DØ, Belle & CMS Grid Computing in India

Naba K Mondal

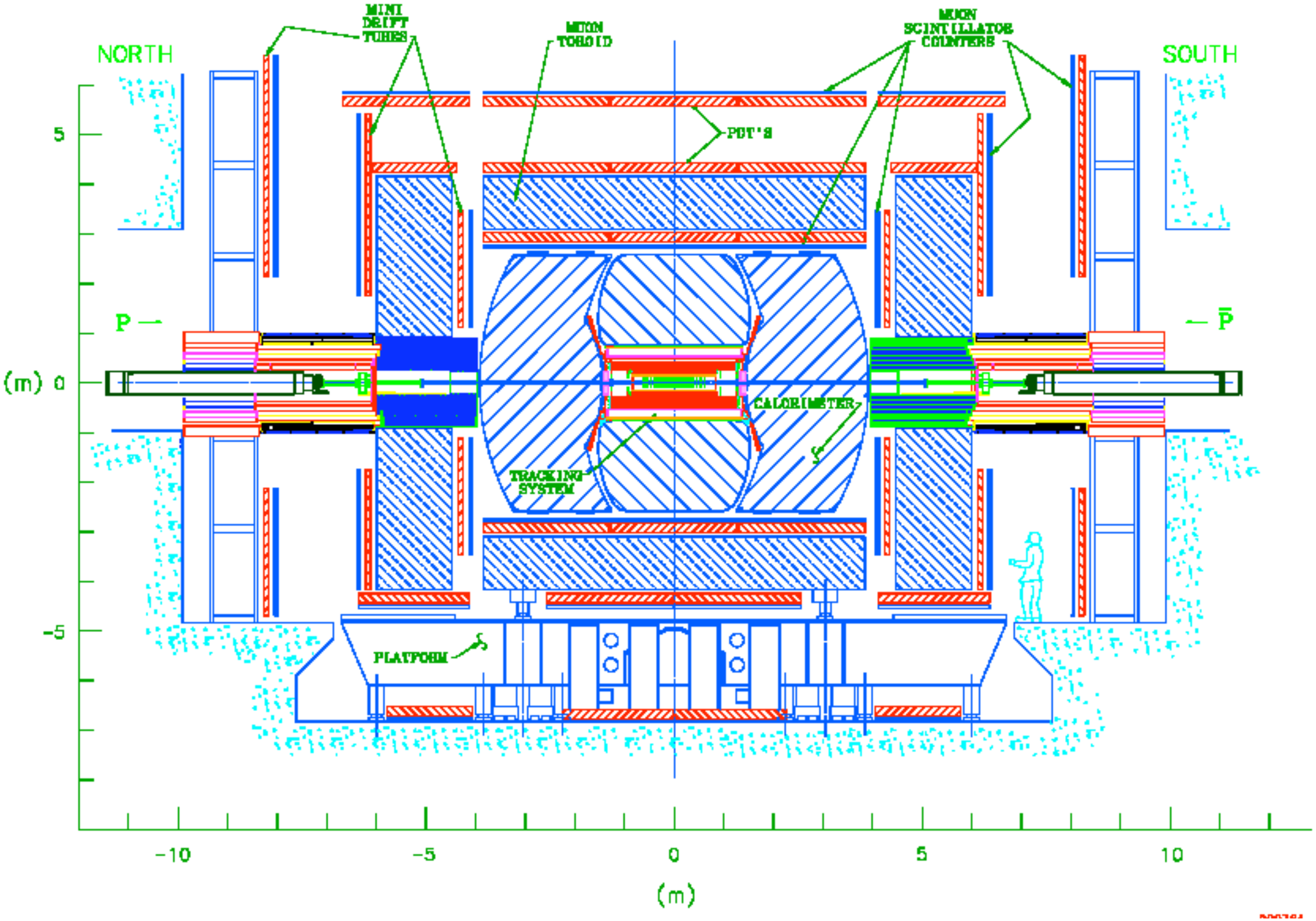
Tata Institute, Mumbai

Supercomputing Relativistic Heavy-Ion Collision Physics, December 5-9, 2005

TIFR Activities on Computing Farm & Grid Computing

- D0 computing Farm and Grid.
- Belle Cluster.
- A pilot project for CMS Grid Computing.
- India-CMS Tier-2 Grid Station: Current Status and Future Projection.

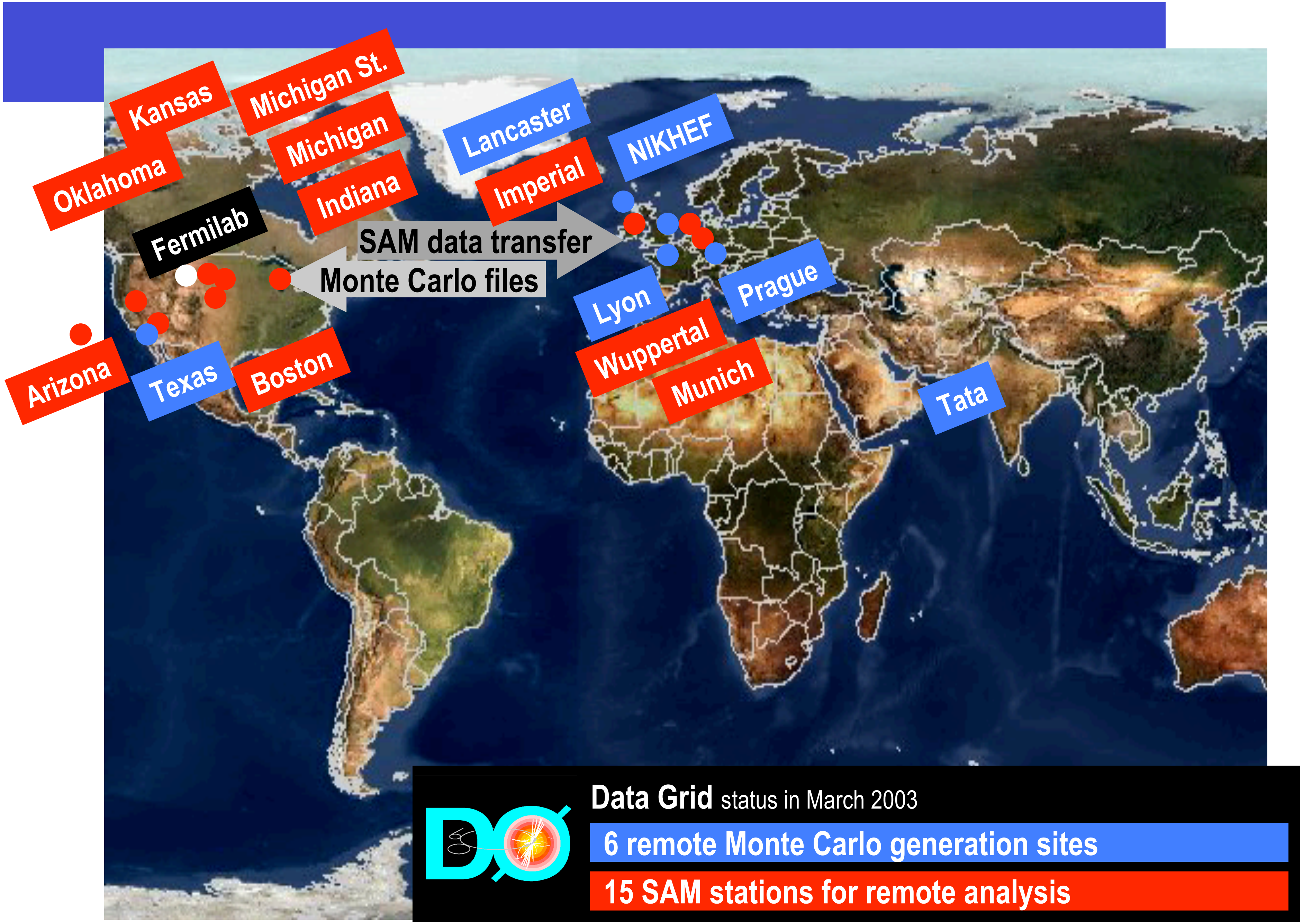
DØ Detector


MINI DRIFT TUBES
MUON TOROID
MUON SCINTILLATOR COUNTERS
NORTH
SOUTH
PDT'S
P →
← P̄
CALORIMETER
TRACKING SYSTEM
PLATFORM
5
(m) 0
-5
-10
-5
0
5
10
(m)

