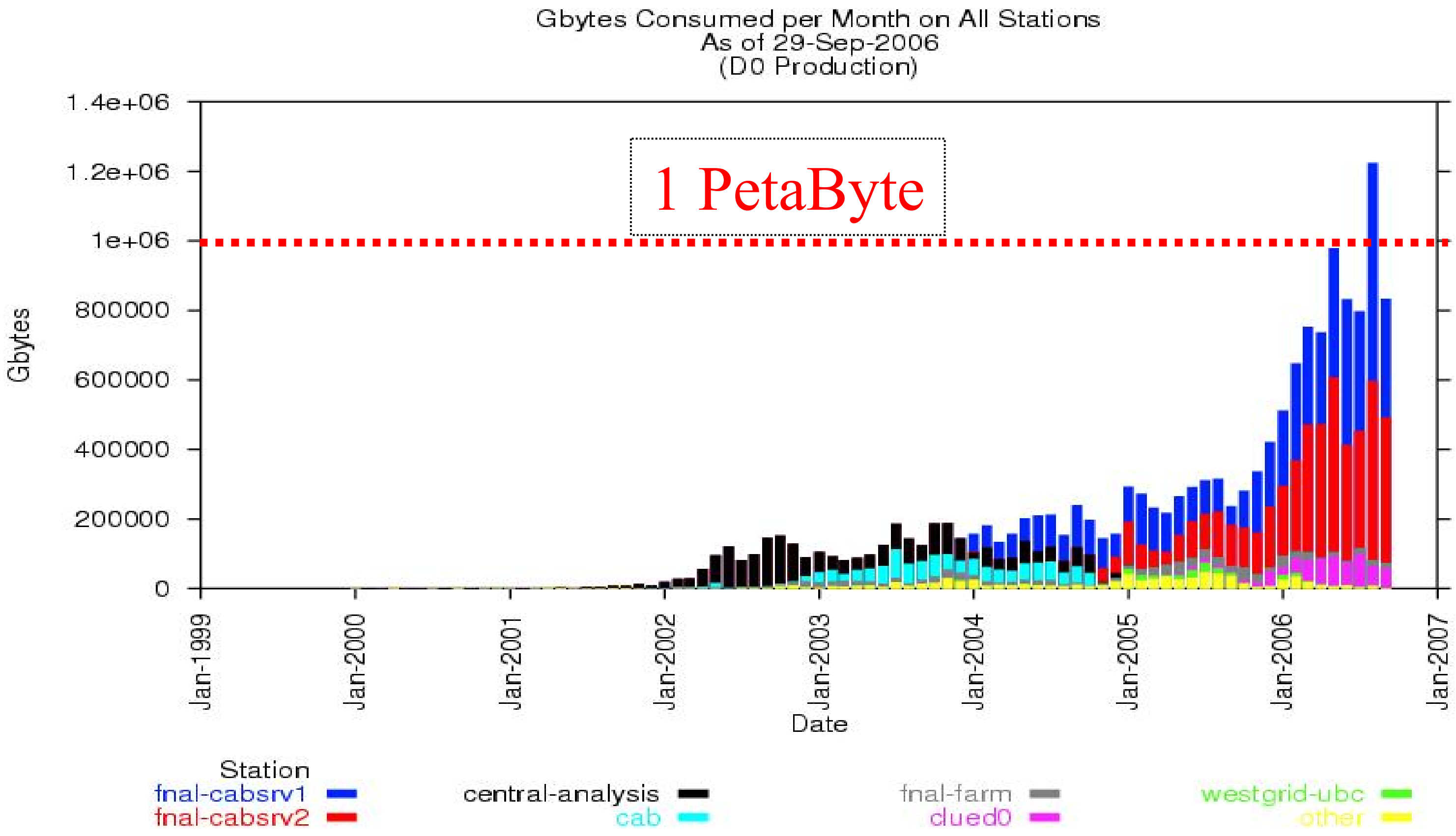
Sept 2005 – Oct 2006

20 TB/day

> 8 PB/year

350 B events/year

Run II Data Consumption/month



SAM –> SAM-Grid

- **SAM performs well → data grid for DØ**

- **BUT ! … more resources needed than available on FNAL farm**
  - e.g. huge amount of MC or to reprocess all old data in parallel with the new data taking, analysis …
  - resources distributed all around a world

- **Grid technology solution:**

  … extend SAM functionalities to the real Computing Grid

  → ***integrating standard Grid tools and protocols***

  → ***developing new solutions for Grid computing*** -

  JIM (Job & Information Manager) project started end of 2001

  → **SAM-Grid = SAM + JIM**

provides common run-time environment and common submission interface as well as monitoring tools

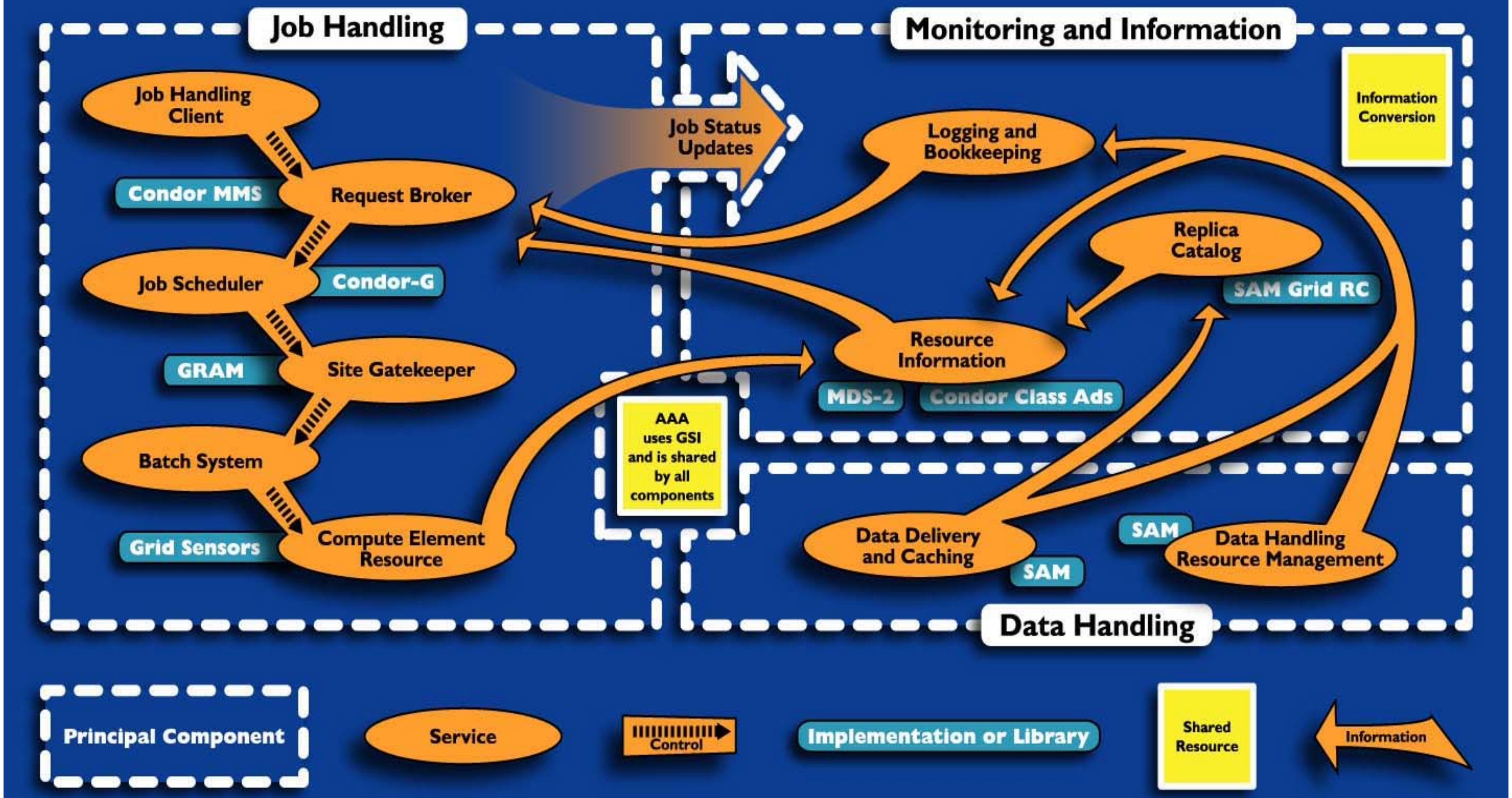
- **does require some DØ specific installations at remote sites**

