How to do Physics Analysis of LHC Data?

Hafeez Hoorani

National Centre for Physics

Outline

- Physics Channels at LHC
- Physics Analysis of LHC data
- Physics Analysis at NCP
- Computing Challenge
- LHC Computing Grid (LCG)
- Analysis Facility at NCP
- How others can join?

Physics Channels

- **Higgs Searches:**
  - $M_H < 140$ GeV $H \rightarrow \gamma\gamma$
  - $140 < M_H < 700$ GeV ($H \rightarrow llll$)
  - $M_H > 500$ GeV ($H \rightarrow lljj$)
- **Studies of CP Violation:**
  - $B^0 \rightarrow J/\psi\, K^0_s$
- **Super-symmetry:**
  - SUSY Higgs Boson
  - Sparticles (sleptons, squarks, gluinos …)

Physics Channels

- Excited Quarks
- Leptoquarks
- Monopoles
- Extra-dimensions
- Compositeness
- Standard Model Physics

Standard Model Physics

- **QCD Studies**
  - Jet Studies
  - α_s and its running
  - Inclusive b production
- **Top Quark Physics**
  - Top pair production (σ, m_t, properties of top)
  - Single top (V_{tb},…)
- **Electroweak Physics** (W & Z cross-section, Drell-Yen, PDF, TGC, …)

