Il Computing di ATLAS

Gianpaolo Carlino
Atlas Italia
Roma, 9 Gennaio 2008

- Referaggio Calcolo LHC
- Analisi nei Tier Italiani
- La Federazione dei Tier-2

Referaggio Computing LHC

Workshop di Referaggio sul Computing di LHC
CNAF: 17 – 18 Gennaio 2008

Descrizione:

Acquisire elementi utili per stabilire come i limitati fondi disponibili nel 2008 si possano utilizzare in modo ottimale per rendere le federazioni T2 pronte e competitive all'appuntamento con LHC.

In particolare:

1) specificando le attività che devono essere svolte e i mezzi indispensabili per realizzarle;
2) valutando il grado di preparazione e i piani di commissioning delle federazioni T2, inclusi i servizi forniti dal T1;
3) chiarendo le questioni tecnico-implementative che possono avere un impatto sui finanziamenti, ovvero:
   - storage: SAN/NAS –
   - tipologia rete locale 10 Gbps –
   - formato macchine per worker nodes

Scopo:

Sblocco finanziamenti sub-judice

Proposte ATLAS

Forti – CSN1 – Sett 07

Atlas	CPU Disponibil ora	Acquisizioni da fare sj 2007	prim. 2008	Totale	DISCO Disponibil ora	Acquisizioni da fare sj 2007	ass. 2008	Totale
	kSI2k	kSI2k	kSI2k	kSI2k	TBN	TBN	TBN	TBN
Roma1	140	63	94	**297**	42	25	43	**110**
Napoli	92	63	94	**249**	37	25	43	**105**
Milano	129	40	62	**231**	32	16.5	29	**77.5**
LNF	41	22	31	**94**	21	7	14	**42**
Tot Atlas	**402**	**188**	**281**	**871**	**132**	**73.5**	**129**	**334.5**

S.J. 2007	CPU kEuro	Disco kEuro	Totale kEuro
Roma1	17	45	**62**
Napoli	17	45	**62**
Milano	11	29	**40**
LNF	6	13	**19**
Tot Atlas	**51**	**132**	**183**

primavera 2008	CPU kEuro	Disco kEuro	Totale kEuro
Roma1	15	60	**75**
Napoli	15	60	**75**
Milano	10	40	**50**
LNF	5	20	**25**
Tot CMS	**45**	**180**	**225**

Referaggio Computing LHC

Sviluppo temporale (secondo i referee):

1. 17-18 Gennaio: Referaggio
2. 28-29 Gennaio: CSN1 – Sblocco del sub-judice
3. Fine Febbraio: approvazione della Giunta INFN
   - Per gli acquisti superiori a 50 k€ è necessario effettuare una procedura di acquisto con firma del presidente
   - Presentazione del capitolato di gara almeno due settimane prima $\Rightarrow$ in contemporanea alla CSN1
   - Procedura di acquisto con il mercato elettronico
4. Metà/Fine Marzo: espletazione della gara con il mercato elettronico
5. Fine Aprile: piena operatività delle risorse

Piano molto ottimistico ma comunque in ritardo rispetto alla data formale del Primo Aprile !

Referaggio Computing LHC

Argomenti di Discussione:

- **Modelli di Calcolo**
  - 20'
  - Riassunto del modello di calcolo aggiornato, con indicazioni delle attività di calcolo previste nelle federazioni T2, descritte in termini quantitativi, con l'aspettato profilo temporale di crescita e in relazione ai ruoli che i vari siti T2 possono svolgere tenuto conto del loro eventuale diverso grado di sviluppo

- **Infrastrutture**
  - 20'
  - riassunto della quantità di spazio rack attrezzato disponibile nel 2008 - stato infrastrutture di rete - sistemi di monitoraggio e sistemi automatici di messa in sicurezza delle macchine di calcolo in situazioni di emergenza (guasti, incendi, black-out, ecc.)

- **Preparazione della collaborazione italiana**
  - 15'
  - coordinamento e struttura operativa al T1 e nelle federazioni T2 - supporto per gli utenti - ricognizione quantitativa sull'effettivo utilizzo dei Tier2 da parte dei fisici dei gruppi Italiani e feedback dagli utenti

Test e Commissioning

- 45'
- stato dell'attività di test e commissioning che dimostrino quantitativamente come i principali elementi del sistema reggano il livello di carico richiesto, possano scalare con il previsto aumento di luminosità e siano in grado di sostenere un uso continuativo (laddove non esistano ancora risultati: piano di lavoro); in particolare: - canali di trasferimento T1<->T2 - event store ai T2, acceduti con i pattern e i rate previsti - servizi di storage al T1 in relazione in particolare al numero di accessi alla libreria di cassette richiesti - valutazioni tecnico-economiche alla base delle scelte di implementazione proposte che presentano implicazioni finanziarie rilevanti (in ordine di importanza): - approccio SAN e NAS nella realizzazione del bulk storage ai T2 - formato macchine worker nodes

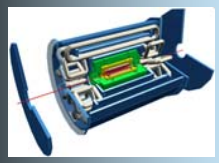
Analisi (Michela Biglietti)

- 20' + 20'
- Dimostrazione di analisi dati dal vivo: preparazione dei dati e sottomissione dei job sulla grid. Lo scopo e' duplice: familiarizzare i referees con gli elementi principali del processo e mostrare la maturità e l'efficacia dei tools e delle procedure impiegate

- Talk 1a -

Il Modello di Calcolo: struttura e replica dei dati (cenni)

Il Computing Model: Cloud Model



~PB/s

Event Builder

10 GB/s

Event Filter

320 MB/s

Tier0

Tier1 ...

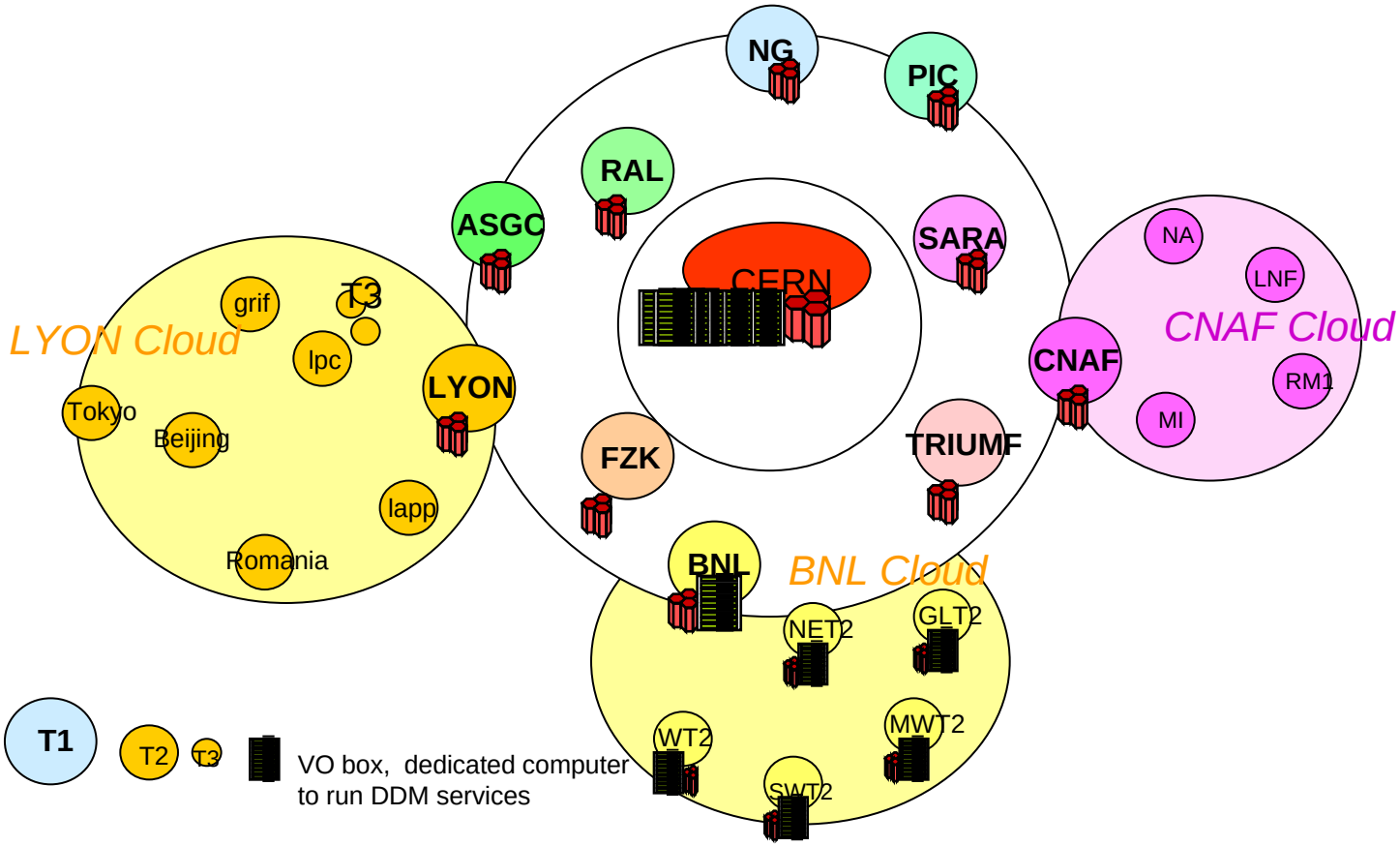
Tier2

Tier3 ...

...

Il Modello di Calcolo per l'offline e l'analisi di ATLAS è un modello gerarchico multi – Tier.

Modello a *cloud*: ad ogni Tier-1 sono associati alcuni (3 o 4) Tier-2 spesso in base a considerazioni geografiche.

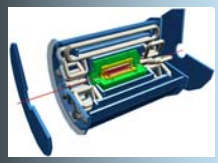


I Tier di ATLAS

10 Tier-1s and 46 Tier-2s

- ASGC
  - AU-ATLAS, TW-FTT,AU-UNIMELB
- BNL
  - AGLT2, BU, MWT2, OU, SLAC, UTA, WISC
- CNAF
  - LNF, MILANO, NAPOLI, ROMA1
- FZK
  - CSCS,CYF,DESY-HH,DESY-ZN,FZU,LRZ, FREIBURG, WUP
- NDGF
- LYON
  - BEIJING, CPPM, LAL, LAPP, LPC, LPNHE, NIPNE_02,NIPNE_07, SACLAY, TOKYO
- PIC
  - IFAE, IFIC, UAM,LIP
- RAL
  - GLASGOW, LANCS, MANC, QMUL, DUR, EDINBURGH, OXF,CAM,LIV,BRUN,RHUL
- SARA
  - IHEP, ITEP, JINR, PNPI, SINP
- TRIUMF
  - ALBERTA, MONTREAL, SFU, TORONTO,UVIC

Il Computing Model: i Tier



Tier-0 (CERN)

- Archivio dei RAW data e distribuzione ai Tier1
- Prompt Reconstruction dei dati in 48 ore
- 1st pass calibration in 24 ore
- Distribuzione output ricostruzione ai Tier-1: ESD, AOD e TAG

Tier-1 (10)

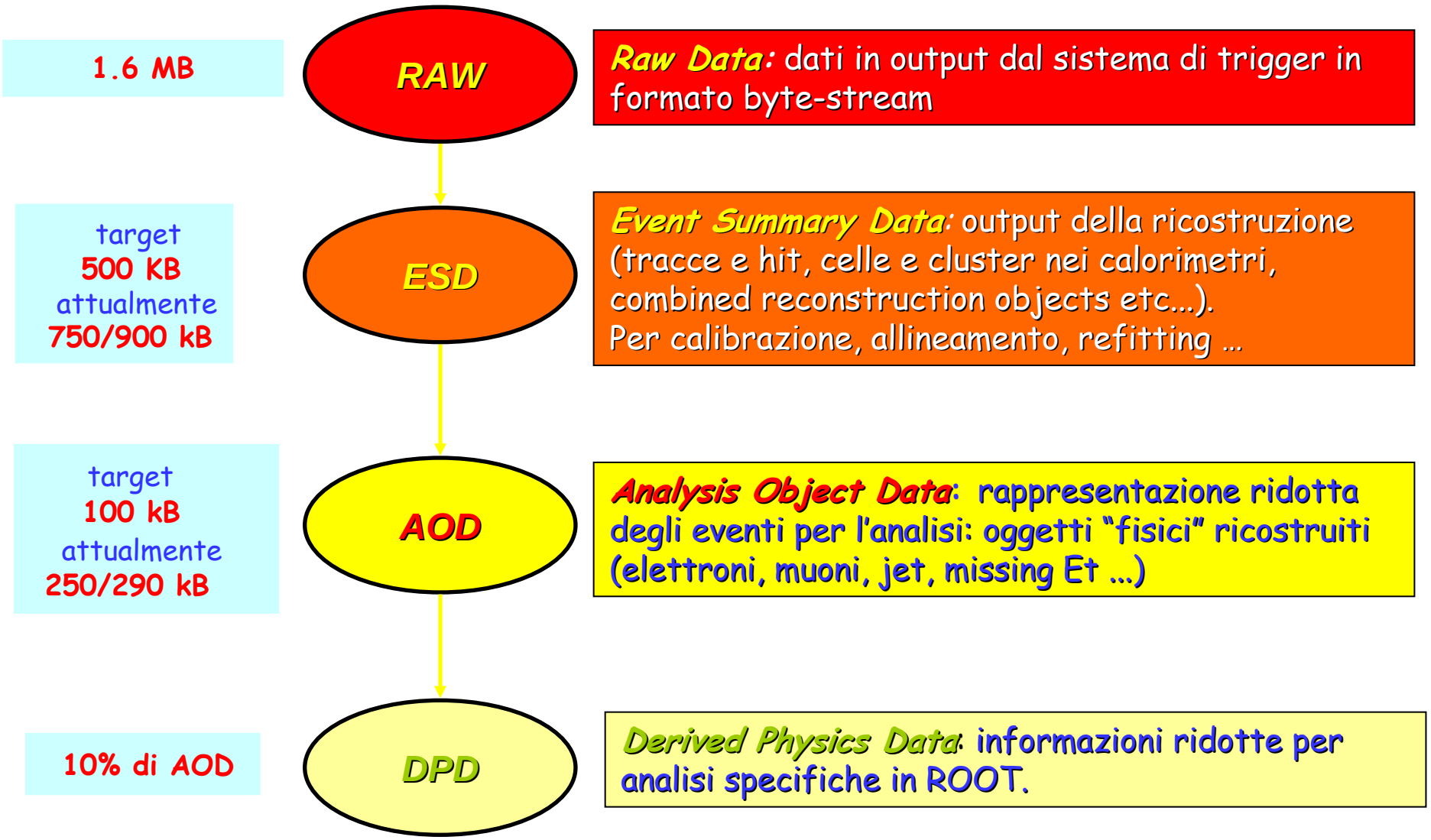
- Accesso a lungo termine e archivio di un subset di RAW data
- Copia dei RAW data di un altro Tier-1
- Reprocessing della ricostruzione dei propri RAW data con parametri di calibrazioni e allineamenti finali 2 mesi dopo la presa dati
- Distribuzione AOD ai Tier-2
- Archivio dati MC prodotti nei Tier-2
- Analisi dei gruppi di fisica

Tier-2

- Simulazione Monte Carlo
- Analisi utenti

Event Data Model: tipi di dati

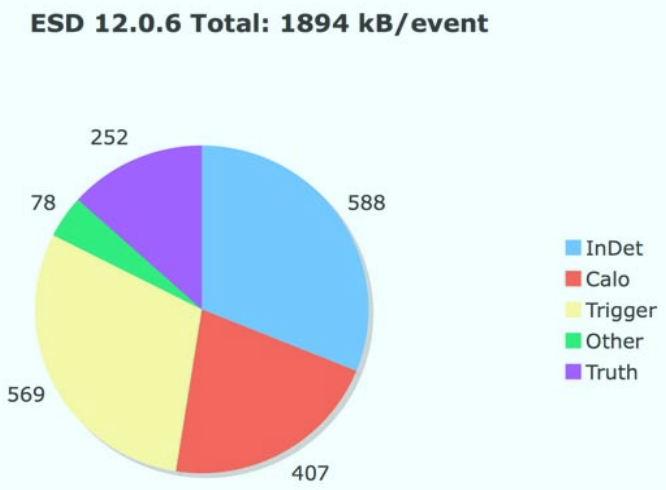
Nelle varie fasi di ricostruzione e analisi ATLAS utilizza diversi formati di dati:

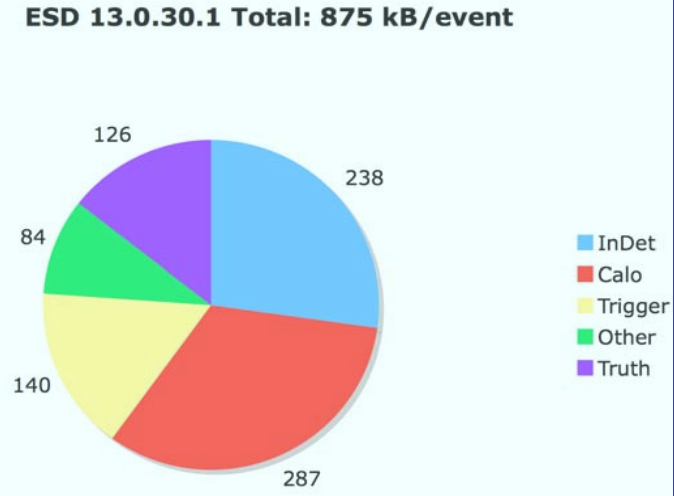


Event Data Model: Event Size

La dimensione degli eventi determina significativamente la necessità di risorse

- Dipende molto dal campione MC analizzato
  - ✓ Si misura la dimensione degli eventi tt in cui sono presenti elettroni, muoni, tau, jet, b-jet, missing ET
  - ✓ Si scala di un fattore 0.7 ottenuto dallo streaming test
- La dimensione dell'evento dipende dalle threshold di trigger e dal menu: menu più grandi e threshold più basse ⇒ maggiore event size
- Valori più accurati dal Full Dress Rehearsal