  (SAM station, DB proxy servers, job manager)

SAM-Grid & Grid Services

- Distributable **sam_client** provides access to:
  - VO **storage service** (sam store command, interfaced to sam_cp)
  - VO **metadata service** (sam translate constraints)
  - VO **replica location service** (sam get next file)
  - Process **bookkeeping services**
- **JIM components provide**:
  - **Job submission service** via Globus Job Manager
  - **Job monitoring service** from remote infrastructure
  - **Authentication services**

SAM-Grid Architecture



Job Handling

Job Handling Client

Condor MMS

Request Broker

Job Scheduler

Condor-G

GRAM

Site Gatekeeper

Batch System

Grid Sensors

Compute Element Resource

Job Status Updates

Monitoring and Information

Information Conversion

Logging and Bookkeeping

Replica Catalog

SAM Grid RC

Resource Information

MDS-2

Condor Class Ads

AAA uses GSI and is shared by all components

Data Delivery and Caching

SAM

SAM

Data Handling Resource Management

Data Handling

Principal Component

Service

Control

Implementation or Library

Shared Resource

Information

Status & Installation of DØ SAM-Grid

- Active execution sites: >10 DØ (1 @ FNAL)

  http://samgrid.fnal.gov:8080/list_of_resources.php?

  http://samgrid.fnal.gov:8080/list_of_schedulers.php?

  - Active Monte Carlo production at multiple sites
  - Reprocessing from raw data 2005 :

    10^9 events ~250 TB of raw data to move

    calibration proxy DB-servers at remote sites

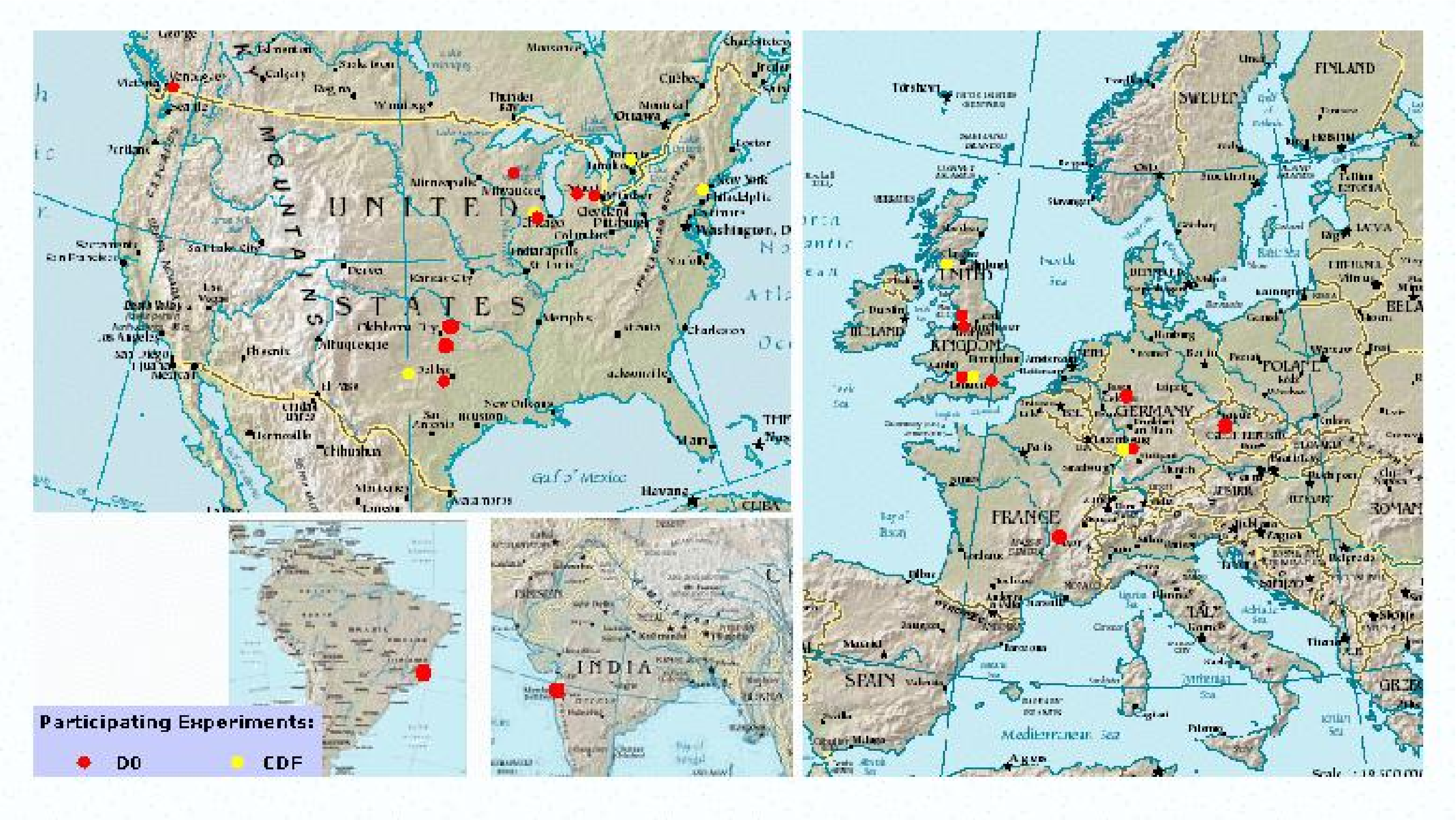
- Installation
  - via ups/upd FNAL products
  - No specific requirements on environment
  - Non invasive system , very flexible

    →Drawback : non trivial configuration

    requires good system understanding

SAM-Grid World

http://samgrid.fnal.gov:8080/



SAM-Grid at CCIN2P3

- **SAM station: ccin2p3-analysis**
- **SAM-Grid** installed in summer 2003 as a
  - **client** (very light-weight) &
  - **submission** &
  - **monitoring** &
  - **execution site**