Analysis overview

Real Life

Machine Events
LHC

⇓

Detector, DataAcquisition
CMS, ATLAS , ALICE

Virtual Reality

Event Generation
PYTHIA, HERWIG

⇓

Detector Simulation
GEANT4, LCG, OSCAR, FAMOS

⇓

Event Reconstruction
ORCA, FAMOS

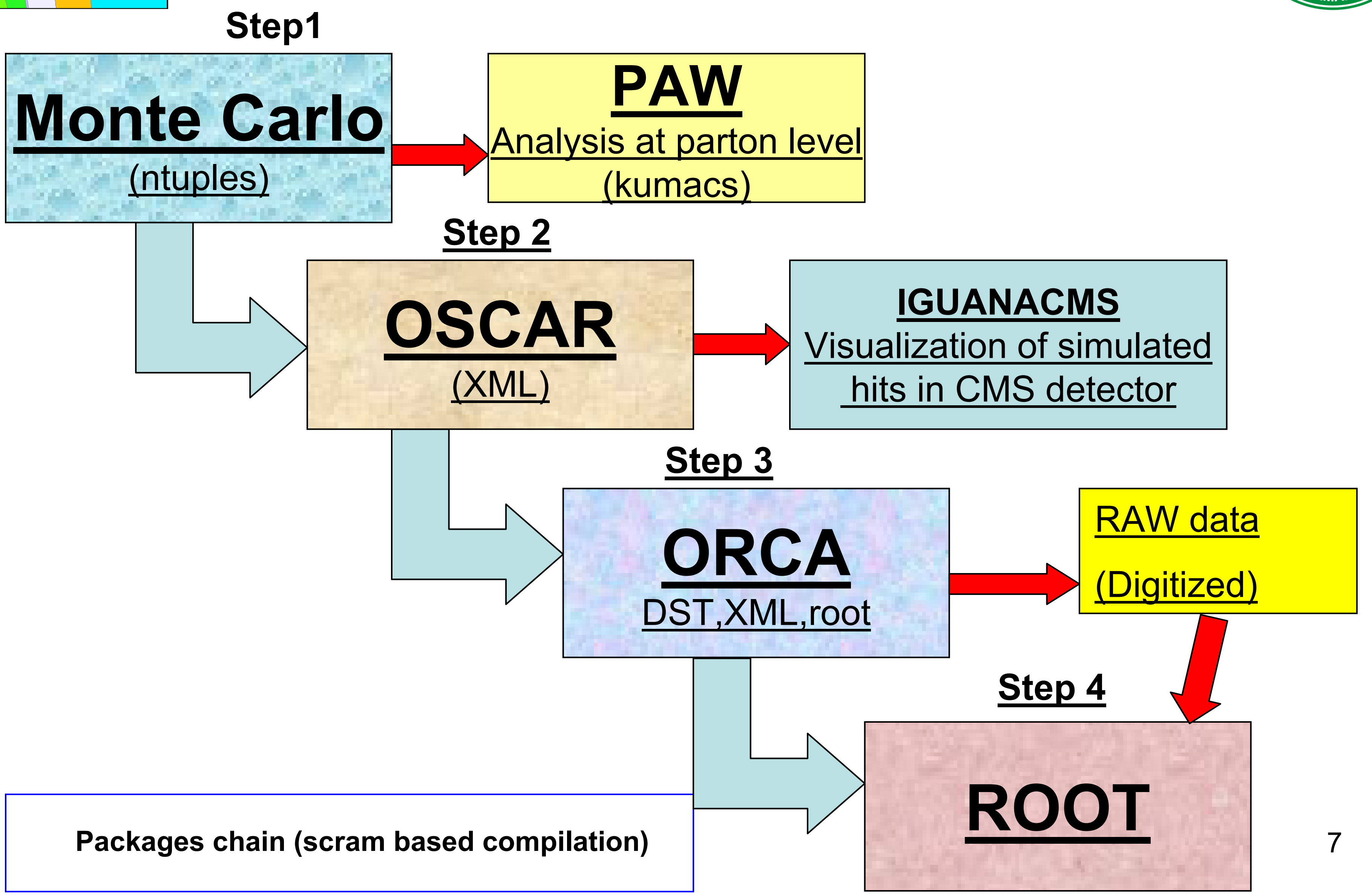
⇓

Physics Analysis
ROOT, PAW

⇓

Plots, Histograms, Graphs, Results & Conclusions

Data Analysis Chain



List of Analysis packages Required

- Generation Step
  - CMKIN(PYTHIA, TOPREX, CompHEP, HERWIG, ISAJET, AlpGen, MadGraph)
    - http://cmsdoc.cern.ch/cms/PRS/gentools/
- Simulation Step
  - OSCAR (Object Oriented Simulation for CMS Analysis and Reconstruction)
    - http://cmsdoc.cern.ch/oscar/
- Reconstruction Step
  - ORCA (Object Oriented Reconstruction for CMS Analysis)
    - http://cmsdoc.cern.ch/orca/
- Analysis Step
  - ROOT (An Object Oriented Data Analysis Framwork)
    - http://root.cern.ch

List of Analysis packages Required

- **Auxiliary Software**
  - COBRA (Coherent Object Oriented based for Reconstruction, Analysis and simulation)
    - http://cobra.web.cern.ch/cobra/
  - IGNOMINY
    - http://ignominy.web.cern.ch/ignominy/
  - Geometry (CMS Geometry Project)
    - http://cmsdoc.cern.ch/cms/software/geometry/index.html
  - FAMOS (CMS FAST simulation Package)
    - http://cmsdoc.cern.ch/famos/
  - IGUANACMS (Interactive Graphics for Users ANAlysis)
    - http://iguanacms.web.cern.ch/iguanacms/
  - SCRAM (Software Configuration, Release and Management)
    - http://cmsdoc.cern.ch/Releases/SCRAM/doc/scramhomepage.html

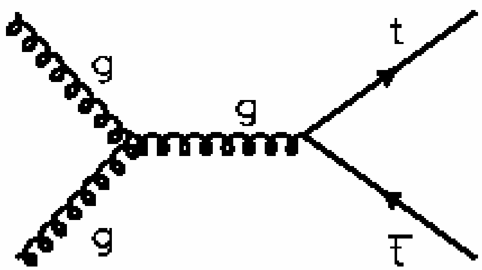
NCP top quark group

- 7 people working on CMS analysis:
  - **Hafeez R. Hoorani:** Supervisor of CMS Analysis Activities
  - Ijaz Ahmed: (PhD) (Semi-leptonic decays)
  - M. Irfan : (PhD) (Single top studies)
  - M. Usman: (PhD) (Higgs Searches)
  - Taimoor Khurshid: (M. Phil)(Semi-leptonic decays)
  - Hamid Ansari: (M. Phil)(Single top)
  - Waqas Mahmood: (M. Phil)(Rare Top decays)

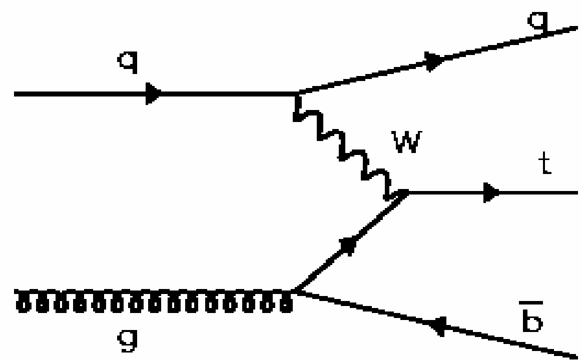
Top quark properties

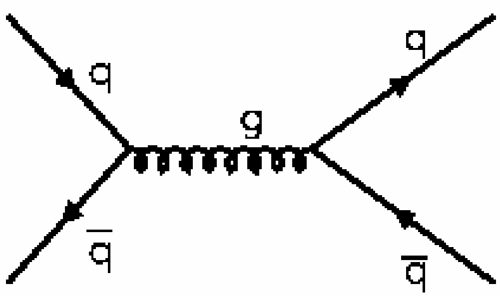
- The heaviest known elementary particle.
- Discovery 1995 at Tevatron
- Pole mass (174 ± 5.1) GeV
- Decay width (1.4 GeV)
- Life time ~ 10^{-25} s
- t → bW (~ 100%)
- Spin and parity $J^P(SM) = 1/2^+$
- Weak iso-spin eigenvalue = I_3 = +1/2
- tt Production cross-section @ 7 TeV = 830 pb
- In one year at LHC if integrated luminosity is 10 fb^{-1} number of tt events produced is 8.3 M

