Data Challenging

Paolo Capiluppi

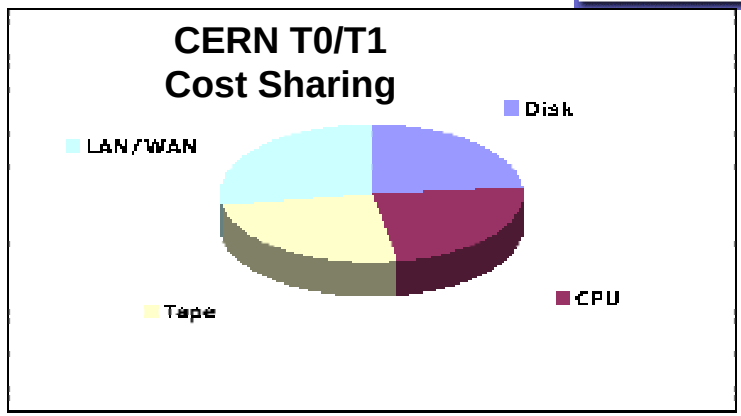
Dipartimento di Fisica e INFN, Bologna

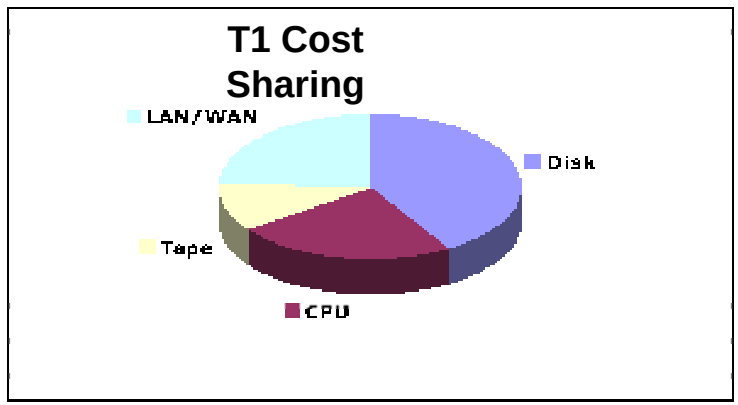
Sommario

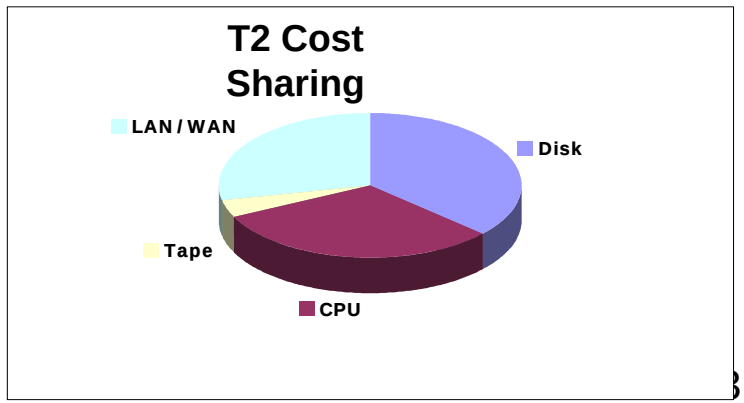
- Perche'?
  - Definire un Computing Model
  - Preparare l'Analisi [→ presentazioni in questo WS]
  - Simulare il comportamento dei detector, trigger compreso [→ idem]
  - Definire l'organizzazione dell'Infrastruttura, risorse e persone
- Come?
  - Attraverso i Data Challenges
  - Componenti specifiche di Esperimento
  - Componenti comuni: LCG e/o Grid
- Dove siamo?
  - Data Challenges di ATLAS e CMS
- Cosa manca?
  - Dimostrare quale scalabilita'
  - Misurare il carico dell'analisi
  - Definire meglio
    - → l'organizzazione
    - → cosa e' comune tra gli esperimenti
  - (I Data Challenges futuri)

Dimensioni del Calcolo ad LHC (~2008)

- **CERN T0/T1**
  - Disk Space [PB] 5
  - Mass Storage Space [PB] 20
  - Processing Power [MSI2K] 20
  - WAN [10Gb/s] ~5?
- **Tier-1s (Sum of ~10)**
  - Disk Space [PB] 20
  - Mass Storage Space [PB] 20
  - Processing Power [MSI2K] 45
  - WAN [10Gb/s/Tier-1] ~1?
- **Tier-2s (Sum of ~40)**
  - Disk Space [PB] 12
  - Mass Storage Space [PB] 5
  - Processing Power [MSI2K] 40
  - WAN [10Gb/s/Tier-2] ~.2?
- **Cost Sharing**
  30% At CERN, 40% T1s, 30% T2's

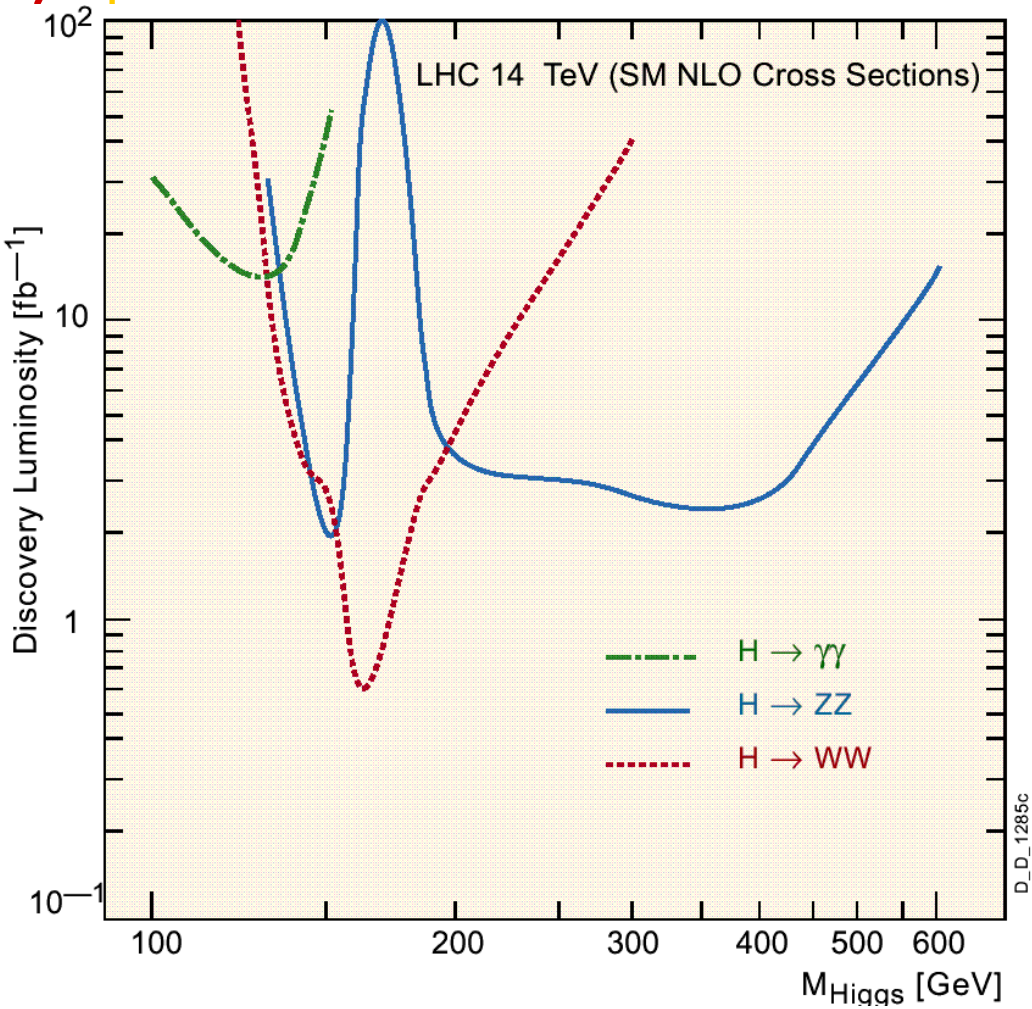






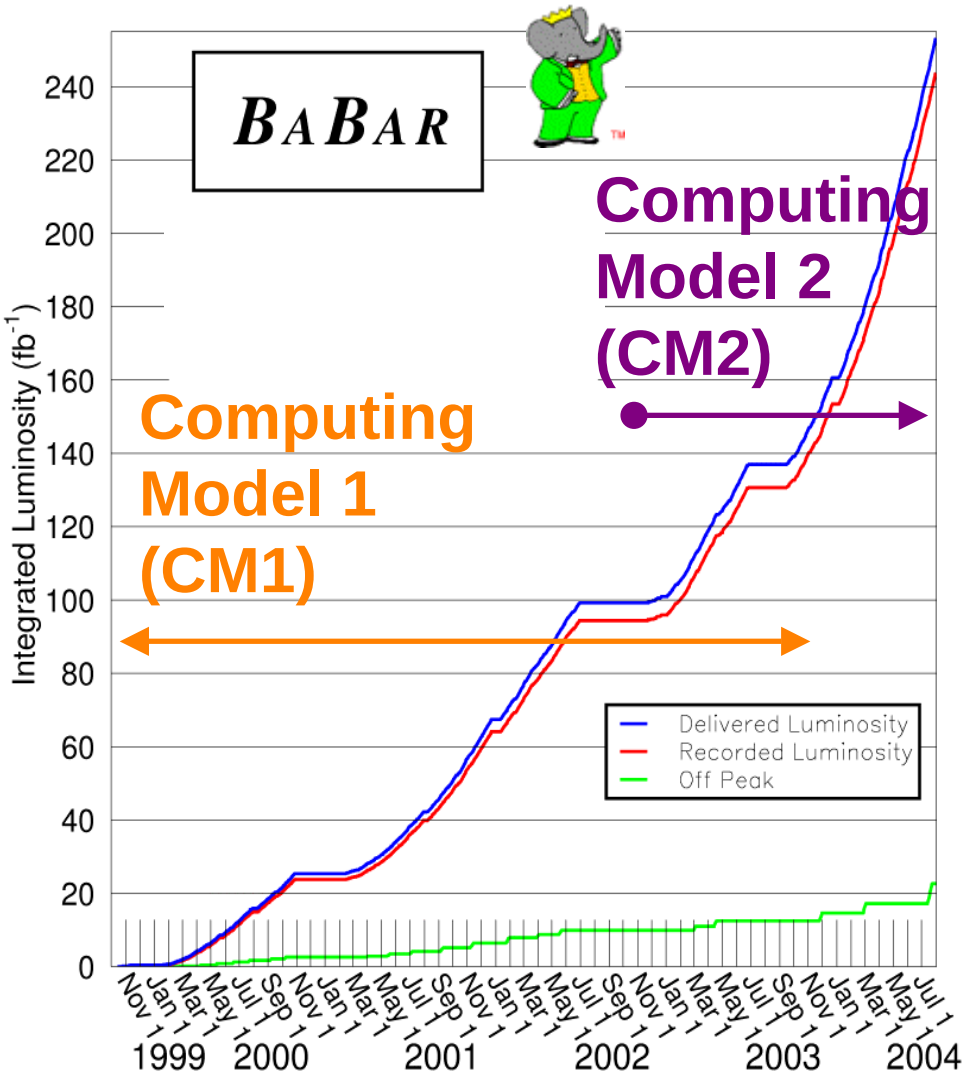
The Goal is the Physics, not the Computing...

