Il calcolo distribuito in ATLAS

Alessandra Doria
INFN Sezione di Napoli

Workshop CCR2003

Paestum 12 Giugno 2003

Argomenti

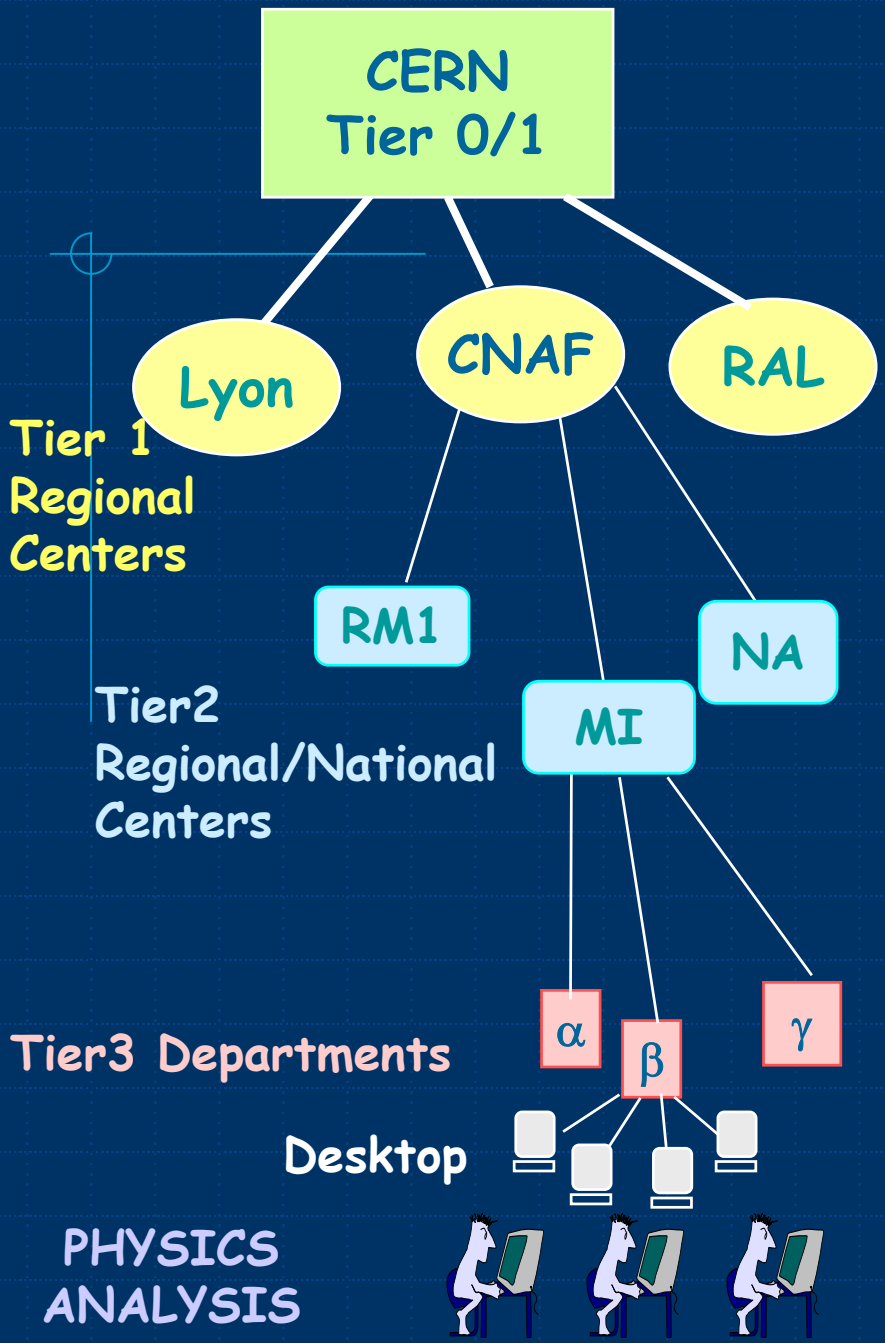
- Modello di calcolo a Tier
- Il sw offline di ATLAS
- ATLAS Data Challenges
- Progetti GRID coinvolti nei DC di ATLAS
- Grid Tools utilizzati
- Piani per i futuri sviluppi

Il modello di calcolo a Tier

- Il modello generale di calcolo per l'offline e l'analisi di ATLAS è quello gerarchico multi-tier adottato dagli esperimenti LHC.
- Il Tier0 risiede al CERN, dove i dati vengono prodotti.
- Tier1 e Tier2 hanno carattere rispettivamente Regionale e Nazionale (dove per Regione si intende una comunità transnazionale) mentre i Tier3 sono locali, a livello di dipartimento o di istituto.