Remote Computing for DØ

D0 Computing group’s views before start of Run II

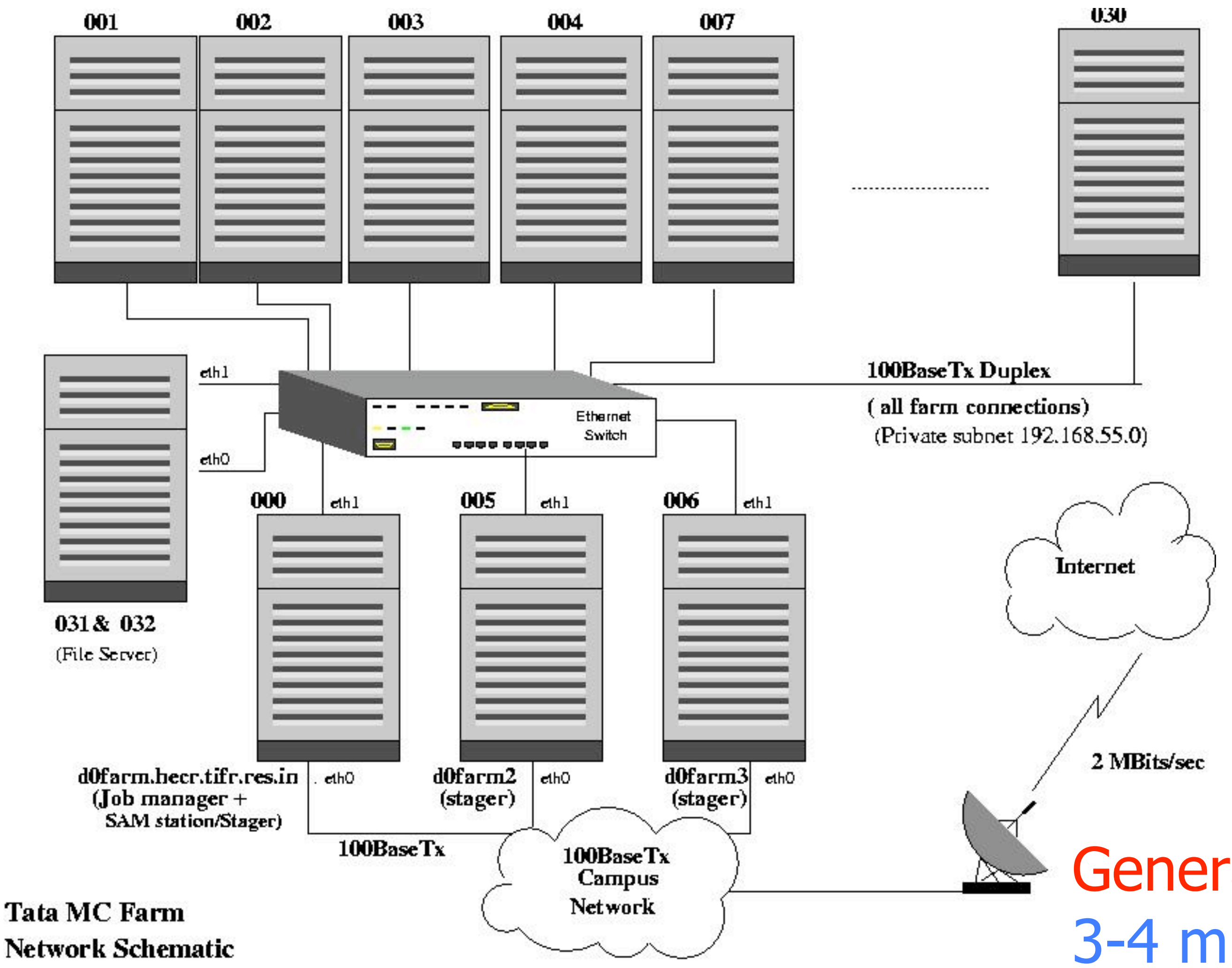
- Around the time RunII starts, 1000 CPU’s of 2GHz will be required for MC studies alone.
- Such a target can be met if computing load is shared by deploying dedicated computing facilities, offsite for both MC Production and data reprocessing.



Data Grid status in March 2003

- 6 remote Monte Carlo generation sites
- 15 SAM stations for remote analysis

DØ Computing Farm at TIFR

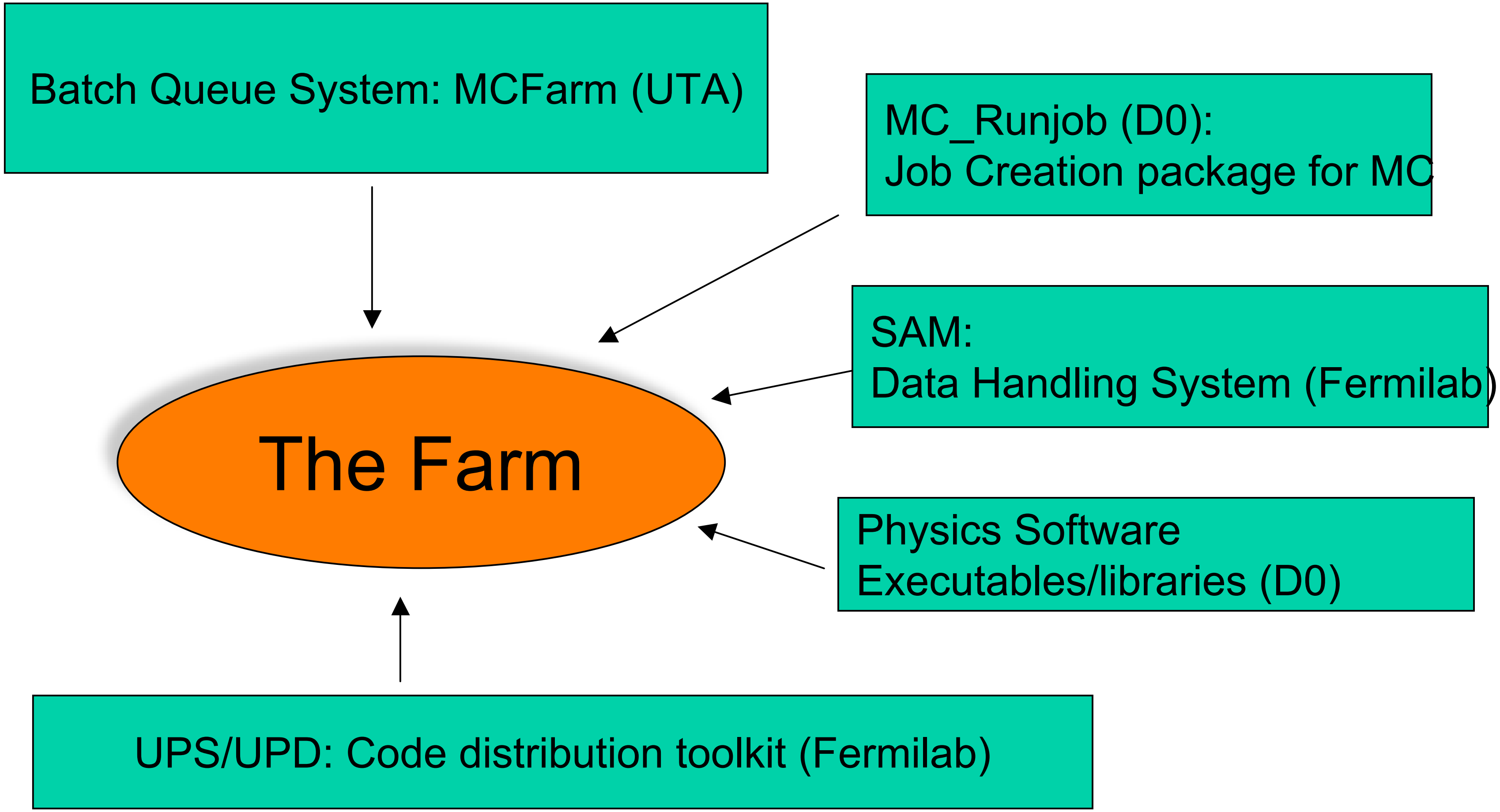


Generation time :

3-4 minutes per event in one processor

DØ Computing Farm at TIFR

Software Components of DØ Farm



Batch Queue System: MCFarm (UTA)

MC_Runjob (D0):
Job Creation package for MC

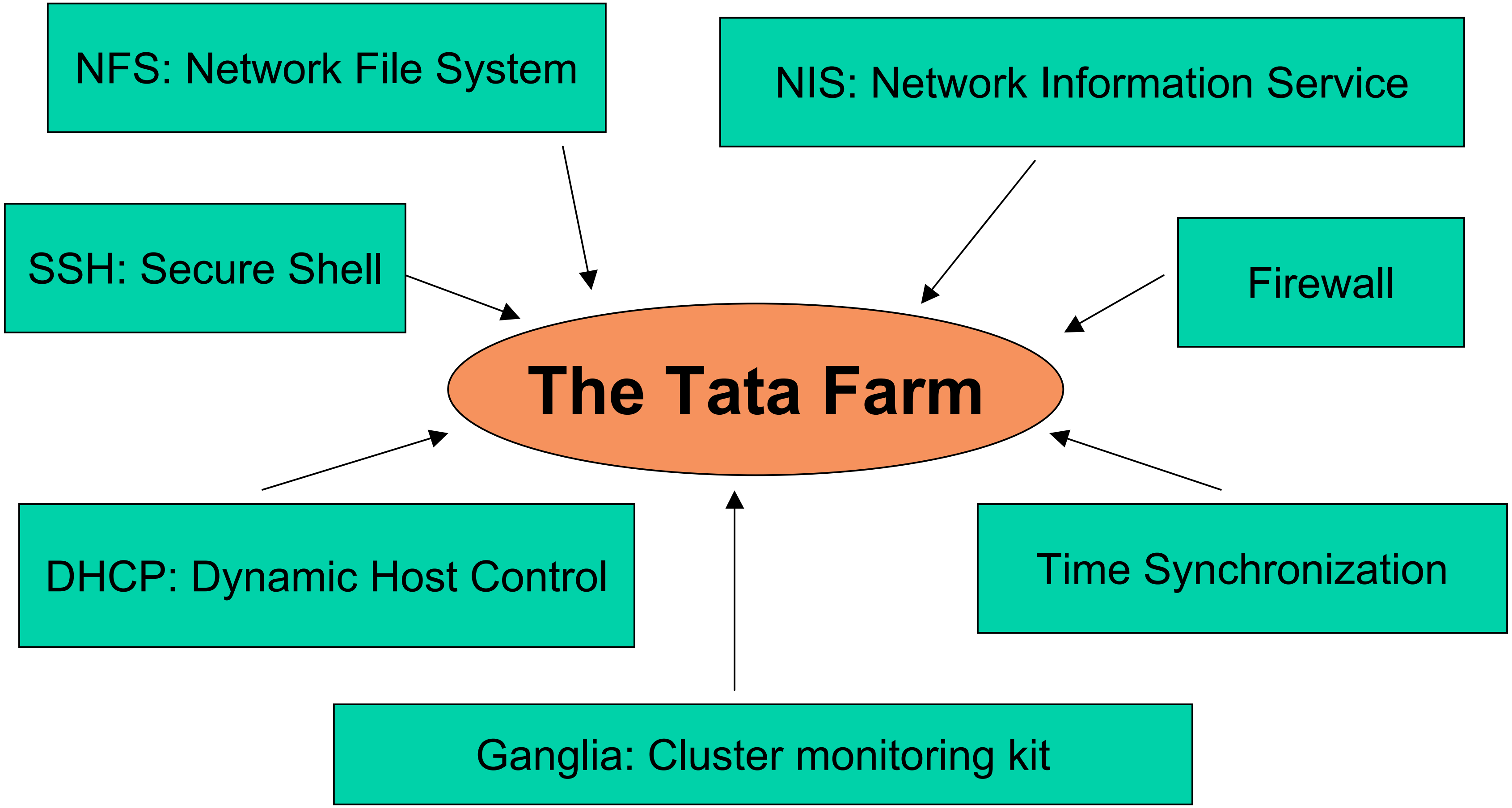
SAM:
Data Handling System (Fermilab)