- Motivation: at $L_0=10^{33}$ $cm^{-2}s^{-1}$,
  - 1 fill (6hrs) ~ 13 pb^{-1}
  - 1 day ~ 30 pb^{-1}
  - 1 month ~ 1 fb^{-1}
  - 1 year ~ 10 fb^{-1}
- Most of Standard-Model Higgs can be probed within a few months
  - Ditto for SUSY
- Turn-on for Detector(s) + Computing and Software will be crucial



Perché un Computing Model?

- **Ogni Esperimento ne ha uno !**
- **Occorre garantire ad ogni partecipante l'accesso ai dati**
  - Per produrre risultati di Fisica sfruttando le capacità di ogni individuo
- **Occorre garantire il mantenimento dei dati**
  - E la consistenza di questi
- **Occorre garantire le priorità e le scelte dell'Esperimento**
  - Salvaguardando l'autonomia di ognuno (e delle Istituzioni)
- **Occorre sfruttare al meglio le risorse**
  - Di hardware ed umane



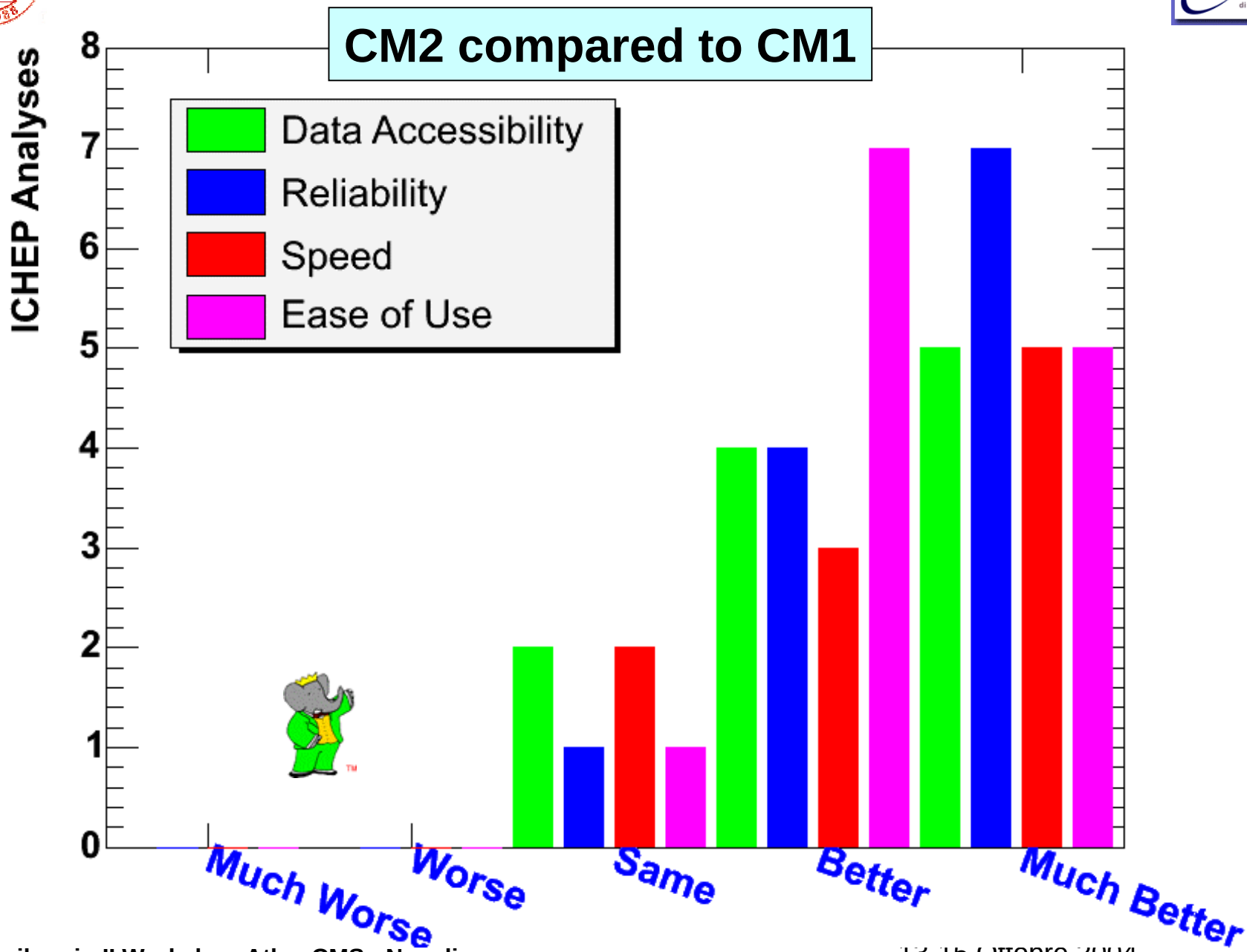
→ Organizzazione dei dati e del loro accesso

Elementi di un Computing Model

- **Componenti**
  - **Data Model**
    - Event data sizes, formats, streaming
    - Data "Tiers" (DST/ESD/AOD etc)
    - Roles, accessibility, distribution,…
    - Calibration/Conditions data
    - Flow, latencies, update freq
    - Simulation. Sizes, distribution
    - File size
  - **Analysis Model**
    - Canonical group needs in terms of data, streams, re-processing, calibrations
    - Data Movement, Job Movement, Priority management
    - Interactive analysis

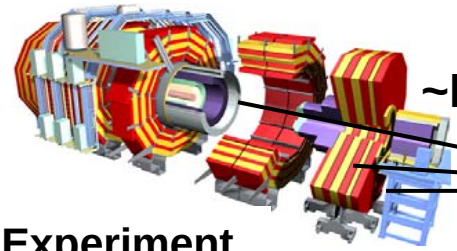
- **Metodologie di implementazione**
  - **Computing Strategy and Deployment**
    - Roles of Computing Tiers
    - Data Distribution between Tiers
    - Data Management Architecture
    - Databases
    - Masters, Updates, Hierarchy
    - Active/Passive Experiment Policy
  - **Computing Specifications**
    - Profiles (Tier N & Time)
      - Processors,
      - Storage,
      - Network (Wide/Local),
      - DataBase services,
      - Specialized servers
    - **Middleware requirements**

Valutazione del CM2 di BaBar



Un Modello Gerarchico (MONARC)

CERN/Outside Resource Ratio ~1:2
Tier0/(Σ Tier1)/(Σ Tier2) ~1:1:1



Experiment

~PByte/sec

Online System

~100-1500 MBytes/sec

Tier 0 +1

CERN Center
PBs of Disk;
Tape Robot

~10k PCs

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

Tier 1

~2.5-10 Gbps

IN2P3 Center ~2k PCs

RAL Center

INFN Center

FNAL Center

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

Tier 2

2.5-10 Gbps

Tier2 Center

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

Tier 3 ~500 PCs

~2.5-10 Gbps

Institute

Physics data cache

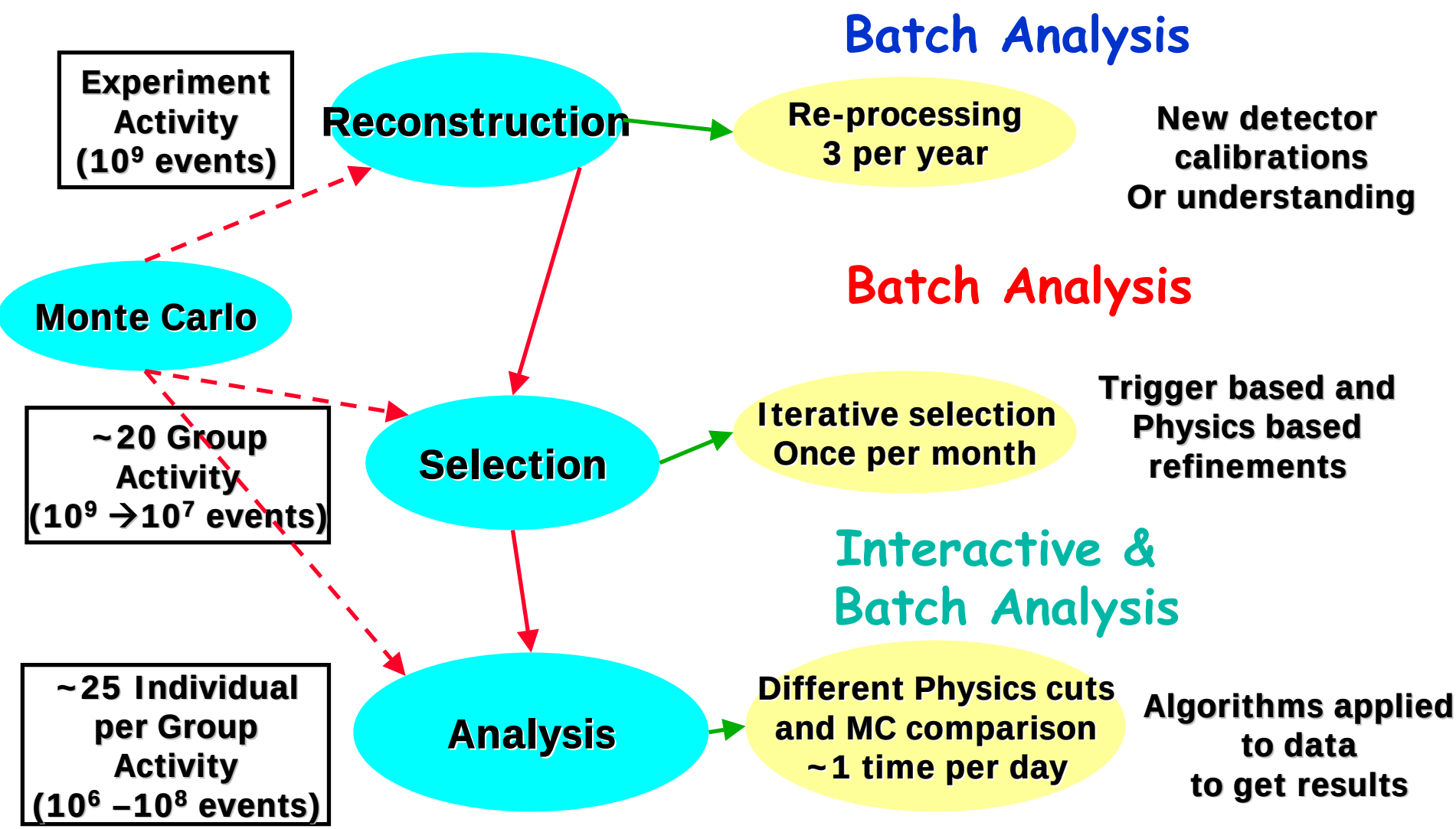
0.1 to 10 Gbps

Workstations

Tier 4

Tens of Petabytes by 2007-8
An Exabyte ~5-7 Years later

Gerarchia di Processi (MONARC)
(Esperimento, Gruppi di Analisi, Individui)



L'organizzazione dell'Infrastruttura

- Tiers e loro dipendenza (quali, quanti, chi fa cosa, quante risorse etc.)
  - Dedicati all'esperimento?
  - Comuni?
- Servizi necessari (databases, supporto sistemistico e agli utenti, accesso e localizzazione dei dati, prestazioni, etc.):
  - SLAs (service level agreements)
  - Personale
  - Priorità/politiche di accesso-autorizzazione
- Software (di ricostruzione, di analisi, di framework, di controllo, di accounting, di autorizzazione, di accesso, etc.)
- Cosa e' comune e cosa no:
  - LCG (contiene le applicazioni=software!)
  - Grid(s)

→ "Sistema" dinamico!