Il modello di calcolo a Tier

- Distinzione tra Tier di livello diverso:
  - nei servizi offerti all'utenza
  - Nella visibilità dall'esterno
  - nel tipo di dati da mantenere localmente.
- Si stima che per l'analisi, nel 2007-2008, le risorse al CERN dovranno soddisfare:
  - CPU 0.4 MSpInt95
  - Tape 6.7 PB
  - Disk 0.5 PB
- Le risorse aggregate di tutti i Tier1 (in ATLAS tra 5 e 10) dovranno essere comparabili a quelle CERN.
- Un Tier2 avrà risorse tra il 5% e il 25% del Tier1.



Tipo di dati da produrre e conservare:

RAW DATA: 2 MB/evento

MC RAW DATA: 2 MB/evento

ESD, Event Summary Data, output ricostruzione:

500 KB/evento

AOD, Analysis Object Data, formato "pubblico" di analisi:

10 KB/evento (MC simulation, Physics Analysis)

DPD, Derived Physics Data, formato "privato" di analisi:

1 KB/evento (Physics Analysis)

Il sw offline di ATLAS

- Tipi diversi di tasks:
  - Simulazione MonteCarlo: CPU intensive
  - Pile-up(sovrapposizione fondo): data intensive
  - Ricostruzione: data intensive, multiple passes
  - Analisi: interattivo, accesso ai dati e carico di lavoro poco prevedibili.
- Atlsim, Dice: applicazioni Fortran/Geant3
- Athena/Gaudi (in collaborazione con Lhcb) : framework OO che fornisce una infrastruttura comune per applicazioni di simulazione, ricostruzione, analisi
- Atlas ha scelto di sviluppare e mantenere il sw di esperimento indipendente da GRID
- Diversi tools fanno da interfaccia verso la griglia

Scopo dei Data Challenges

- Test della catena di sw specifico dell'esperimento.
- Produzione di un campione di 10^7 eventi simulati per studi di High Level Trigger.
- Integrazione progressiva del sw di esperimento con il middleware di Grid.
- Validazione generale del modello di calcolo adottato.

Fasi dei DC

- DC0 - primi mesi 2002 - al CERN. Test della catena di sw (non OO), produzione di piccoli campioni di eventi.
- DC1 - Apr 2002-Maggio 2003 - lavoro distribuito . Simulazioni ancora Fortran/Geant3, sovrapposizione pileup, ricostruzione in Athena.
- DC2 - Luglio 2003-Luglio 2004 - Validazione dell' *Event data model* e di Geant4, POOL, Simulazioni in Athena, ricostruzione completa, uso di LCG-1 *as much as possible*. Richiesta 2-3 volte la CPU utilizzata nel DC1.

Risorse utilizzate nel DC1

- Simulazione:
  - 39 istituti in 18 paesi, con 3200 CPU (110 KSpecInt95)
  - 10^7 eventi completi e $3*10^7$ particelle singole , 30 TB output, 71000 CPU days (PIII, 500MHz)
- Pileup:
  - 56 istituti in 21 paesi
  - $3*10^6$ eventi, 25 TB , 10000 CPU days
- Ricostruzione:
  - solo nei centri con ~100 CPU
  - ricostruiti 10^6 eventi.

Il DC1 in Italia

- Alle simulazioni hanno preso parte il CNAF (40 CPU) e le sezioni di Roma1(46), Milano (20), Napoli (16) e Frascati(10) (+ contributi di Genova e Pisa) .
- Il contributo italiano è circa il 5% del totale.
- Tutti i dati simulati e con pileup sono raccolti al Tier1 al CNAF, $2*10^6$ eventi completi, 5 TB
- Per la ricostruzione al CNAF sono state usate 70 CPU.

Data Challange e progetti GRID

- Le produzioni del DC1 sono state fatte in parte con metodi “tradizionali” (sottomissione tramite script con sistema batch PBS)
- ATLAS ha già usato diverse Grid nella produzione, sia per la fase di simulazione che di ricostruzione :
  - NorduGrid
  - US Grid
  - EDG
- Atlas intende utilizzare LCG-1 il pù possibile, appena sarà disponibile, senza tuttavia abbandonare subito l’uso delle altre Grid.