→ full grid functionality

→used for official MC-production – from 2004

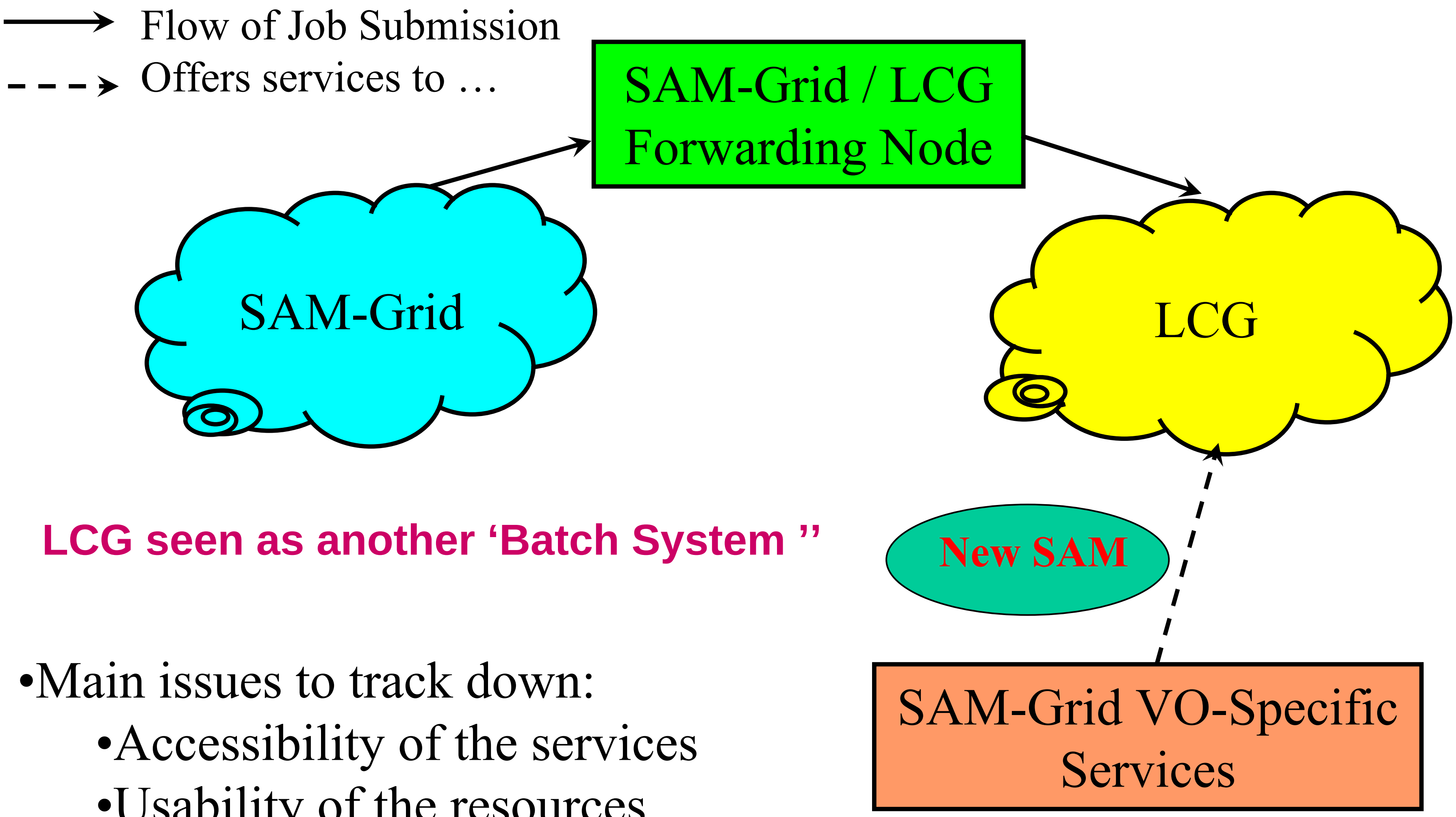
→reprocessing from raw data – 2005

- production & merging individual thumbnails

The SAM-Grid/LCG Interoperability

- **Motivation & Goals**
  - resources and manpower drifting towards LHC
  - make LCG resources available to DØ via SAM-Grid
  - integration project, no massive code changes expected
- **Limitations & Problems**
  - most of the LCG resources w/o SAM-Grid gateway node
  - firewall problems : station interfaces use callbacks
  - SAM/LCG batch adapter to be developped
  - security : authentication → agreement on a set of CA
    authorization to use LCG resources