Data Challenges

- **Test dei Computing Model**
- **Preparazione alla Analisi**

- **Verifica progressiva della maturita' di:**
  - **Software**
  - **Infrastruttura**
- **Physics o Data Challenges?**
  - **Entrambi! per tutti gli esperimenti LHC, in tempi e modalita' diverse**
  - **I "challenges" correnti (tutti ne hanno gia' fatti negli anni scorsi):**
    - → ATLAS: DC2 (2004)
    - → CMS: DC04 (2003-2004)
    - → ALICE: PDC04 (2004)
    - → LHCb: DC'04 (2004)

Argomenti "comuni" nei Test dei Computing Models: DCs

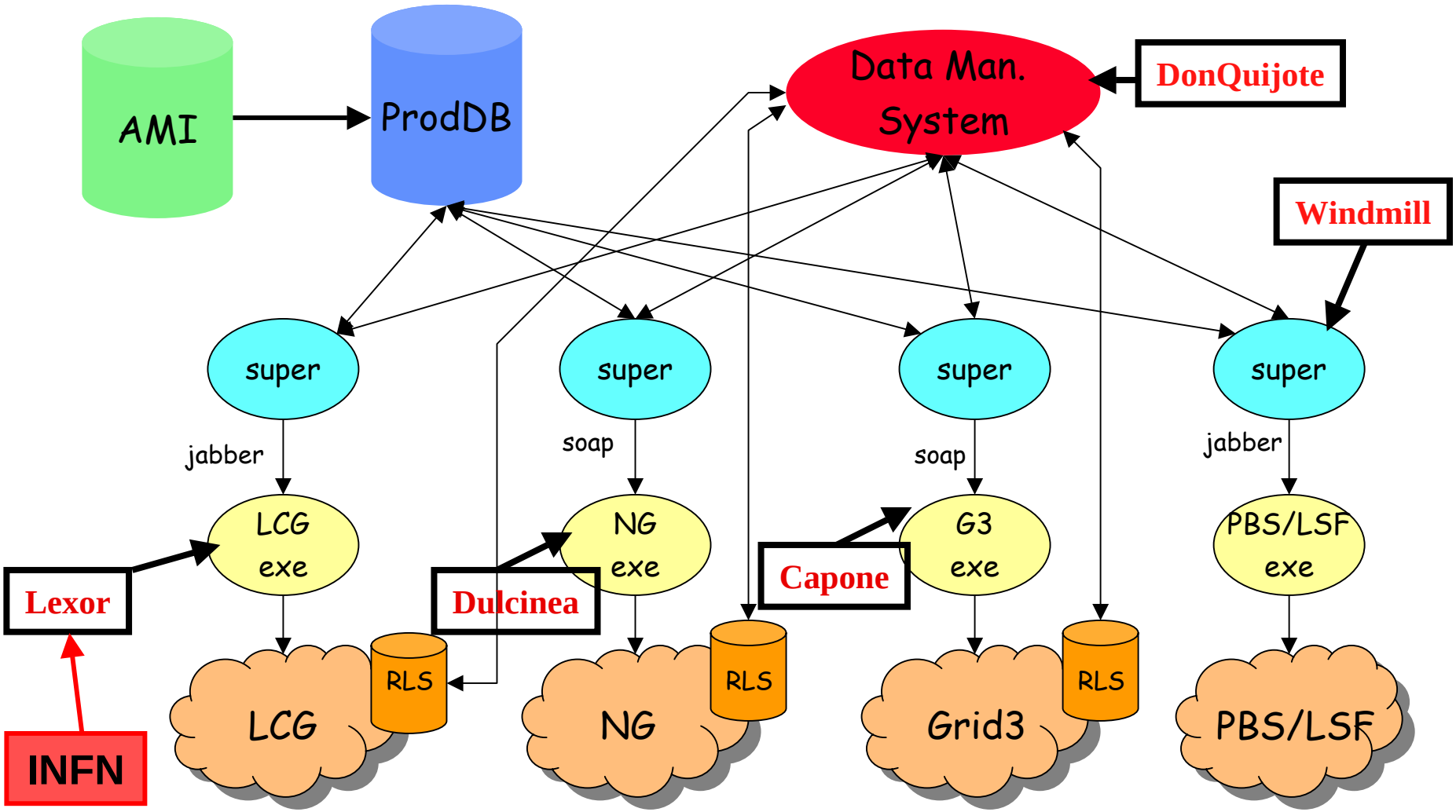
- Move a copy of the raw data away from CERN in "real-time"
  - Second secure copy
    - → 1 copy at CERN
    - → 1 copy spread over N sites
  - Flexibility.
    - → Serve raw data even if Tier-0 saturated with DAQ
  - Ability to run even primary reconstruction offsite
- Streaming online and offline
  - (Maybe not a common theme yet)
- Simulation at T2 centers
  - Except LHCb, if simulation load remains high, use Tier-1

- ESD Distributed n copies over N Tier-1 sites
  - Tier-2 centers run complex selections at Tier-1, download skims
- AOD Distributed to all (?) Tier-2 centers
  - Maybe not a common theme.
    - → How useful is AOD, how early in LHC?
    - → Some Run II experience indicating long term usage of "raw" data
- Horizontal Streaming
  - RAW, ESD, AOD,TAG
- Vertical Streaming
  - Trigger streams, Physics Streams, Analysis Skims

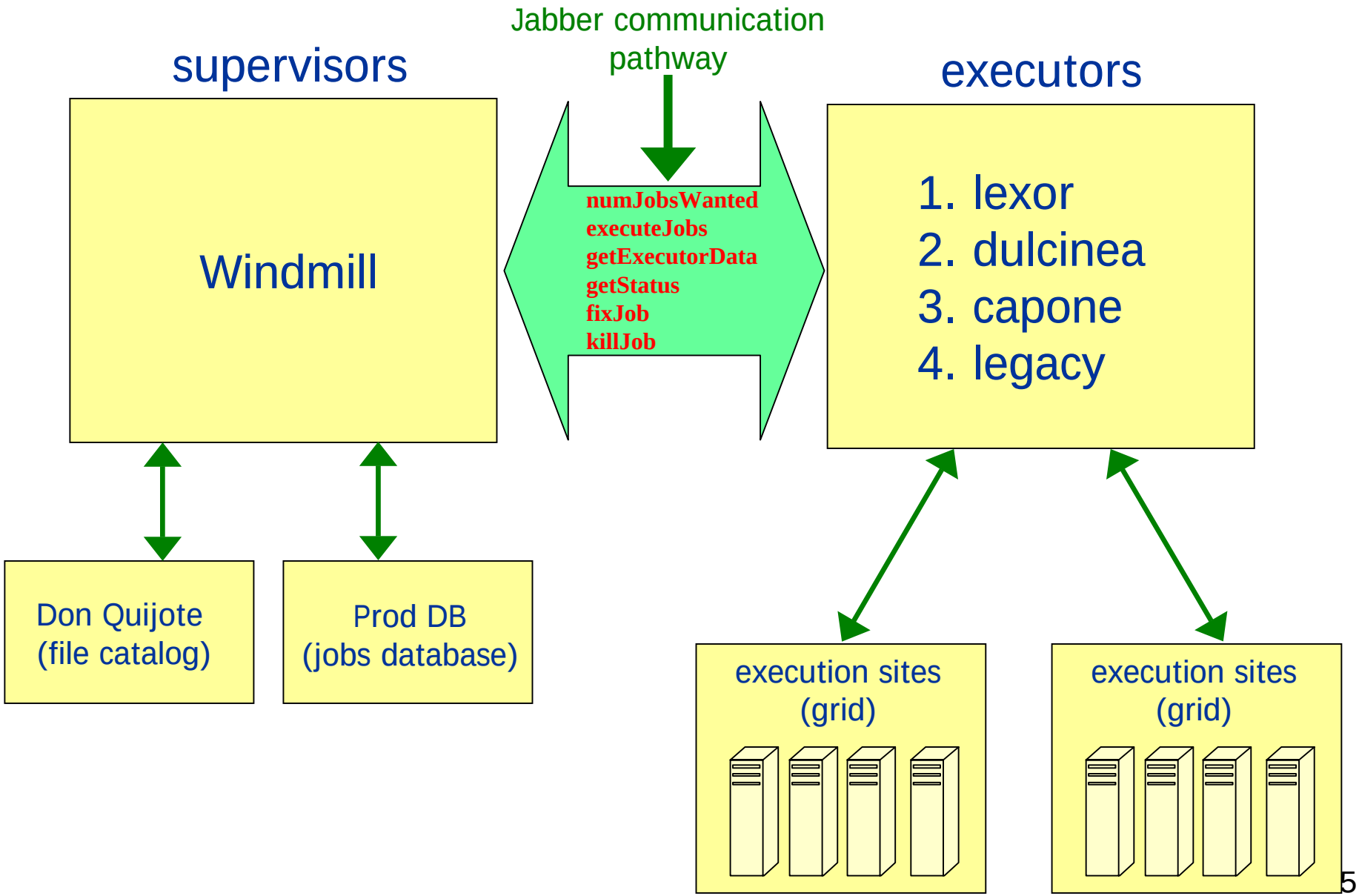
ATLAS DC2

- **Consider DC2 as a three-part operation:**
  - part I: production of simulated data (July-September 2004)
    - → running on "Grid"
    - → Worldwide
  - part II: test of Tier-0 operation (November 2004)
    - → Do in 10 days what "should" be done in 1 day when real data-taking start
    - → Input is "Raw Data" like
    - → output (ESD+AOD) will be distributed to Tier-1s in real time for analysis
  - part III: test of distributed analysis on the Grid
    - → access to event and non-event data from anywhere in the world both in organized and chaotic ways
- **Requests**
  - ~30 Physics channels (10 Millions of events)
  - Several millions of events for calibration (single particles and physics samples)