Nordugrid

- Progetto di 9 università e istituti dei paesi del Nord Europa
  - 4 cluster di test dedicati (3-4 CPU)
  - alcuni cluster universitari (20-60 CPU)
  - 2 grandi cluster in Svezia (200-300 CPU)
- Tutti i servizi sono presi da Globus o scritti usando librerie e API di Globus
- Topologia *mesh* (a maglia), senza *Resource Broker* centralizzato, ogni sito ha un *gatekeeper*
- I tools ed il middleware si sono dimostrati molto affidabili, per il DC1 non sono state usate risorse esterne a Nordugrid.

US ATLAS grid testbed

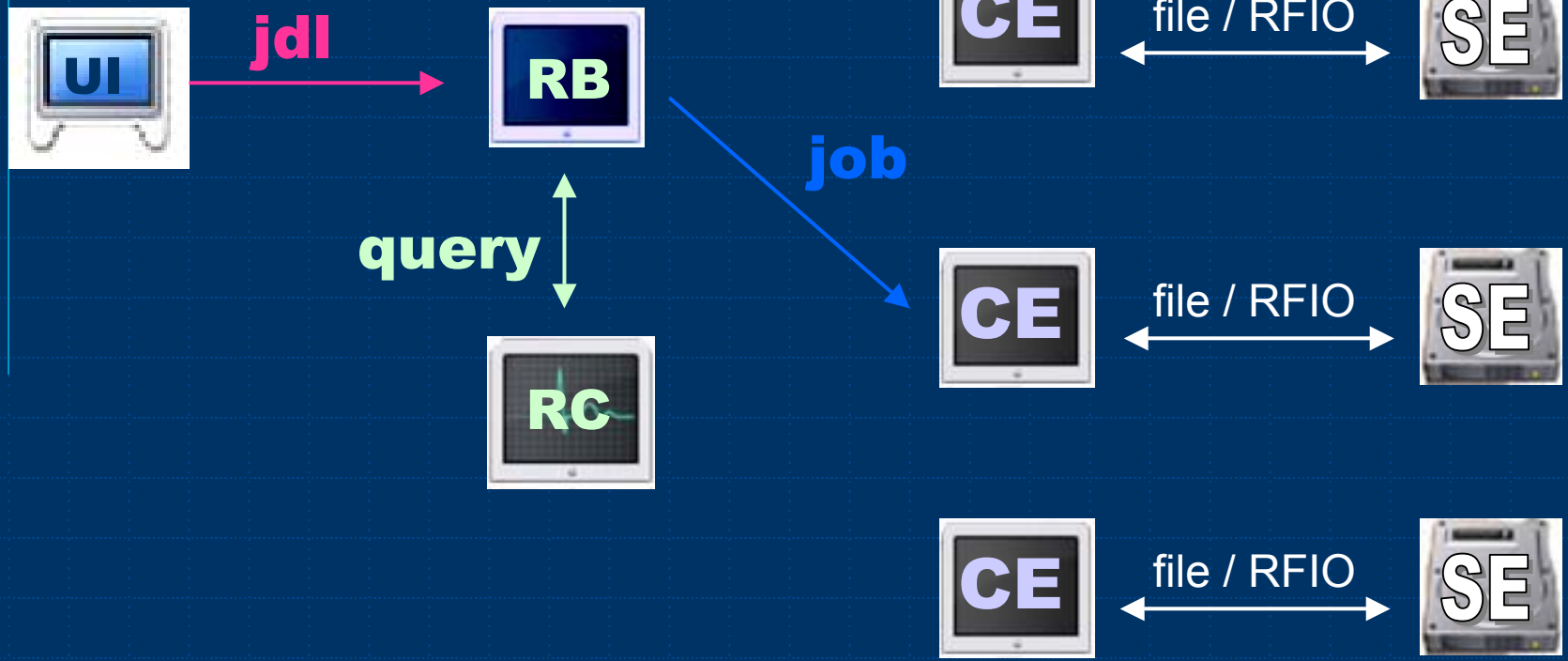
8 di siti US nel grid testbed. I principali sono:

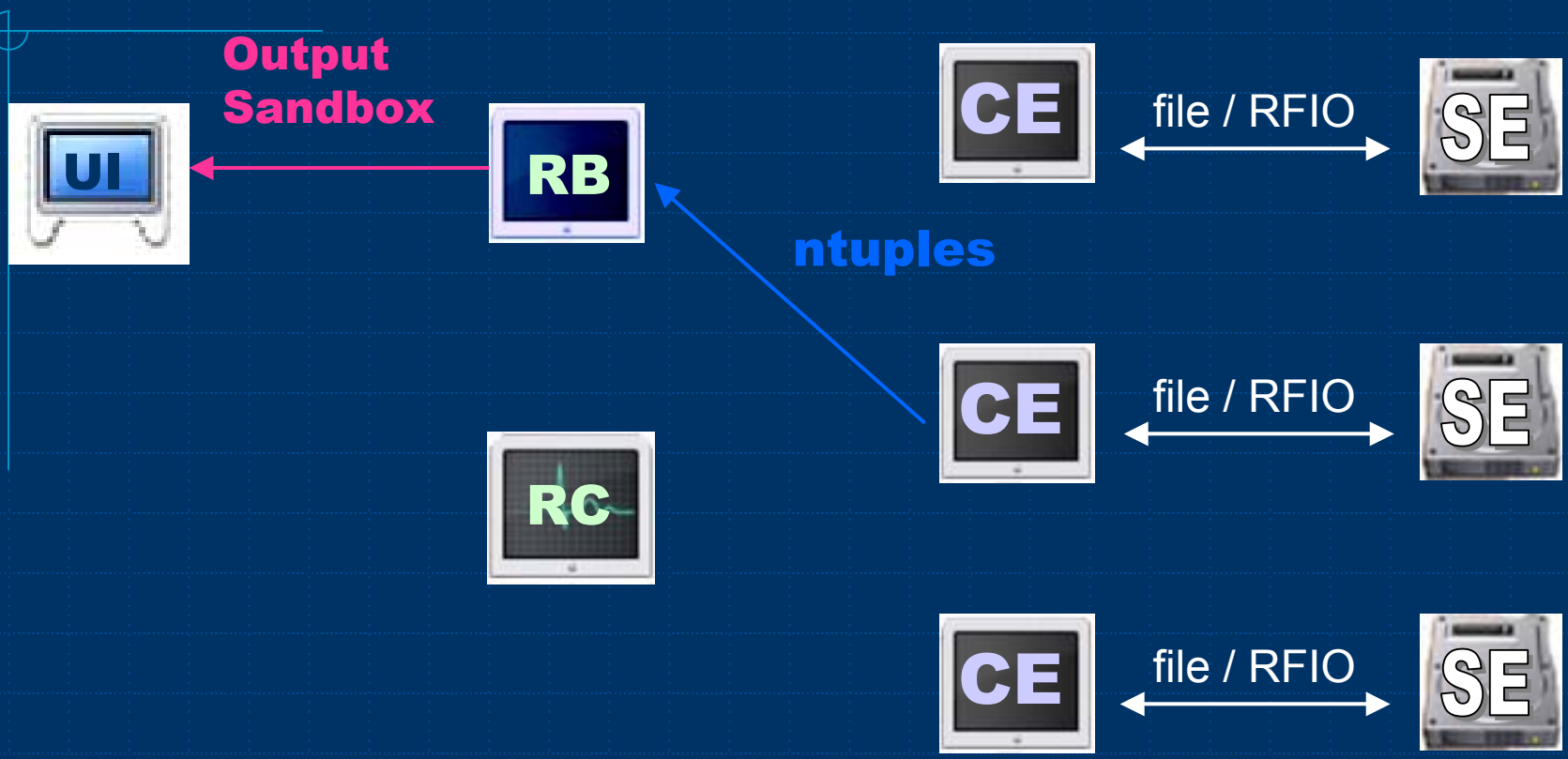
- BNL - U.S. Tier 1, 2000 nodes, 5% ATLAS (100), 10 TB
- LBNL - PDSF cluster, 400 nodes, 5% ATLAS (20) , 1 TB
- Boston U. - prototype Tier 2, 64 nodes
- Indiana U. - prototype Tier 2, 32 nodes

- Topologia *mesh*
- Parte del DC1 eseguito a Brookhaven in modo tradizionale
- Incontrati molti failure nelle produzioni con Grid (tuttavia 80% efficienza)
- Utilizzati in produzione GRAT, GRAPPA, Chimera...