Leading ttbar pair production diagrams

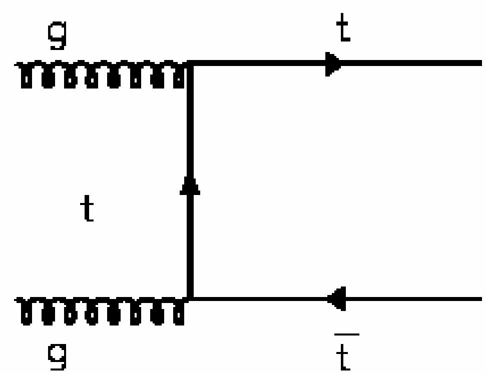


Gluon fusion



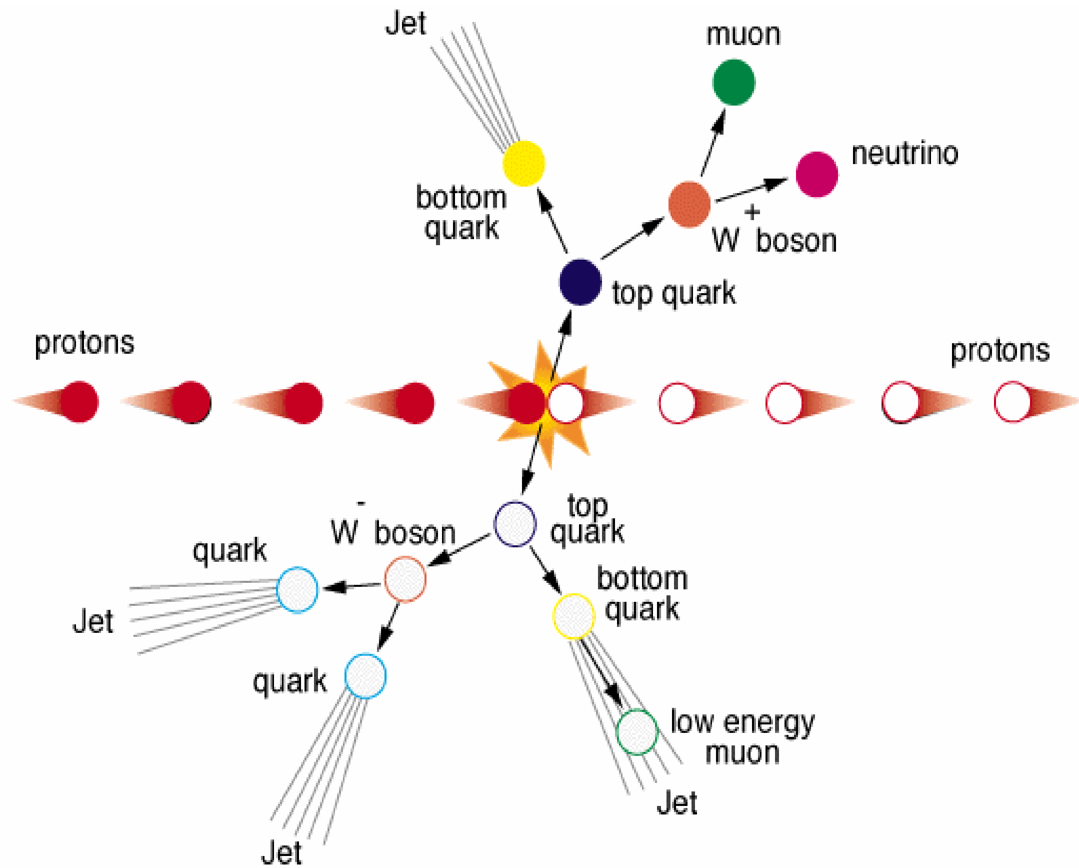


Quark annihilation



Associated productions

Topology of the Semi-Leptonic Top Decay Channel

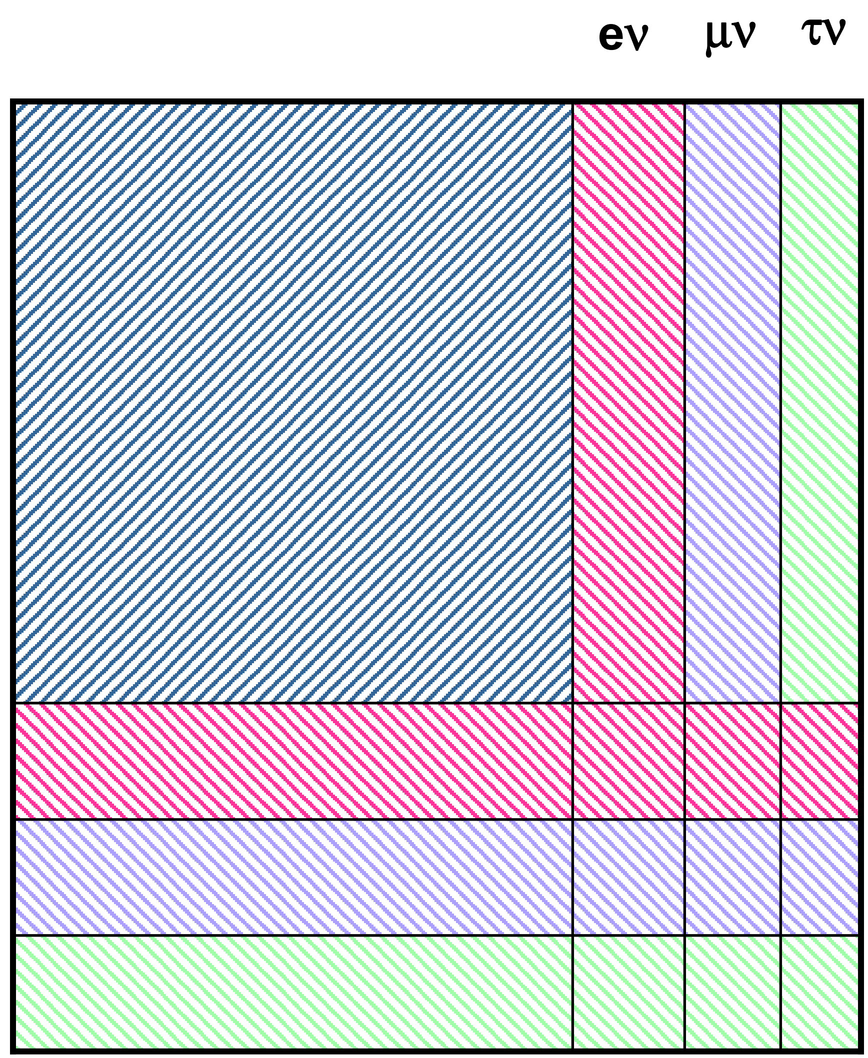


A Top Antitop Quark Event

Events depend on W decay modes
Leptons plus jets:
one W decays to jets(67%)
other into leptons (33%)