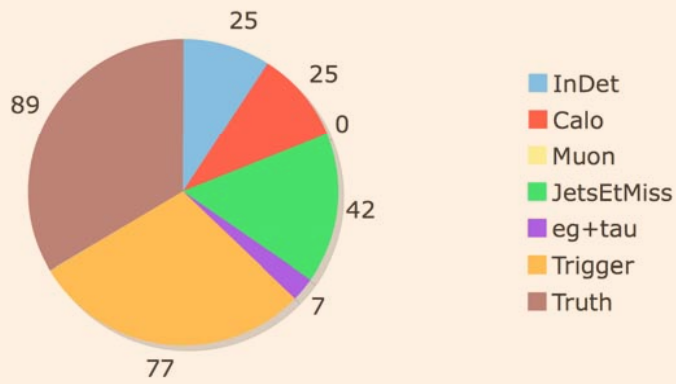
ESD

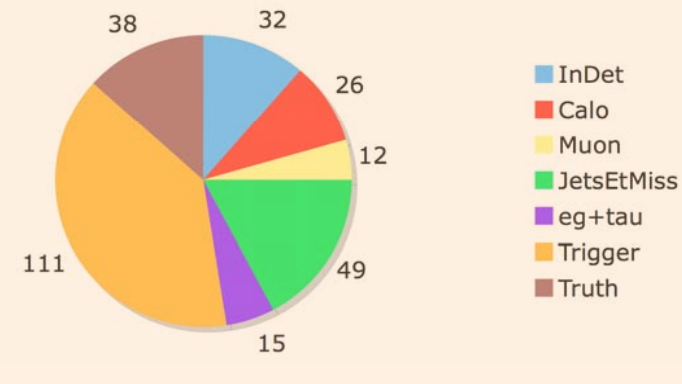
- Computing Model: 500 kB/event
- v12: ~ 1.9 MB/event, v13: ~ 0.9 MB/event
- Diminuzione dell'event size dalla release 12 alla 13:
  - ✓ grande collaborazione tra sviluppatori e utenti
  - ✓ miglioramenti tecnici (seperazione T/P, class merging)
  - ✓ ottimizzazione info sul trigger e MCtruth
- ulteriori guadagni marginali a meno di perdita di info significative

Event Data Model: Event Size

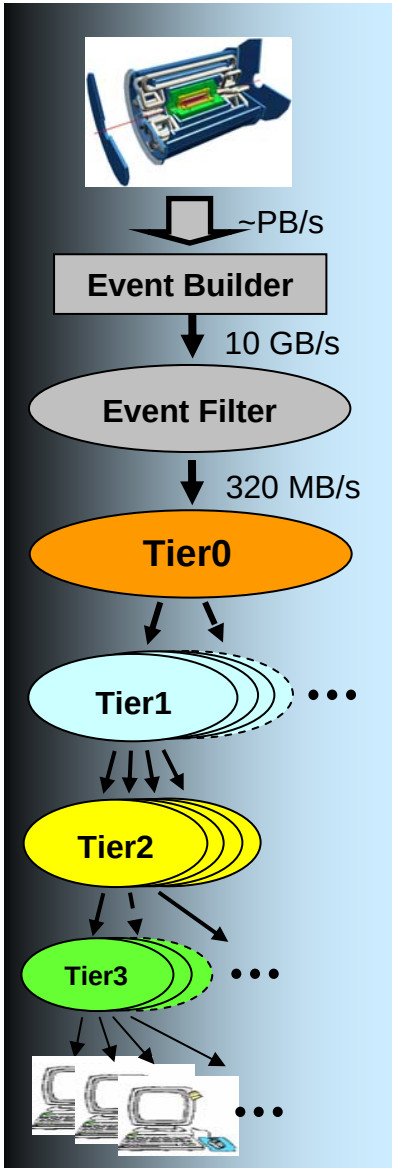
AOD

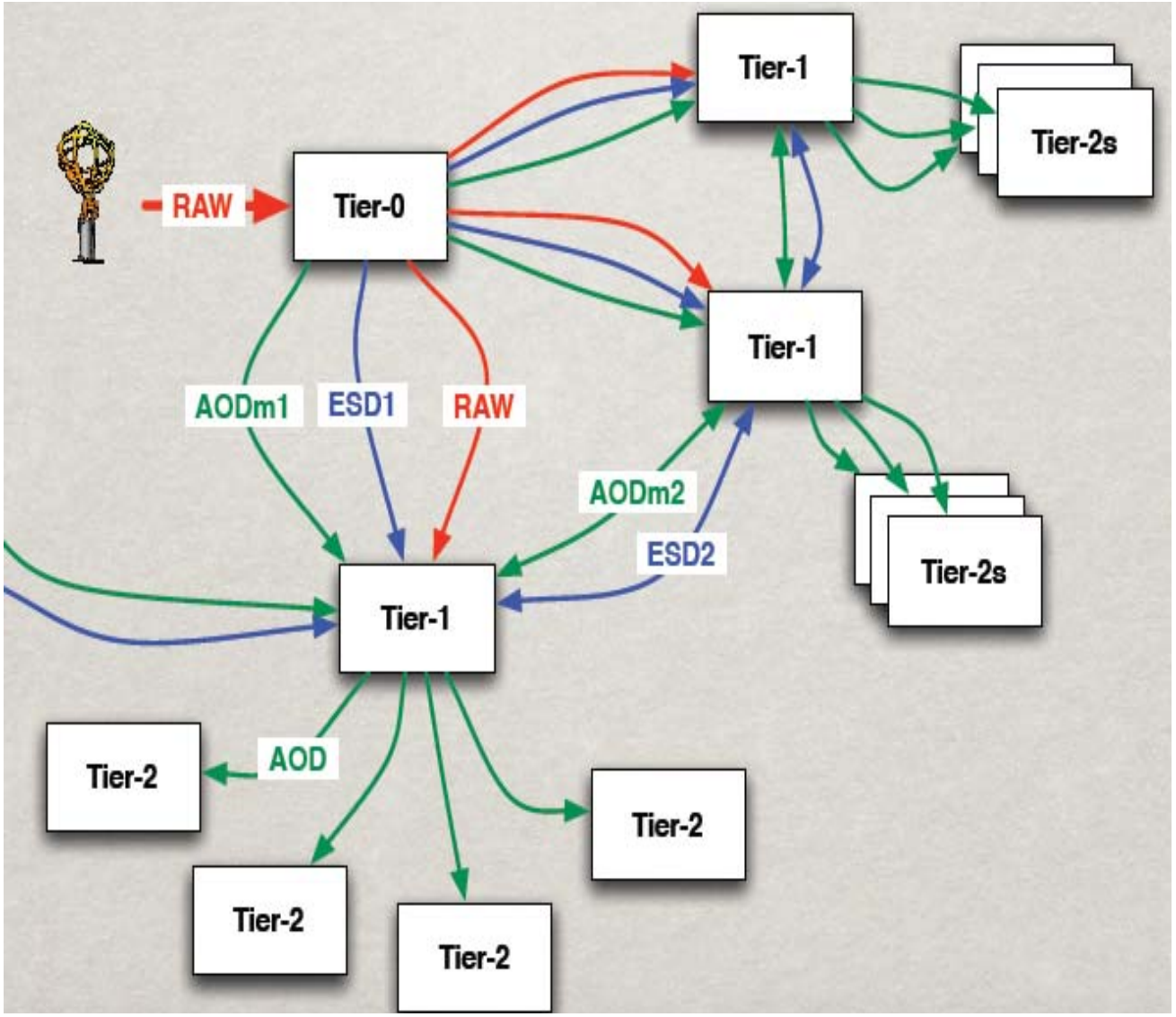
- Computing Model: 100 kB/event
- v12: ~ 270 kB/event, v13: ~ 290 kB/event
- MCtruth ridotta
- Aumento della dimensione per l'aggiunta delle celle del calorimetro per i muoni e i candidati egamma e le tracce associate
- Aumento Trigger size per menu più completo e threshold più basse.
  - Frazione di trigger da 22% (dijet) a 38% (top).
  - Trigger decision (per l'analisi): 10 kB. Il resto utile per studi dettagliati sulle trigger performance (necessario anche all'inizio del data taking)
- Riduzione significativa possibile diminuendo le collezioni di jet



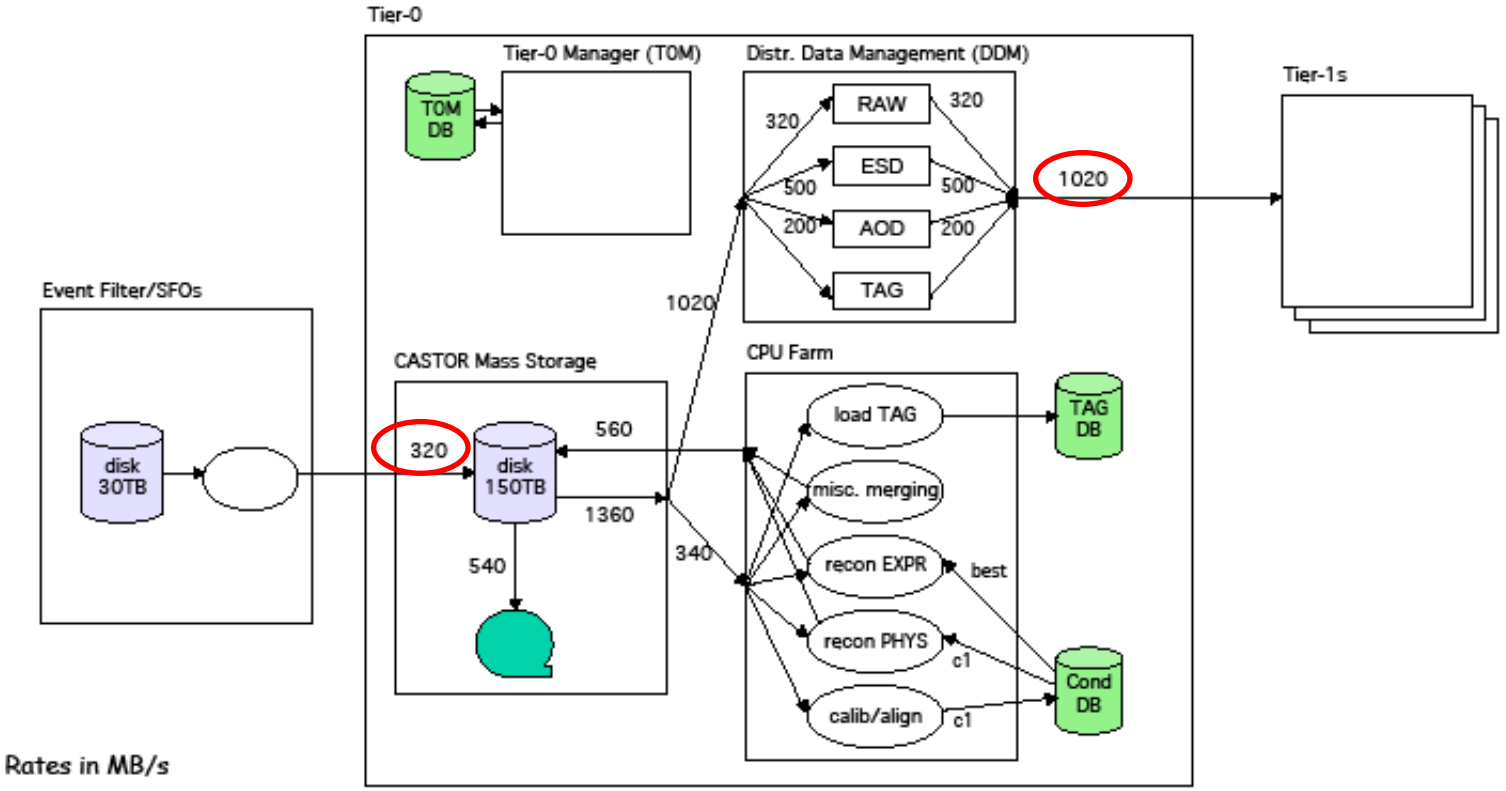


Il Computing Model: distribuzione dei dati





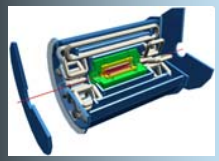
Tier-0 throughput schema



Output rate dal TDAQ = 320 MB ⇒ 200 Hz (trigger rate) · 1,6 MB (event size)

Collegamento in fibra dedicato Tier-0 ↔ Tier-1s = 10 Gbps

Il Computing Model: il Tier-0



~PB/s

Event Builder

10 GB/s

Event Filter

320 MB/s

Tier0

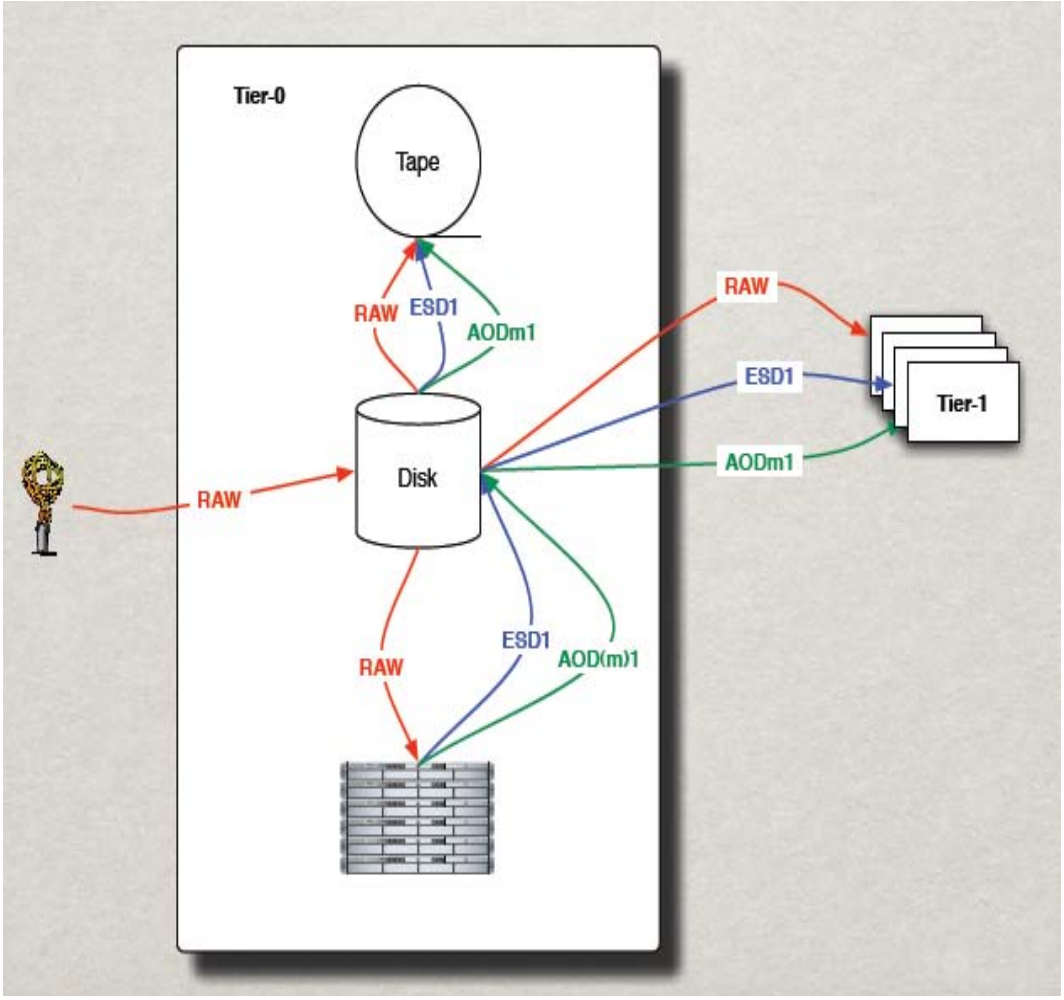
Tier1 ...

Tier2

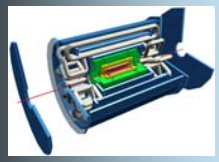
Tier3 ...

...