New ATLAS DC2 Production System

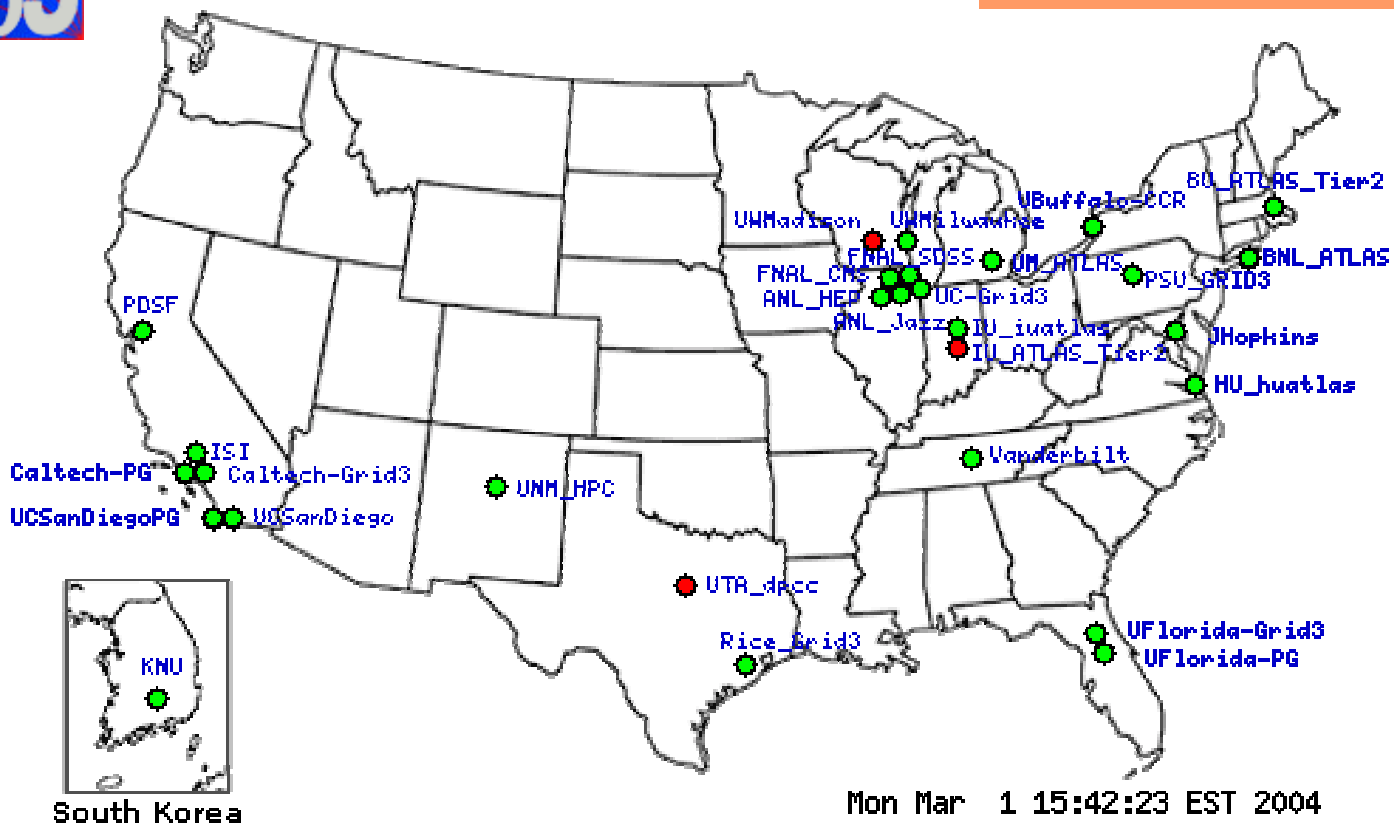


ATLAS DC2 Supervisor -Executors



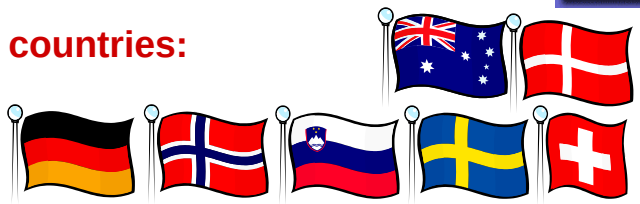
ATLAS Grid3 DC2 Sites

- 28 sites, multi-VO
- shared resources
- ~2000 CPUs
- dynamic – roll in/out



ATLAS DC2 NorduGrid & Co. Resources

◆**7 countries:**



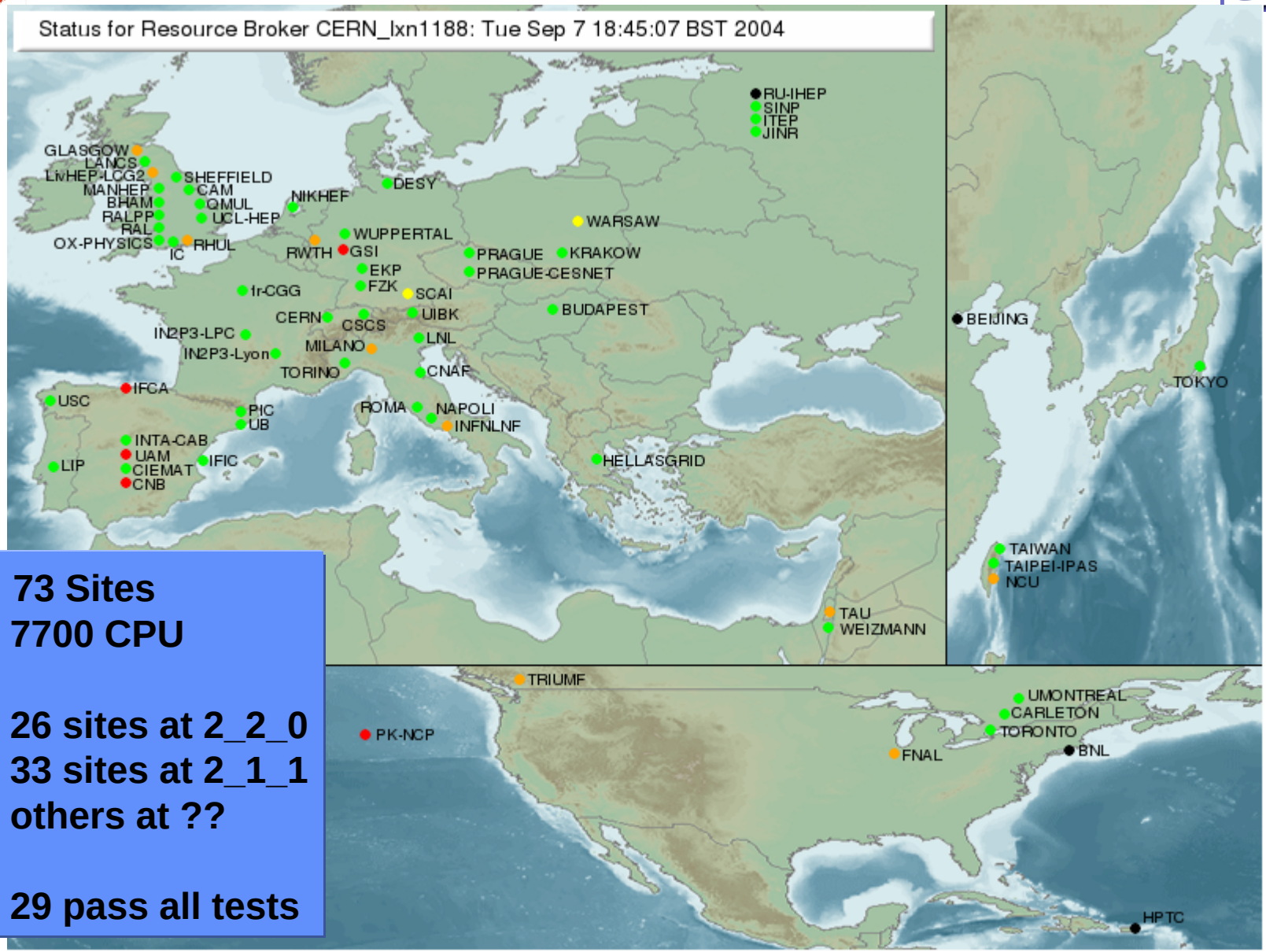
◆**Sites for ATLAS: 22**
- **Dedicated: 3, the rest is shared**

◆**CPUs for ATLAS:** ~3280
- **Effectively available:** ~800

◆**Storage Elements for ATLAS: 10**
- **Capacity:** ~14 TB**, all shared**

Cluster	running	waiting	finished	failed	(%)	total
atlas.hpc.unimelb.edu.au	28	86	641	20	(3%)	828
brenta.ijs.si	50	30	3200	217	(7%)	3562
bluesmoke.nsc.liu.se	48	70	1949	145	(7%)	2354
lxsrv9.lrz-muenchen.de	6	56	695	70	(10%)	1051
hypatia.uio.no	56	18	835	106	(13%)	1011
hagrid.it.uu.se			3550	508	(14%)	5325
benedict.aau.dk	46	41	2050	326	(16%)	2292
grid.uio.no	13	22	580	90	(16%)	726
sigrid.lunarc.lu.se	16	84	2542	441	(17%)	3510
sg-access.pdc.kth.se		58	2736	491	(18%)	2876
lheppc10.unibe.ch	12	14	455	82	(18%)	576
fire.ii.uib.no	10	12	838	163	(19%)	1073
farm.hep.lu.se	45	70	911	214	(23%)	1120
ingrid.hpc2n.umu.se	7		3507	886	(25%)	3774
fe10.dcsc.sdu.dk			1052	342	(33%)	1058
genghis.hpc.unimelb.edu.au		8	608	336	(55%)	653
morpheus.dcgc.dk	17	17	456	289	(63%)	490
charm.hpc.unimelb.edu.au			718	456	(64%)	916
atlas.fzk.de	15	23	77	52	(68%)	115
hive.unicc.chalmers.se			34	26	(76%)	34
lscf.nbi.dk	16	17	188	147	(78%)	221
grid.fi.uib.no			1	1	(100%)	1
TOTAL	385	626	27623	5408	(20%)	33566

ATLAS DC2 LCG-2 sites: 7/9/04

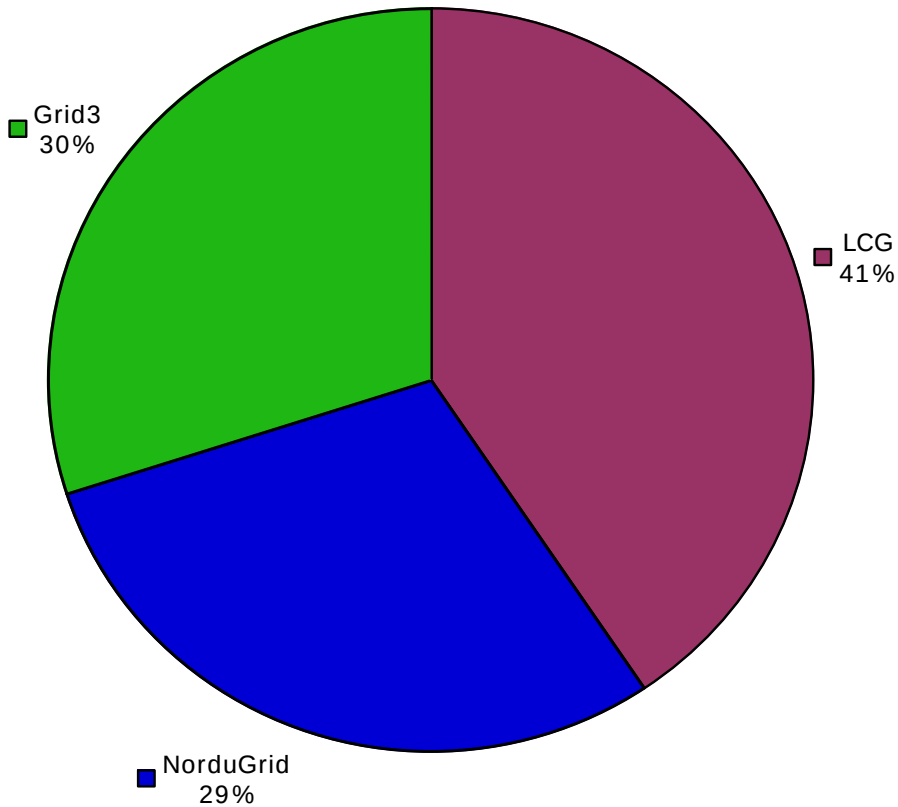


- 73 Sites
- 7700 CPU

- 26 sites at 2_2_0
- 33 sites at 2_1_1
- others at ??