$t\bar{t}$ Decay Modes

Decay mode	Branching ratio
$tt \rightarrow W^{+}W^{-}bb \rightarrow bbqqqq$	36/81
$tt \rightarrow W^{+}Wbb \rightarrow bbqq'e\nu$	12/81
$tt \rightarrow W^{+}Wbb \rightarrow bbqq'\mu\nu$	12/81
$tt \rightarrow W^{+}Wbb \rightarrow bbqq'\tau\nu$	12/81
$tt \rightarrow W^{+}Wbb \rightarrow e\nu\mu\nu bb$	2/81
$tt \rightarrow W^{+}Wbb \rightarrow e\nu\tau\nu bb$	2/81
$tt \rightarrow W^{+}Wbb \rightarrow \mu\nu\tau\nu bb$	2/81
$tt \rightarrow W+Wbb \rightarrow e\nu e\nu bb$	1/81
$tt \rightarrow W+Wbb \rightarrow \mu\nu\mu\nu bb$	1/81
$tt \rightarrow W+Wbb \rightarrow \tau\nu\tau\nu bb$	1/81



Major Background Processes

Process	Cross-section (pb)
$t\bar{t}(signal)$	830
$b\bar{b} \rightarrow l\upsilon + jets$	2.2×10^{6}
$W + jets \rightarrow l\nu + jets$	7.8×10^{3}
$Z + jets \rightarrow l^{+}l^{-} + jets$	1.2×10^{3}
$WW \rightarrow l\upsilon + jets$	17.1
$ZZ \rightarrow l^{+}l^{-} + jets$	3.4
$WZ \rightarrow l\upsilon + jets$	9.2

Computing Challenge

- CMS Detector has 15 million channels.
- Typical detector occupancy is 10 – 15%
- Average event size is 1 MB
- Event rate is 100 Hz
- In a given year LHC will run for 10 million seconds

Total Data Size = 10^6 x 100 x 10^7 = 1 PB

Computing Challenge

Year	Beam Time (sec/year)	Lumi. cm^{-2} s^{-1}	RAW Data (MB)	RECO (MB)	AOD (MB)
2007	5×10^6	5×10^{32}	1.5	0.25	0.05
2008	10^7	2×10^{33}	1.5	0.25	0.05
2009	10^7	2×10^{33}	1.5	0.25	0.05
2010	10^7	2×10^{34}	1.5	0.25	0.05

LHC Computing Grid (LCG)

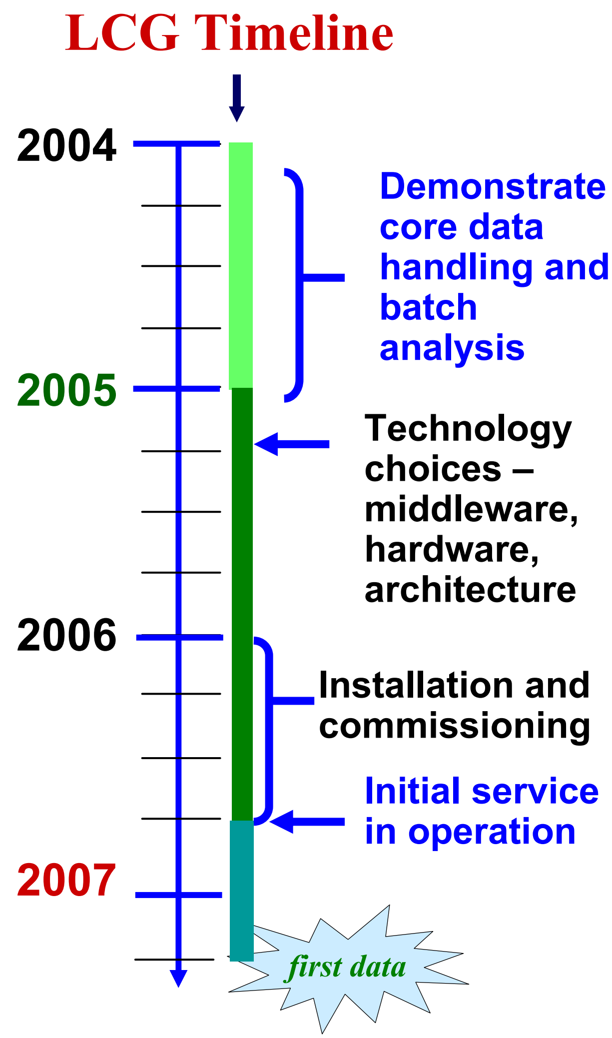
LCG – 1 Service opened September 15, 2003.

- It is common computing facility for all LHC experiments.
- As of May 31, 2005:
  - Number of Nodes: 148
  - Number of CPUs: 13,268
  - Total Storage: 5 PB (5,000,000 GB)

LHC will produce **15 PB** of data per year.