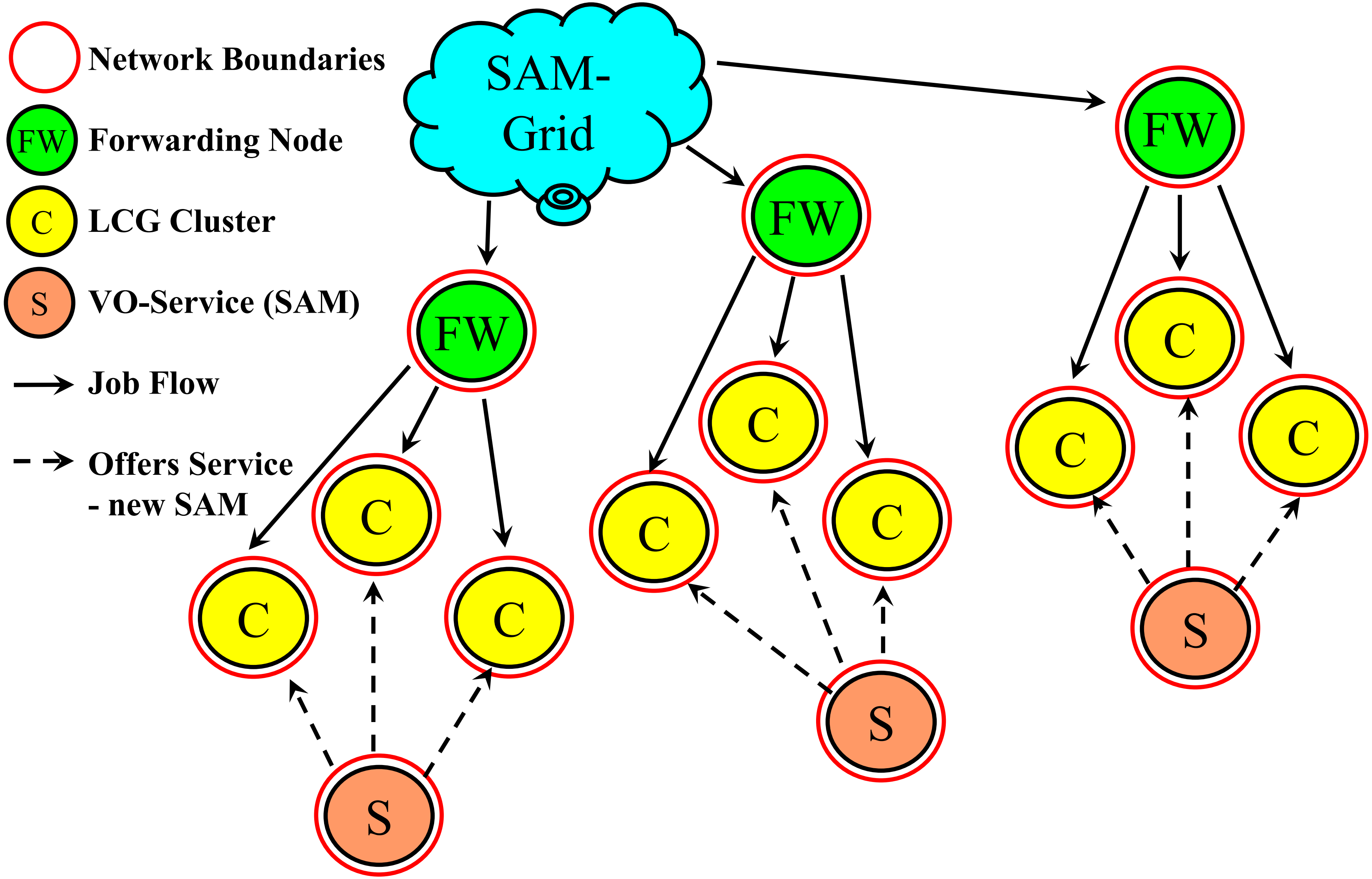
SAMGrid/LCG - Basic Architecture



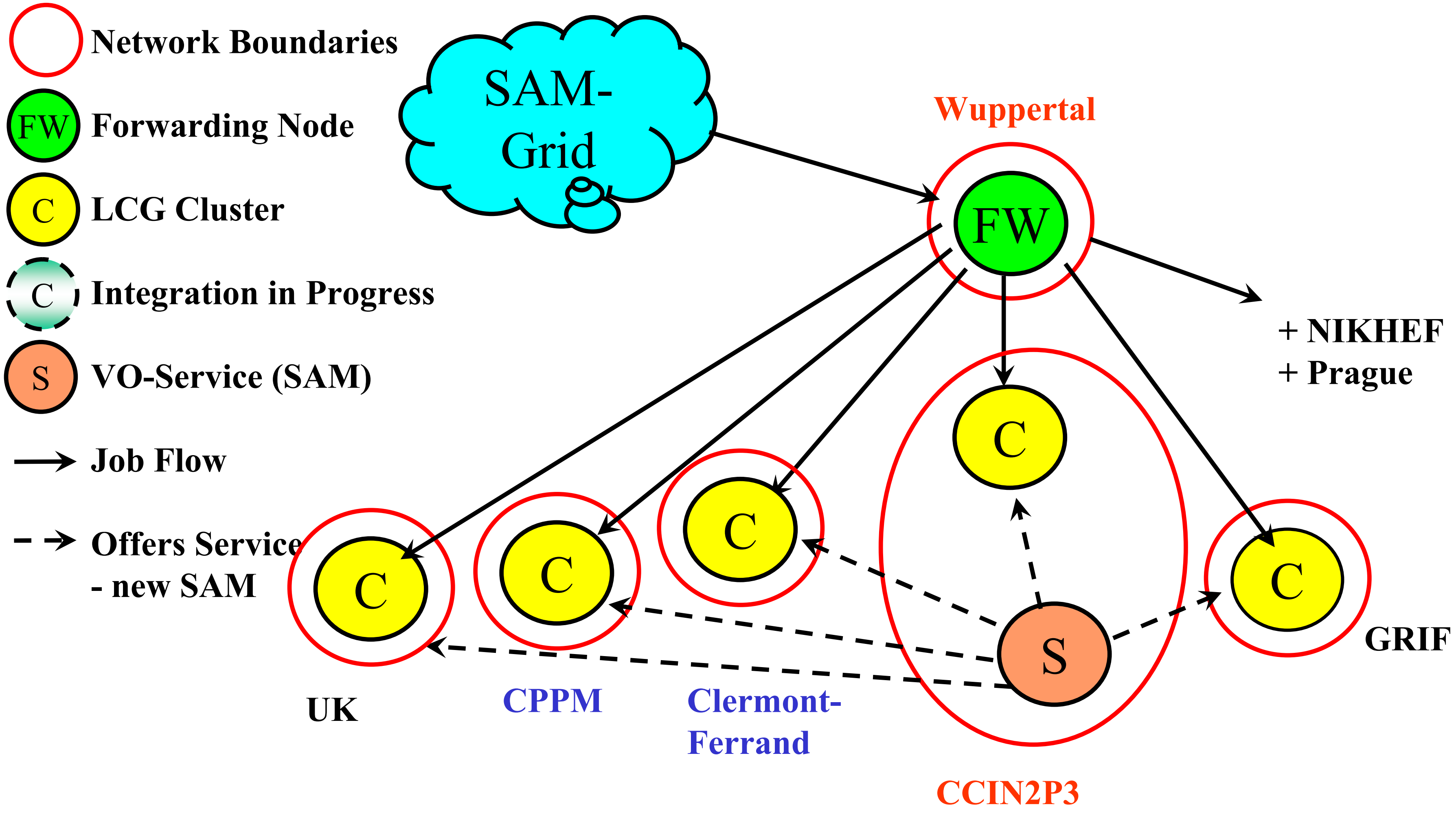
LCG seen as another ‘Batch System ’’

- Main issues to track down:
  - Accessibility of the services
  - Usability of the resources
  - Scalability