- 29 pass all tests

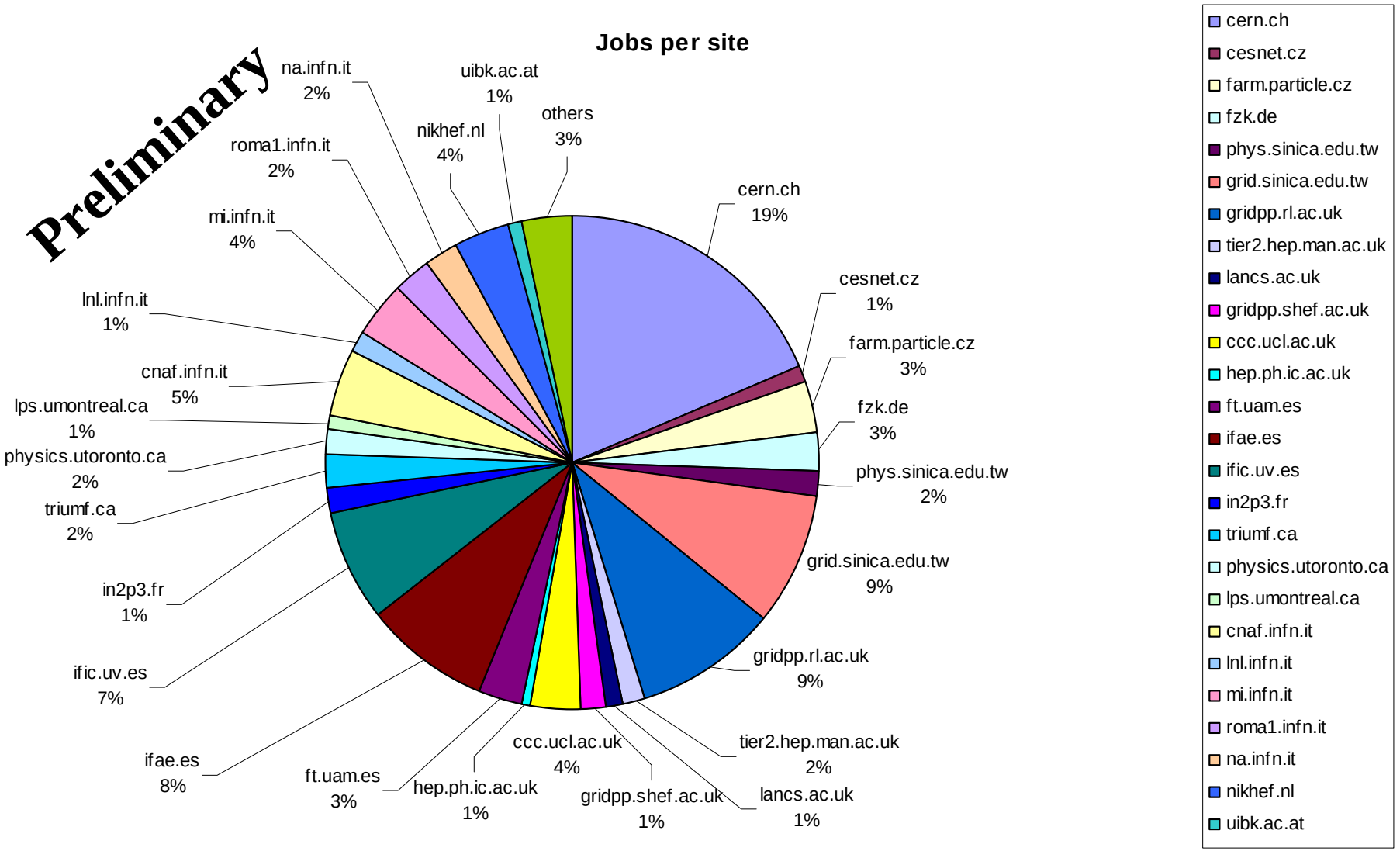
ATLAS DC2 status (CPU usage for simulation)



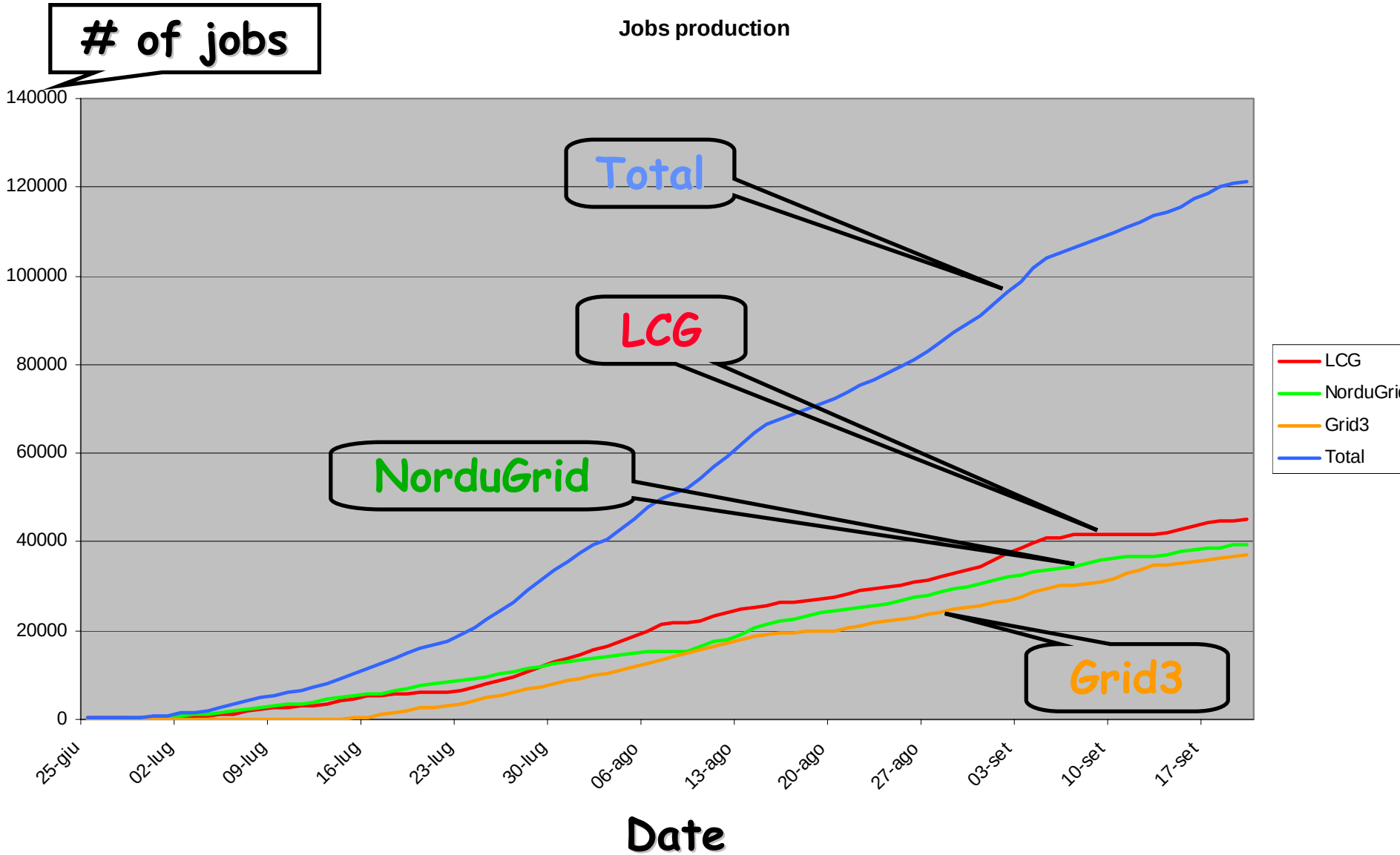
Total

~ 1470 kSI2k.months
~ 100000 jobs
~ 7.94 million events (fully simulated
~ 30 TB

Jobs distribution on LCG for ATLAS DC2



ATLAS DC2 Production Jobs



ATLAS DC2: Statistiche e problemi LCG

- 8 M eventi prodotti con Geant4
  - 100 k jobs da 24 ore circa
  - 30TB di output e 1470 kSpI2k*months
- LCG
  - Sommario quantitativo dei problemi trovati da 1-8 a 7-9 2004
  - 750 jobs falliti per misconfigurazione siti (E1)
  - 1985 " per WLMS e servizi collegati (E2)
  - 4350 " per Data Man. e servizi collegati (E3)

Jobs finiti bene nello stesso periodo 29303 (OK)

- Efficienza LCG = OK/(OK+E1+E2+E3)= **81%**

Ma l'efficienza globale è più bassa, ci sono stati problemi anche nella parte ATLAS (circa 7000 non molto diverso da LCG) e circa 5000 di difficile assegnazione

- Efficienza DC2(parte LCG)=OK/(OK+FAILED)= **62%**

Atlas DC2 sommario

- **Major efforts in the past few months**
  - Redesign of the ATLAS Event Data Model and Detector Description
  - Integration of the LCG components (G4; POOL; …)
  - Introduction of the Production System
    - → Interfaced with 3 Grid flavors (and "legacy" systems)
- **Delays in all activities have affected the schedule of DC2**
  - Note that Combined Test Beam is ATLAS 1st priority
  - And DC2 schedule was revisited
    - → To wait for the readiness of the software and of the Production system

- **DC2**
  - About 80% of the Geant4 simulation foreseen for Phase I has been completed using only Grid and using the 3 flavors coherently;
  - The 3 Grids have been proven to be usable for a real production
- **BUT**
  - Phase I progressing slower than expected and all the involved elements need improvements:
    - → Grid middleware; Production System; deployment and monitoring tools over the sites
  - It is a key goal of the Data Challenges to identify these problems as early as possible.