The Farm

Physics Software
Executables/libraries (D0)

UPS/UPD: Code distribution toolkit (Fermilab)

Network Services on Farm



The Tata Farm

NFS: Network File System

NIS: Network Information Service

SSH: Secure Shell

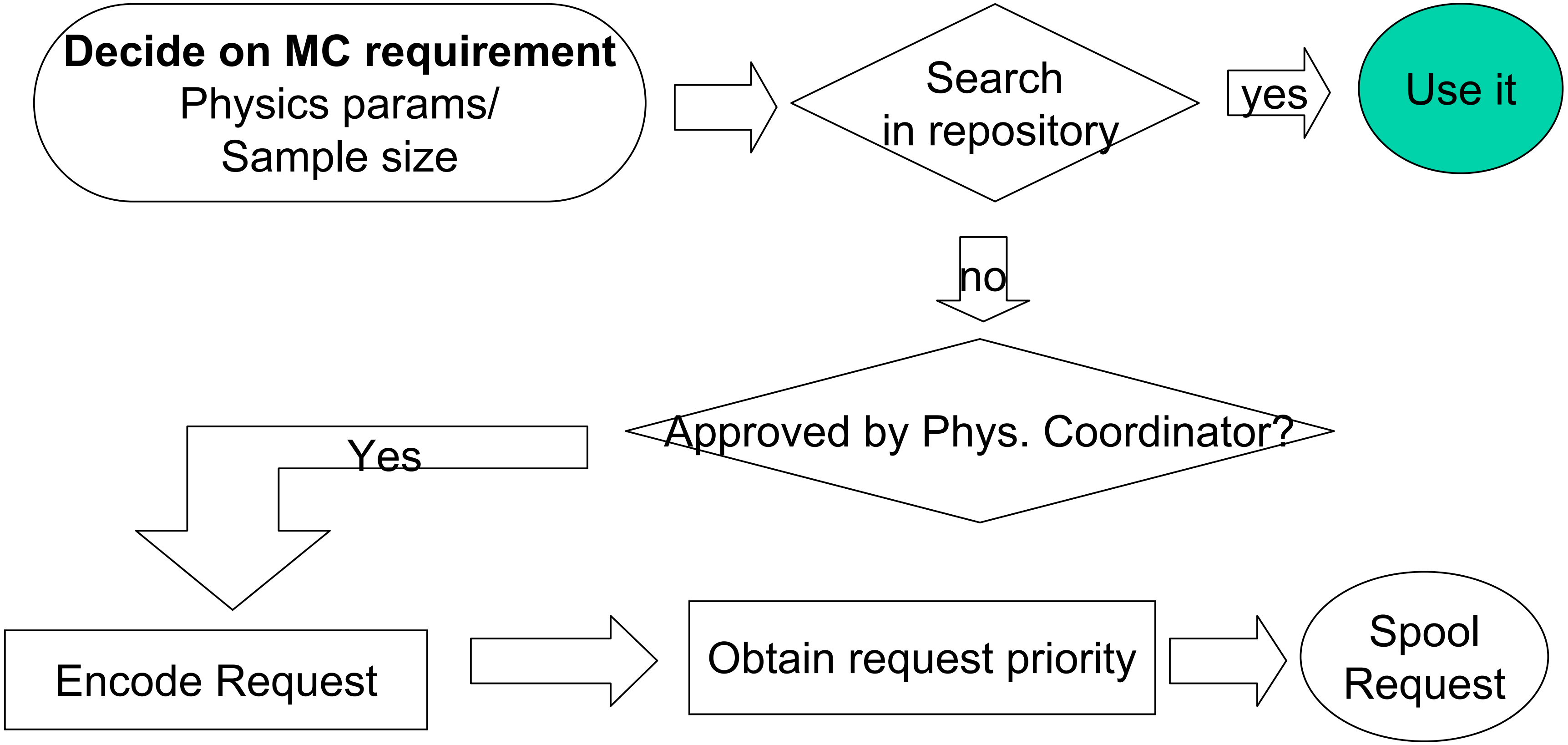
Firewall

DHCP: Dynamic Host Control

Time Synchronization

Ganglia: Cluster monitoring kit

Life Cycle of a MC request from a Physicist



Decide on MC requirement
Physics params/
Sample size

Search
in repository

yes

Use it

no

Approved by Phys. Coordinator?