Data Challenge in EDG

- In Aprile già fatti dei test con CNAF, RAL, Lione, in EDG 1.4.8 con la vecchia versione del SW di ATLAS su RH 6.2.
- In questo momento sta terminando la ricostruzione in EDG di campioni a bassa priorità per il DC1/EDG.
  - Non più test, ma vere produzioni (resp. Guido Negri, CNAF).
  - Utilizzata la nuova versione sw 6.0.3 che richiede RH 7.3 sui WN.
  - Siti coinvolti: CNAF, MI, Roma1,Cambridge, Lyon (Resource Broker del CNAF)
  - Informazioni in https://classis01.roma1.infn.it/grid/


UI
jdl
RB
query
RC
job
CE
file / RFIO
SE
CE
file / RFIO
SE
CE
file / RFIO
SE


Output
Sandbox
UI
RB
RC
CE
CE
CE
file / RFIO
file / RFIO
file / RFIO
SE
SE
SE
ntuples

DC1 in EDG

- Creati i profili LCFGng per installare WN con RH7.3
- Copiati input files (125000 eventi, 250 input files) da Castor al CERN (senza grid tools, scp-ftp) agli SE dei siti.
- Registrati input files nel Replica Catalogue
- Creati script di sottomissione dei job (site independent).
- Qualche problema dall' Information System (MDS)
- Tutto dovrebbe essere terminato per questa settimana.

DC1 Bookkeeping: AMI Atlas Metadata Interface

- In termini di grid → ***application meta data base***
- **Registra le informazioni che descrivono i contenuti logici dei data files. Query su criteri fisici.**
- Implementato in java e basato su database relazionale.
- C++ APIs per Athena, GUI, Web interface
- L'interfaccia web è collegata a Magda per ottenere la locazione effettiva dei dati a partire dai *logical file names*.
- Per DC1 sono stati registrati 60000 files (500 dataset)
- Attualmente 31 MB, previsto 1.2 GB nei prossimi 3 anni

DC1 Bookkeeping: MAGDA Manager for Grid-based Data

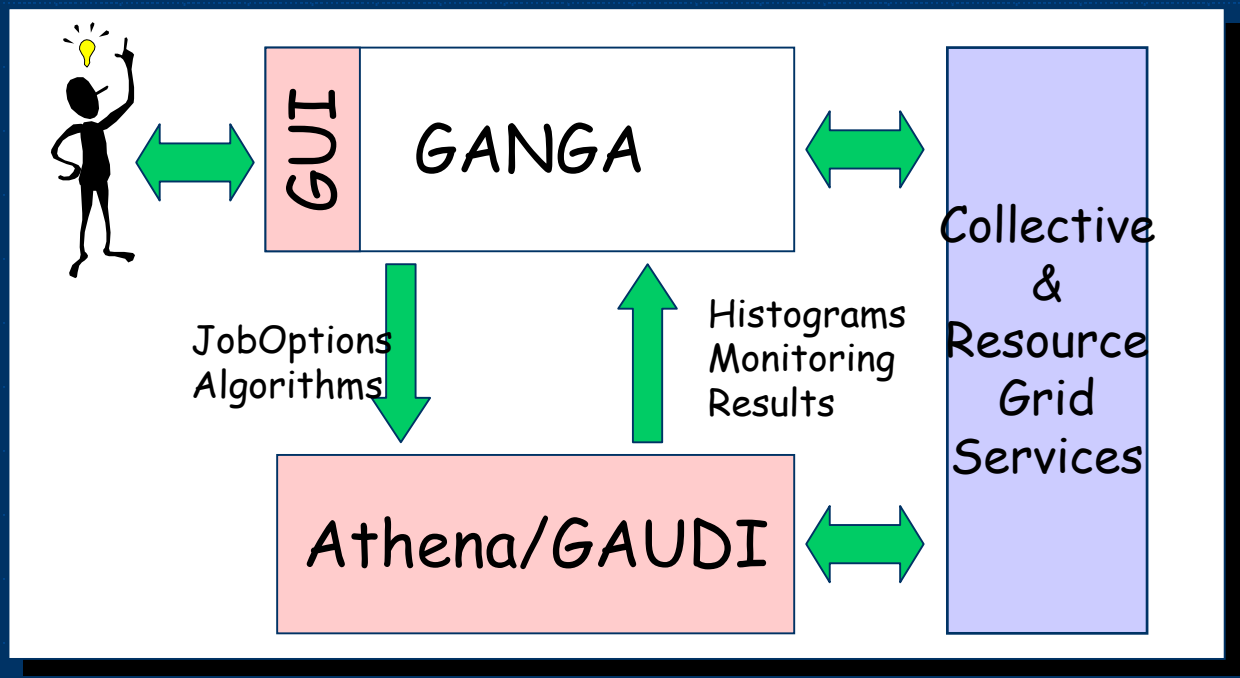
- Storage di dati su siti distribuiti e replicazione verso host su cui girano le applicazioni.
- Risolve un *logical file name* in una istanza fisica del file
- Replicazione possibile tra tipi diversi di data stores (es. disco-disco, Castor-disco, Castor-HPSS …)
- Core del sistema DB *MySql* + infrastruttura di gestione *perl,* trasferimento files *gridftp, bbftp, scp*
- Processi *spider* tengono aggiornato il catalogo.
- Interfaccia web, command line, C++ e Java APIs
- Atlas Magda server per DC1 a BNL, registrati 45000 files, 35 TB + repliche.