I dati originali (RAW data e AOD e ESD primari) risiedono al Tier-0



Il Computing Model: i Tier-1



~PB/s

Event Builder

10 GB/s

Event Filter

320 MB/s

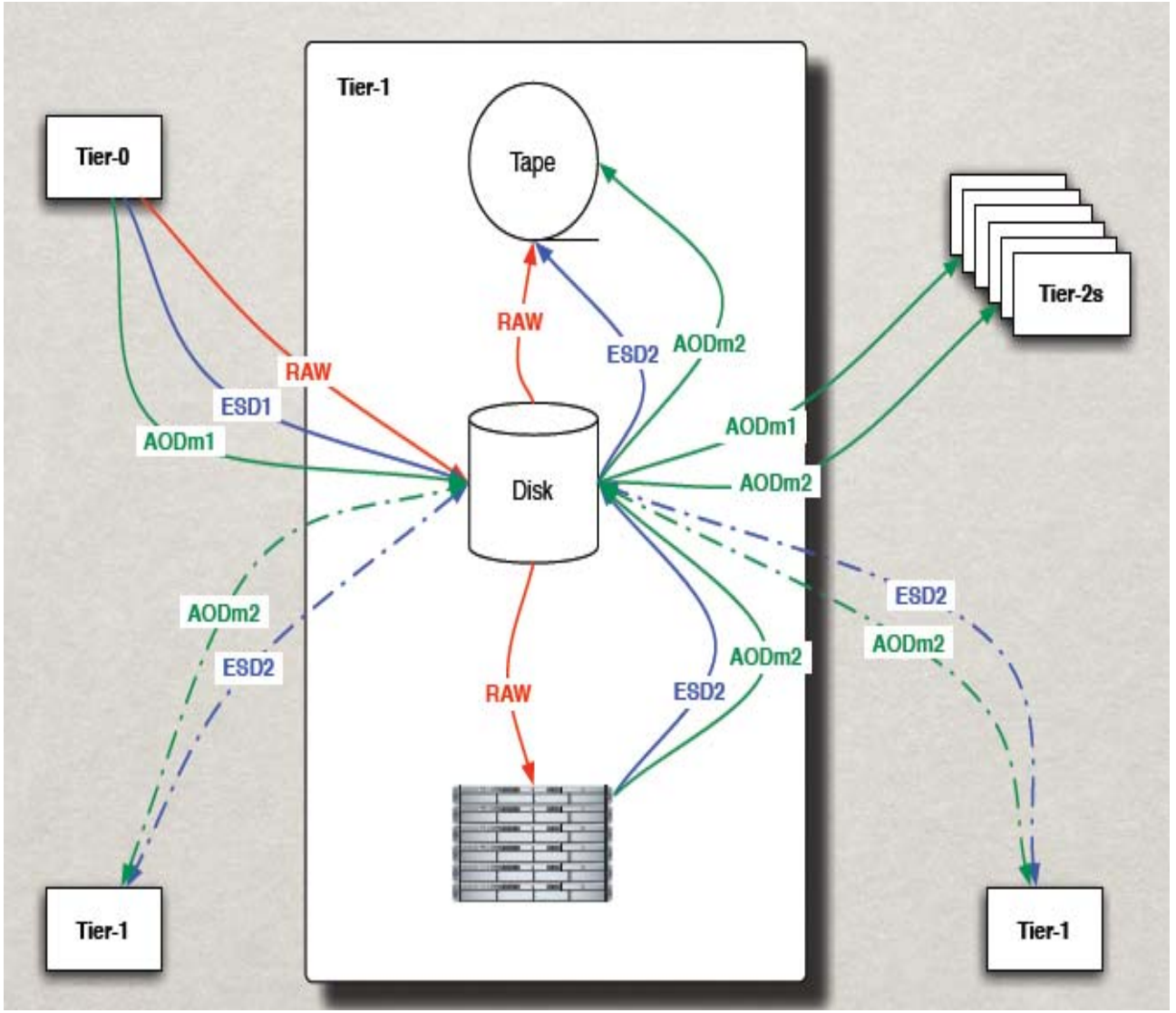
Tier0

Tier1 ...

Tier2

Tier3 ...

...



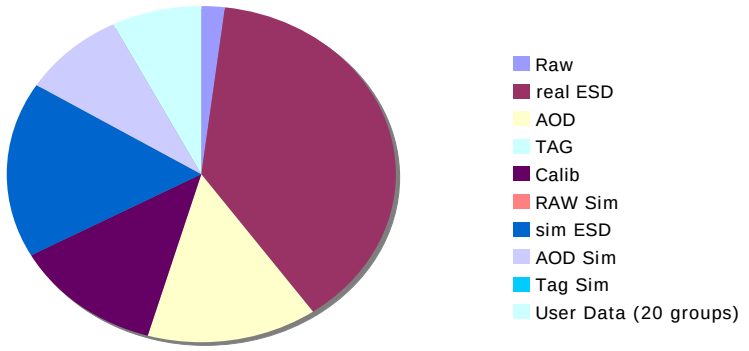
Il Computing Model: i Tier-1

Accesso schedulato per analisi centralizzate e produzione

Replica dei dati
- Una copia di RAW data nell'insieme dei Tier-1 (10% su disco)
- ESD replicati in due copie ai Tier-1
- Una copia di AOD e TAG in ogni Tier-1
- Ogni Tier-1 riprocessa i suoi RAW e li replica secondo lo stesso schema
- Una copia di DPD di gruppo per ogni Tier-1

Attività di calcolo (2008)
- 25% Reprocessing
- 25% Simulazione (Gen + Reco)
- 50% Analisi Centrale

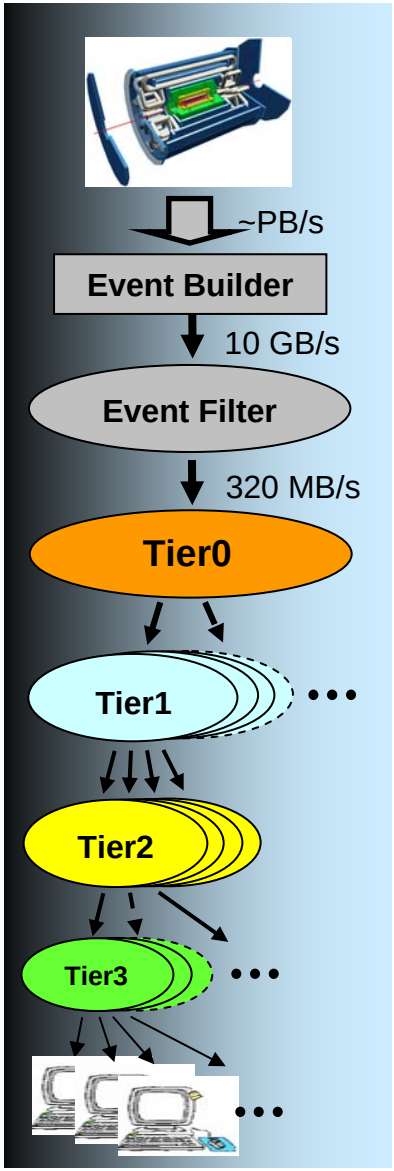
Tempi di processamento
Simu = 400 kSI2k·sec
Reco = 15 kSI2k·sec

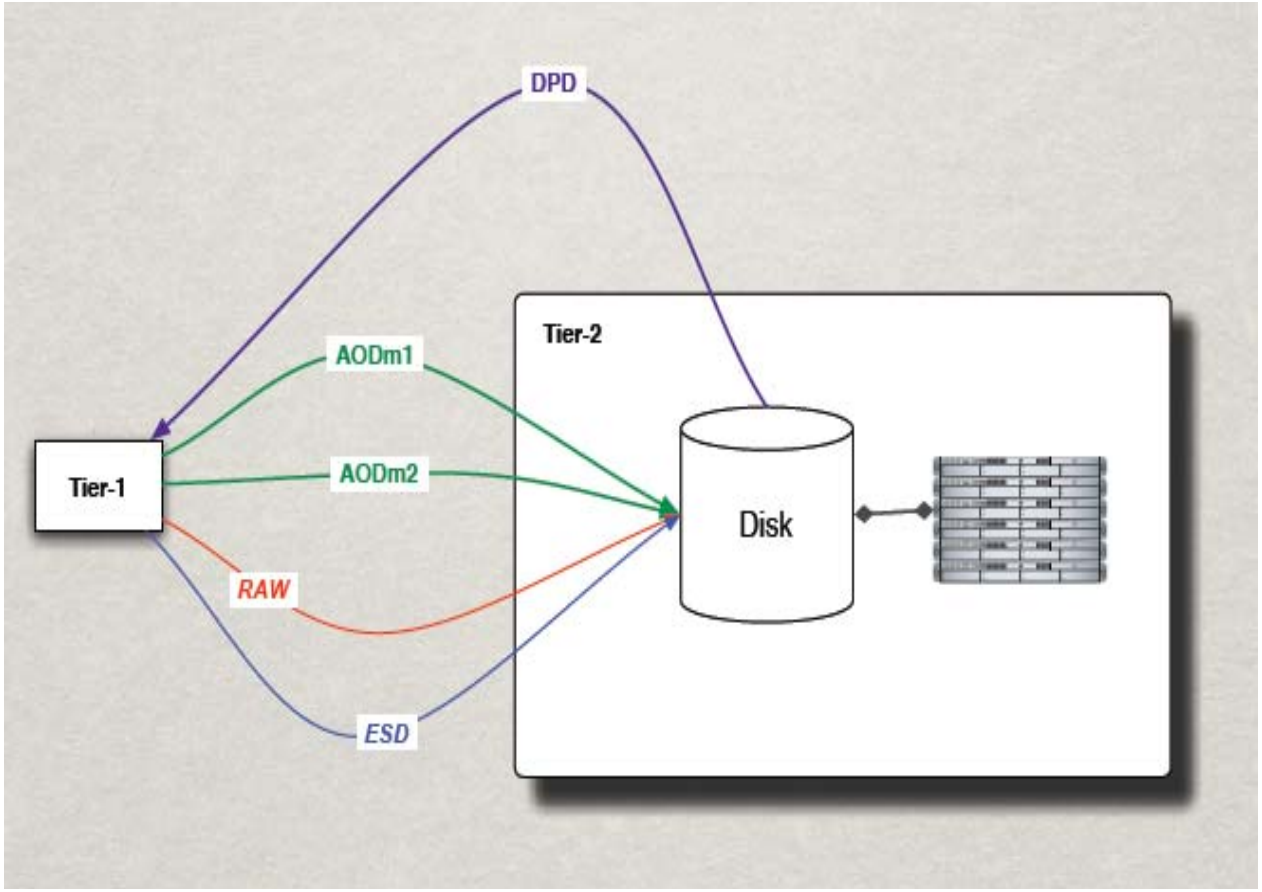


Atlas @ CNAF nel 2007
CPU: 400 kSI2k
Disco : 130 TBn

- Raw = 2%
- ESD (real + MC) = 55%
- AOD (real + MC) = 25%
- User = 7%
- Calib = 11%

Il Computing Model: i Tier-2



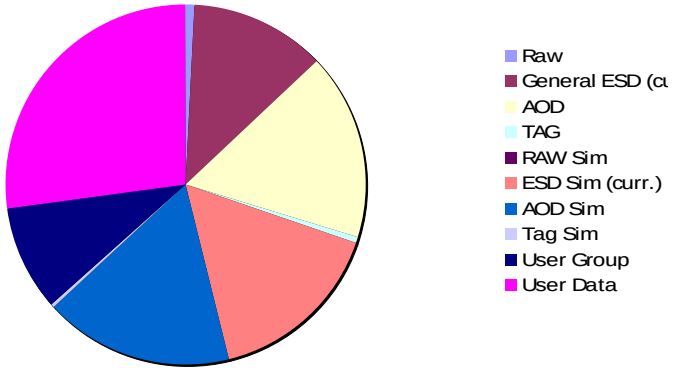


Il Computing Model: i Tier-2

Simulazione e accesso "caotico" per analisi utenti

Replica dei dati
- Una copia di AOD e TAG in ogni cloud di Tier-2
- Copie di DPD di gruppo e di utenti
- RAW data: 30% nel 2008 e 10% nel 2009 in tutte le cloud di Tier-2
- ESD: 150% nel 2008 e 30% nel 2009 in tutte le cloud di Tier-2

Attività di calcolo (2008)
- 15% Ricostruzione
- 0% Reprocessing
- 37% Simulazione
- 48% Analisi Utenti



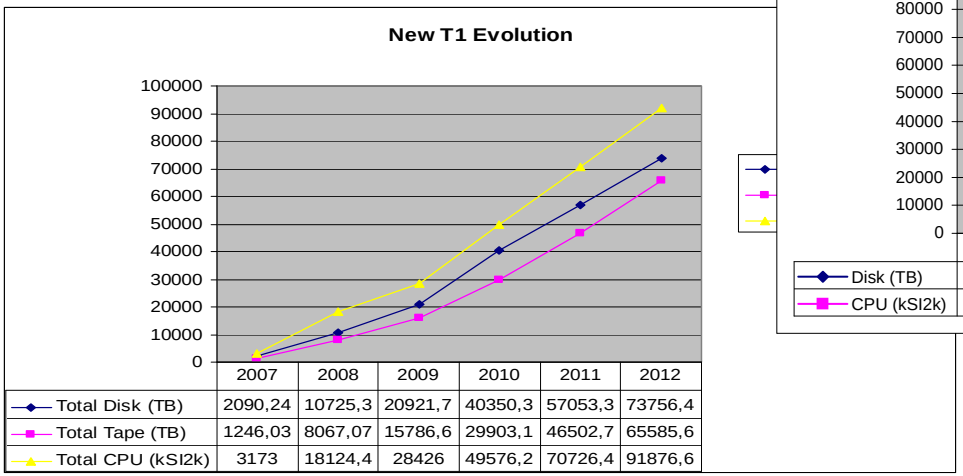
Tier2 italiani (2007)

CPU: ~ 400 kSI2k

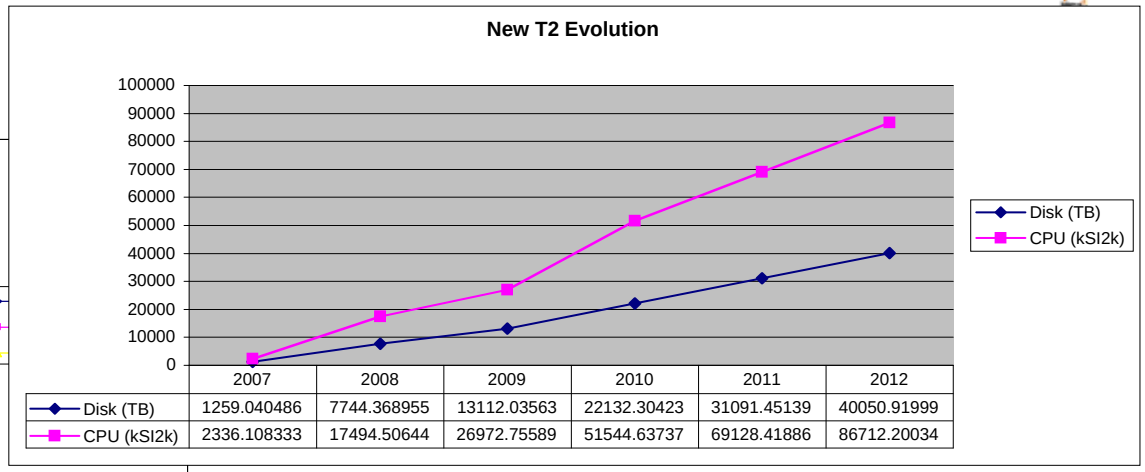
Disco : 165 TBr

- Raw = 1%
- ESD (real + MC) = 28%
- AOD (real + MC) = 34%
- User Group = 10%
- Users = 27%

Sviluppo Risorse



	2007	2008	2009	2010	2011	2012
Total Disk (TB)	2090,24	10725,3	20921,7	40350,3	57053,3	73756,4
Total Tape (TB)	1246,03	8067,07	15786,6	29903,1	46502,7	65585,6
Total CPU (kSI2k)	3173	18124,4	28426	49576,2	70726,4	91876,6



	2007	2008	2009	2010	2011	2012
Disk (TB)	1259.040486	7744.368955	13112.03563	22132.30423	31091.45139	40050.91999
CPU (kSI2k)	2336.108333	17494.50644	26972.75589	51544.63737	69128.41886	86712.20034

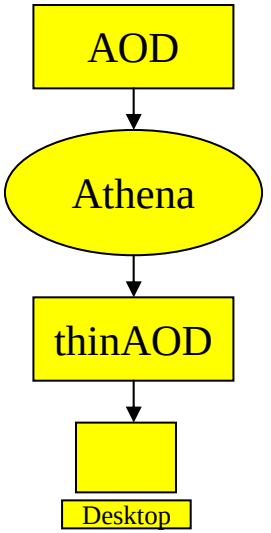
	CPU (MSi2k)		Disk (PB)		Tape (PB)	
	2008	2010	2008	2010	2008	2010
Tier-0	3.7	6.1	0.15	0.5	2.4	11.4
CERN Analysis Facility	2.1	4.6	1.0	2.8	0.4	1.0
Sum of Tier-1s	18.1	50	10	40	7.7	28.7
Sum of Tier-2s	17.5	51.5	7.7	22.1	10.5	41.1
Total	41.4	112.2	18.9	65.4		

1/30 quota media singolo T2

Analysis Model

Recente evoluzione dell'Analysis Model motivata dal bisogno di:

- Aver accesso diretto ai dati nel formato AOD con la velocità tipica dell'analisi in ROOT e la possibilità di sfruttare la potenza di Athena
  - Nuovo formato di AOD permesso dalla separazione tra I dati in formato transiente utilizzato in Athena e persistente (file) che può essere letta in Root
- Ridurre il numero di formati di DPD e la loro dimensione
  - Questo nuovo formato di AOD opportunamente ridotto (procedura di Skimming e Thinning che seleziona gli eventi interessanti e riduce la quantità e dimesione delle informazioni) e con l'aggiunta delle informazioni specifiche dell'analisi costituisce il nuovo e unico formato di DPD (size ~10% dell'AOD)
  - Discussione in corso per decidere le procedura e la localizzazione della produzione dei DPD