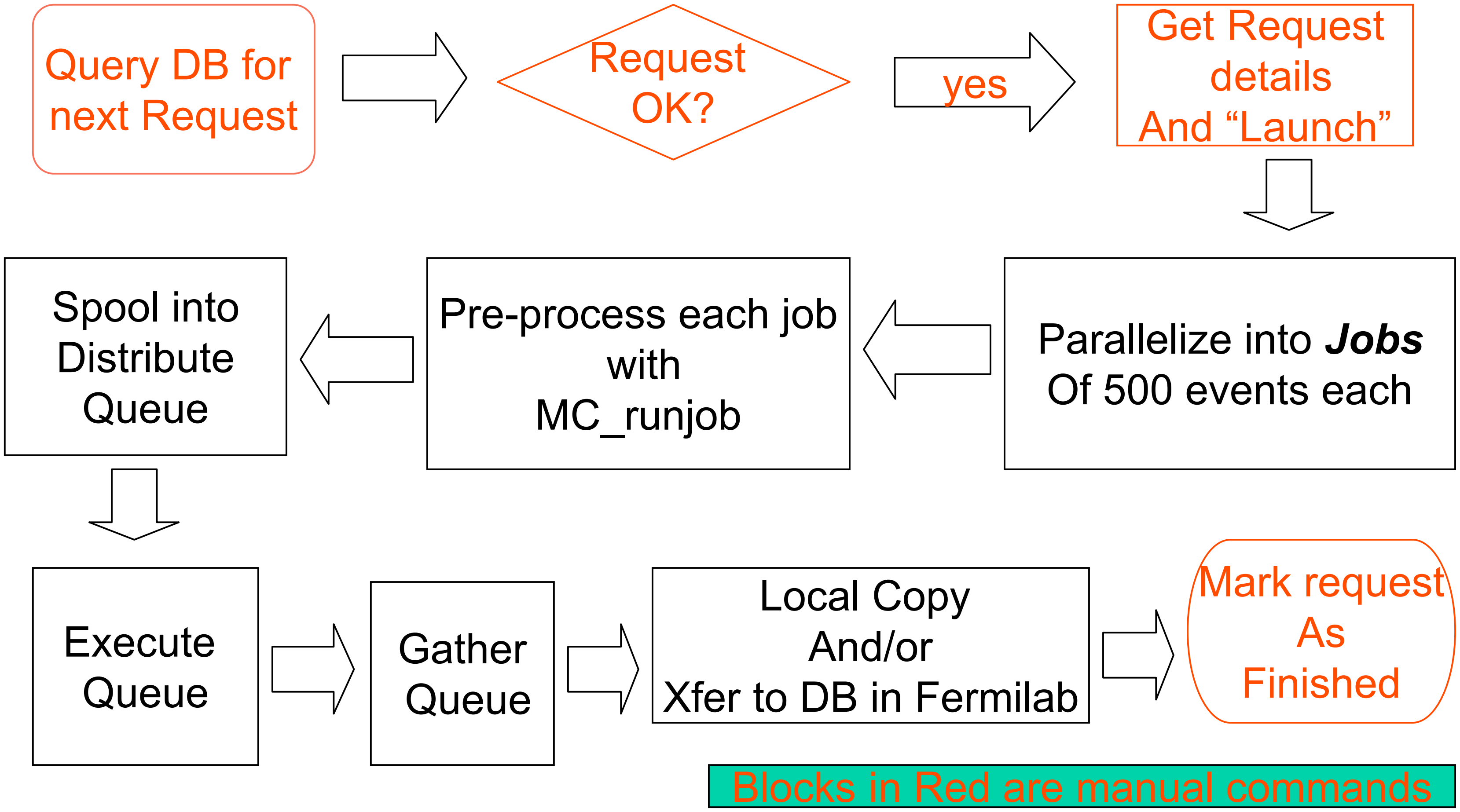
Yes

Encode Request

Obtain request priority

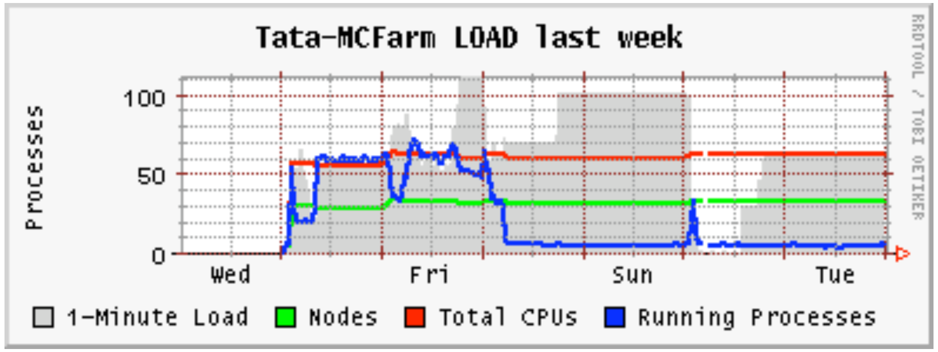
Spool
Request

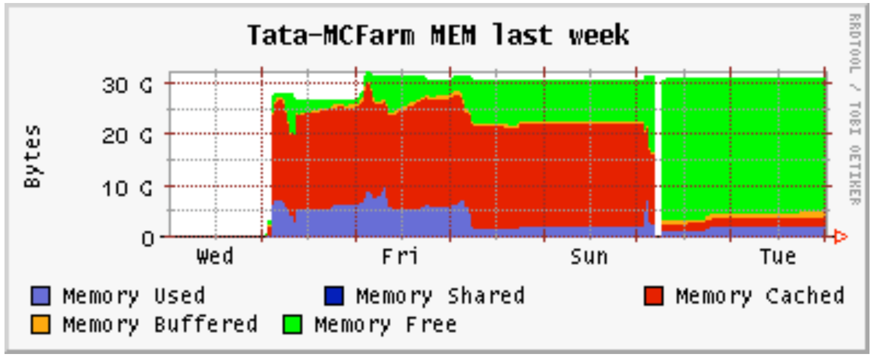
Life Cycle of a MC request execution by a Farmer


Query DB for next Request
Request OK?
yes
Get Request details And “Launch”
Parallelize into *Jobs* Of 500 events each
Pre-process each job with MC_runjob
Spool into Distribute Queue
Execute Queue
Gather Queue
Local Copy And/or Xfer to DB in Fermilab
Mark request As Finished
Blocks in Red are manual commands

Grid Report for Wed, 16 Apr 2003 00:15:37 -0500

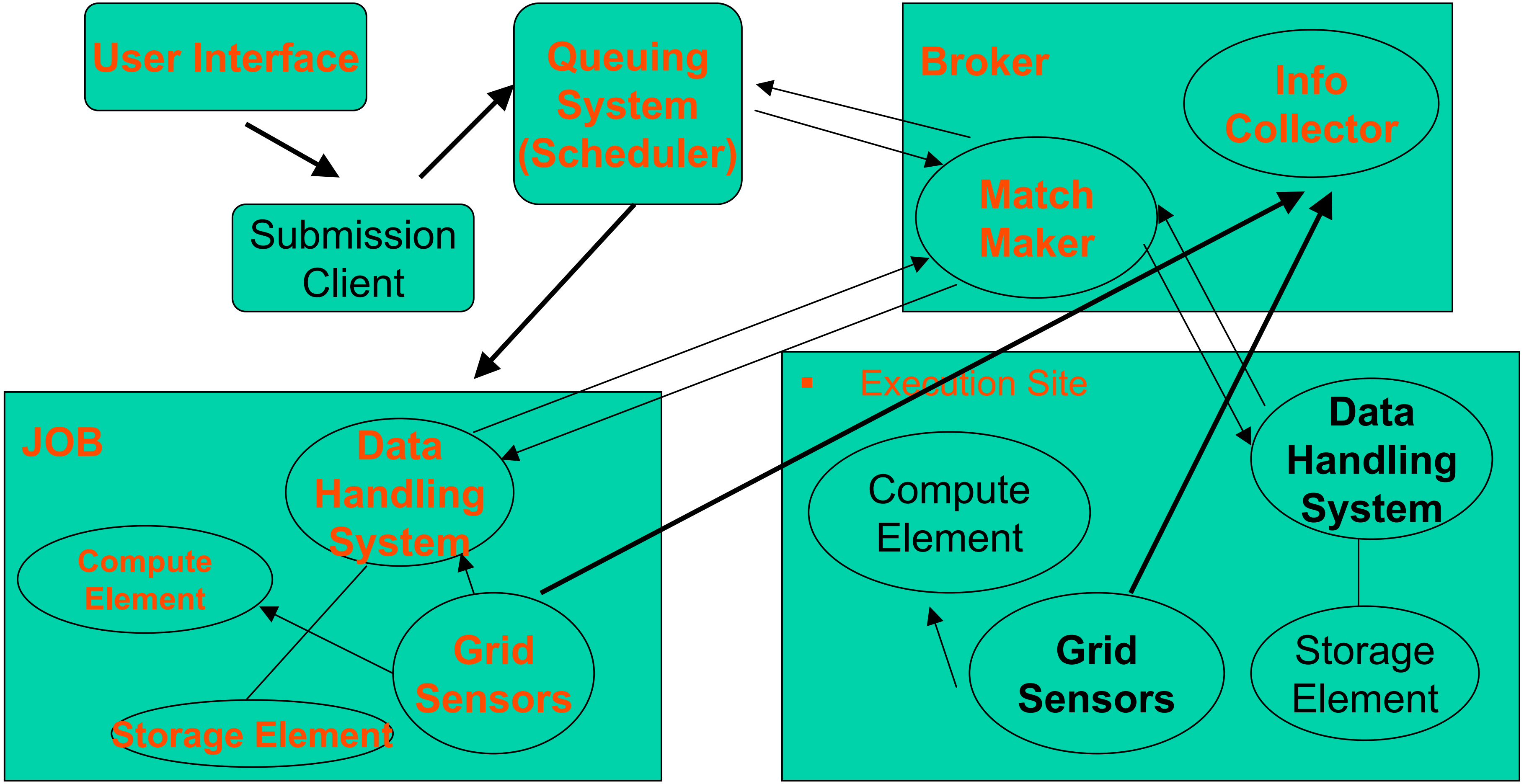
Metric Last Sorted





CSE-FARM **LTU Physics** **OUHEP** **SWIFT-FARM** **Tata-MCFarm**

Switching over to SAMGrid



SAMGrid: Basic Action

- UI Can be any desktop/laptop, with client software installed
- User must have *digital certificate* signed by one of accepted *Certificate Authorities*
- For the Grid, the user **job** is the set of various items: request id, software versions, data inputs, job size, and any other control parameters (***input sandbox***)
- The ***Grid Job*** is received by the *Scheduler*
- The Scheduler queries the *Resource Broker* for free resources
- As advised by the RB, the Scheduler pushes the job to an appropriate *Execution Site*

SAMGrid Basic Action (Contd)

- The ***Gatekeeper*** at the chosen Exec site receives the job,
- Authentication is performed, permissions are checked
- If all OK, Gatekeeper transfers the job to ***Job managers***
- The single grid job is decomposed into many local jobs, runnable in the local ***Batch Queue System***

SAMGrid Jobs: Execution and completion

- Status of jobs in the batch system are monitored (*waiting, active, errored, held,..)*
- Gathering of standard output, diagnostics and other output files (***output sandbox***)
- Each sandbox has unique ID.
- All sandboxes are “tarred” along with log files, and the single tar file is transferred to the Grid (could be local Storage Element)

Current Status

Worker nodes :	44	PIII 933 MHz
	8	Xeon 2.8 GHz
	16	Xeon 3.0 GHz
	18	Xeon 3.2 GHz

Total Computing power: 71k SI2k

Servers : Dual CPU server (Xeon 2.8GHz)