Scopo del Data Challenge 04 di CMS

Aim of DC04:

- Reach a sustained 25Hz reconstruction rate in the Tier-0 farm (25% of the target conditions for LHC startup)
- Register data and metadata to a catalogue
- Transfer the reconstructed data to all Tier-1 centers
- Analyze the reconstructed data at the Tier-1's as they arrive
- Publicize to the community the data produced at Tier-1's
- Monitor and archive of performance criteria of the ensemble of activities for debugging and post-mortem analysis

Not a CPU challenge, but a full chain demonstration!

Pre-challenge production in 2003/04

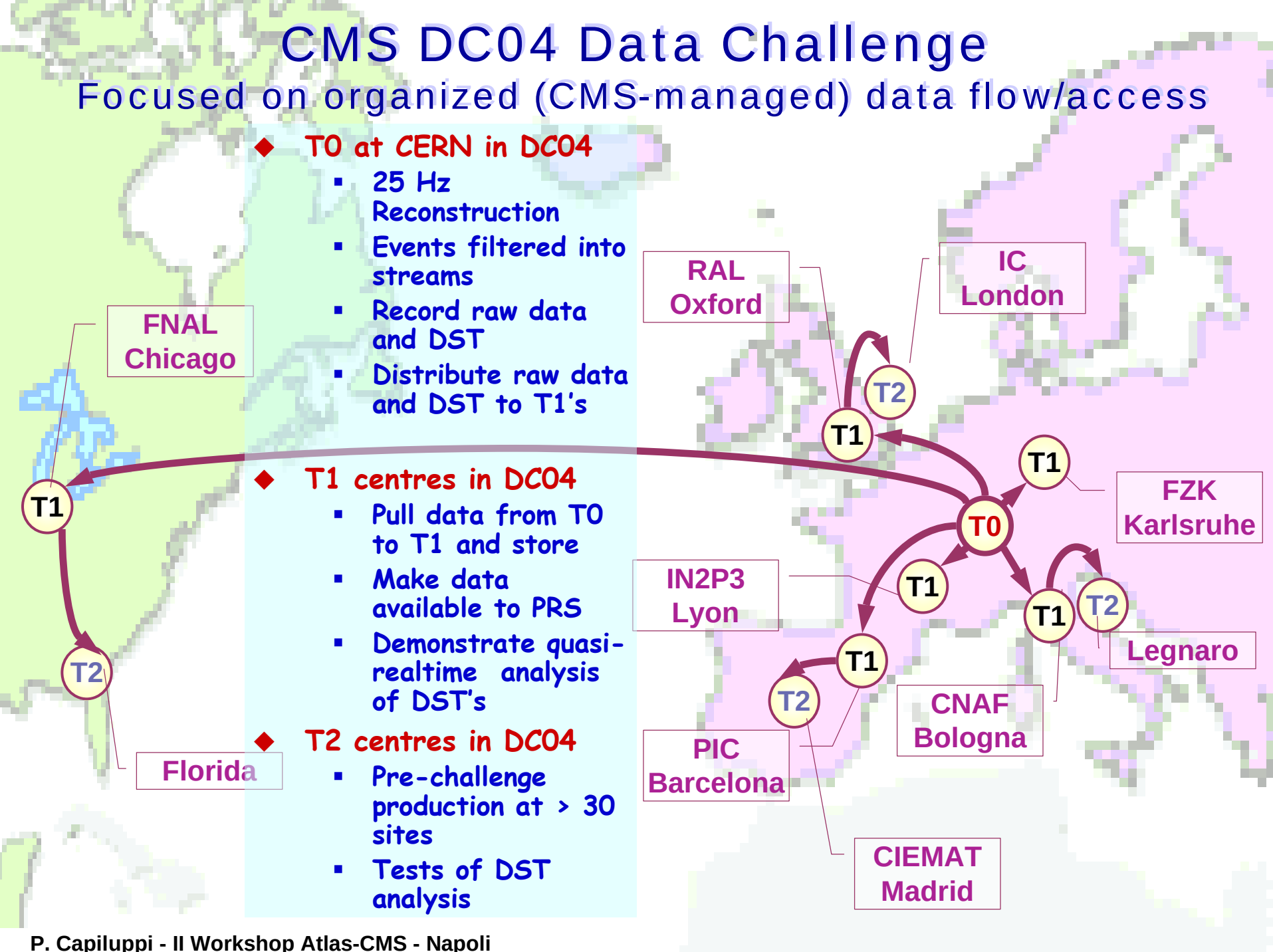
- 70M Monte Carlo events (30M with Geant-4) produced
- Classic and grid (CMS/LCG-0, LCG-1, Grid3) productions

Era un "challenge", e ogni volta che si e' trovato un limite di scalabilita' di una componente, e' stato un Successo!

CMS DC04 Data Challenge

Focused on organized (CMS-managed) data flow/access

- **T0 at CERN in DC04**
  - 25 Hz Reconstruction
  - Events filtered into streams
  - Record raw data and DST
  - Distribute raw data and DST to T1's
- **T1 centres in DC04**
  - Pull data from T0 to T1 and store
  - Make data available to PRS
  - Demonstrate quasi-realtime analysis of DST's
- **T2 centres in DC04**
  - Pre-challenge production at > 30 sites
  - Tests of DST analysis



CMS Data Challenge 04: numbers

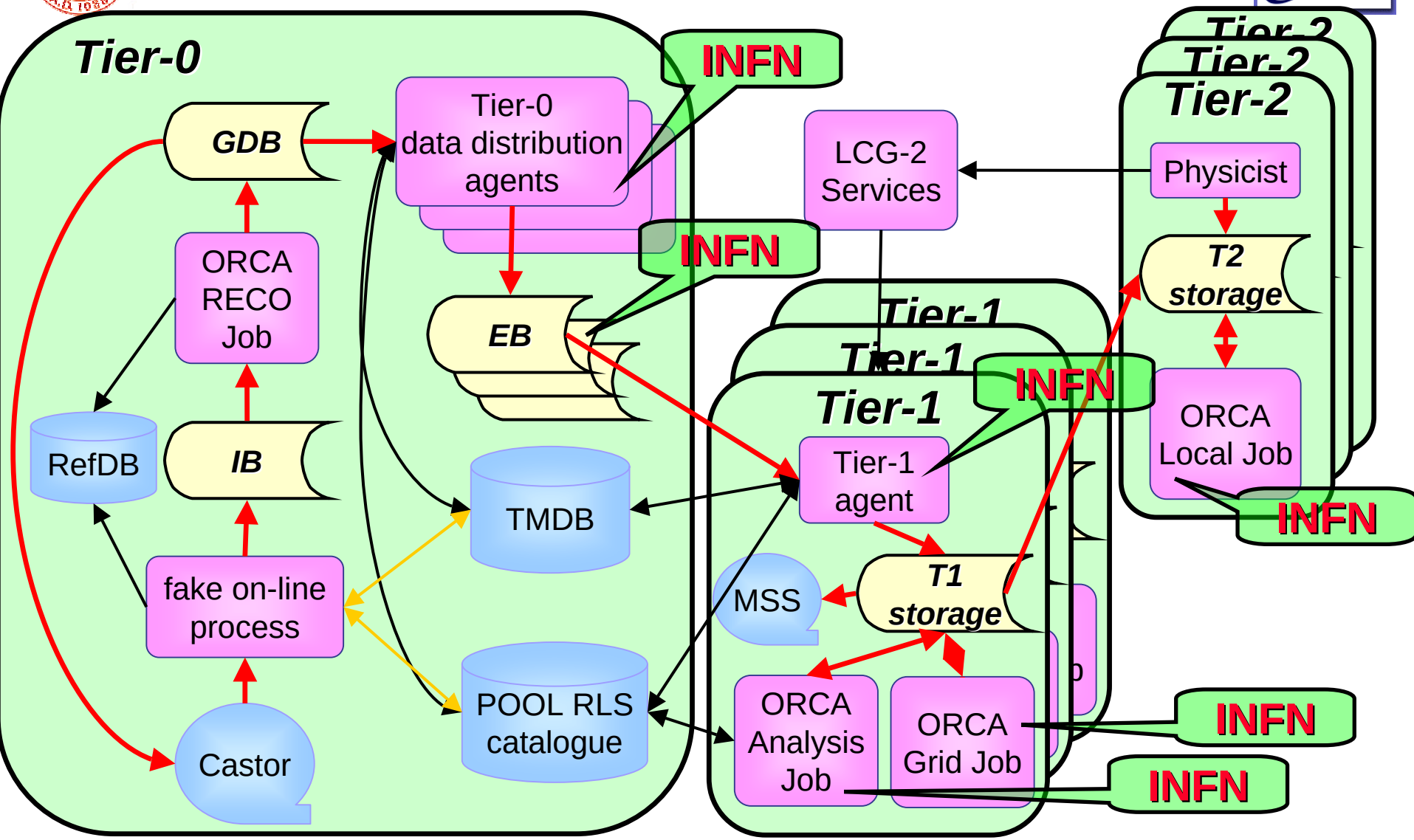
◆Pre Challenge Production (PCP04) [Jul03-Feb04]

- Eventi simulati : 75 M events [750k jobs, ~800k files, 5000 KSI2000 months, 100 TB of data] *(~30 M Geant4)*
- Eventi digitizzati (raw): 35 M events [35k jobs, 105k files]
- Dove: INFN, USA, CERN, …
- In Italia: ~ 10-15 M events *(~20%)*
- Per cosa (Physics and Reconstruction Software Groups): "Muons", B-tau", "e-gamma", "Higgs"

◆Data Challenge 04 [Mar04-Apr04]

- Eventi ricostruiti (DST) al Tier0 del CERN: ~25 M events [~25k jobs, ~400k files, 150 KSI2000 months, 6 TB of data]
- Eventi distribuiti al Tier1-CNAF e Tier2-LNL: gli stessi ~25 M events e files
- Eventi analizzati al Tier1-CNAF e Tier2-LNL: > 10 M events [~15 k jobs, ognuno di ~ 30min CPU]