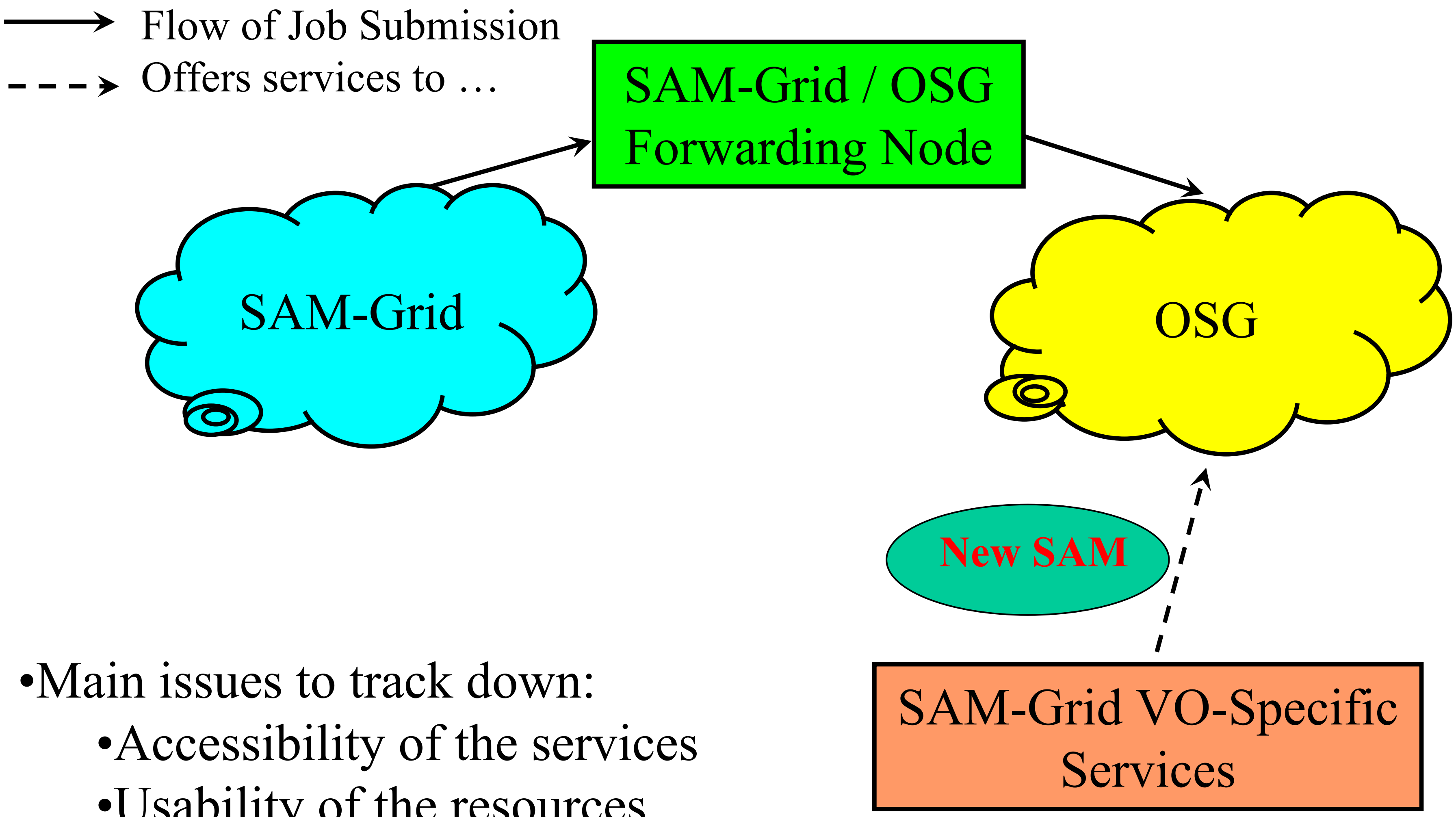
Service/Resource Multiplicity



SAMGrid/LCG - Current Configuration

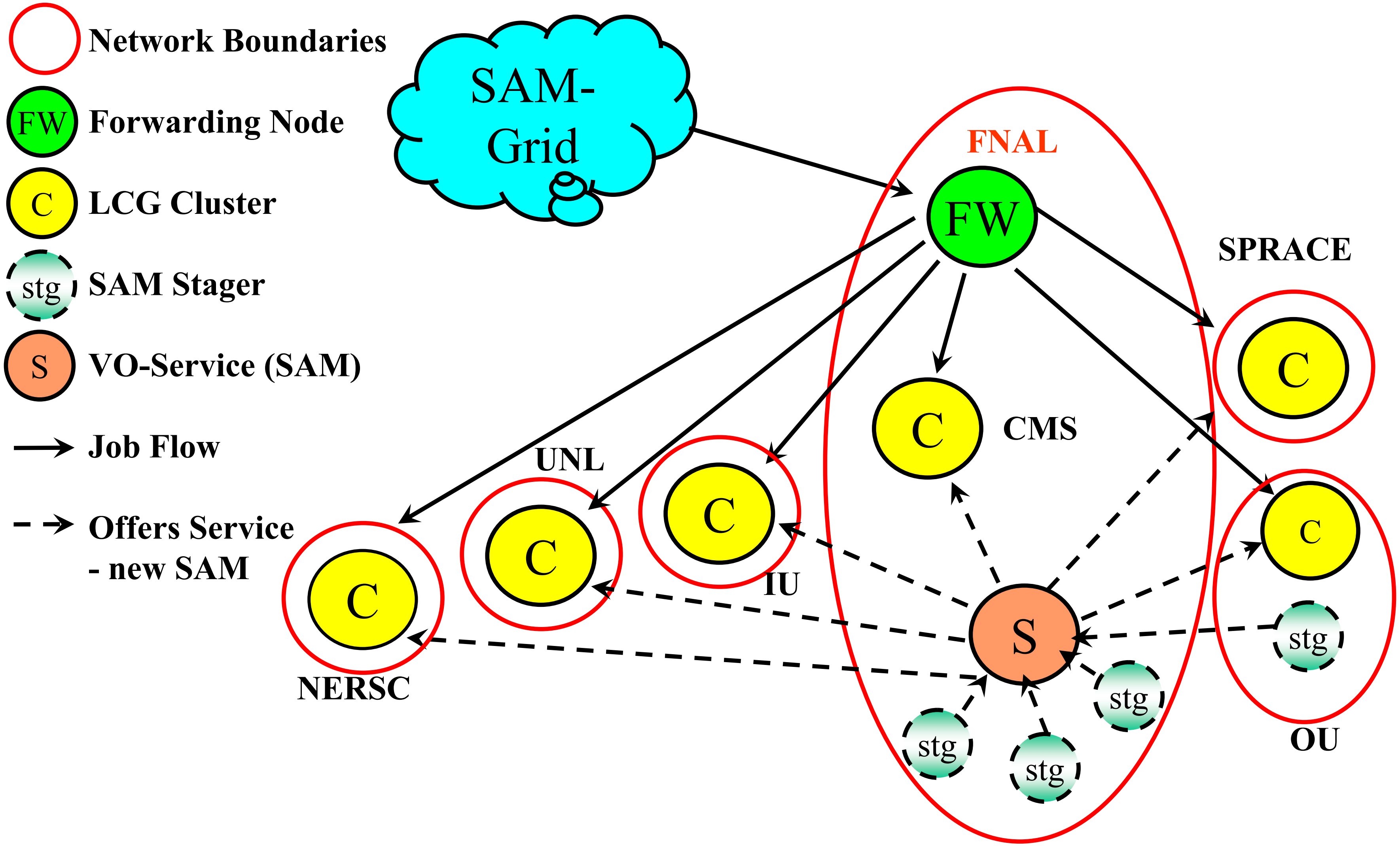


Basic Architecture



- Main issues to track down:
  - Accessibility of the services
  - Usability of the resources
  - Scalability

SAMGrid/OSG - Current Configuration



○ Network Boundaries

FW Forwarding Node

C LCG Cluster

stg SAM Stager

S VO-Service (SAM)

→ Job Flow

- - → Offers Service - new SAM

SAM-Grid/LCG Integration Status

- we can submitt DØ real data reconstruction & MC-jobs to LCG clusters via SAMGrid FW-node at Wuppertal

- jobs rely on the SAM station at CCIN2P3 Lyon to handle input (binaries and data) & output

- jobs are running on any LCG cluster with V0- Dzero

- Durable location for output files at Manchester

- Final results – merged files stored at FNAL

Job Definition File

- job_type = dzero_monte_carlo
- runjob_requestid = 35966
- runjob_numevts = 50000
- events_per_file = 250
- d0_release_version = p17.09.06
- jobfiles_dataset = sg_p17.09.06-v2_mcr06-05-22-v2_cf00-09-07
- minbias_dataset = Zerobias_p17_06_03MC_set1
- sam_experiment = d0
- sam_universe = prd
- group = dzero
- check_consistency = true
- instances = 1
- station_name = ccin2p3-grid1
- lcg_requirement_string = clrlcgce02.in2p3.fr:2119/jobmanager-lcgpbs-dzero

Operation Status

- Up to now in production - for refixing (113 mil. Events)
  Lancaster, Clermont-Ferrand, Prague,
  Imperial College, NIKHEF, Wuppertal

- MC – tests and certification requests

- started first MC-production on LCG clusters
  - Clermont-Ferrand : 3 CEs ~380 CPUs
  - Marseille : 1 CE ~64 CPUs

  → September ’06 production on UK-clusters started

Problems - Lessons - Questions

- Scratch space …. $TMPDIR
- Sites Certification
- Job Failure Analysis / Operation support
- Jobs Resubmission
- SAM & Network Configuration

SAM & Network Configuration

- SAM can only use TCP-based communication
  (as expected, UDP does not work in practice on the WAN)

- call-back interface was replaced by the pull-based one
  - SAM had to be modified to allow service accessibility for jobs within private networks

- For future : SAM should be modified to provide port range control
  - currently sam-client is using dynamic range
    → all ports have to be open
  - sites hosting SAM must allow incoming network traffic from the FW node & from all LCG clusters (WNs) to allow data handling & control transport

Summary (1)

- **DØ – running HEP experiment:**
  - handles PetaBytes of data
  - computing resources distributed around the world
- **SAM – Distributed Data Handling System**
  - reliable data management & worldwide file delivery to the users
- **SAM-Grid – full Grid functionality**
  - standard Grid middleware + specific products
  - MC-production running (all MC remotely produced)
  - Reprocessing , fixing

Summary (2)

- **SAM-Grid/LCG interoperability**

  - running MC-production
- **working on interoperability SAM-Grid/OSG**

  → continuation of a global vision for the best use of available resources

  …. About to start next reprocessing of Run IIb data
- **Remote, distributed computing**

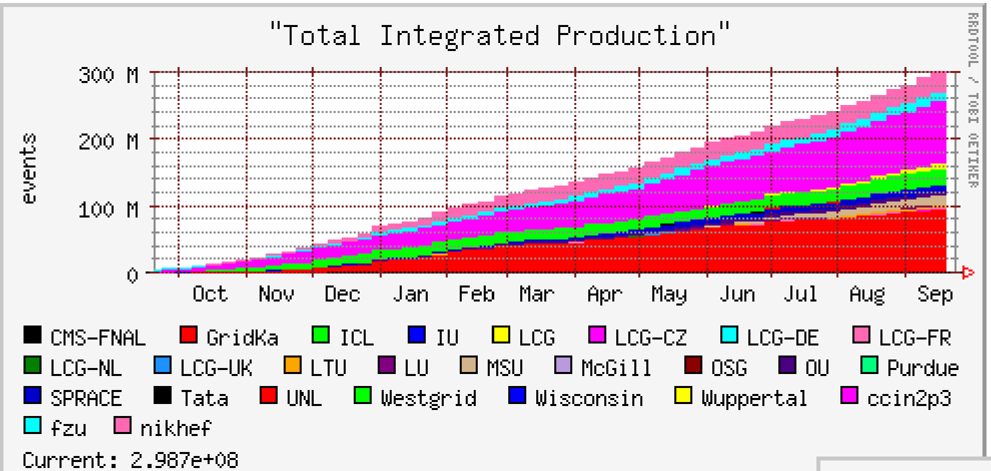
  – huge profit to DØ experiment

  → excellent physics results !

- **CCIN2P3: major contribution to the DØ computing**

… backup slides….