LHC Computing Grid

- LCG is based on concept of Regional Centres.
- Uses concepts of VO.
- Hierarchical Structure
  - Tier – 0, Tier – 1, …, Tier – 4
- CERN is a Tier – 0, all RAW Data from LHC will be stored at CERN.
- Model based on 1/3 (CERN), 2/3 (Outside)
- **NCP is LCG Grid Node.**



Requirements for LCG Node

For Grid Node requirements are:

- Public Key Infrastructure (PKI) for use with Grid authentication middleware.
- High bandwidth network connectivity
- Grid node hardware elements:
  - **Computing Element** **(CE)**
  - **User Interface** **(UI)**
  - **Resource Broker** **(RB)**
  - **Berkley Database Index Information** **(BDII)**
  - **Proxy Server** **(PS)**
  - **Storage Element** **(SE)**

PK-Grid-CA

- **PK-Grid-CA**
  - Part of **EU-Grid PMA**, authorized International Policy Management Authority working under IGF
  - Compliant with **RFC-2527**
  - Based on Cryptographic toolkit OpenSSL
  - Online Certificate Request for User/Host Certificate

PK-Grid-CA

- **The use of PKI enables**
  - A secure exchange of digital signatures
  - Encrypted documents
  - Authentication
  - Authorization
  - Other functions in open networks where many communication partners are involved

Analysis Facility at NCP

- NCP has a network connectivity of 2Mbps, used for:
  - LCG Node and CMS Data Production
  - Sending/receiving emails and web surfing
- LCG Node, Web and Mail Servers are maintained by NCP staff
- Increased hardware resources recently for the LCG Node

Analysis Facility at NCP

- Linux based PC Cluster, running Scientific Linux 3.0.4
- PC Cluster is based on:
  - CPU **22 P-IV 3.2 GHz**
  - Disk 2.66 TB
  - Memory **1 GB each**
  - Servers 5 Intel Xeon 3.2 GHz
    Dual Processor
    Hot pluggable SCSI Drives
    Redundant Power Supplies

Analysis Facility at NCP

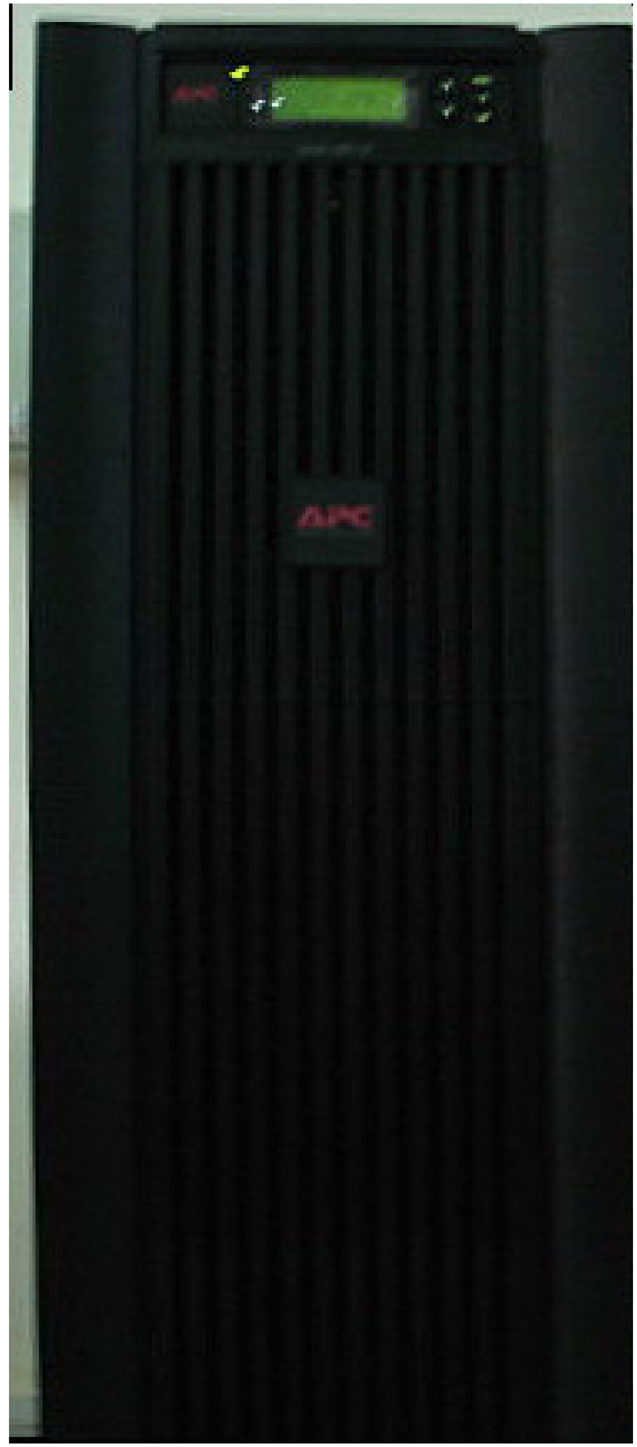
- Mass Storage: **RAID – 5** with hot pluggable SCSI
- Magnetic Tape Storage: **40/80 GB tapes**
- Total Storage Capacity: ~ **10 TB**

- All user accounts (NIS), files (NFS) and software (CVS) is centralized
- Job scheduling done using open PBS