Additional server : One P1V (1.7Ghz), to share load of data transfers

Firewall : One PIV server acting as Firewall and Network Address Translator

Storage : 4TB RAID storage, served by a dual Xeon server via FiberChannel

BELLE CLUSTER

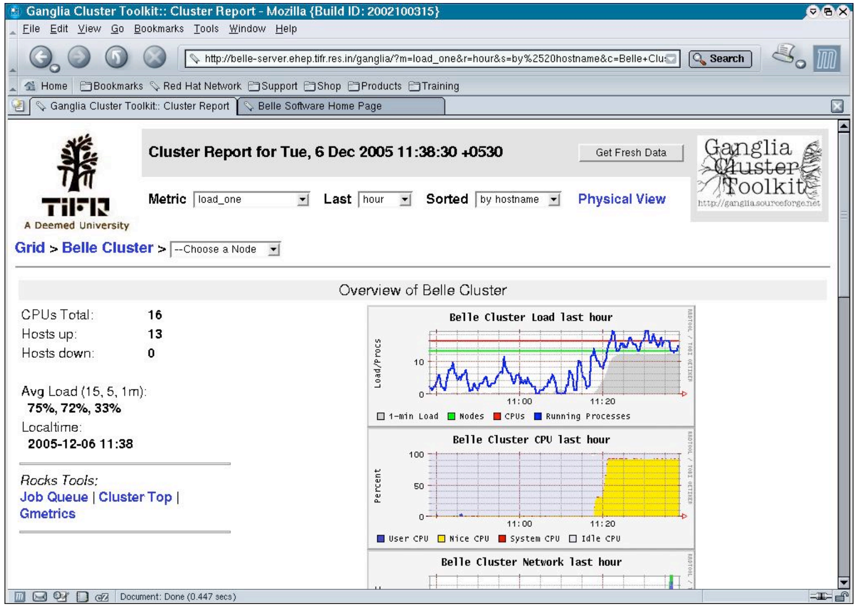
- 14 Pentium Class nodes
- 2 TB Disk Space, 80 GB Mass storage
- Software:
- ROCKS (3.3.0)
- Batch Processing(Platform LSF, Open PBS, Condor)
- LIBs for Parallel Processing
- (MPI, PVM, MAUI, HDF, LAM etc)
- C3 (Cluster Command and Control Suite)
- Ganglia (Web Monitoring Tool)

Performance

- 200, 000 Generic BELLE Events Simulated per day
- About One Million BELLE events per week
- About 12 GB/Day transfer to KEK Possible

When BELLE Goes for Grid Computing
we are ready to merge with BELLE Grid

Belle Cluster



Ganglia Cluster Toolkit:: Cluster Report - Mozilla {Build ID: 2002100315}
File Edit View Go Bookmarks Tools Window Help
http://belle-server.ehep.tifr.res.in/ganglia/?m=load_one&r=hour&s=by%2520hostname&c=Belle+Clus
Search
Home Bookmarks Red Hat Network Support Shop Products Training
Ganglia Cluster Toolkit:: Cluster Report
Belle Software Home Page
TIFR
A Deemed University
Ganglia Cluster Toolkit
http://ganglia.sourceforge.net

Cluster Report for Tue, 6 Dec 2005 11:38:30 +0530

Get Fresh Data

Metric load_one **Last** hour **Sorted** by hostname **Physical View**

Grid > Belle Cluster > --Choose a Node

Overview of Belle Cluster

CPUs Total:	**16**
Hosts up:	**13**
Hosts down:	**0**

Avg Load (15, 5, 1m):
75%, 72%, 33%

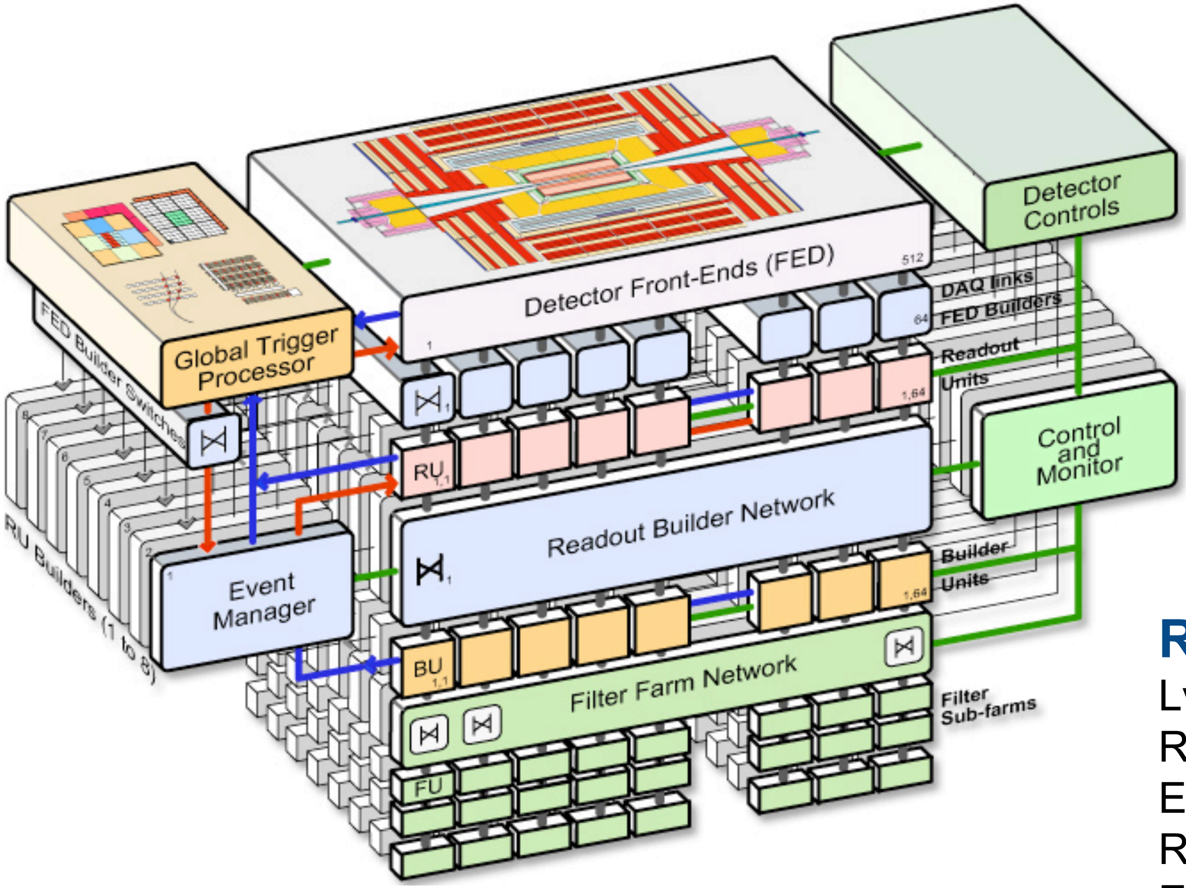
Localtime:
2005-12-06 11:38

Rocks Tools:
Job Queue | Cluster Top | Gmetrics

Belle Cluster Load last hour
Load/Procs
10
0
11:00
11:20
□ 1-min Load 🟩 Nodes 🟥 CPUs 🟦 Running Processes
Belle Cluster CPU last hour
Percent
100
50
0
11:00
11:20
🟦 User CPU 🟨 Nice CPU 🟥 System CPU □ Idle CPU
Belle Cluster Network last hour
RRDTOOL / TOBI OETIKER

Document: Done (0.447 secs)

CMS – DAQ implementations and scaling



Data to surface:

Average event size	1 Mbyte
No. FED S-link64 ports	700
DAQ links (2.5 Gb/s)	512+512
Event fragment size	2 kB
FED builders (8x8 dual)	64
Technology(2004)	Myrinet

Readout Builders (x8):

Lv-1 max. trigger rate	12.5 kHz
RU Builder (64x64)	125 Tbit/s
Event fragment size	16 kB
RU/BU systems	64
Event filter power	10^5 SI95
EVB technology (2006)	Open

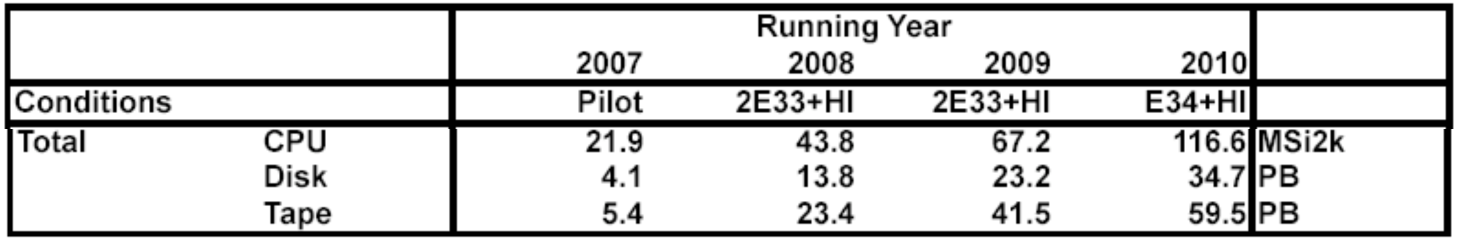
CMS Online Data Rates

Level –1 Trigger rate	100 kHz
Event Size	1 MB
Event Builder Bandwidth (10^5 Hz X 1 MB)	100 GB
# of Events to be written in tape	100 Hz
Rejection factor for High Level Trigger (HLT)	1000
CPU power required for HLT decision using Pentium III processors running at 1 GHz (41 SI95)	~ 300 msec
Total Event Filter Computing Power	1.2 X 10^6 SI95
Data Production	10 TB/day

DAQ system must provide the means to feed data from 700 front-end modules To ~1000 commercial processors at a sustained bandwidth of 100 GB/s