MC Production

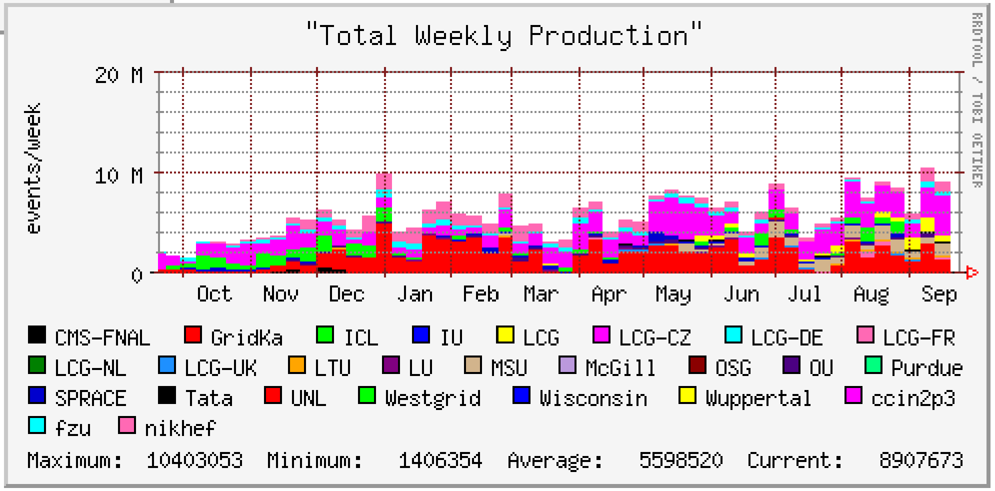


CCIN2P3

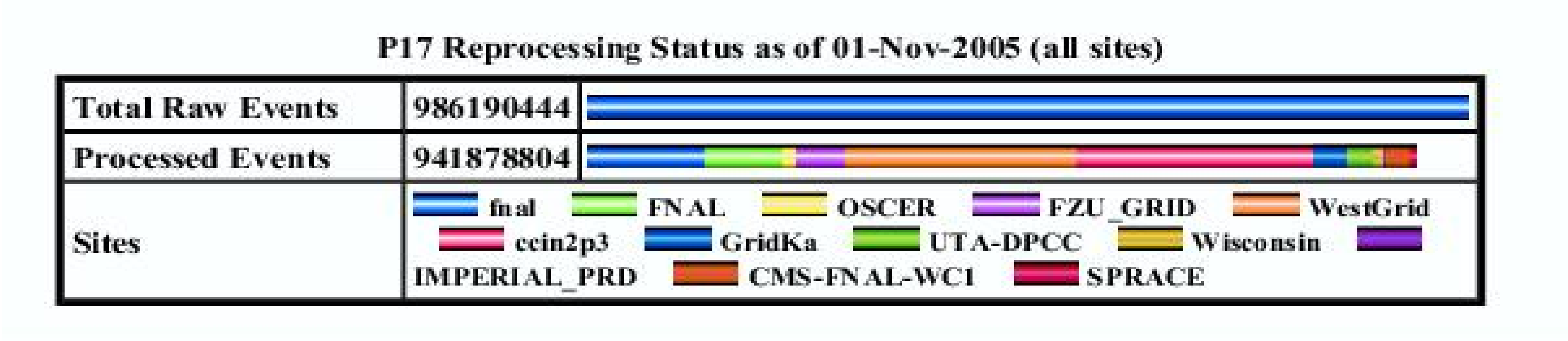
100 M events

last year

All MC produced on remote sites



Grid Reprocessing 2005



P17 Reprocessing Status as of 01-Nov-2005 (all sites)

Total Raw Events	986190444	
Processed Events	941878804	
Sites	🟦 fnal 🟩 FNAL 🟨 OSCER 🟪 FZU_GRID 🟧 WestGrid 🟥 ccin2p3 🟦 GridKa 🟩 UTA-DPCC 🟨 Wisconsin 🟪 IMPERIAL_PRD 🟧 CMS-FNAL-WC1 🟥 SPRACE	

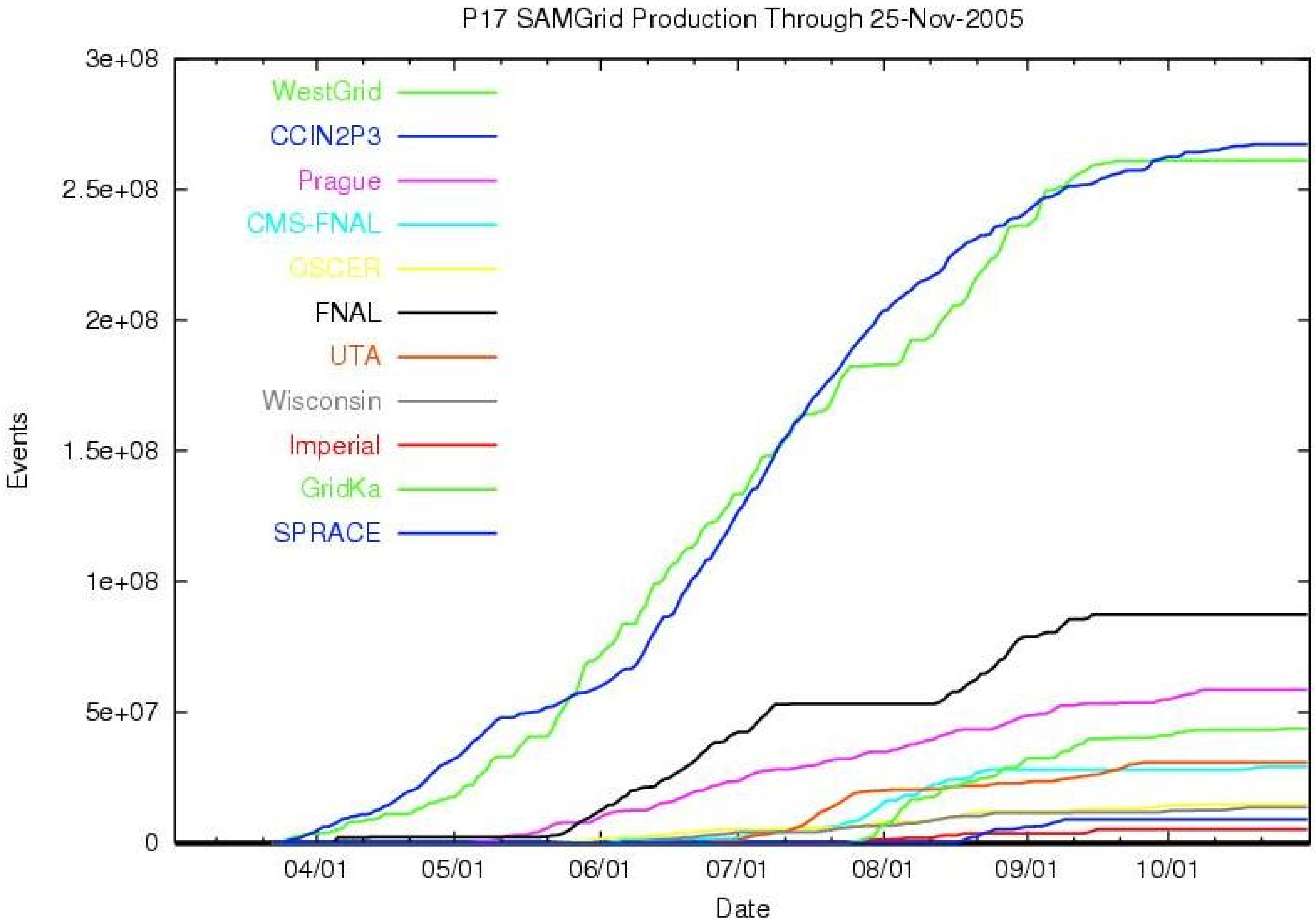
Declared Available Resources Total **3430 CPUs** (1 GHz PIII)

Total # Events to be reprocessed **986.2 M**

…few examples …. not all sites!