UI
RB
CE


PK-GRID-CA
SIGNING SERVER


WORKER NODES
UI
RB
CE
WORKER NODES


APC

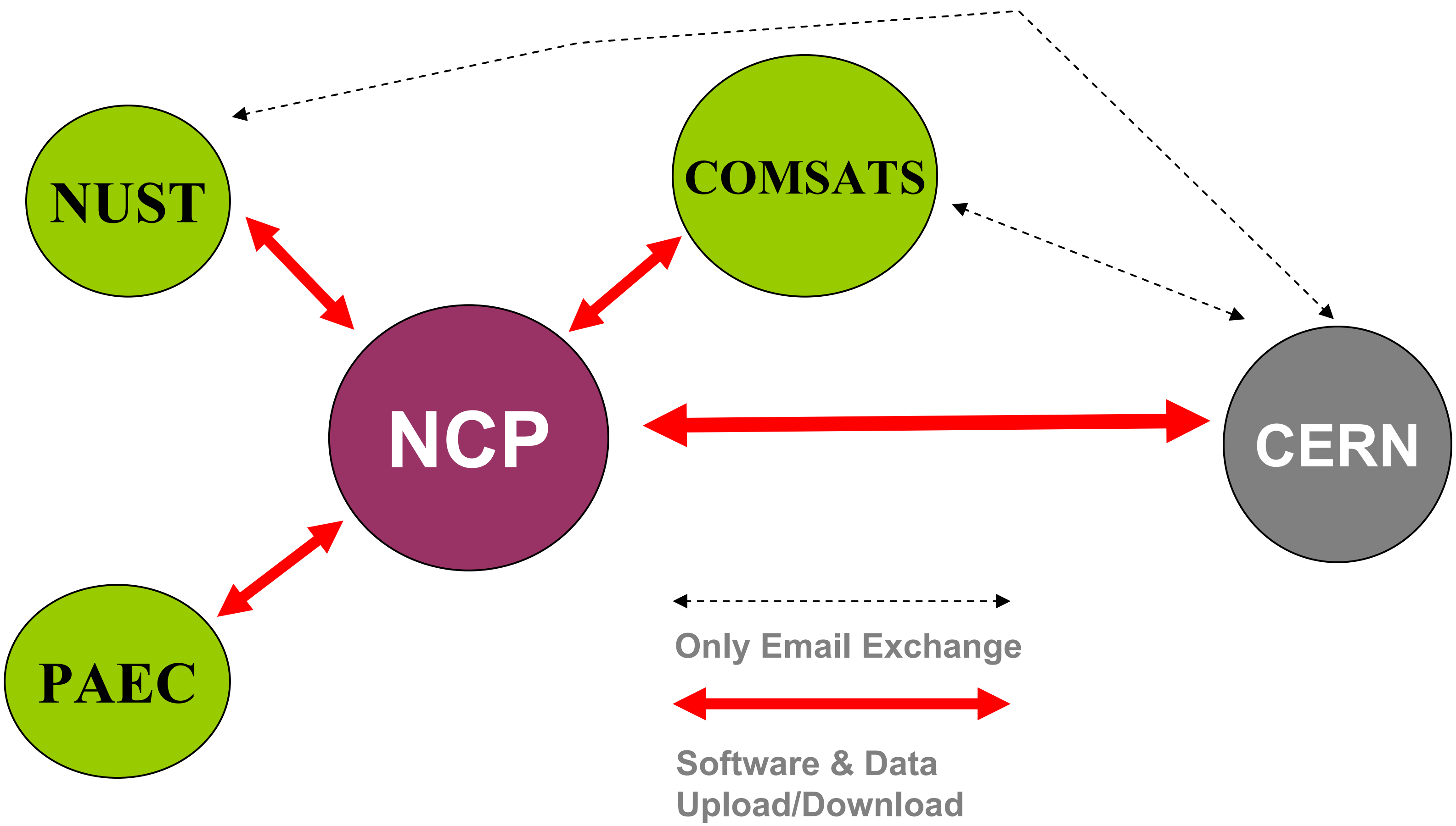
NCP Computing Model

- NCP computing model is based on pooling and sharing the resources of other institutes.
- In return providing them Linux training, access to GRID & CERN Software.
- Following are our partners in Pakistan:
  - PAEC
  - NUST
  - COMSATS
- NCP is establishing PK–SNT–GRID.

Computing Resources

INSTITUTE	PCs	Bandwidth (kbps)	FTE
NCP	28	2048	3
NUST	08	1024	8
PINSTECH	14	1024	3
KANUPP	14	Dial Up	4
HMC-III	05	128	3

Small GRID


NUST
COMSATS
NCP
CERN
PAEC
Only Email Exchange
Software & Data
Upload/Download

Accomplishments

- Organization of Grid Technology Workshop in October, 2003 in collaboration with CERN
- Became Regional Centre for CMS Production in August 2003
- Produced 1.38 million events for CMS (~ 280 GB)
- In May 2004, NCP became a fully operational LCG Grid Node in Pakistan
- Run CMS Data Production on Grid, more than 7 M CMKIN and 2.1 M OSCAR events have been produced
- **NCP is accredited GRID-CA, since 09/2004**

How others can take part?

- Access to LHC data is open to anybody.
- To become part of this activity requirements are:
  - Group of Physicists
  - Computing Infrastructure
  - Funds for traveling and stay at CERN
  - Authorship of papers
    - No charge for research students
    - Faculty & Ph.D authors contribute towards M&O

How others can take part?