CMS Computing Model

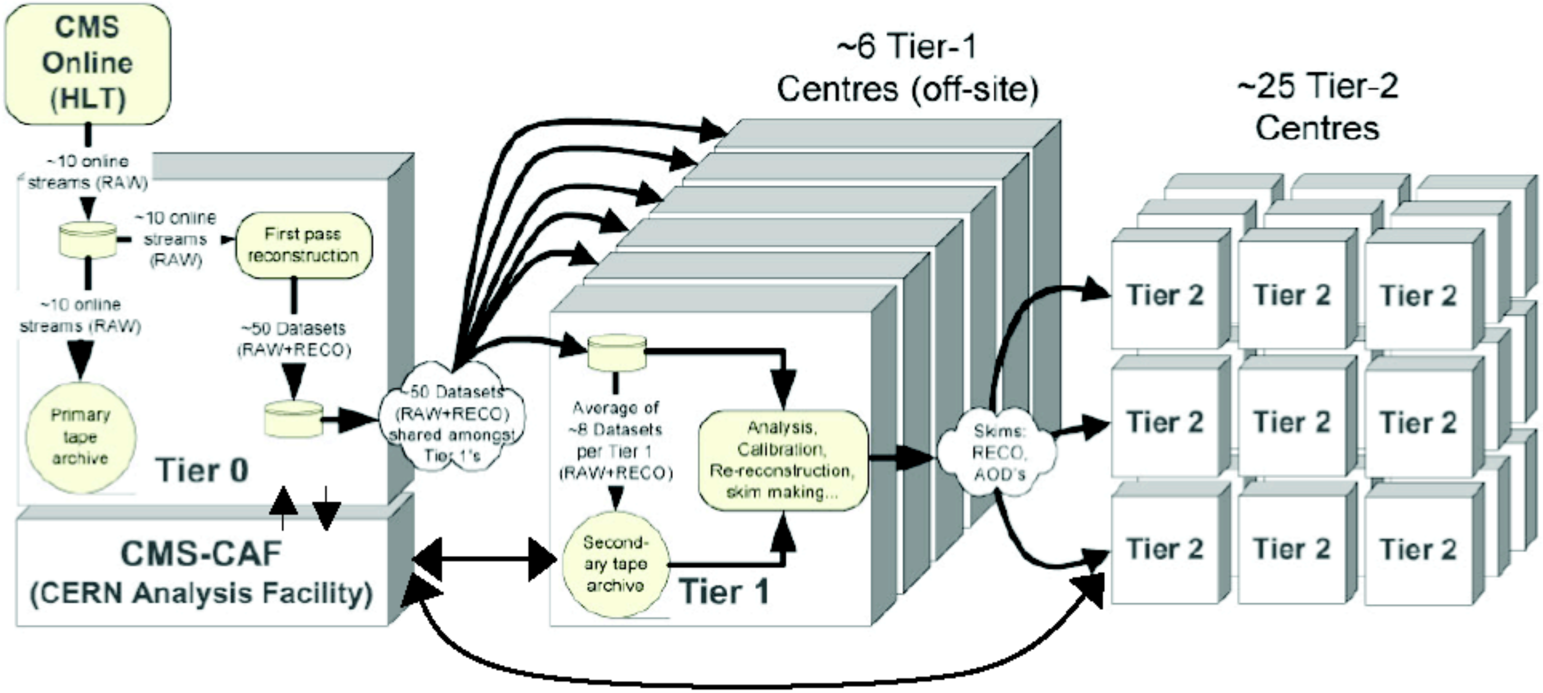
- Distributed model for computing in CMS
  - Cope with computing requirements for storage, processing and analysis of data provided by LHC



		Running Year				
		2007	2008	2009	2010	
Conditions		Pilot	2E33+HI	2E33+HI	E34+HI	
Total	CPU	21.9	43.8	67.2	116.6	MSi2k
	Disk	4.1	13.8	23.2	34.7	PB
	Tape	5.4	23.4	41.5	59.5	PB

  - Computing resources need to be geographically distributed, interconnected via high throughput networks and operated by means of Grid software
  - CMS computing TDR released in June 2005

Tiered Architecture



Tier-0:

- Accepts data from DAQ
- Prompt reconstruction
- Archives data and distributes them to Tier-1's

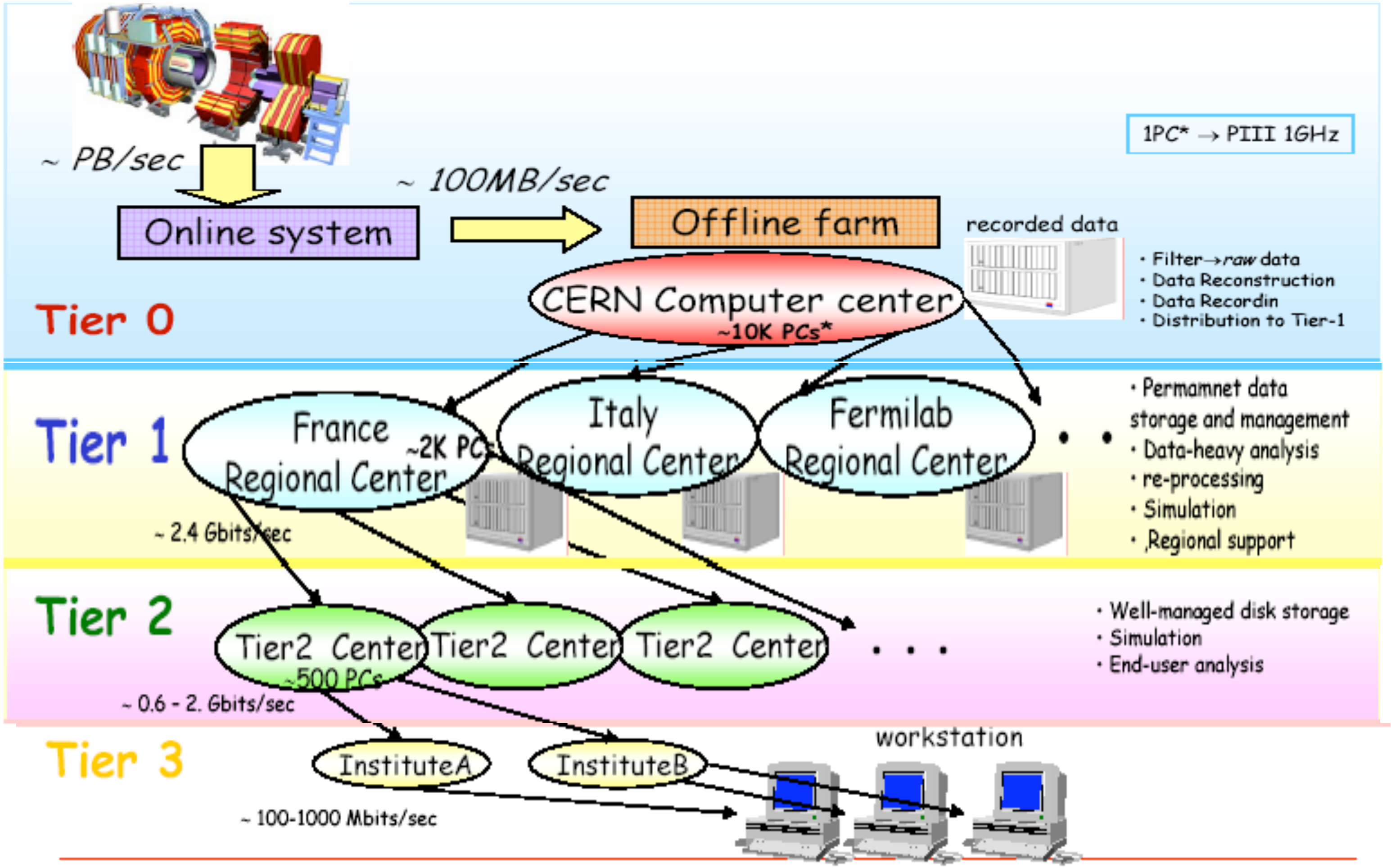
Tier-1's:

- Real data archiving
- Re-processing
- Calibration
- Skimming and other data-intensive analysis tasks
- MC data archiving

Tier-2's:

- Data Analysis
- MC simulation
- Import datasets from Tier-1 and export MC data

CMS Grid Structure



~ PB/sec

~ 100MB/sec

Online system

Offline farm

recorded data

1PC* → PIII 1GHz

Tier 0

CERN Computer center
~10K PCs*

- Filter→*raw* data
- Data Reconstruction
- Data Recordin
- Distribution to Tier-1

Tier 1

France Regional Center ~2K PCs

Italy Regional Center

Fermilab Regional Center

. .

- Permamnet data storage and management
- Data-heavy analysis
- re-processing
- Simulation
- ,Regional support

~ 2.4 Gbits/sec

Tier 2

Tier2 Center ~500 PCs

Tier2 Center

Tier2 Center

. . .