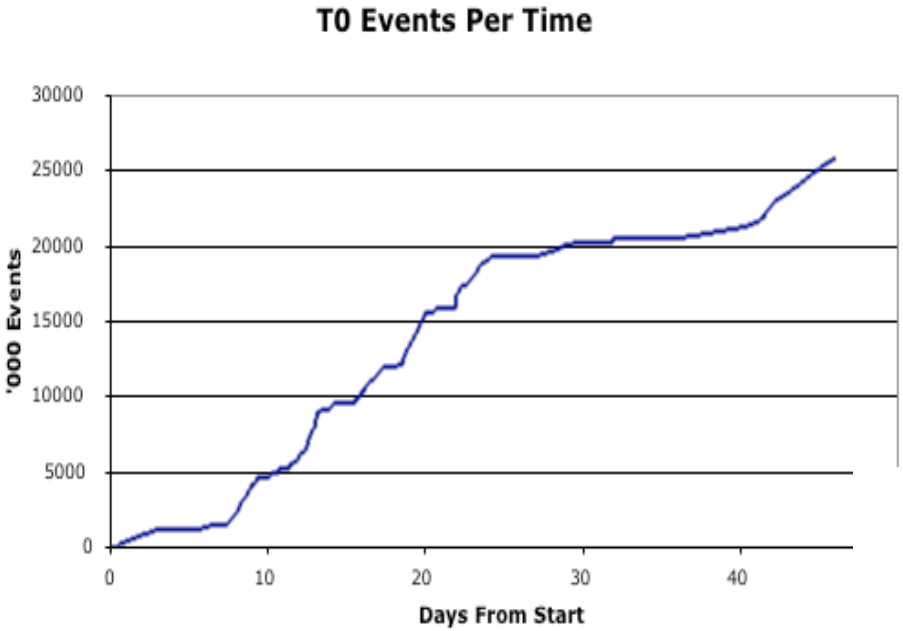
CMS Data Challenge 04: layout



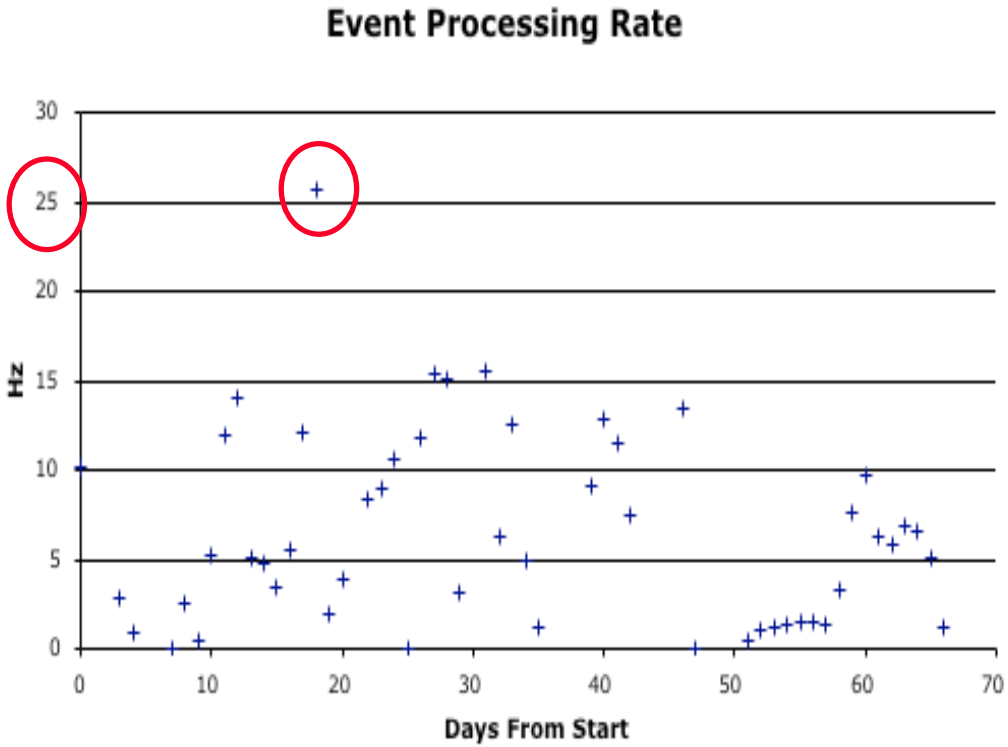
Full chain (but the Tier-0 reconstruction) done in LCG-2, but only for INFN and PIC

Not without pain…

CMS Data Challenge 04 Processing Rate

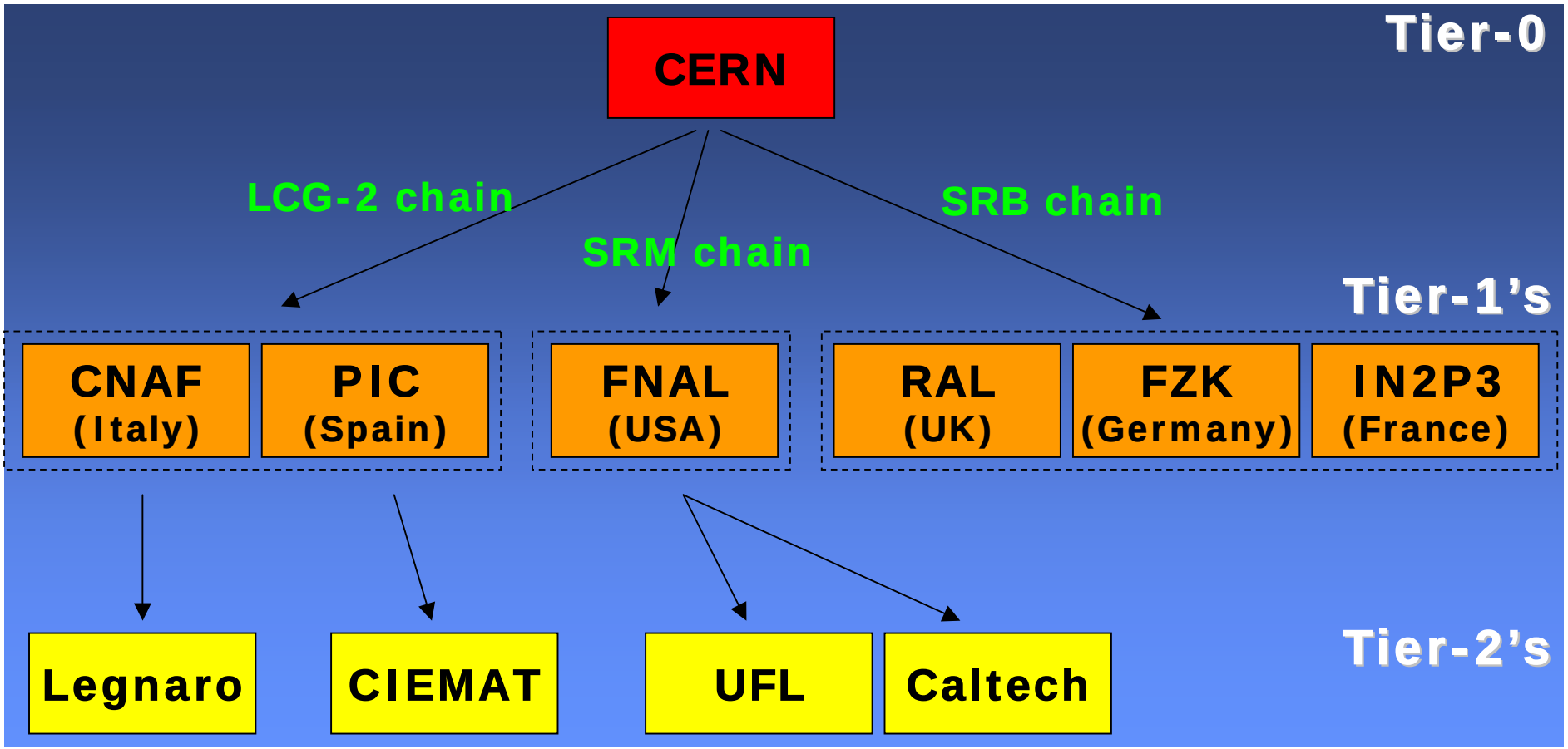


- Processed about 30M events
  - But DST "errors" make this pass not useful for analysis
  - Post-DC04 3rd version ready for production in next weeks
- Generally kept up at T1's in **CNAF**, FNAL, PIC



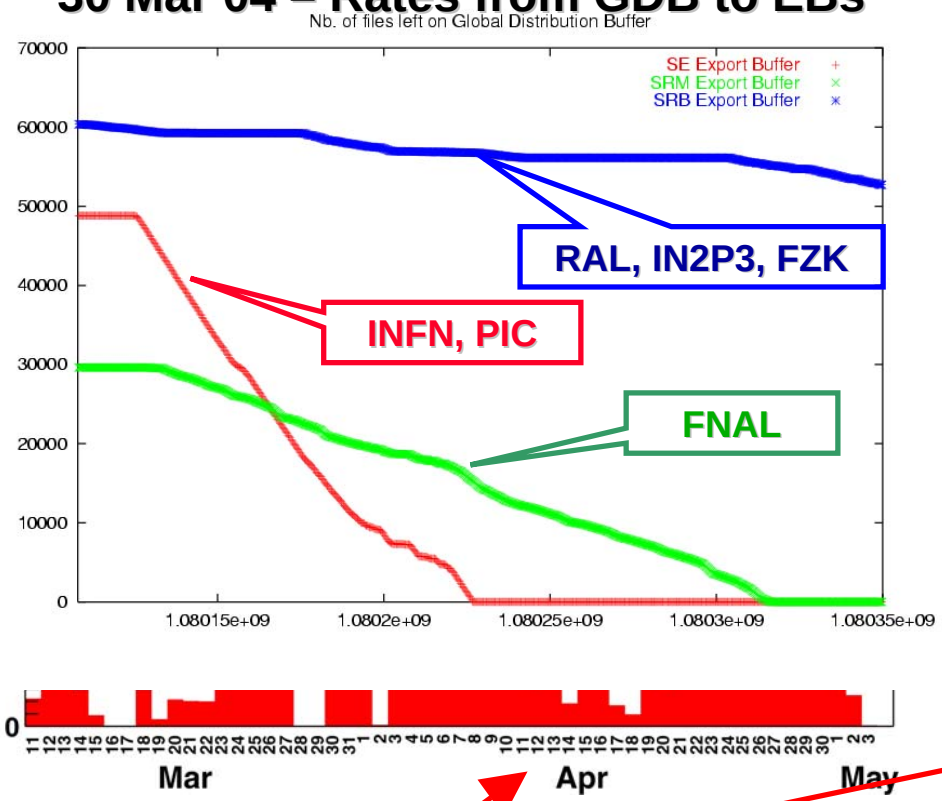
- Got above 25Hz on many short occasions
  - But only one full day above 25Hz with full system
- RLS, Castor, overloaded control systems, T1 Storage Elements, T1 MSS, ...