- Write proposal to a funding agency.
- Training can be provided by NCP:
  - Data Analysis
  - Computing
  - Workshop such as this is a good opportunity
- NCP will also provide main computing infrastructure:
  - Large data storage
  - Software update.

System Requirements

- **Supported Platforms** CERN Scientific Linux (slc3_ia32_gcc323)
- Others (CERN Red Hat)

Distribution	Arch	Version	Perl-Tk	Installation	Toolchecker	OSCAR Verify	ORCA Verify
Scientific Linux CERN	i386	3.0.x	perl-Tk from distribution	ok	ok	ok	ok
Scientific Linux FERMI	i386	3.0.x	perl-Tk from distribution	ok	ok	ok	ok
CERN RedHat	i386	7.3.x	perl-Tk from distribution	ok	ok	ok	ok

System Requirements

- P-IV PCs are enough
- Can start analysis with couple of PCs
- Storage; minimal-requirement is 1TB
- Good connectivity is required
- At least one machine on live IP

Prerequisite for CMS software Installation (XCMSI)

- http://cmsdoc.cern.ch/cms/oo/repos_standalone/download/
- GOAL
  - Provide complete CMS software environment for development and data analysis
  - Desirable properties of experiment software installations
  - No root privilege required
  - Relocatable packages
  - Optional network download
  - Batch mode installable
  - Save able and re-useable setup

Prerequisite for CMS software Installation (XCMSI)

- Included validation procedure
- Concise configuration for less experienced users
- Multiple platform support
- Multiple installation possible
- Suitable packages in the form of RPMs
- PERL version 5.6.0 or higher perl-Tk rpm

* List of all available RPMs and download tags are available on this page, just follow the instructions
* Documentation: http://cmsdoc.cern.ch/cms/oo/repos_standalone/download/doc/xcmsi/
* Description of the packages: http://cmsdoc.cern.ch/cms/oo/repos_standalone/download/desc.php

Monte Carlo Generators

- **Why Generators?**
  - Generators acts like accelerators(LHC,LEP,TEVATRON)
  - Discovery of Top, Higgs, Supper-symmetry
  - Allow theoretical and experimental studies of complex multi-particle physics
  - Vehicle of ideology to disseminate ideas from theorists to experimentalists
  - Predict the event rates and topology (Kinematics of particles resulted from collisions)
  - Simulate possible backgrounds
  - Study detector requirements

Monte Carlo Generators

- Study detector imperfections
- Evaluation of acceptance corrections
- Estimation of cross-sections ,branching ratios and decay Widths
- PDF uncertainties
- Hard processes and resonance decays
- ISR and FSR
- LO and NLO calculations

Event Generation Structure

- **Initialization step**
  - Select process to study
  - Modify physics parameters
  - Set kinematic constraints
  - Modify generator settings
  - Initialize generator
  - Book histograms
- **Generation loop**
  - Generate one event at a time
  - Analyze it
  - Add results to histograms
  - Print a few events
- **Finishing step**
  - Print cross-sections/BR
  - Print/save histograms

OSCAR

- *Full CMS simulation based on the Geant4 toolkit*
- **Geant4**
  - Physics processes describing in detail electro-magnetic and hadronic interactions tools for the CMS detector geometry implementation interfaces for tuning and monitoring particle tracking
- CMS changed from **CMSIM/GEANT3(Fortran)** to **OSCAR/GEANT4(C++)** at the end of 2003
- OSCAR used for substantial fraction of DC04 production
- It is being used for physics TDR production

OSCAR

- **CPU**
  - OSCAR ≤ 1.5 x CMSIM - with lower production cuts
- **Memory**
  - ~110 MB/evt for pp in OSCAR ≈ 100 MB in CMSIM
- **Robustness**
  - From ~1/10000 crashes in pp events (mostly in hadronic physics) in DC04 production to 0 crashes in latest stress test (800K single particles, 300K full QCD events)

OSCAR

Interfaces and Services

- Application steering handled by CMS framework
- Detector geometry construction automated via Detector Description Database which converts input from XML files managed by Geometry project Generator input (via RawHepEvent CMS format and recently HepMC) converted to G4Event
- Specific generator type and event format run-time configurable Interface from CMS magnetic field services to G4
- Field selection run-time configurable

OSCAR

- Propagation parameters via DDD/XML Infrastructure for physics lists and production cuts via DDD/XMLUser actions (monitoring, tuning) via dispatcher-observer pattern for observable entity Persistency
- Histogramming, monitoring etc transparently through CMS framework (COBRA)
- Time spent in magnetic field query (P4 2.8 GHz) for 10 minimum bias events (with delta=1mm) 13.0 vs 23.6 s for G3/Fortran field, new field ~1.8-2 times faster than FORTRAN/G3GEANT4 volumes can be connected to corresponding magnetic volumes ⇒ avoid volume finding ⇒ potential ~2x improvement with G4, possible to use local detector field managers

ORCA

- Framework for reconstruction and is intended to be used for final detector optimizations, trigger studies or global detector performance evaluation
- Object oriented system for which C++ has been chosen as programming language
- Design is based on CARF (**C**MS **A**nalysis and **R**econstruction **F**ramework), which was developed to prototype reconstruction methods, initially for testbeam applications
- A database of digitized events which are created given as an input from OSCAR
- Contains: MC info, SimTrack, SimVertex
- Simulated hits (sub-detector info)