Metodi di Analisi

- Athena – metodo usuale, batch o interattivo, che permette un accesso totale ai tool e ai servizi del framework
- Athena Root Access (ARA) – metodo innovativo, interattivo o batch, che utilizza gran parte dei servizi di Athena (no DB o metadata info). Leggero, veloce e facile da usare

- Talk 1b -

Attività di computing nel 2008, ovvero la verifica del CM

- FDR e CCRC
- Run di Cosmici (Mx)
- Dati LHC

Final Dress Rehearsal (FDR)

Lo scopo è testare l'intero computing system come se si trattasse di dati reali per trovare in tempo i problemi che si potrebbero verificare durante il data taking

Esercizio completo dell'intera catena, dall'on-line/trigger all'analisi distribuita, per integrare i test svolti fino ad ora in modo indipendente:

- Simulazione di 1 giorno di presa dati
- Immissione dei dati nel TDAQ e running a partire dagli SFO
- Completo utilizzo del Tier-0
  - merging, scrittura su tape, calibrazione, reprocessing etc
- Esecuzione del Computing Model in maniera completa
  - distribuzione dei dati, re-processing, analisi
- Simulazione MC completa in parallelo
  - running ai Tier-2, trasferimento dati e ricostruzione ai Tier-1

2 Run: FDR-1 in Febbraio e FDR-2 in Aprile/Maggio

CCRC (Common Computing Readiness Challange) in contemporanea per dimostrare che le infrastutture e servizi sono in grado di supportare le attività contemporanee dei 4 esperimenti LHC

Final Dress Rehearsal (FDR)

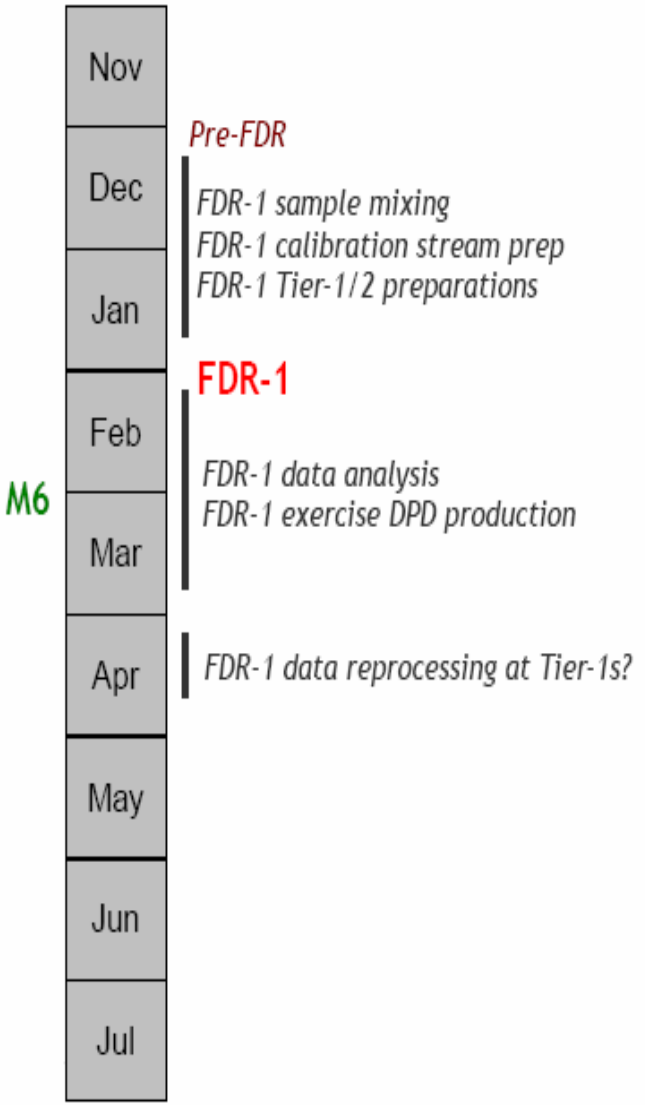
- Generazione e simulazione di eventi MC e mix di tutti i canali di fisica, in proporzione alle sezioni d'urto, per riprodurre un campione il più possibile simile ai dati reali
- Riproduzione della tipologia di dati in output all'HLT: simulazione del trigger, produzione del byte stream e streaming degli eventi. Tabelle di Trigger realistiche
- Input dei dati al P1 come dati reali
- Trasmissione dei RAW data dal P1 al Tier-0
- Data quality monitoring, calibrazioni e allineamento al Tier-0
- Ricostruzione in tempo reale al Tier-0 → produzione di ESD, AOD, TAG
- Distribuzione di ESD, AOD, TAG ai Tier-1 e Tier-2
- Produzione del TAG database e dei DPD
- Riprocessamento dei RAW data ai Tier1 e redistribuzione di AOD
- Processamento dell'analisi distribuita
- Simulazione continua in parallelo ai Tier-2 (~ 100k jobs/day)

In rosso gli step sincroni come durante il data taking

Final Dress Rehearsal (FDR)

Round 1:

- Utlizzo dati RDO simulati con la v12
  - ~ 150 TB (copiati al CERN in nov/dic 07)
- Mixing dei dati in formato bytestram (detector-like)
  - alcune settimane necessarie per il mixing
- Simulazione di un fill (10 hr) a 10^{31}
  - Luminosità istantanea decrescente durante il fill
  - Menu di Trigger a 10^{31} fisso durante il fill
  - ~ 400 nb^{-1} in tatale
  - Rate 200 Hz, 10 h. di run $\rightarrow$ 7.2 M eventi/giorno
  - ~ 12 TB al giorno (7.2 M ev · 1.6 MB/ev)
- Simulazione precisa delle condizioni di run di LHC
  - Trasferimento dati da SFO a Castor in 10 h (max rate) e 14 h rimanenti per calibrazione e data processing
- Introduzione delle express e calibrations streams
- Replica di questo fill nei 20 giorni successivi simulando le diverse condizioni di data taking
- SFO disponibili solo una settimana (dal 4 Feb) le rimanenti 3 settimane datastream da Castor



Final Dress Rehearsal (FDR)

Round 1 – Data Volumes:

Al Tier-0:

RAW = 1.6 MB/ev
ESD = 1 MB/ev
AOD = 0.2 MB/ev

	Per hour (TByte)	Per day (10 hrs)	Per FDR (20 days)
RAW	1.152	11.52	230.0
ESD	0.72	7.20	140.0
AOD	0.144	1.44	28.8

TIER-0 FDR DATA VOLUME PER DATA TYPE

Tier-1	Share (%)	RAW (TB)	ESD (TB)	AOD (TB)
BNL	25	2.88	7.2	1.44
IN2P3	15	1.73	2.16	1.44
SARA	15	1.73	2.16	1.44
RAL	10	1.15	1.44	1.44
FZK	10	1.15	1.44	1.44
CNAF	5	0.58	0.72	1.44
ASGC	5	0.58	0.72	1.44
PIC	5	0.58	0.72	1.44
NDGF	5	0.58	0.72	1.44
TRIUMF	5	0.58	0.72	1.44

STORAGE NEEDED PER DATA TYPE PER DAY

Tier-1	Share (%)	T1D0 RAW (TB)	T0D1 ESD+AOD (TB)
BNL	25	58	173
IN2P3	15	35	72
SARA	15	35	72
RAL	10	23	58
FZK	10	23	58
CNAF	5	12	43
ASGC	5	12	43
PIC	5	12	43
NDGF	5	12	43
TRIUMF	5	12	43

DATA STORAGE NEEDED FOR THE WHOLE TEST

Ai Tier-1:

Volumi variabili in base ai diversi share e richieste particolari
2 copie di ESD ai Tier-1
Copia completa di AOD per Tier-1

Ai Tier-2:

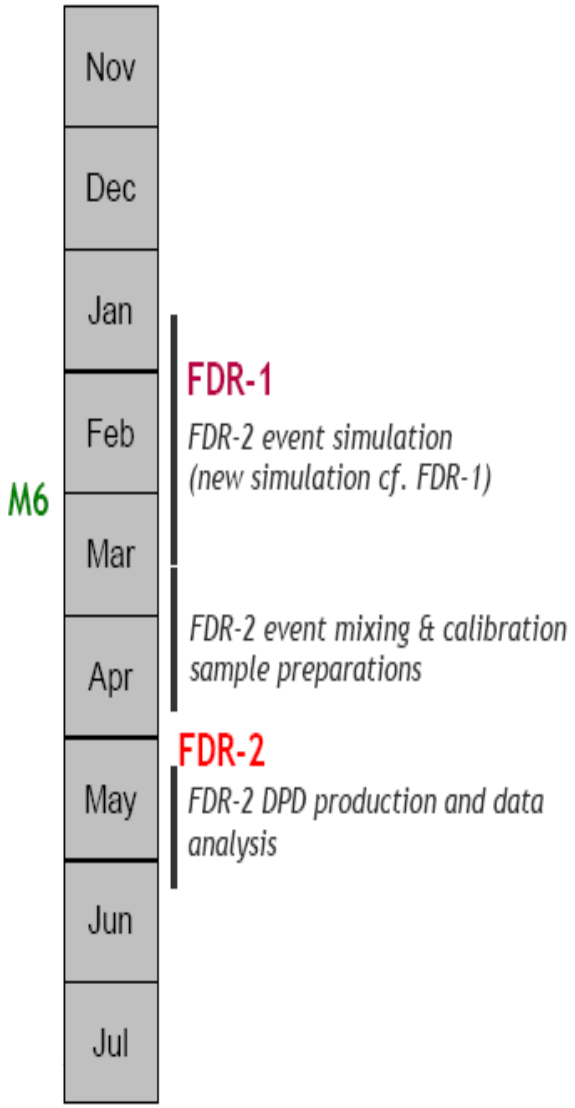
Replica completa degli AOD divisi secondo le usuali quote

Final Dress Rehearsal (FDR)

Round 2:

come il Round 1 ma a luminosità più alta

- Dati simulati con la v13
  - Nuova produzione!
  - ~ 120 Milioni di eventi
  - Validazione della release fine gennaio
  - Produzione a partire da febbraio per 3 mesi
  - Simulazione nei Tier-2
- Fill con luminosità 10^{33}
  - Rate 200 Hz, 12 TB al giorno
  - Menu di trigger sempre più complicati e fisica più ricca
- Ripetizione del fill più volte
- L2 muon calibration stream, calibrazione ai Tier-2
- Produzione centrale di DPD mediante le procedure di slimming degli AOD
- Tuning dei tool di analisi distribuita
- Analisi dai DPD attaverso il framework ARA



Final Dress Rehearsal (FDR)

Round 2 – Data Volumes:

- Dati sample
  - 0.5 M minimum bias e cavern events
  - 10 M eventi di fisica
  - > 100 M eventi fakes per ottenere un mixing realistico
- Production Rate ai Tier-2
  - 1.5 M ev al giorno (30k job al giorno)
  - 3 mesi di produzione
  - Fattore 3 rispetto ai rate attuali
  - Storage buffer = 1 – 10 TB
- Upload al Tier-1 e mixing con eventi di background, RDO files (2.5 MB/ev)
- Ricostruzione e produzione di ESD e AOD
  - 80 + 80 TB di spazio (share CNAF 5%)
- Upload al Tier-0 e mixing per ottenere il formato bytestream
  - 4 settimane di mixing, output ~ 12 TB
- Partenza dell'FDR-2
  - Durata (e conseguente volume di dati) ancora da decidere

Common Computing Readiness Challange (CCRC)

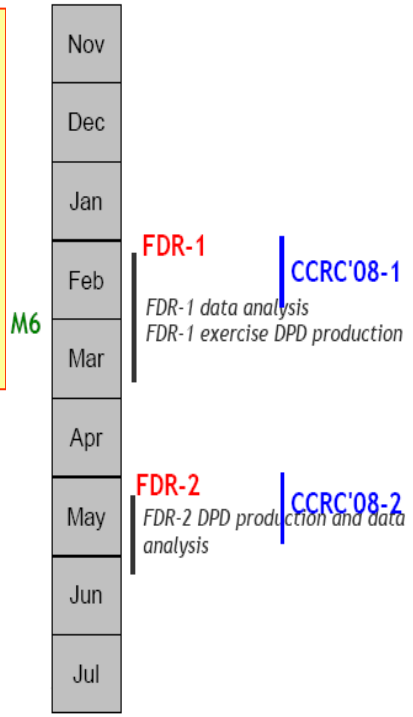
Nel 2008:

- LHC finalmente sarà operativo e tutti gli esprimenti prenderanno dati
- Tutti gli esperimenti useranno le infrastrutture di computing simultaneamente
- Il Tier-0, molti Tier-1 e alcuni Tier-2 gestiscono l'attività di più esperimenti e devono garantire le funzionalità previste dai singoli Computing Model

per cui …

Un challenge combinato deve dimostrare la capacità delle infrastrutture di computing a funzionare anche in situazioni di concorrenza tra tutti gli esperimenti LHC prima dell'inizio della presa dati ad una scala comparabile ai volumi previsti nel 2008

Tutto deve essere svolto in tempo per evidenziare imperfezioni, bottlenecks e permettere le neccessarie correzioni



Le due fasi del CCRC coincidono con quelle degli FDR.

Run di Cosmici (Mx)

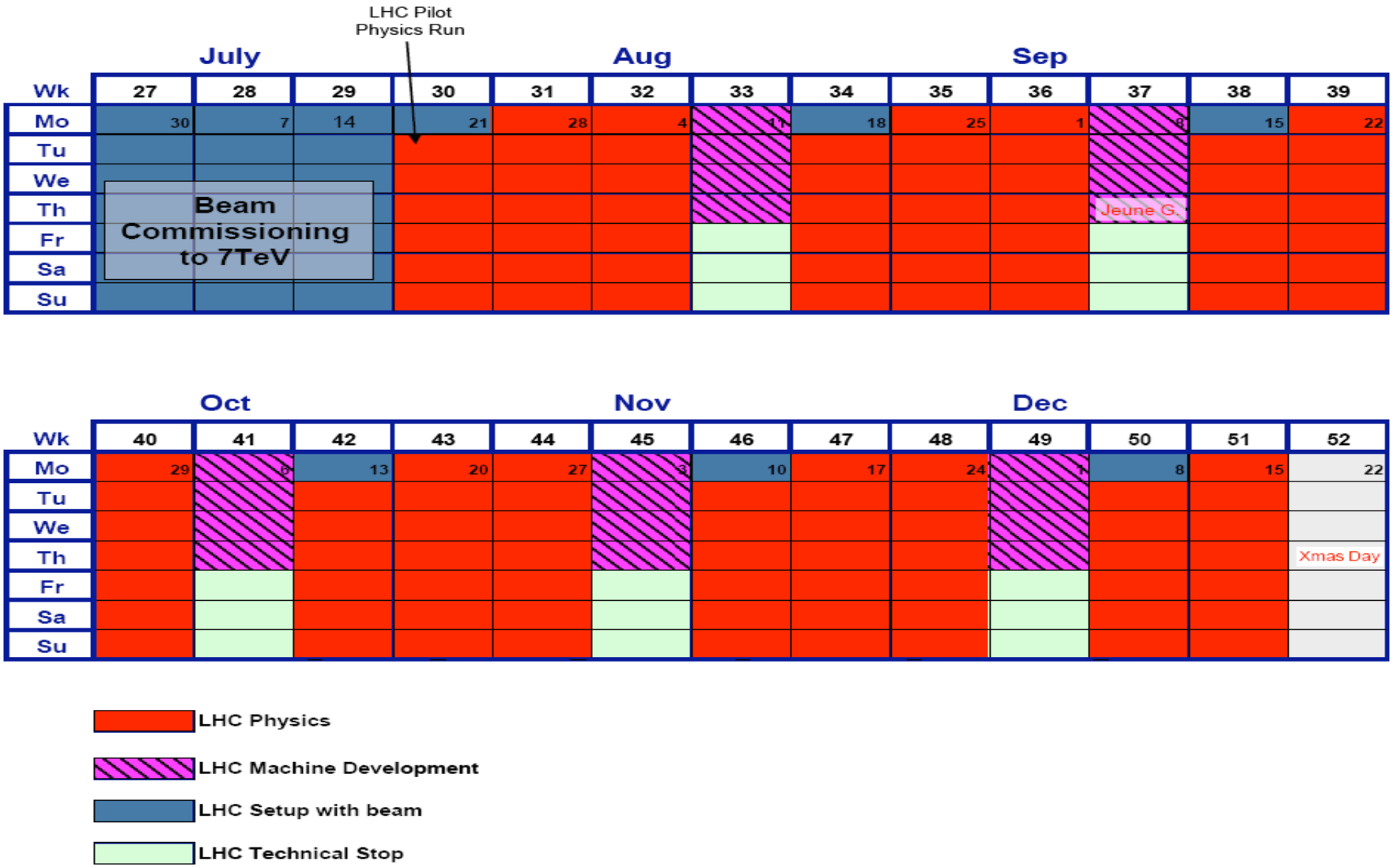
Non sono dei veri esericizi di computing in quanto lo scopo primario è l'integrazione di rivelatori

- M6 Febbraio/Marzo, M7 solo se il fascio ritarderà
- Tipicamente della durata di 2 settimane
  - La prima per l'integrazione dei sottorivelatori
  - La seconda per la raccolta e distribuzione dei dati
- non abbastanza stabili e lunghi per essere utilizzati come test di throughput
- Non abbastanza completi per testare in maniera significativa il Computing Model
  - No AOD o DPD
  - Uso dei RAW data per l'analisi poiché ESD non adeguati

Dal punto di vista del computing utile test di distribuzione dei dati (efficienza)

LHC Data Taking

Schedula ufficiale Jun 2007 non modificata



LHC Data Taking

Physics Run = $15 \cdot 10^6$ sec, eff = 30% $\Rightarrow$ $5 \cdot 10^6$ sec

Rate = 200 Hz $\Rightarrow$ Raw Data = 10^9 eventi

Dati simulati ~ 40% dei dati reali: $4 \cdot 10^8$ eventi

Totale = $1.4 \cdot 10^9$ eventi

Risorse Calcolo necessarie:

✓ CPU simulazione:

- $4 \cdot 10^8$ eventi · 400 kSI2k·sec/ev = $1.6 \cdot 10^{11}$ kSI2k·sec

$\Rightarrow$ (per il periodo di presa dati · 10% italia) ~ 1 MSI2k

✓ CPU analisi:

- 15 kSI2k/utente

$\Rightarrow$ (per 100 utenti equivalenti) ~ 1.5 MSI2k

Risorse Storage necessarie:

RAW (1.6 MB/ev) = 2.2 PB - ESD (0.9 MB/ev) = 1.3 PB - AOD (0.2 MB/ev) = 280 TB

✓ Cloud Tier-2:

- 1% RAW + 5% ESD (* 2 vers.) + 100% AOD (* 2 vers.)