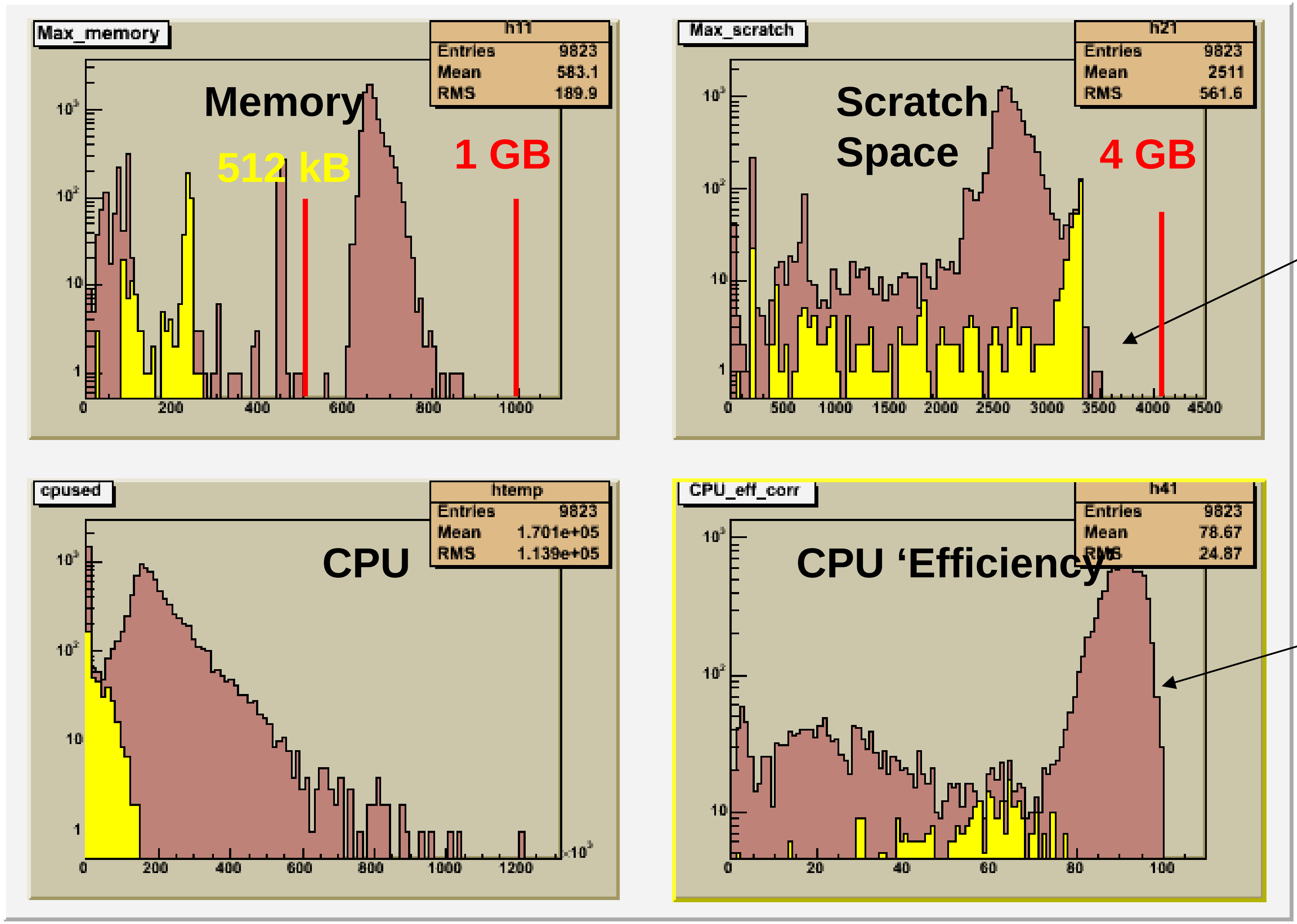
Institution	**Available Resources**		**# Events Reprocessed**		**QF**
UK (4 sites)	750	(21.9 %)	3.2M	(0.3 %)	0.01
WestGrid Vancouver	600	(17.5 %)	261.0M	(26.5 %)	1.51
GridKa Karlsruhe	500	(14.6 %)	39.0M	(4.0 %)	0.27
CCIN2P3	**400**	**(11.7 %)**	**267.3M**	**(27.1 %)**	**2.32**
FNAL	340	(9.9 %)	218.7M	(22.0%)	2.22
FZU-GRID Prag	200	(5.8 %)	54.9M	(5.6 %)	0.97
CMS-Farm FNAL	100	(2.9 %)	29.2M	(3.0 %)	1.03

Reprocessing Statistics



Jobs Characteristics

→ Optimal BQS class selection for different applications



Merging jobs

I/O oriented
Class P

Reconstruction jobs

CPU oriented
Class T, W

SAM-Grid Reprocessing Lessons

- **Data availability → bottleneck:**
  - data prestaging on remote sites for efficient operation

- **Scalability problems not to underestimate**
  - central FNAL servers, local head nodes, access to the input binaries

- **Deployment & operation requires close collaboration between SAM-Grid and local experts**
  - each new site is a new adventure with unique problems and constraints

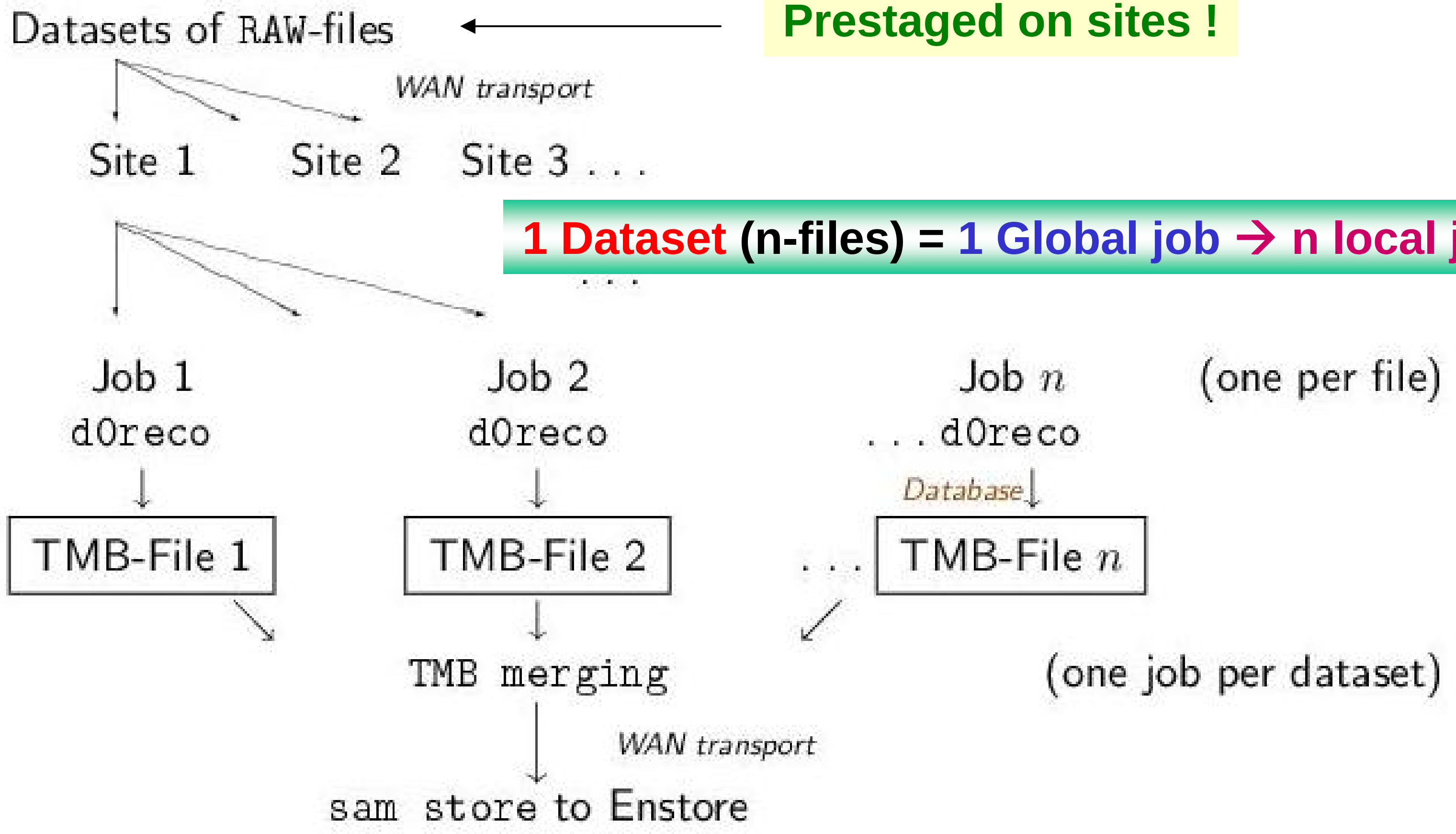
- **Manpower needs**
  - entire operation still manpower intensive ~1 FTE for each remote site

- **Available CPU – be careful with numbers !**
  - hundreds of declared CPU don't mean automatically high production yield (efficiency)

Reconstruction & Merging

Application flow

Overview



Datasets of RAW-files

WAN transport

Site 1 Site 2 Site 3 . . .

. . .