Ganga
Gaudi/Athena and Grid Alliance

- GUI o script: l'utente interagisce solo con GANGA, a tutti i livelli dell'analisi (sottomissione, monitoring, risultati)
- Funzionalità già implementate con EDG 1.4
- Dovrebbe collaborare con LCG per la Physicist Interface

Progetto Atlas/LHCb

Interfaccia a LSF, non ancora PBS



CHIMERA

- Catalogo virtuale per dati derivati (non raw), associati alle procedure di derivazione.
- VDL "*Linguaggio virtuale dei dati*" per descrivere la "ricetta" per produrre un dataset (provenienza, parametri di produzione, geometrie…)
- Accoppiato ai servizi GRID permette di ritrovare dati già prodotti o di effettuarne la produzione in base alle "ricetta".
- Interfaccia verso Condor. Può generare DAG.

Distribuzione del sw

- Il sw di ATLAS è accessibile sotto AFS al CERN.
- Per il DC1 è stato realizzato un kit di RPM per installare la distribuzione completa nei siti remoti (sia manualmente che con LCFG).

- EDG si basa su LCFGng per la configurazione dei nodi di calcolo, compreso sw di esperimento.
- US ATLAS grid utilizza PACMAN per la distribuzione del sw.

Prossimi sviluppi del calcolo di ATLAS Grid

- ATLAS ha incoraggiato lo sviluppo ed il test tools diversi, in ambienti Grid diversi.
- Il sistema generale di produzione è stato mantenuto più semplice possibile.
  - si sono evitate dipendenze di Athena da particolari middleware
  - si è evitato di "rincorrere" mw in rapida evoluzione con compicate interfaccie
- La pianificazione finale del sistema di produzione ed analisi sarà fatta nel framework di LCG
  - Poco interesse per soluzioni *ad interim*
  - Tutti gli sforzi devono andare verso EDG V2

Piani per il DC2

- Luglio 2003-Luglio 2004 (simulazione distribuita da Aprile 2004)
- Stessa quantità di eventi (10^7) del DC1, maggiore utilizzo di CPU (Geant4)
- Sedi coinvolte per la produzione CNAF, Milano, Roma, Napoli, per l'analisi tutte.
- Non si prevedono massicci trasferimenti di dati, non sono necessari incrementi di rete.

DC2: Time scale

- **End-July: Release 7**
- **Mid-November: pre-production release**
- **February 1st: "production" rel.**
- **April 1st**
- **June 1st: "DC2"**
- **July 15th**

- **Put in place, understand & validate**:
  - Geant4
  - POOL persistency & LCG App.
  - Event Data Model
  - Digitization; pile-up; byte-stream
  - Conversion of DC1 data to POOL and run reconstruction
- **Testing and validation**
  - Run test-production
- **Start final validation**
- **Start simulation**
- **Pile-up & digitization**
- **Transfer data to CERN**
- **Start Reconstruction on "Tier0"**
- **Distribution of ESD & AOD**
- **Calibration; alignment**
- **Start Physics analysis**
- **Reprocessing**

DC2 resources (based on Geant3 numbers)

Process	No. of events	Time duration	CPU power	No. of CPU 500SI2k ε=80%	Volume of data	At CERN	Off site
		months	kSI2k		TB	TB	TB
Simulation	10^7	2	260		24	8	16
Pile-up digitization	10^7	2	175		(75)	(25)	(50)
Byte-stream	10^7	2			15	15	10
Total	10^7	2	435	870	39 (+75)	23 (+25)	26 (+50)
Reconst.	10^7	0.5	600	1200	5	5	5