$\Rightarrow$ (20 + 130 + 560) TB ~ 700 TB + 2 TB/user (+ calib. + dati temp. + ...)

-Talk 1c -

Attività e Risorse nei Tier-2

Attività nei Tier-2 italiani

Attività nella cloud dei Tier-2 italiani:

✓ Partecipazione alle attività di computing di ATLAS
- Produzione e Ricostruzione dati MC
- Analisi utenti

Fino ad ora

✓ Share risorse e attività ~10% di ATLAS

✓ Nessuna differenziazione significativa tra i vari siti
- stesse quote di dati replicati (~ 25%) anche se differenze nella produzione legate alle risorse disponibili
- Roma1 è uno dei 3 Muon Calibration sites

Nel 2008

✓ share dei pledge 2008 = 5%
- le risorse sono tali però che si possa contribuire comunque con uno share del 10% tutelando contemporaneamente l'analisi italiana (vedi dopo)

✓ Differenziazioni tra i siti
- quote di dati replicati maggiori nei Tier-2 approvati e maggiore attività di calcolo
- Possibilità di definire un Tier-2 più grande degli altri con risorse e maggiori e attività più intense

Risorse disponibili inizio 2008

	CPU		Disco		
	I - 2007 (kSI2k)	II - 2007	I - 2007	II – 2007	Totale inizio 2008
Milano	142	250 CI06	34 TBr 27 TBn	50 TBr 40 TBn	84 TBr 67 TBn
Napoli	90	20 kSI2k ~ 400 CI06	40 TBr 32 TBn	36 TBr 29 TBn	76 TBr 61 TBn
Roma1	118	1300 CI06	30 TBr 24 TBn	36 TBr 29 TBn	66 TBr 53 TBn
LNF	41	1000 CI06	16 TBr 12 TBn	32 TBr 26 TBn	48 TBr 38 TBn

Note:

- I – 2007 e II – 2007 indicano risorse acquisite nella prima e seconda parte (soprattutto con lo sblocco del sub judice) del 2007
- per i nuovi processori la potenza di calcolo viene indicata in CINT2006_Rate (CI06) e il fattore di conversione a SI2k dipende dal processore stesso, in alcuni casi non esiste e dobbiamo ricavarlo noi
- Sono state considerate le dismissioni di macchine obsolete

Proposta dei referee

Forti – CSN1 – Sett 07

- Sbloccare i SJ 2007
  - Vedi piano alle slide successive
- Per il 2008:
  - Assegnare 1/3 * 1.5M€ SJ ai risultati di un workshop da tenere a gennaio che chiarisca:
    - Attività degli esperimenti
    - Scelte architetturali (disco e rete)
    - Piano dettagliato degli acquisti
  - E' essenziale che il SJ 2008 (detto primavera 2008) possa venire sbloccato nella riunione di fine Gennaio, per permettere agli esperimenti di acquistare il materiale in tempo per l'estate
  - Riservare 2/3 * 1.5M€ in una tasca indivisa da assegnare quando la schedule di LHC è più chiara
- Note:
  - CMS è forse un po' più pronto di Atlas, ma non ci sembra ci siano ancora gli estremi per spendere i fondi 2008
  - Per LHCb non ci sono invece dubbi infrastrutturali. Si propone un'assegnazione di 35 k€ su BO per CPU al CNAF.

Proposte ATLAS

Forti – CSN1 – Sett 07

Atlas	CPU Disponibil ora	Acquisizioni da fare sj 2007	prim. 2008	Totale	DISCO Disponibil ora	Acquisizioni da fare sj 2007	ass. 2008	Totale
	kSI2k	kSI2k	kSI2k	kSI2k	TBN	TBN	TBN	TBN
Roma1	140	63	94	**297**	42	25	43	**110**
Napoli	92	63	94	**249**	37	25	43	**105**
Milano	129	40	62	**231**	32	16.5	29	**77.5**
LNF	41	22	31	**94**	21	7	14	**42**
Tot Atlas	**402**	**188**	**281**	**871**	**132**	**73.5**	**129**	**334.5**

S.J. 2007	CPU kEuro	Disco kEuro	Totale kEuro
Roma1	17	45	**62**
Napoli	17	45	**62**
Milano	11	29	**40**
LNF	6	13	**19**
Tot Atlas	**51**	**132**	**183**

primavera 2008	CPU kEuro	Disco kEuro	Totale kEuro
Roma1	15	60	**75**
Napoli	15	60	**75**
Milano	10	40	**50**
LNF	5	20	**25**
Tot CMS	**45**	**180**	**225**

Finanziamenti 2008

- ✓ 1.5 M€ in totale per ATLAS e CMS
- ✓ ~ 1/3 s.j. da sbloccare con il workshop = 225 k€
  - ✓ 1/4 CPU = 45 k€
  - ✓ 3/4 Disco = 180€
- ✓ Suddivisione tra i siti:
  - ✓ 30% per Milano, Napoli e Roma1
  - ✓ 10% per Frascati

Possibili variazioni delle percentuali CPU/disco o delle suddivisioni tra i siti per livellare eventuali differenze

In tal caso i Tier-2 approvati avrebbero a disposizione circa 100 TB ognuno

Tier-2 pilota nel 2008

I Referee chiedono che un Tier-2 sia dotato di maggiori risorse per testare la funzionalità delle soluzioni adottate al crescere del carico di lavoro.

E' da verificare soprattutto il sistema di storage

- ✓ Architettura hardware: DAS vs SAN
- ✓ Middleware di Gestione (SRM): DPM vs STORM/GPFS

Richiedono che venga fatto un test di scalabilità.

- ✓ Dopo le acquisizioni della prima tranche del 2008 ogni Tier-2 approvato avrebbe un volume di storage di ~100 TB.
- ✓ In Atlas siti che adottano lo stesso sistema di storage dei Tier-2 italiani (Glasgow) hanno dimostrato la scalabilità e la perfetta funzionalità del sistema.
- ✓ E' necessario quindi effettuare test su volumi maggiori.

Tier-2 pilota nel 2008

Proposte iniziali dei Referee:

1. Destinare tutte le risorse ad un unico Tier-2
2. Proporre ai Tier-2 di Roma1, ATLAS e CMS, di uniformare le scelte di storage

Per diversi motivi nessuna è per noi accettabile o praticabile

Possibile soluzione:

- ✓ Fornire ad un Tier-2 le (notevoli) risorse di calcolo dismesse dal CNAF per i noti problemi di potenza frigorifera della sala calcolo ma non obsolete
- ✓ Sbloccare una parte dei finanziamenti previsti per la seconda metà del 2008 per garantire una crescita significativa del volume di storage
  - ✓ La quantità di risorse di calcolo (e quindi rack) da installare dipenderà dal volume totale di storage a disposizione in modo da conservare il corretto bilanciamento CPU/dischi
- ✓ Tale Tier-2 effettuerà nel primo semestre del 2008 i test che riterremo significativi per definire l'architettura del sistema di storage e di rete da adottare

Sito proposto: Milano

- ✓ Maggiore flessibilità infrastrutturale per ospitare un notevole aumento di risorse
- ✓ Interesse del Servizio Calcolo

- Talk 3a -

La Federazione Italiana dei Tier-2

La Federazione Italiana dei Tier-2

Organizza la partecipazione INFN alle attività di servizio del computing di ATLAS con lo scopo di ottimizzare le risorse disponibili nel Tier-1 e nei Tier-2 e fornire supporto all'analisi in Italia. Rappresenta la comunità computing ATLAS Italia verso WLCG (CB) e verso ATLAS per le attività di servizio

- ✓ Attività primaria nel Production System (sviluppo, gestione, operazione e shift)
  - Tutorial a Milano nel settembre 2006 (primo in ATLAS)
- ✓ Monitoring dei siti (servizi di Grid e servizi locali)
- ✓ Accounting dei siti (HLR)
- ✓ Messa a punto e controllo di DDM e componenti in Italia
- ✓ Distribuzione dei dati in Italia: interfaccia per le sottoscrizioni e gestione trasferimenti
- ✓ Analisi distribuita: test e utilizzo di GANGA
  - Tutorial a Milano, feb 2007 (secondo in ATLAS) e scuola Grid, nov 2007

Struttura della Federazione

- ✓ Responsabile della Federazione
  - ✓ Gianpaolo Carlino dal 2008, in sostituzione di Laura Perini dal 2006
  - ✓ in origine si prevedeva una rotazione della responsabilità tra i Tier-2. La proposta attuale è di far coincidere questa figura con quella del Coordinatore Nazionale del Calcolo di ATLAS visto l'effettiva sovrapposizione dei compiti e per fare in modo che possa essere eletto dall'intera comunità di ATLAS Italia.
- ✓ Deputies
  - ✓ Laura Perini: in qualità di responsabile uscente
  - ✓ Alessandro De Salvo: technical coordinator
  - ✓ .....: collegamento con l' Atlas Distributed Computing (ADC) group
  - ✓ nominati dal Responsabile della Federazione
- ✓Tier-2:
  - ✓ Milano:
    - ✓ Responsabile: Laura Perini
    - ✓ Responsabile operativo: Attilio Andreazza
  - ✓ Napoli:
    - ✓ Responsabile: Gianpaolo Carlino
    - ✓ Responsabile operativo: Alessandra Doria
  - ✓ Roma 1:
    - ✓ Responsabile: Lamberto Luminari
    - ✓ Responsabile operativo: Alessandro De Salvo
  - ✓ Frascati (proto Tier-2):
    - ✓ Responsabile: Mary Censa Ferrer
    - ✓ Resposabile Operativo: Elisabetta Vilucchi

Struttura operativa al Tier-1

- ✓ .....

Personale attivo nella Federazione

Frascati:
- Mary Censa Ferrer (responsabile Tier2)
- Elisabetta Vilucchi (gestione del Tier2, ADC shifter, DB calibrazione)
- Wen Mei (ADC senior shifter, DDM)
- Agnese Martini (produzione: shifter)
- Claudio Soprano (amministratore Tier-2)

Napoli:
- Gianpaolo Carlino (responsabile Tier2)
- Alessandra Doria (responsabile operativo Tier2, ADC senior shifter, monitoring)
- Leonardo Merola (coordinamento Tier2 in PoN SCoPE)
- Michela Biglietti (analisi distribuita)
- Francesco Conventi (analisi distribuita)
- Elisa Musto (ADC shifter, installazione software dal 2008)
- Sergio Ricciardi (monitoring e infrastruttura rete)

CERN:
- Simone Campana (ATLAS production coordinator)
- Alessandro Di Girolamo (DDM operations e monitoraggio storage cloud italiana, SAM test)

CNAF:
- Claudia Ciocca (DDM Italian Cloud Manager)
- Lorenzo Rinaldi (dal 2008, DDM)

Milano:
- Laura Perini (responsabile Tier2)
- Silvia Resconi (produzione e ADC senior shifter, analisi distribuita)
- Guido Negri (produzione e ADC senior shifter, responsabile GGUS)
- David Rebatto (produzione: sviluppatore Lexor e sottomissione)
- Luca Vaccarossa (gestione del Tier2)
- Elisabetta Molinari (experimental services WMS per la produzione)
- Leonardo Carminati (analisi distribuita)
- Tommaso Lari (analisi distribuita)

Roma I:
- Lamberto Luminari (responsabile Tier2)
- Alessandro De Salvo (responsabile op. Tier2; VO manager; installazione e validazione sw)
- Alex Barchiesi (gestione Tier2)
- Daniela Anzellotti (gestione Tier2 e amministratore DB calibrazione)

Bologna:
- Franco Brasolin (SIT, dal 2008)

Roma III:
- Fulvio Galeazzi (Tier-3 task force)

Struttura della Federazione

Rapporti con i gruppi ATLAS

- ✓ E' necessario rafforzare il legame tra la Federazione dei Tier-2 e i vari gruppi in quanto l'attività di computing è trasversale e il buon funzionamento dell'intera struttura è nell'interesse di tutti.
  - ✓ Aumentare la diffusione delle informazioni sul computing nei gruppi italiani
    - Esempio: utilizzo e distrubuzione dei dati di M5
  - ✓ Il personale dei Tier-2 non è in grado di gestire da solo contemporaneamente i Tier-2 e tutte le operazioni di computing in italia
    - manpower molto limitato
    - attività molto impegnative e intense

- ✓ E' necessario:
  - ✓ aumentare la collaborazione del personale "non Tier-2"
    - ✓ per attività di computing
      - esempio: definizione benchmark per il test dei processori (Bologna)
    - ✓ per attività strettamente connesse all'analisi
      - analisi distribuita
      - distrubuzione dei dati

Struttura della Federazione

✓ Definizione di una mailing-list con almeno un rappresentante per gruppo/sezione:

- ✓ persone interessate e sufficientemente competenti di computing
- ✓ rappresentano e coordinano le attività di computing nei gruppi

Richiesta per i capi gruppo di identificare e indicarmi queste persone

✓ Riunioni

- ✓ Riunioni telefoniche bisettimanali sulle attività della federazione e lo stato dei siti e delle operazioni.
  - Riunioni abbastanza tecniche
  - Mi aspetto una partecipazione dei rappresentanti dei gruppi

- ✓ Proposta di avere riunioni dedicate su argomenti di interesse più generale con scadenza da definire in base alle necessità contingenti
  - Prima possibilità un incontro sullo stato dell'analisi distribuita in italia (Milano ?)

- ✓ Presentazione dello stato generale delle attività o di attività specifiche in Atlas Italia

- Talk 3b -

Analisi nei Tier-2 italiani

Analisi nei Tier-2

1. I Tier-1 e i Tier-2 di ATLAS sono risorse comuni disponibili per l'intera collaborazione
2. Il Computing Model prevede che le risorse dei Tier-2 siano dedicate al 50% per la simulazione e al 50% per l'analisi

Bisogna trovare un modo per garantire l'uso delle risorse dei Tier-2 italiani alla comunità italiana impedendo che le attività centrali di ATLAS o gli utenti non italiani le usino in maniera predominante

1. Creazione di un gruppo **atlas/it** a livello di VO
2. Job Priority Mechanism:
   - definizione di quote dedicate per le varie attività (p.es. produzione 50% e analisi 50%) e i vari gruppi (atlas e atlas.it)
3. Fair-Share Mechanism per ottimizzare l'uso delle risorse
   - bilanciamento temporale dell'uso delle risorse per impedire che rimangano inutilizzate quando non viene utilizzata completamente la quota dedicata ad una precisa attività o a un gruppo

Analisi nei Tier-2

- Aspetto Politico: OK
  - purché non venga (potenzialmente) limitato in maniera eccessiva l'utilizzo dei Tier-2 ai non italiani
  - altre nazioni adottano già soluzioni di questo tipo

- Aspetto Tecnico: OK .....
  - soluzione già prevista a livello di VO. Deve essere implementata
  - mapping delle credenziali VOMS dei gruppi/ruoli con quelle locali degli scheduler
  - associazione delle quote a queste credenziali

..... però richiede un lavoro non banale di riconfigurazione dei sistemi di sottomissione dei job nei siti

La proposta è di dedicare inizialmente il 30% delle risorse di calcolo per l'analisi agli utenti italiani.
Monitoraggio accurato nel 2008 per verificare che questa quota soddisfi le necessità della comunità.

Analisi nei Tier-2

Suddivisione delle attività nei Tier-2

- Notevoli miglioramenti tecnici nel sistema di distribuzione dei dati e nei tool di analisi distribuita:
  - aumento dell'efficienza del sistema di sottoscrizioni
  - Ganga permette
    - di gestire in maniera efficiente l'analisi su dataset non completi attraverso un nuovo sistema di definizione dei sub-job
    - di definire una serie di siti preferenziali (cloud italiana) su cui eseguire i job
- i dati saranno divisi in stream inclusive e la maggior parte delle analisi necessita dei dati appartenenti a stream diverse

- per l'utente sarà indiffente il sito su cui lanciare le proprie applicazioni
- suddivisione dei dati nei Tier-2 solo in base alla percentuale delle risorse disponibili
- per le generiche attività di analisi non è necessario definire un rapporto preciso tra gruppi e siti (comunità di riferimento)
- rimane una corrisponenza per attività specifiche come calibrazioni e studi di rivelatori che richiedono dati in formato RAW o ESD.