- Well-managed disk storage
- Simulation
- End-user analysis

~ 0.6 - 2. Gbits/sec

Tier 3

InstituteA

InstituteB

workstation

~ 100-1000 Mbits/sec

CMS Computing Requirements at Tier-n

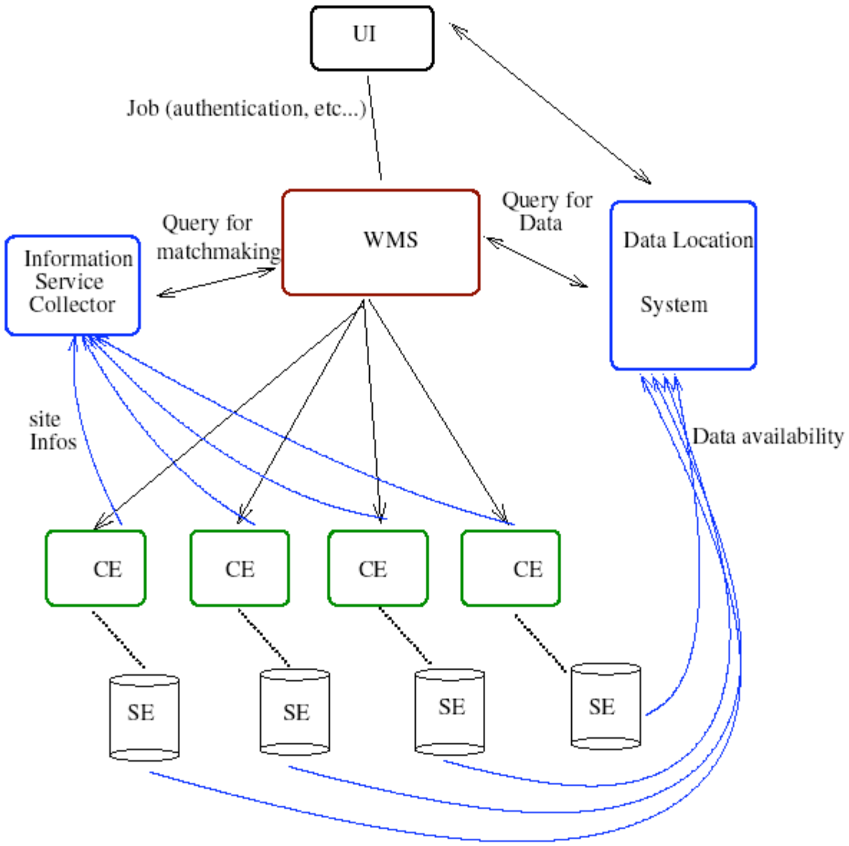
		Running Year				
Year		2007	2008	2009	2010	
conditions		Pilot	2E33+HI	2E33+HI	E34+HI	
A Tier-0	CPU	2.3	4.6	6.9	11.5	MSi2k
	DISK	0.2	0.4	0.4	0.6	PB
	TAPE	1.9	3.8	8	11	PB
	WAN	5	10	14	22	Gb/s
A Tier-1	CPU	1.1	2.1	3.1	5.8	MSi2k
	DISK	0.6	1.1	1.7	2.5	PB
	TAPE	0.9	1.8	3.7	5.5	PB
	WAN	5	9	14	21	Gb/s
A Tier-2	CPU	0.4	0.8	1.4	2.2	MSi2k
	DISK	0.1	0.2	0.4	0.7	PB
	WAN	0.6	1	1.7	2.5	Gb/s

Workload and Data Management Systems

Design philosophy:

- Use Grid Services as much as possible and also CMS-specific services
- Baseline system with minimal functionality for first physics
- Keep it simple!
- Optimize for the common case:
    - Optimize for read access (most data is write-once, read-many)
    - Optimize for organized bulk processing, but without limiting single user
- Decouple parts of the system:
    - Minimize job dependencies
    - Site-local information stays site-local
- Use explicit data placement
    - Data does not move around in response to job submission
    - All data is placed at a site through explicit CMS policy
- Grid interoperability (LCG and OSG) We expect to operate in a hetrogeneous GRID enviornment but required the details of local GRID implementations to be largely invisible to CMS physicists.

WMS & DMS Services Overview



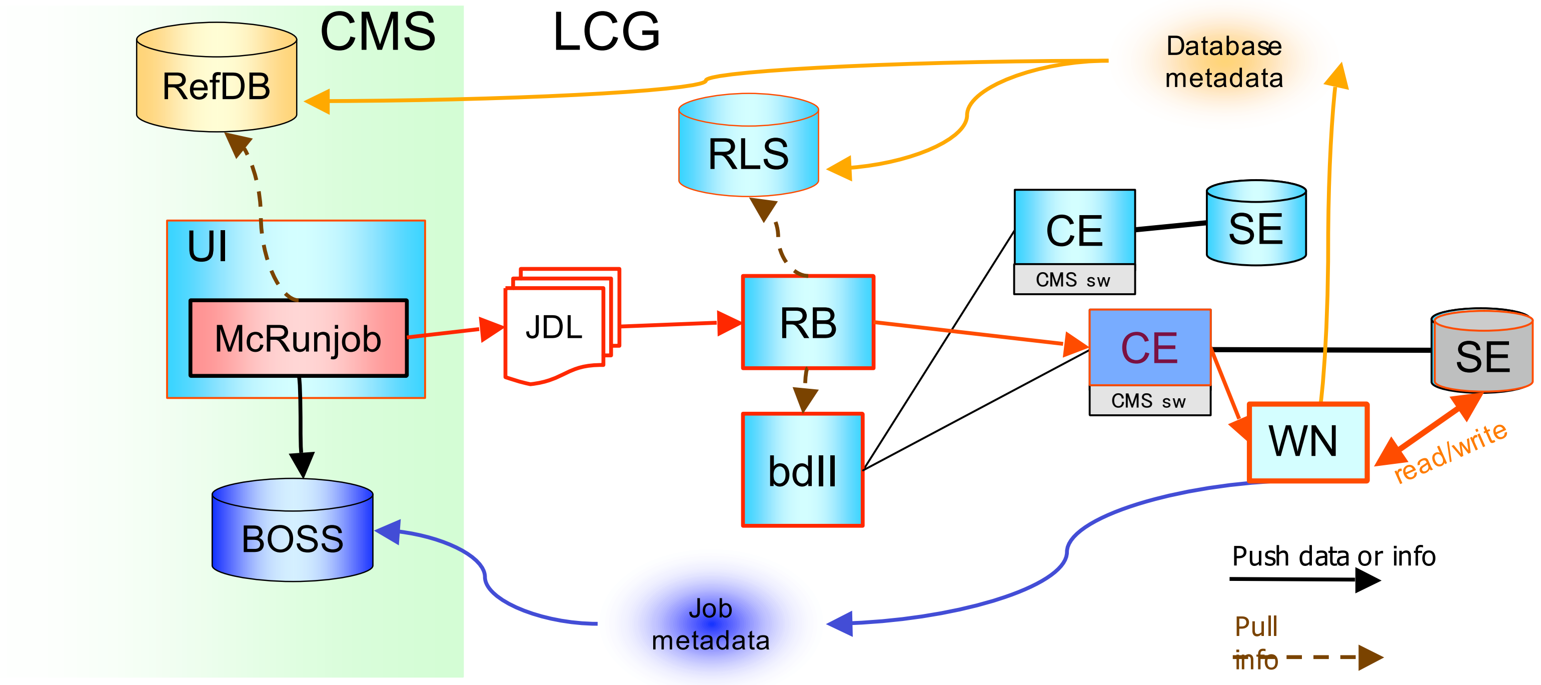
DMS

- No global file replica catalogue
- Track and replicate data with a granularity of 'file blocks'
- Dataset Bookkeeping System(DBS)
  - "What data exist?"
- Data Location Service (DLS)
  - "Where are data located?"
- Local File catalogue
- Data Access and Storage
  - SRM and posix-IO
- Data Transfer and placement system

WMS

- Rely on Grid Workload Management
  - Reliability, performance, monitoring
- Hierarchical task queue in future
- Grid and CMS-specific job monitoring and bookkeeping

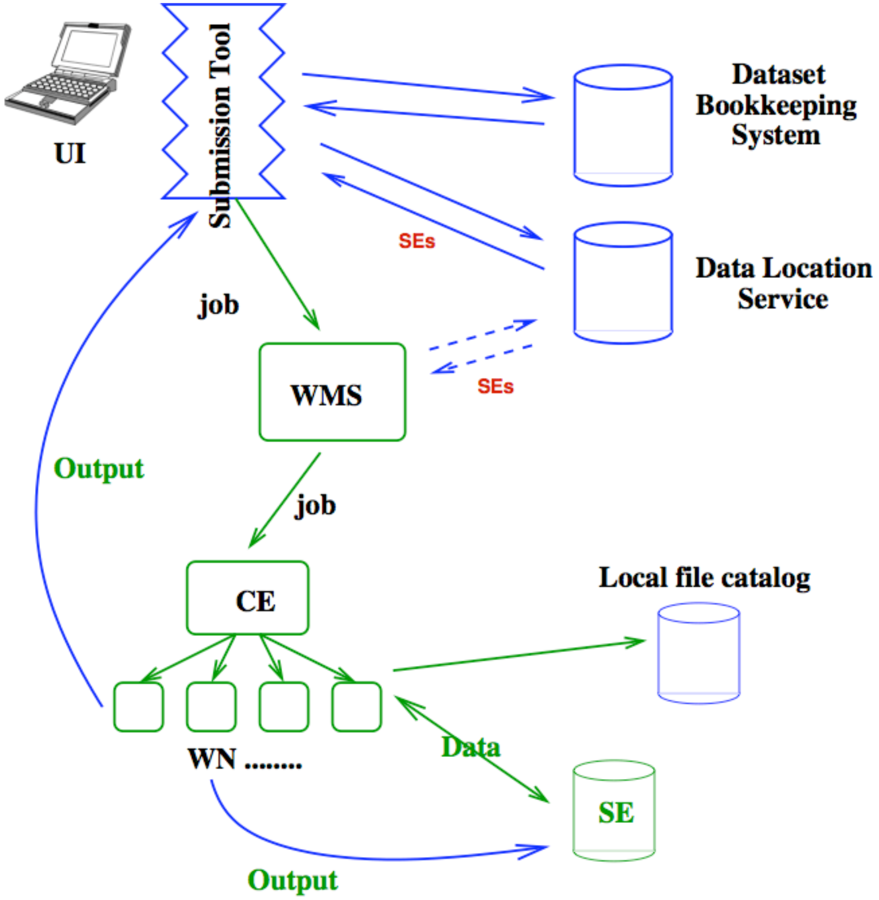
LCG Production Workflow



- Physics groups submit data production request to a central system (RefDB)
- Quasi-real-time job monitoring through BOSS (Batch Object Submission System)
- Normally experiment software pre-installed