Hierarchy of Tiers in CMS DC04 and data distribution chains



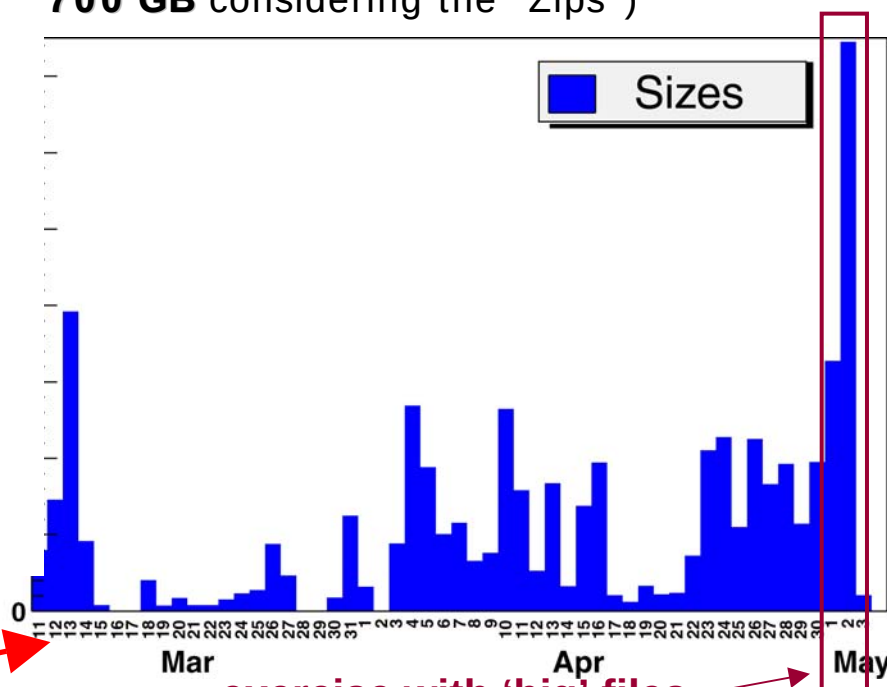
CMS Data Challenge 04: data transfer from CERN to INFN

30 Mar 04 – Rates from GDB to EBs



transferred CERN T0 → CNAF T1

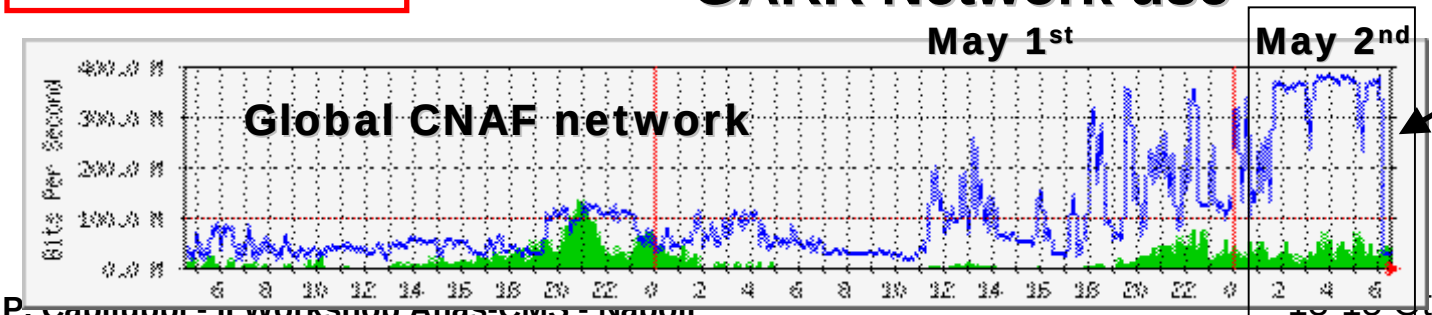
700 GB considering the "Zips")



exercise with 'big' files

CNAF - Tier1

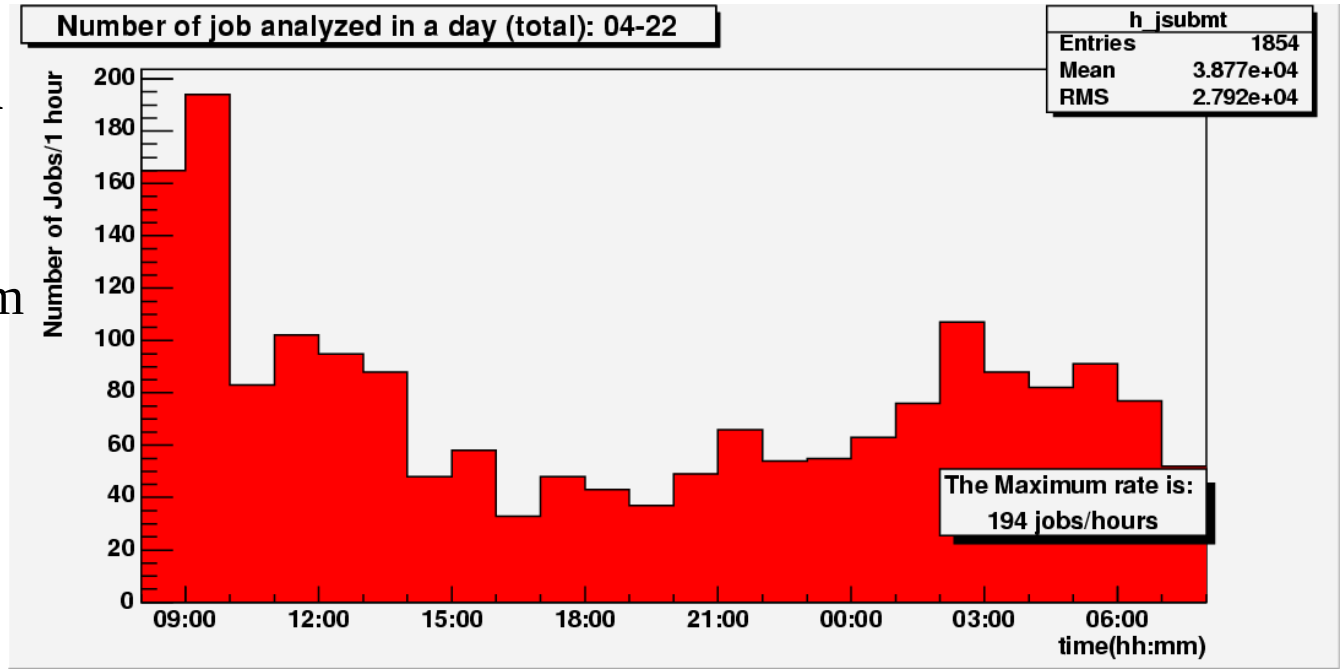
GARR Network use



~340 Mbps (>42 MB/s) sustained for ~5 hours (max was **383.8 Mbps**)

CMS DC04 Real-time Analysis on LCG

- Maximum rate of analysis jobs: **194 jobs/hour**
- Maximum rate of analysed events: **26 Hz**
- Total of **~15000** analysis jobs via **Grid** tools in ~2 weeks **(95-99% efficiency)**
- **20 min latency** from delivery by Tier0 to start of Analysis Job at the Tier1/2



➢ Datasets examples:

- $B^0_S \rightarrow J/\psi\ \varphi$
  Bkg: mu03_tt2mu, mu03_DY2mu
- **tTH, H → bbbar t→ Wb W → lν T → Wb W → had.**
  Bkg: bt03_ttbb_tth
  Bkg: bt03_qcd170_tth
  Bkg: mu03_W1mu
- **H → WW → 2μ 2ν**
  Bkg: mu03_tt2mu, mu03_DY2mu

CMS Data Challenge 04: componenti MW e SW, un esempio

◆CMS specific

- Transfer Agents per trasferire i files di DST (al CERN, ai Tier1)
- RefDb, Database delle richieste e "assignment" di datasets (al CERN)
- Cobra, framework del software di CMS (CMS wide)
- ORCA, OSCAR (Geant4), ricostruzione e simulazione di CMS (CMS wide)
- McRunJob, sistema per preparazione dei job (CMS wide)
- BOSS, sistema per il job tracking (CMS wide)
- SRB, sistema di replicazione e catalogo di files (al CERN, a RAL, Lyon e FZK)
- MySQL-POOL, backend di POOL sul database MySQL (a FNAL)

◆US specific

- Monte carlo distributed prod system (MOP) (a FNAL, Wisconsin, Florida, …)
- MonaLisa, sistema di monitoring (CMS wide)
- Custom McRunJob, sistema di preparazione dei job (a FNAL e…forse Florida)

◆LCG "common"

- User Interfaces including Replica Manager (al CNAF, Padova, LNL, Bari, PIC)
- Storage Elements (al CNAF, LNL, PIC)
- Computing Elements (al CNAF, a LNL e a PIC)
- Replica Location Service (al CERN e al Tier1-CNAF)
- Resource Broker (al CERN e al CNAF-Tier1-Grid-it)
- Storage Replica Manager (al CERN e a FNAL)
- Berkley Database Information Index (al CERN)
- Virtual Organization Management System (al CERN)
- GridICE, sistema di monitoring (sui CE, SE, WN, …)
- POOL, catalogo per la persistenza (in CERN RLS)
- Mass Storage Systems su nastro (Castor, Enstore, etc.) (al CERN ai Tier1)
- ORACLE database (al CERN e al Tier1-INFN)

CMS DC04 Sommario

◆Concentrated on the Organized, Collaboration-Managed, aspects of Data Flow and Access

- Functional DST with streams for Physics and Calibration
  - → DST size OK; further development now underway
- Tier-0 farm reconstruction
  - → 500 CPU. Ran at 25Hz. Reconstruction time within estimates.
- Tier-0 Buffer Management and Distribution to Tier-1's
  - → TMDB- CMS built Agent system OK
- Tier-2 Managed Import of Selected Data from Tier-1
  - → Meta-data based selection OK.
- Real-Time analysis access at Tier-1 and Tier-2
  - → Achieved 20 minute latency from T0 reconstruction to job launch at T1/T2
- Catalog Services, Replica Management
  - → Significant performance problems found and being addressed?!

◆Demonstrated that the system can work for well controlled data flow and analysis, and for a few expert users

- Next challenge is to make this useable by average physicists and demonstrate that the performance scales acceptably

◆BUT: Physics TDR requires physicist access to DC04 data !

- Re-reconstruction passes
- Alignment studies
- Luminosity effects
  - → Estimate 10M events/month throughput required

◆Therefore use requirements of Physics TDR to build understanding of analysis model, while doing the analysis

- Make it work for Physics TDR

Cosa manca?

- Organizzazione!
- Definire le attivita' di Computing dei primi 100 giorni
  - Calibrazioni/allineamenti
  - Trigger rate e suo uso
  - Etc.
- Dimostrare la scalabilita' del "SISTEMA"
- Valutare l'impatto dell'Analisi e delle calibrazioni
  - Incluso il Condition database

- In fin dei conti: Manca un (sia pur preliminare) Computing Model completo che possa essere misurato
  - Nelle prestazioni
  - Scalabilita'
  - Affidabilita'
  - Facilita' di uso nell'accesso ai dati
  - (Il software e il middleware NON sono il problema)

Conclusioni

- **Ma non siamo messi cosi' male**
  - Sia ATLAS che CMS hanno misurato alcune delle componenti essenziali (alcune purtroppo ancora mancano)
  - Sia ATLAS che CMS stanno andando verso un "sistema continuo" di produzione ed analisi
  - Entro il 2005 avremo i Computing TDR (e quindi un Computing Model)
  - Da quest'ultimi nasceranno (stanno nascendo…) i Computing MoUs
- **I prossimi Data Challenges saranno quelli "finali" prima della presa dati reali: DC3-ATLAS, DC06-CMS, entrambi nel 2006**

Quelli successivi saranno sui dati veri !