See tutorials on http://cmsdoc.cern.ch/orca/

ORCA

- Digitized events and associations
- All Triggering levels (LV1,LV2,HLT)
- Reconstruction starts from above data as:
  - Data unpacking
  - Apply calibration scheme
  - Reconstruction of clusters or hits
  - Reconstruction of Tracks
  - Reconstruction of Vertices
  - Particle identification, ($e,\gamma,\mu/\pi$,Jets,b)

DSTs

- Store a complete record of all physics objects created during reconstruction process by ORCA
- Provide compact information for Analysis
- Consist of a set of homogeneous collections of Reconstructed (~50 objects)
- Objects looks like tracks, vertices, muons, electrons, light jets, b-jet, taus

http://cmsdoc.cern.ch/orca/testdata.html

Public Datasets

- **PubDB link**
  - http://cmsdoc.cern.ch/cms/production/www/PubDB/GetPublishedCollectionInfoFromRefDB.php
- **Dataset Name (link to RefDB) Owner Name (link to PubDB)**
- **Contains info about:**
  - **ORCA version**
  - **Compiler info**
  - **Luminosity**
  - **Number of Events**
  - **Run numbers**
  - **Regional centre(RC)**
  - **Grid station(server) link**
  - **Name Hits, Digi (pile-up + without pile-up), DST**

Public Datasets

- For example if one wants to use the DST for $t\bar{t}$ inclusive channel
  - ***dataset name (shows required channel)***
- *Name:* *jm03b_TTbar_inclusive*
- *Dataset owner:* *jm_DST871_2x1033PU_g133_OSC*

- Background : jm03_Wjets_150_250 (dataset)
  jm_DST871_2x1033PU_g133_OSC (owner)

How to access datasets? (CRAB)

- http://cmsdoc.cern.ch/cms/ccs/wm/www/Crab/
- **Prerequisites:**
  - CERN AFS account on LXPLUS
  - CMS published data (PubDB)
  - GRID station name (e.g; CERN,FNAL,CNAF,IN2P3,LNL,BA etc.)
  - Working area should be UI
  - Valid GRID certificate and valid GRID-proxy on UI
  - Virtual organization must be CMS
  - CMS software's must be installed on UI
  - CRAB is a Python program intended to simplify the process of creation and submission of CMS analysis jobs into Grid environment.
  - Parameters and card-files for analysis should be provided by the user changing the configuration file *crab.cfg*.

CRAB: Job creation + Job submission + Job monitoring + Output retereival

CRAB

- CRAB generates scripts and additional data files for each job.
- The produced scripts are submitted directly to the Grid.
- CRAB is aimed to give access to all data produced on any GRID station in the world without any knowledge of LCG at all.
- Get it from CVS repository
- How to get a certificate from the CERN CA?

CRAB is a tool for the CMS analysis on the Grid environment. It is based on the ideas from CMSprod, a production tool

CRAB

- The CERN CA requires three conditions:
  - to have an LXPLUS account
  - to have a valid CERN access card
  - to be present at CERN
- **How to get a certificate from the another CA**
- If you cannot be at CERN anytime soon, you should request a certificate from Certificate Authority
  - https://lcg-registrar.cern.ch/pki_certificates.html
- **NCP is also a Certificate Authority**, issuing digital certificates to grid users/hosts
  - PK-Grid-CA, the only certification authority in Pakistan
  - You can submit an online request for user/host certificate at: http://www.ncp.edu.pk/pk-grid-ca

Physics Analysis using CRAB

- ***Requirements*:**
  - **Owner and dataset name** to be analyzed
  - **Executable (ExRootAnalysis, ExDSTStatistics…)**
  - ORCA executable name (e.g: EXDigiStatistics): CRAB finds the executable in the user scram area (e.g:/afs/cern.ch/user/i/iahmed/ORCA_8_7_3/bin/Linux__2.4/here)
  - **output_file name**
  - name of outputs produced by ORCA executable (comma separated list). Empty entry means no output produced
  - **total_number_of_events, job_number_of_events, first_event**
  - total number of events to be analyzed, number of events for each job and first event number to be analyzed

CRAB

- **orcarc_file**
  - ORCA card to be used. This card will be modified by CRAB according to the job splitting. Use the very same card you used in your interactive test: CRAB will modify what is needed.
- **data_tier**
  - possible choices are ``DST, Digi, Hit'' (comma separated list, mind the case!) If set, the job will be able to access not only the data tier corresponding to the dataset/owner asked, but also to its ``parents''. This requires parents published in the same site of the primary dataset/owner. If not set, only the primary data tier will be accessible
- **Create a proxy**, before submitting jobs:
  - At CERN, you can use ``lxplus'' as a UI by sourcing the file
  - In this case you would not need to move to a CRAB working directory.
  - The executable file is crab.py
  - CRAB uses initialization file *crab.cfg* which contains configuration parameters. This file is written in the Windows INI-style. The default filename can be changed by the *-cfg* option.

For Further Information:

https://lcg-registrar.cern.ch/pki_certificates.html
http://lcg.web.cern.ch/LCG/catch%2Dall%2Dca/
http://service-grid-ca.web.cern.ch/service-grid-ca/
http://service-grid-ca.web.cern.ch/service-grid-ca/help/user_req.html
http://service-grid-ca.web.cern.ch/service-grid-ca/help/renew.html
http://cmsdoc/peopleCMS.shtml
https://edms.cern.ch/file/454439//LCG-2-UserGuide