Tier-3 Task Force

- Dalle pagine Twiki: *"The Task Force was created to help document requirements to facilitate setting up Tier-3 for ATLAS use."*
- Costituita durante la Glasgow Week, composta da 12 elementi, coordinati da Stephen Gowdy: SLAC, Nikhef, Munich, BNL, UTA, Roma3 (Fulvio Galeazzi), Valencia, Annecy, Lancaster, Desy, Cern, Bogota
  - HN Forum dedicato e Riunioni settimanali ogni venerdi
- Rapporto iniziale sulle attivita' verra' presentato alla Atlas Week di Febbraio...
  - ...subito dopo il "Tier3 Workshop" organizzato a fine Gennaio
- Scopo della Task Force
  - Individuare physics analysis use-cases, ipotizzando anche siti di dimensioni diverse
  - Predisporre raccomandazioni e documentazione su come installare e gestire un Tier-3 o una Analysis Facility
    - Questo includera' anche le stime di necessita' CPU, disco, software, personale
- Cosa è stato fatto fino ad ora
  - Serie di presentazioni per mostrare quello che esiste o e' in corso di realizzazione (es. Roma3) nei vari siti
  - Iniziata discussione su vari argomenti: tipo di storage, modalita' di copia dei dati, Xrootd (Scalla), PROOF, ecc.
- Orizzonte temporale: ~fine dell'estate
  - Il futuro della Task Force e' di trasformarsi in un Working Group all'interno del gruppo "Atlas GRID, Tools and Services"

- Talk 3c -

Supporto per gli utenti

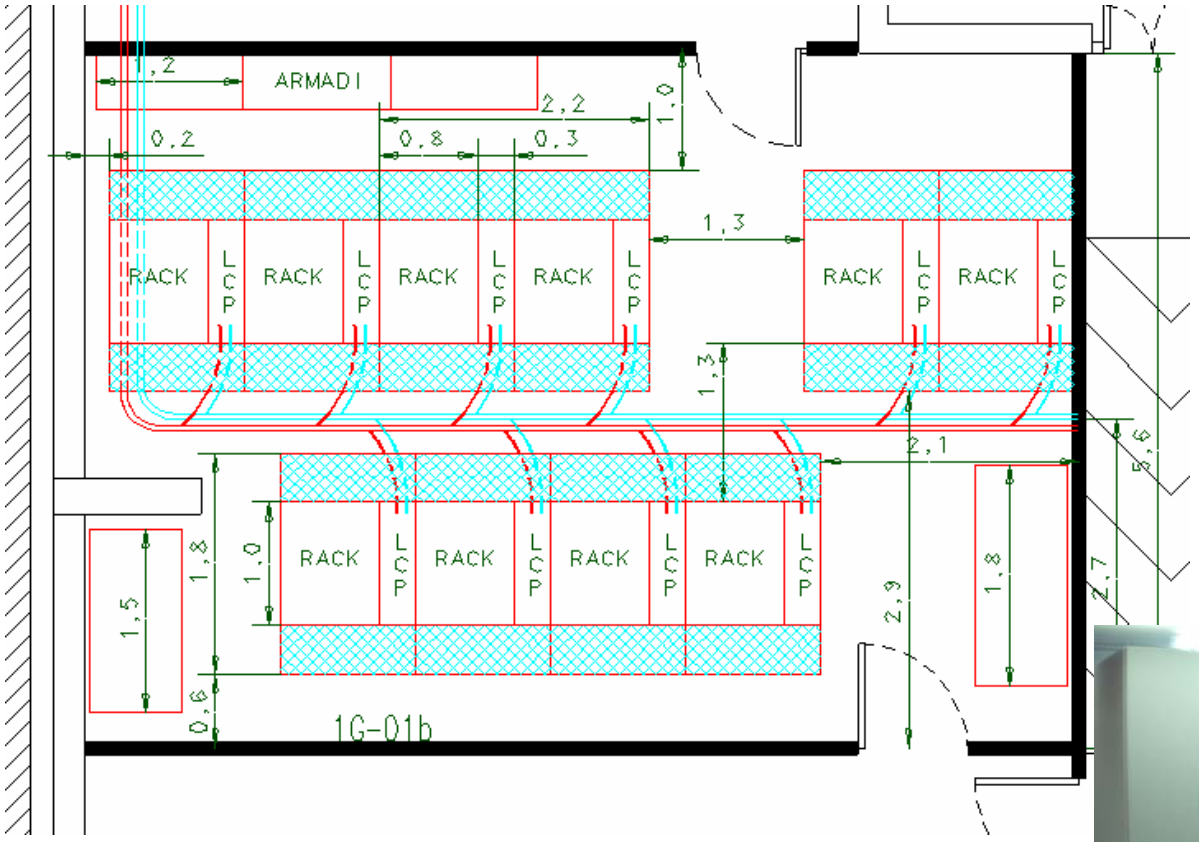
Supporto e Feedback utenti

,,,,,

- Talk 2a -

Infrastrutture dei Tier-2 italiani

Il Tier-2 di Napoli



Sala ATLAS INFN
- ✓ Superficie 44 m2

Nella seconda metà 2008 disponibile Sala PON SCoPE
- ✓ Superficie 120m2
- ✓ Capacità 120 Rack (10 Tier-2)

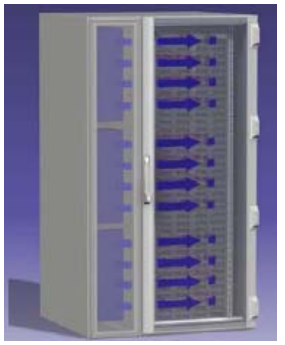
4 Rack installati attualmente:
- ✓ 2 Tier-2 ATLAS e 2 PON SCoPE

Espansione fino a 10 Rack
- ✓ Impianti dimensionati per tale capacità

Il Tier-2 di Napoli

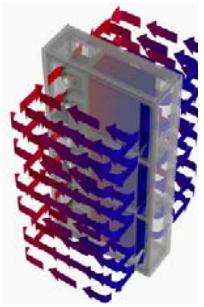
Impianto Elettrico:

- ✓ Max potenza disponbile: 250 kW
- ✓ 2 Gruppi di continuità da 60 kVA in parallelo. Autonomia a pieno carico 7'. In corso installazione sistema di videosorveglianza
- ✓ Monitoraggio remoto dei parametri elettrici dell'armadio di zona
- ✓ Ad ogni rack arriva una linea elettrica trifase da 22KW
- ✓ Gruppo elettrogeno verrà installato entro la metà del 2008



Impianto termico:

- ✓ Chiller con capacità di raffreddamento di 90 kW, due compressori indipendenti
- ✓ Rack autoraffreddanti RIMatrix della Rittal con potenza dichiarata di 12kW espandibile a 20 KW modificando la temperatura e i flussi dell'acqua
- ✓ Raffreddamento ambientale della sala garantito da due unità da 6 KW



Impianto Antincendio:

- ✓ Doppio sistema antincedio:
  - ✓ Protezione dei rack
    - Centralina che attraverso una coppia di rivelatori per rack (in AND) attiva la scarica all'interno dei rack stessi
  - ✓ Protezione della sala
    - Analogo funzionamento ma i sensori sono distribuiti nella sala dove avviene la scarica

Il Tier-2 di Roma1

Nuova sala disponibile da fine Novembre 2007

- ✓ Dimensione sala 60 m2 espandibile fino a oltre 120 m2
- ✓ 4 rack attualmente installati (2 per ATLAS e 2 per CMS), 3 ordinati e in consegna a marzo 2008
- ✓ Capacità della sala: 14 rack con gli attuali impianti, fino a 21 modificando la rete idraulica (progettata per questa eventualità)

Impianto termico:

- ✓ Rack autocondizionati ad acqua della Knuerr
- ✓ Max potenza per rack: 17kW
- ✓ 2 chiller da 80 KW ognuno con doppia pompa indipendente

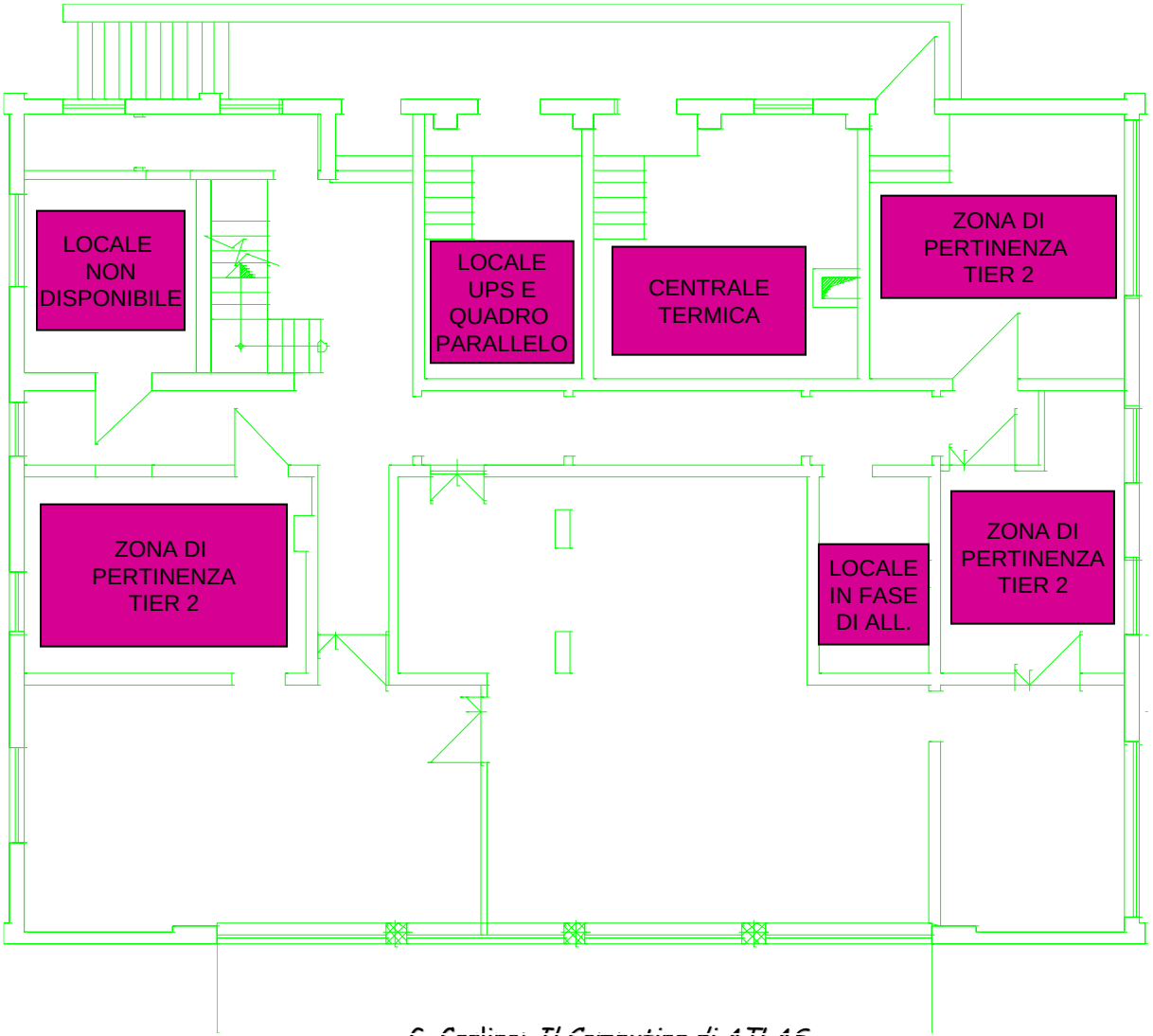
Impianto Elettrico:

- ✓ Max potenza disponibile: 360
- ✓ UPS da 120 KVA, un secondo simile in consegna a marzo 2008 con autonomia di 10' a pieno carico

Impianto Antincendio:

- ✓ Impianto a gas inerte (non tossico per le persone) che agisce sull'intera sala macchine e all'interno dei rack.
- ✓ Sensori posti sia nella sala che all'interno dei rack
- ✓ La centralina di controllo è situata al'interno della sala macchine verrà collegata con un sistema di allarmistica alla vigilanza dello stabile (DeltaPol)

La Sala Macchine e gli spazi per il Tier-2



Il Tier-2 di Milano

Impianto termico:

- ✓ Il sistema di condizionamento realizzato per l'intera sala è costituito da due macchine che asportano via il calore in grado di smaltire 90 kW termici ognuna
- ✓ Modifiche al sistema di distribuzione dell'aria sono già previste per ottimizzarlo
- ✓ Max potenza per rack: 17kW
- ✓ Espandibilità fino a XX Rack, YY ordinati e in consegna a marzo 2008
- ✓ 2 chiller da ZZ KW con doppia pompa indipendente, funzionanti in modalità fail-over

Impianto Elettrico:

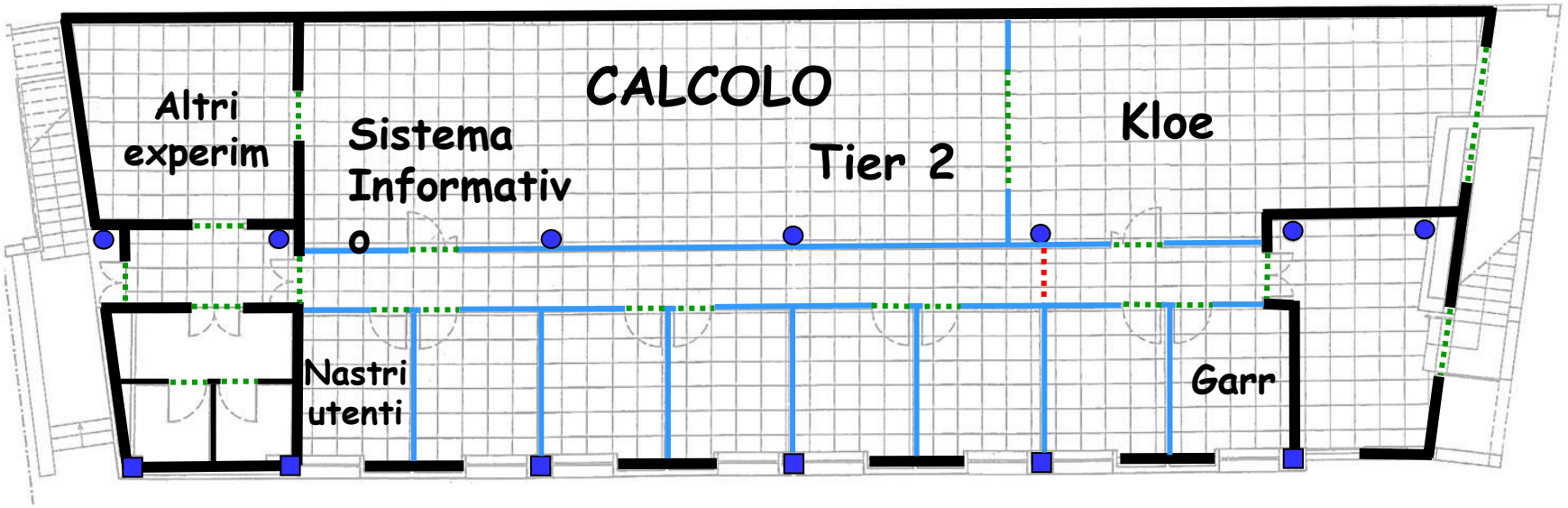
- ✓ Gruppo di continuta da 200 KVA corrispondenti a 160 KW, autonomia 15'.
- ✓ Ordinato un gruppo elettrogeno da 400 KVA in esclusivo uso della sala macchine, in grado di sopperire alle esigenze della parte elettrica e del sistema di raffreddamento. Autonomia 11 ore.

Impianto Antincendio:

- ✓ Il sistema attualmente installato non copre tutte le zone previste, nel prossimo anno è prevista la sua revisione e la sostituzione dell'estinguente attualmente non più a norma

Il proto-Tier-2 di Frascati

- La sala che ospita attualmente il proto-Tier2 e' situata al pian terreno di un edificio a due piani che ospita il servizio di calcolo dei LNF, una libreria a nastro dell'esperimento Kloe, il sistema informativo dell'INFN ed il POP GARR dell'area di Frascati.
- Superficie 97 m2.
- Il Tier-2 occupa attualmente due rack e può essere espanso con altri 4 rack



Il proto-Tier-2 di Frascati

Impianto elettrico:

- ✓ Potenza attualmente necessaria: 15 kW (Atlas) + 40 kW (altre risorse)
- ✓ UPS da 160 KVA, autonomia 30'
- ✓ Gruppo elettrogeno da 120 kW dopo un minuto

Impianto termico:

- ✓ L'impianto di raffredamento esistente e' a circolazione d'acqua ricavato deviando una parte del condizionamento di Dafne

Impianto Antincendio:

- ✓ Impianto a gas inerte (FM200) dimensionato tenendo conto della destinazione d'uso e dimensione dei vari ambienti

Il proto-Tier-2 di Frascati

Strategia del LNF riguardo al proto Tier-2 di ATLAS

Il Direttore dei Laboratori ha espresso interesse per avere un centro di calcolo scientifico di cui il Tier-2 di ATLAS farà parte, e ha chiesto al coordinatore di Gruppo I di formare in proposito un gruppo di lavoro.

C'è l'impegno del coordinatore a fornire conclusioni preliminari entro due mesi.