Data Analysis on the Grid

- Data samples for the CMS Physics TDR distributed in Tier-1 sites (~80 million events)
- End-to-end analysis via LCG Grid
- Simple analysis scenario where data is pre-located and jobs are sent to the data
- CMS remote Analysis Builder (CRAB) tool for job preparation, submission, execution and basic monitoring



India-CMS Grid

- India-CMS grid will have the following structure
  - A Tier-2 station at Tata Institute, Mumbai
  - Tier-3 stations at other participating institutes - Panjab University, Delhi University, BARC. (Vishwabharati will join later)

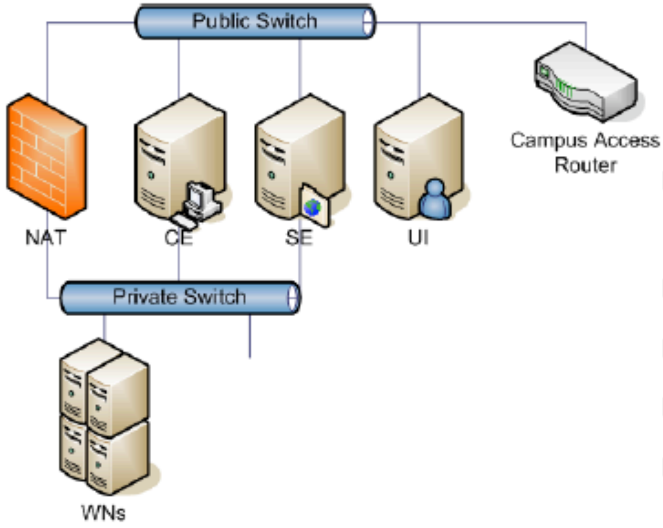
  - **Projected Resource Requirements at CMS Startup**

Item	Tier 2 Centre	Tier 3 Centres
CPU Requirement (SI95)	24 k	3 X 32 k
Disk Storage (TB)	330	3 X 30
Tape Storage (TB)	200 (+1000)	3 X30
Network bandwidth (Mbps)	622 (34)	34 (2)

CMS Tier-2 Pilot Project

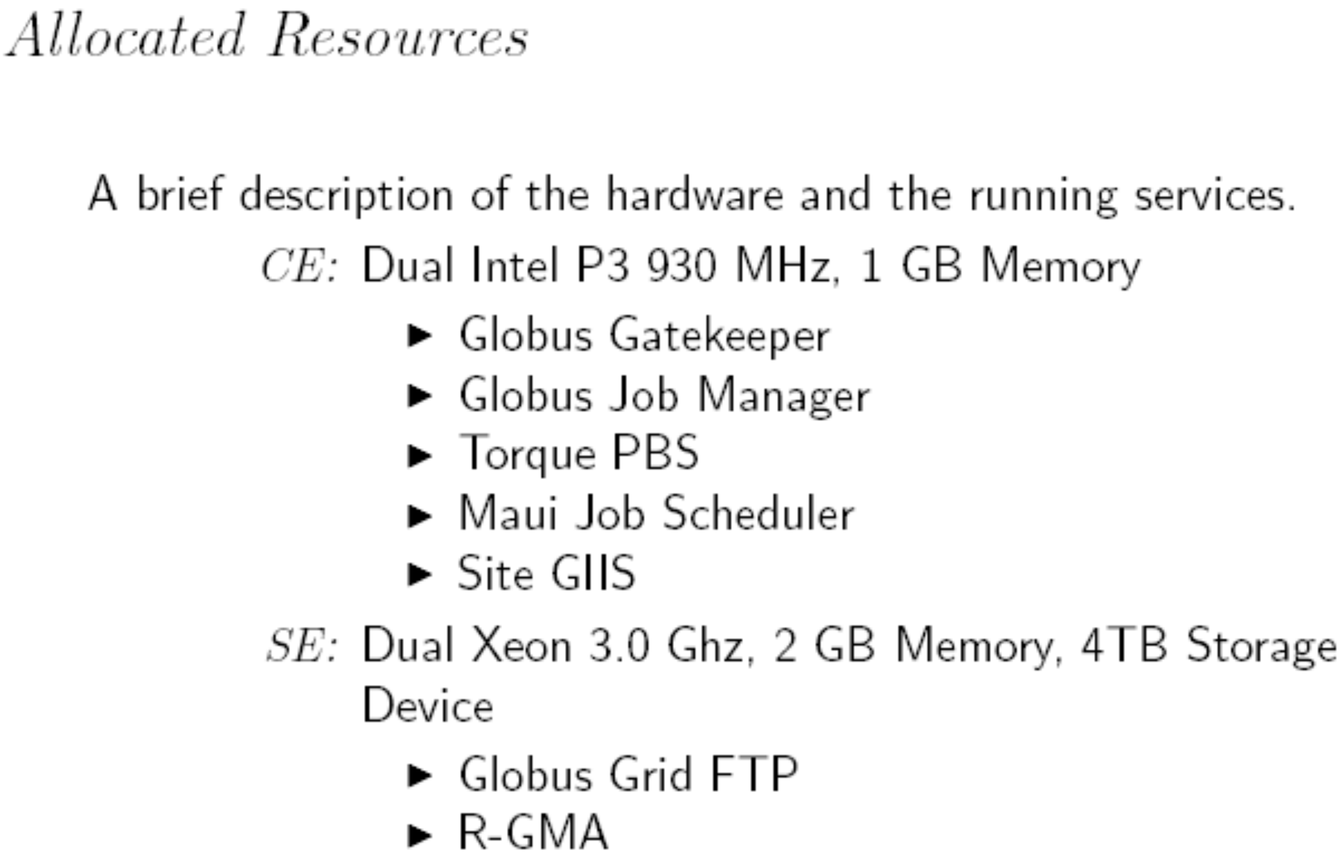
- TIFR-CMS LCG2 Grid site :
  - The idea is to familiarize ourselves before setting up the official CMS Tier-2 Grid station with all its funtionality.
  - Supported VOs are CMS and dteam
  - Runs CMS production MC jobs submitted in the Grid

Logical Organisation of the Site



- Computing Element x 1 (CE)
- Storage Element x 1 (SE)
- NAT x 1
- Worker Nodes x 3 (WN)
- User Interface x 1 (UI)
  Doubles up as a worker node.

Allocated Resources



Allocated Resources

A brief description of the hardware and the running services.

CE: Dual Intel P3 930 MHz, 1 GB Memory

- ▶ Globus Gatekeeper
- ▶ Globus Job Manager
- ▶ Torque PBS
- ▶ Maui Job Scheduler
- ▶ Site GIIS

SE: Dual Xeon 3.0 Ghz, 2 GB Memory, 4TB Storage Device

- ▶ Globus Grid FTP
- ▶ R-GMA

Allocated Resources

NAT: Dual Intel P3 930 MHz, 1 GB Memory

- Network Address Translation: Route outgoing requests from Worker Nodes to public network.

WN: Dual Xeon 3.0 Ghz, 2GB memory

- Torque Clients
- Experimental Software: Currently OSCAR, ORCA and CMKIN.
- Networked in a private subnet, not visible to outside world.
- Use the NAT to connect to the internet. (Required for the CMS experimental software)

Status of India-CMS Tier 2 project

- Present status
- Hardware
- 4 Grid managing servers.
  - User Interface (UI)
  - Computing Element (CE)
  - Storage Element (SE) with 1 TB storage disk
  - DNS Server
- 36 Intel Pentium-IV worker nodes.
- 34 Mbps internet connection.
-
- Software
- Scientific Linux 3.0.5 O.S.
- LCG-2_6_0 middleware installed.
- Portable Batch processing System, PBS installed.
- CMS software is installed
- Immediate Future:
- **CPU Power : 80k SI2k**
- **Storage Device : 50 TB**

Status of India-CMS Tier 2 project

- **Site information:**

Site name:	INDIACMS-TIFR
Site address:	http://www.indiacms.res.in
Email:	support@indiacms.res.in
User Interface:	ui.indiacms.res.in

- Presently site is up and in a process of testing by *Site Functional Testing (SFT)* team. Person having certificate can submit job to "**ce.indiacms.res.in:2119/jobmanager-lcgpbs-dteam/cms**" job scheduler queues.

Summary

CMS:

- CMS has adopted a distributed computing model which makes use of Grid technologies
- Production CMS services on the Grid in place
  - Data Management and Workload Management systems
    - Data transfer and placement system
    - Monte Carlo production
    - Data Analysis
- Steadily increase in scale and complexity

India:

- Over the last several years we are stadily developing Farm & Grid infrastructure and successfully running two computing farms one dedicated to D0 and other to Belle.
- D0 farm now being replaced by SAMgrid.
- A pilot project with full functionality of a typical CMS Tier-2 Grid station but with limited resources has been implemented.
- The setting up of the India-CMS Tier-2 Grid station at TIFR is on track.
- India-CMS Tier-3 stations are at various stages of implementations
- Lot of work ahead for all of us.