Job 1 d0reco

Job 2 d0reco

Job n . . . d0reco

(one per file)

Database

TMB-File 1

TMB-File 2

. . . TMB-File n

TMB merging

(one job per dataset)

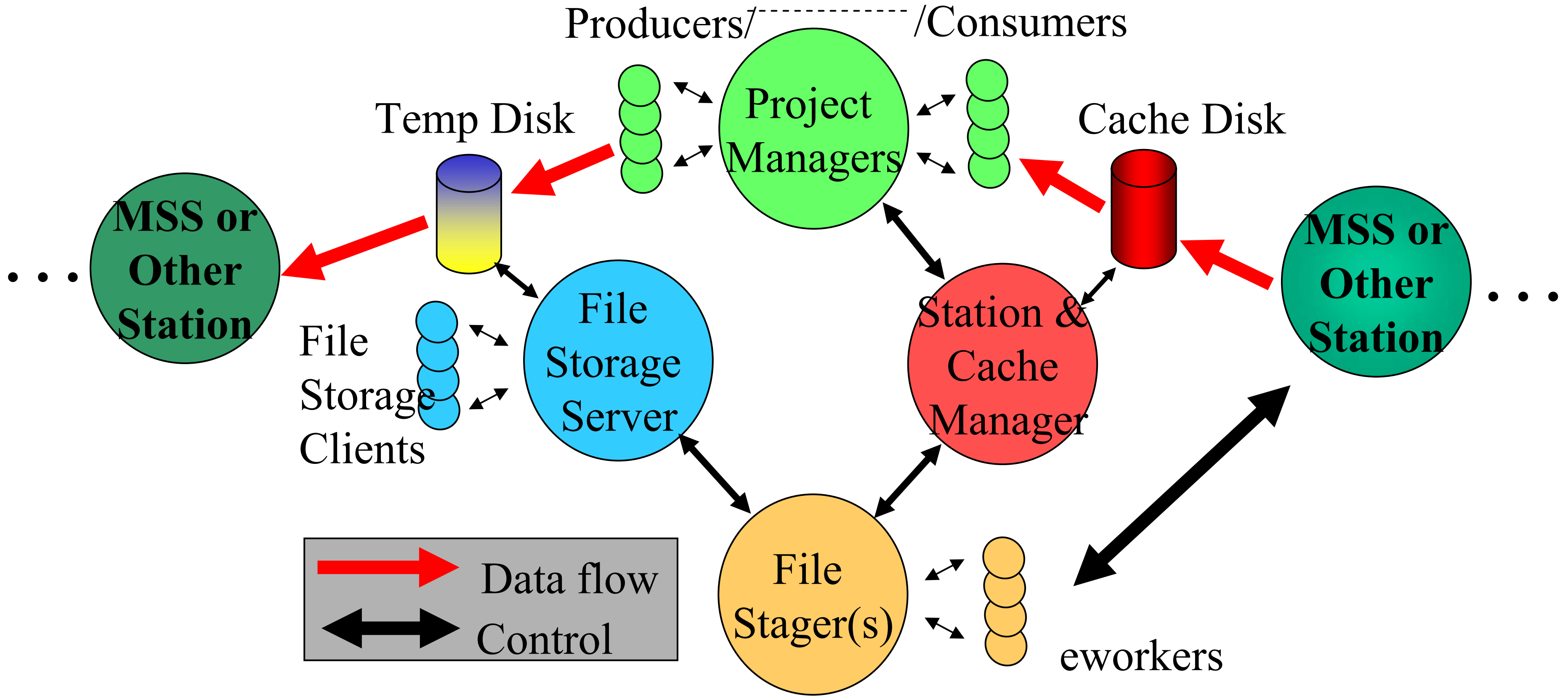
WAN transport

sam store to Enstore

Prestaged on sites !

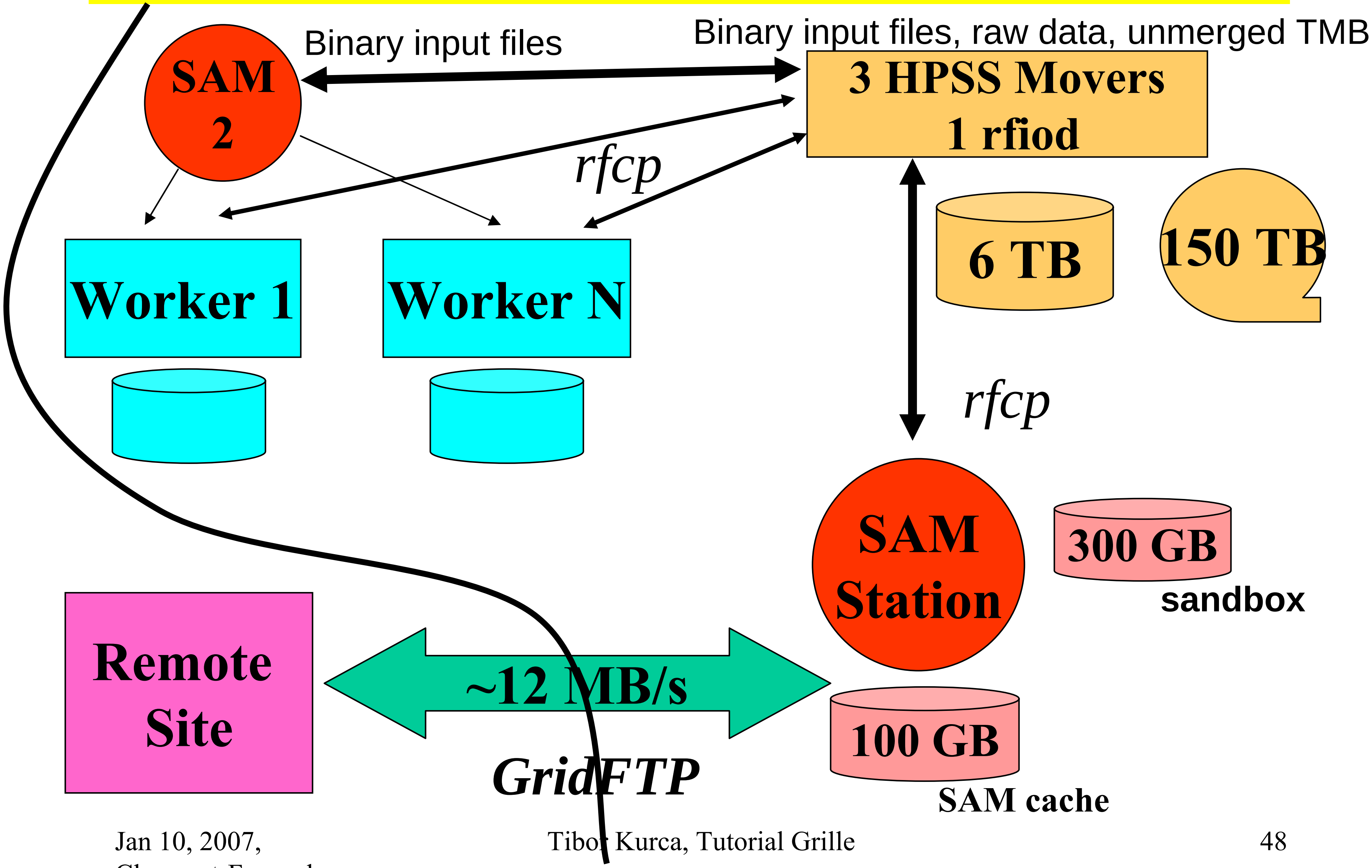
1 Dataset (n-files) = 1 Global job → n local jobs

Components of a SAM Station



- SAM: distributed data movement and management service: data replication by the use of disk caches during file routing
- SAM is a fully functional meta-data catalog.

SAMGrid @ CCIN2P3



Tevatron Upgrade - Run II

- Higher energy:

  $\sqrt{s}$ = 1.8 TeV → 1.96 TeV

  ➢ Higher cross sections

  (30 % for top)

- More(anti)protons/bunch

  (New Main Injector & Recycler)

- More bunches:

  6x6 → 36x36 bunches

  (3.5 μs → 396 ns)

  ➢ Higher luminosity

  Run I : 2×10^{31} cm^{-2}s^{-1}

  ? Run II : 2×10^{32} cm^{-2}s^{-1}