- Talk 2b -

Sistemi di Monitoraggio e Gestione delle emergenze

Monitoring di infrastutture e servizi

Sistemi di monitoring, allarmistica e gestione:

- Monitoring dei servizi grid, allarmi: **SAM test**
- Monitoring risorse e servizi, allarmi: **Nagios**
- Monitoring risorse e servizi: **Ganglia**
- Monitoring ambientale:
- Gestione Emergenze: script automatici di spegnimento e/o accendimento delle farm

Test SAM

Obiettivo: Controllo dei servizi di GRID

- Test centralizzati
- Tipologia dei test
  - sottomissione di job ai siti
  - replica di dati
  - verifica certificati e versioni del middleware
  - periodicità circa 2 ore
- Test sia Atlas specifici sia per le VO di test (dteam/ops)
- In caso di fallimenti invia e-mail agli amministratori dei siti
  - in caso di non risoluzione del problema il sito viene inserito in una blacklist

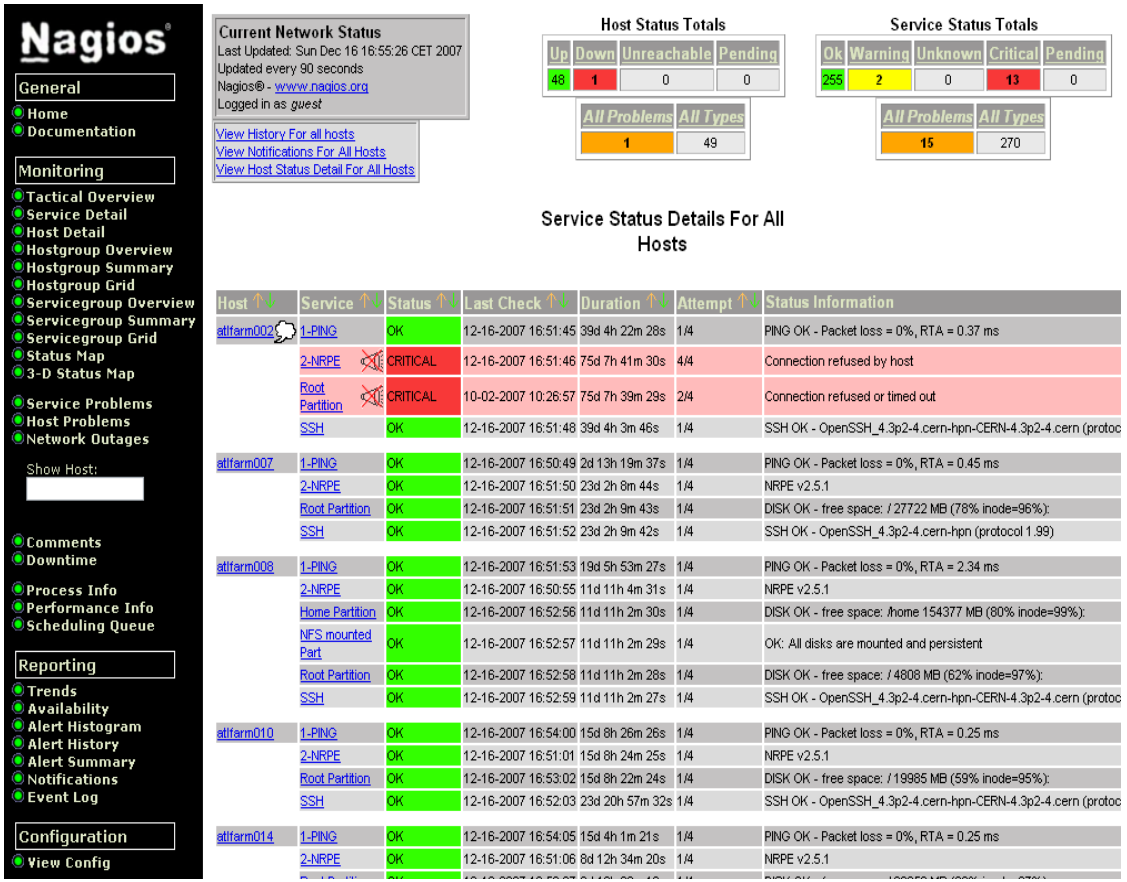
Nagios

Obiettivo: Monitorare **servizi** locali e **risorse** hw/sw

- Nagios è un sistema di monitoraggio *non grid aware* ma che consente di monitorare ogni aspetto del sito grazie a plugin lanciati periodicamente sugli host da monitorare
- Possono essere configurati controlli a piacere ed azioni da intraprendere in caso di fallimenti
- Permette di notificare agli amministratori del sito in caso di situazione anomale (invio e-mail, sms ..)

Monitoring Risorse:

- Stato degli host up/down (ping)
- Carico delle CPU
- Carico della memoria centrale e swap
- Spazio dischi interni e arraw
- Numero degli utenti sulla macchina
- Temperatura interna della macchina (ove possibile)



Monitoring Servizi:

- SSH e NFS
- Area del sw di esperimento montata e disponibile ai WN
- demone SQL
- Code: job running e in coda

Ganglia

Obiettivo: Monitorare l'andamento corrente e storico di job e risorse

- utilizzato largamente per visualizzare lo stato del cluster e il suo andamento nel corso del tempo
- Conserva i dati dell'ultimo anno
- Permette di scrivere plugin e metriche ad-hoc
- Non consente di inviare notifiche in caso di situazioni anomale

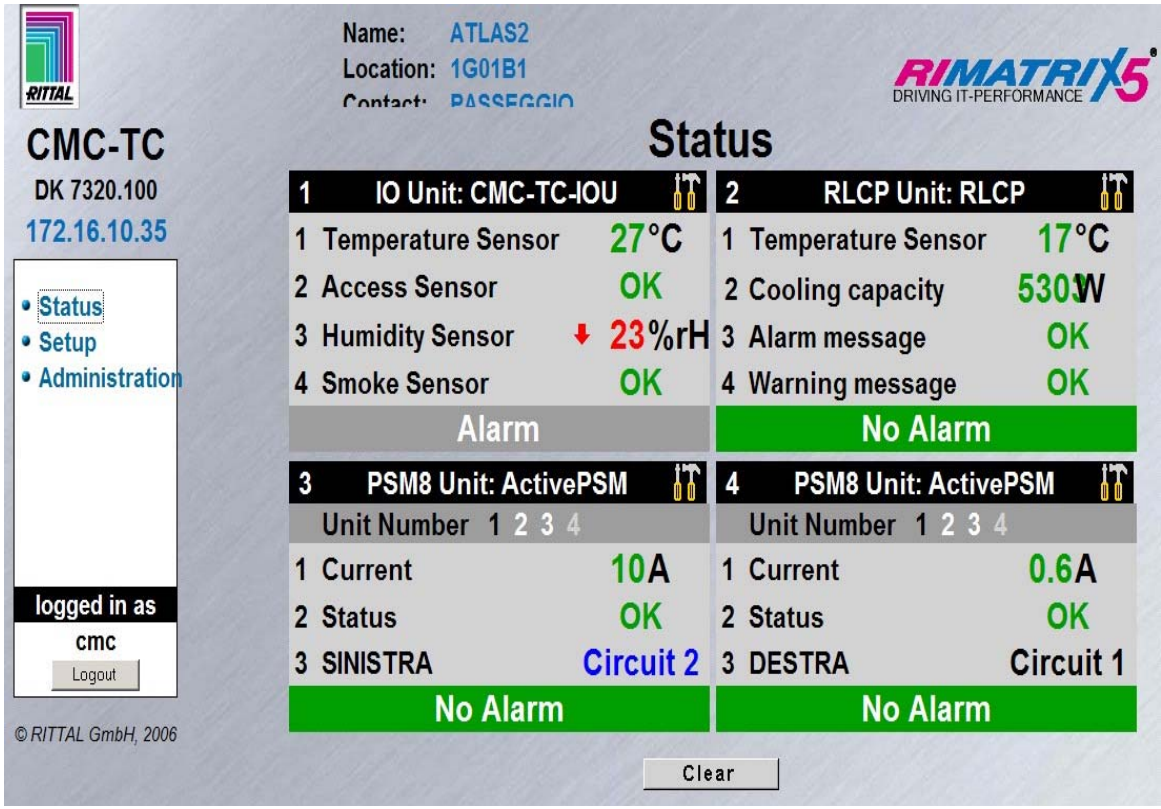
Monitoraggio ambientale

Obiettivo: Monitoraggio dei parametri ambientali e elettrici

Esempio del sistema di controllo CMC dei rack Rittal di Napoli:

Parametri monitorati:

- ✓Temperatura dell'acqua in ingresso ai moduli di raffreddamento
- ✓ Portata dell'acqua
- ✓Temperatura aria in ingresso e uscita
- ✓ Umidità nei rack
- ✓ Presenza di fumi, fiamme, allagamento
- ✓ Corrente assorbita dalle singole prese intelligenti



Tutti questi parametri possono essere letti e monitorati da remoto grazie alle unità CMC in grado di mandare avvisi o allarmi in vario modo come email, sms, trap snmp nonché ovviamente avvisi sonori e ottici

Gestione emergenze

Obiettivo: spegnimento e accensione automatici di farm e sistemi di calcolo con gestione delle emergenze

- Sono in fase avanza di sviluppo delle procedure di gestione delle emergenze che si basano sull'azione di script automatici per lo spegnimento dei sistemi di calcolo
- le procedure di riaccensione delle farm possono essere automatiche attraverso gli stessi script o manuali

Le procedure possono essere inizializzate da:

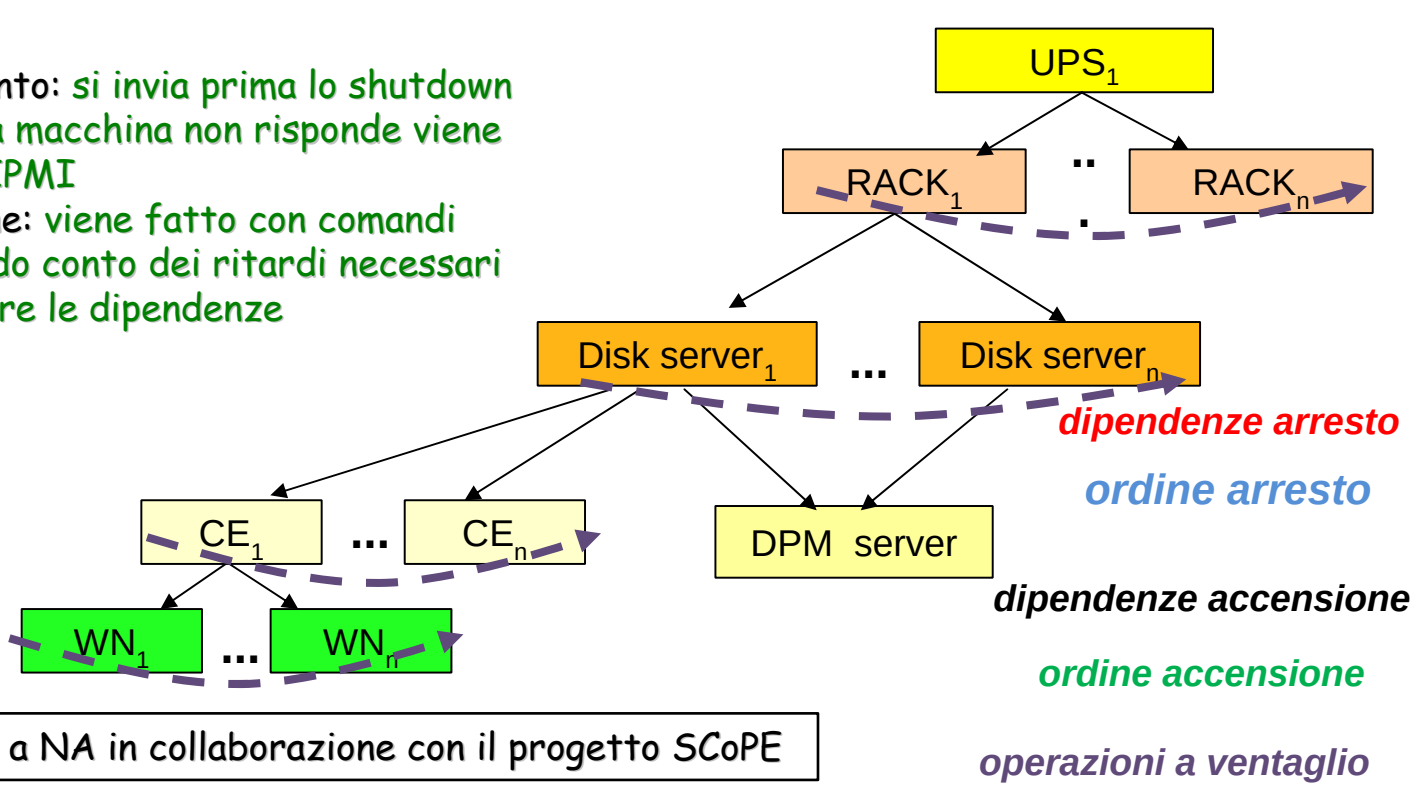
1. Mancanza / Ritorno corrente agli UPS
2. Valori fuori range dei sensori ambientali
3. Operazioni manuali di manutenzione

Gestione emergenze

Powerfarm: esegue le azioni di spegnimento e accensione

- Spegnimento totale o parziale degli elementi e riaccensione quando le condizioni lo permettono
- esegue le azioni nell'ordine e nei tempi opportuni rispettando le dipendenze funzionali tra i dispositivi
- L'esecuzione della procedura può essere invertita in tutta sicurezza al sopraggiungere dell'opportuna condizione (ritorno alimentazione)
- ottimizzazione dei tempi (operazioni parallele "a ventaglio" ove possibile)

- Spegnimento: si invia prima lo shutdown via ssh, se la macchina non risponde viene spenta con IPMI
- Accensione: viene fatto con comandi IMPI tenendo conto dei ritardi necessari per rispettare le dipendenze



Sviluppato a NA in collaborazione con il progetto SCoPE

Gestione emergenze

Servers check utility: Ping Test Ssh Test

PING TEST Sistema del Centro di Calcolo di Milano

Shutdown automatico

Group level 0	Group level 1	Group level 2	Group level 3	Group level 4	Group level 5	Group level 6
backupt2	pcgamma2	pcalcolo	mannolo	grid001	grid020	magutt
pcatlpixel	pcica9	zdf	cucciolo	t2-ce-01.mi.infn.it	ds20bib	grid006
prod-hlr-01	helena	webced	dotto	obelix	pennyblack	grid013
t2-se-01.mi.infn.it	pcfluka	studenti	catorcio	nercurio	smtp1	jag
dpmdisk01	pcfluka2	web	encrypt	lxmi	smtp2	coco
egee-rb-01.mi.infn.it	pcica7	palantir	wgate	ptt	smtp3	minerva
atlfarm001.mi.infn.it	hector	labmaster	zufola	biancaneve	bibo	Group Ended
atlfarm003.mi.infn.it	nasone	ckptmi	m04	pl01	bibi	
atlfarm006.mi.infn.it	pcjolly	linuxbox	innominato	pl02	Group Ended	
atlfarm010.mi.infn.it	ds20gr3	bbq	news	pl03		
atlfarm011.mi.infn.it	pcbabar9	pcbackup	m05	pl04		
atlfarm013.mi.infn.it	etsfmi	forecast	aramis	pl05		
atlfarm014.mi.infn.it	dstrare	pga	atlfarm008	pl06		
atlfarm015.mi.infn.it	pcteserver	pista	pcalcolo2	pl07		
grid002.mi.infn.it	mason	athos	test-host-down	pl08		
grid005.mi.infn.it	portos	Group Ended	Group Ended	panoramix		
grid008.mi.infn.it	Group Ended			[illegible]		
grid009.mi.infn.it				pl09v01		
grid012.mi.infn.it				Group Ended		
grid015.mi.infn.it						
grid016.mi.infn.it						
grid017.mi.infn.it						
grid018.mi.infn.it						
grid019.mi.infn.it						
grid021.mi.infn.it						
grid022.mi.infn.it						
grid023.mi.infn.it						
grid024.mi.infn.it						
grid025.mi.infn.it						
grid026.mi.infn.it						
t2-wn-02.mi.infn.it						
t2-wn-03.mi.infn.it						
t2-wn-04.mi.infn.it						
t2-wn-05.mi.infn.it						
t2-wn-06.mi.infn.it						
t2-wn-07.mi.infn.it						
t2-wn-08.mi.infn.it						
t2-wn-09.mi.infn.it						
t2-wn-13.mi.infn.it						
t2-wn-14.mi.infn.it						
t2-wn-15.mi.infn.it						
t2-wn-16.mi.infn.it						
t2-wn-17.mi.infn.it						
t2-wn-18.mi.infn.it						
t2-wn-19.mi.infn.it						
t2-wn-21.mi.infn.it						
t2-wn-22.mi.infn.it						
t2-wn-23.mi.infn.it						
t2-wn-24.mi.infn.it						
Group Ended						

- Server di shutdown dedicato (blindato) con chiave ssh su ogni macchina;
- La procedura, attivabile anche manualmente, e' basata sulle risposte a interrogazioni snmp all' UPS (200 kVA);
- I nodi (125) sono divisi in gruppi. Lo shutdown (ed il restart) avvengono per gruppi per salvaguardare eventuali dipendenze (es. mounting NFS);
- Lo shutdown comincia dopo 20 minuti di interruzione della fornitura di energia dalla linea primaria. L'ultimo gruppo inizia lo spegnimento dopo 30 minuti.
- La procedura tiene conto dell'attuale carico sull'impianto elettrico e del fatto che non e' ancora installato il gruppo elettrogeno.
- Allarmistica di down elettrico sia via mail che sms;
- Programma con interfaccia web per controllo sullo status complessivo delle macchine inserite nella procedura;
- Restart manuale per gruppi (in futuro tramite protocollo ipmi o wake on lan). Controllo della procedura attraverso il programma sopracitato.

- Talk 4a -

Attività di Commisioning del Computing nel 2007

Sistema di Distribuzione dei Dati (DDM)

Il sistema di Distributed Management (DDM) di ATLAS, Don Quijote (DQ2), implementa tutte le funzionalità previste dal Computing Model relative alla:

- Distribuzione di dati raw e ricostruiti, reali e simulati, tra i vari Tier

Il sistema, ha un'organizzazione basata sui datasets:

- Cataloghi di dataset centrali, suddivisi in vari DB per facilitare l'accesso
  - Dataset Repository, Dataset Content Catalog, Dataset Location Catalog, Dataset Subscription Catalog
- Cataloghi di file distribuiti (locali)
  - mapping nome logico↔ nome fisico: **LFC** (LCG File Catalog) al Tier1

Trasferimento dei file attraverso il Sistema di Sottoscrizione:

- T0 → T1 e T1 → T1 (trasferimenti tra clouds)
- T1 → T2 e T2 → T1 (trasferimenti nella cloud)

Upgrade versione 0.3 nel Giugno 2007

Sistema di Distribuzione dei Dati (DDM)

…… Fondamentale la collaborazione con gli utenti !!

Ma gli utenti devono essere rassicurati che il sistema di trasferimento dei dati può funzionare con efficienza e velocità e non è necessario reperire i dati con mezzi alternativi

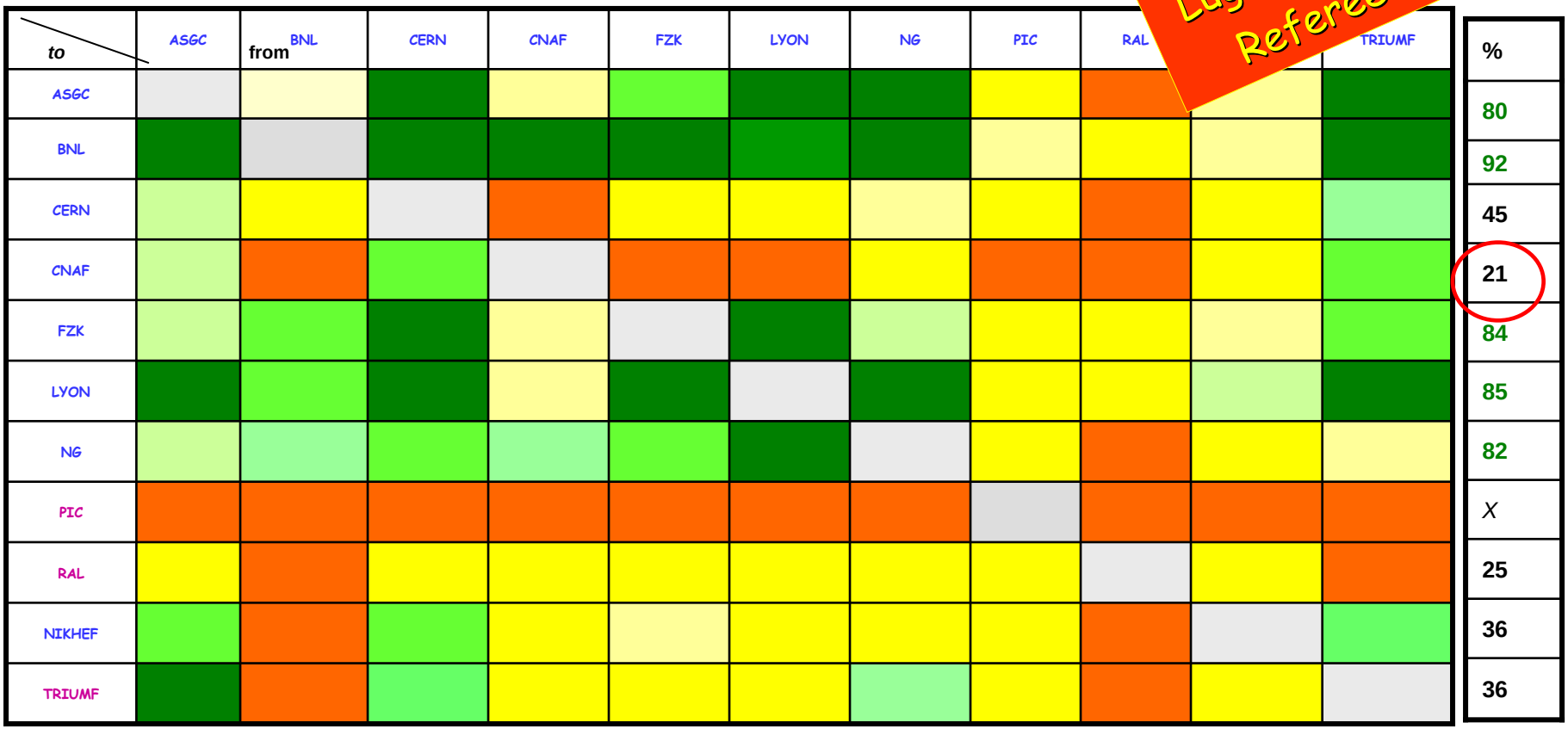
Breve riassunto dei progressi negli ultimi mesi:

- Fino a Luglio 2007 – situazione drammatica
- Estate 2007 – miglioramenti grazie a nuove versioni di DQ2 e del sistema di storage al CERN e al CNAF (Castor) e soprattutto all'attenzione continua sia in Atlas Italia che al CNAF
- Autunno 2007 – test del sistema con risultati più che incoraggianti, passaggio a un nuovo sistema di storage al CNAF (STORM)

Sistema di Distribuzione dei Dati (DDM)

- Distribuzione tra i Tier1 di AOD e Ntuple
- Data Replication Period Feb – Jun 2007, DQ2 0.2
- Data Volume: 3200+ datasets, 570+ Kfiles, 23+ TB
- Target: efficienza 100%

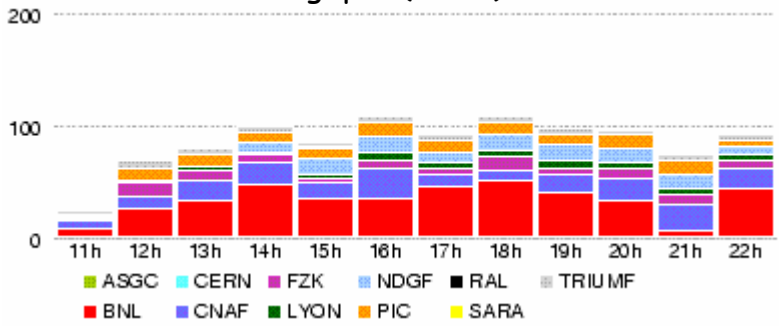
Luglio 07 – incontro Referee ATLAS



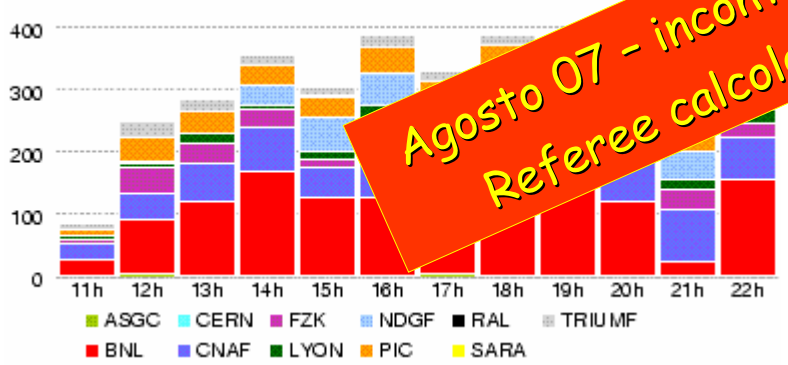
no AODs consolidation within the cloud or/and replication was stopped

Sistema di Distribuzione dei Dati (DDM)

Throughput (MB/s)



Data Transferred (GB)



Agosto 07 - incontro Referee calcolo

	Transfers			
Cloud	Efficiency	Throughput	Files Done	Datasets Done
ASGC	7%	0 MB/s	442	0
BNL	93%	35 MB/s	13781	1796
CERN	0%	0 MB/s	0	0
CNAF	89%	17 MB/s	11476	204
			Click on the site name to go to the	
CNAFDISK	88%	9 MB/s	6937	43
CNAFTAPE	98%	3 MB/s	2569	0
LNF	48%	0 MB/s	101	55
MILANO	100%	4 MB/s	1466	64
NAPOLI	0%	0 MB/s	0	0
ROMA1	52%	1 MB/s	403	42
FZK	51%	8 MB/s	4588	117
LYON	42%	5 MB/s	3403	103
NDGF	92%	9 MB/s	7580	19

Picchi di trasferimento al CNAF

8 agosto

Eff. ~ 90%

Throughput 17 MB/s

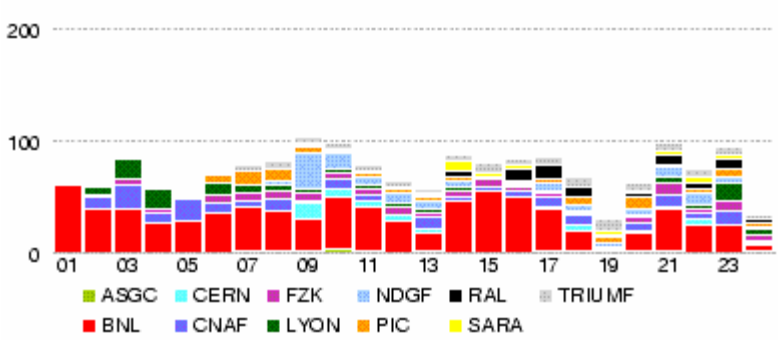
23 agosto

Eff. ~ 95%

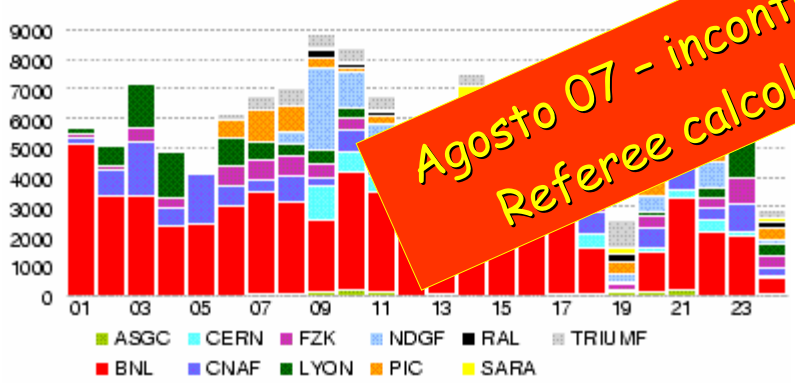
Throughput 25 MB/s

Sistema di Distribuzione dei Dati (DDM)

Throughput (MB/s)



Data Transferred (GB)



Agosto 07 - incontro Referee calcolo

Activity Summary ('2007-08-01 08:20' to '2007-08-24

Click on the cloud name to view list of sites

		Transfers			Services	
Cloud	Efficiency	Throughput	Files Done	Datasets Done	DQ	Grid
ASGC	15%	0 MB/s	24562	0		
BNL	95%	33 MB/s	640700	91865		
CERN	67%	2 MB/s	77527	38		
CNAF	58%	8 MB/s	217546	33046		
FZK	57%	5 MB/s	145219	10307		
LYON	56%	6 MB/s	147093	10344		
NDGF	61%	6 MB/s	158223	214		
PIC	53%	4 MB/s	118913	3181		
RAL	54%	3 MB/s	89888	231		
SARA	17%	2 MB/s	54209	9799		
TRIUMF	74%	5 MB/s	148947	4200		

trasferimento al CNAF - Agosto
Eff. ~ 58%

Miglioramento rispetto
al passato ma molto ancora da
migliorare.

L'inefficienza è causa anche
delle sorgenti

Computing Operations – fine 2007

Test del sistema di distribuzione dei dati:

1. **Functional Test**: Simulazione del data flow previsto dal CM a basso rate
   - Obiettivo: replicare completamente i dataset alle cloud
     - Definizione di un insieme di dataset di ~ 30 files di dimensioni variabili
     - Trasferimenti T0 → T1 e di seguito T1 → T2 della cloud secondo lo share previsto dal Computing Model
       - ✓ CNAF ~10% del totale
       - ✓ ogni Tier2 italiano 25% dei dati del CNAF
     - Trasferimenti T1 → T1 dei dati riprocessati
   - Studio dell'efficienza dei trasferimenti in termini di numero di dataset replicati correttamente e velocità di arrivo dei file, numero di retry

2. **T0 Throughput exercise**: Test di throughput
   - Obiettivo: mantenere con stabilità i throughput di trasfermento tra il Cern e le clouds previsti dal Computing Model:
     - ✓ T0 → ΣT1 = 1 GB/s
     - ✓ T0 → CNAF = 100 MB/s (2/3 su disco e 1/3 su nastro)

3. **Run di cosmici M5**
   - Trasferimento dei dati (RAW e ESD) al CNAF e nei Tier2 secondo le percentuali previste dal Computing Model

Functionl Test – Oct 2007

Trasferimenti Tier-1 ↔ Tier-1

	ASGC	BNL	CNAF	FZK	LYON	NGDF	PIC	RAL	SARA	TRIUMF
ASGC										
BNL										
CNAF										
FZK										
LYON										
NDGF										
PIC										
RAL										
SARA										
TRIUMF										

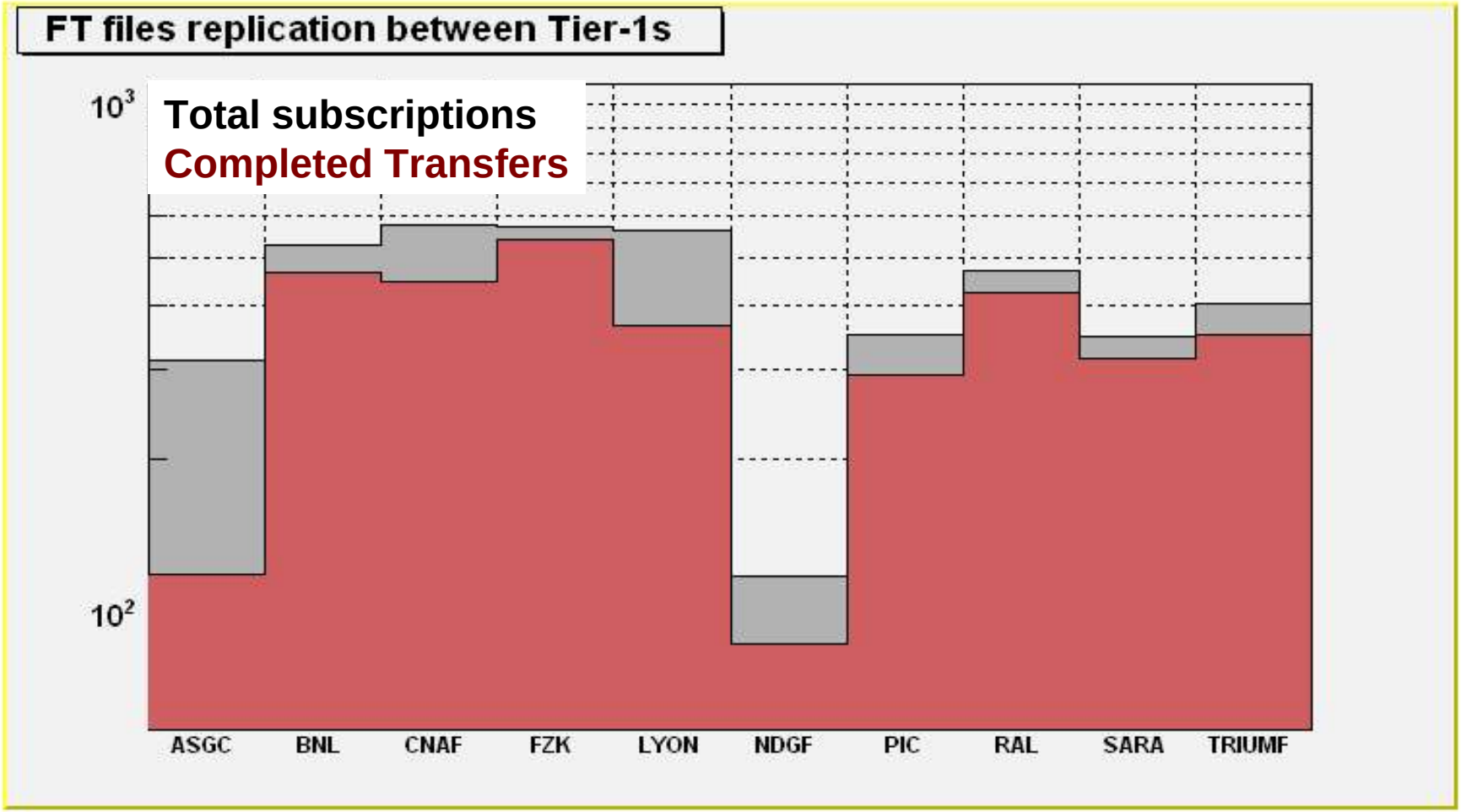
Trasferimenti CERN → Tier-1s

	ASGC	BNL	CNAF	FZK	LYON	NDGF	PIC	RAL	SARA	TRIUMF
CERN										

100% , 90+% , 50%, less than 50%, of data transferred within 24h

Functionl Test – Oct 2007

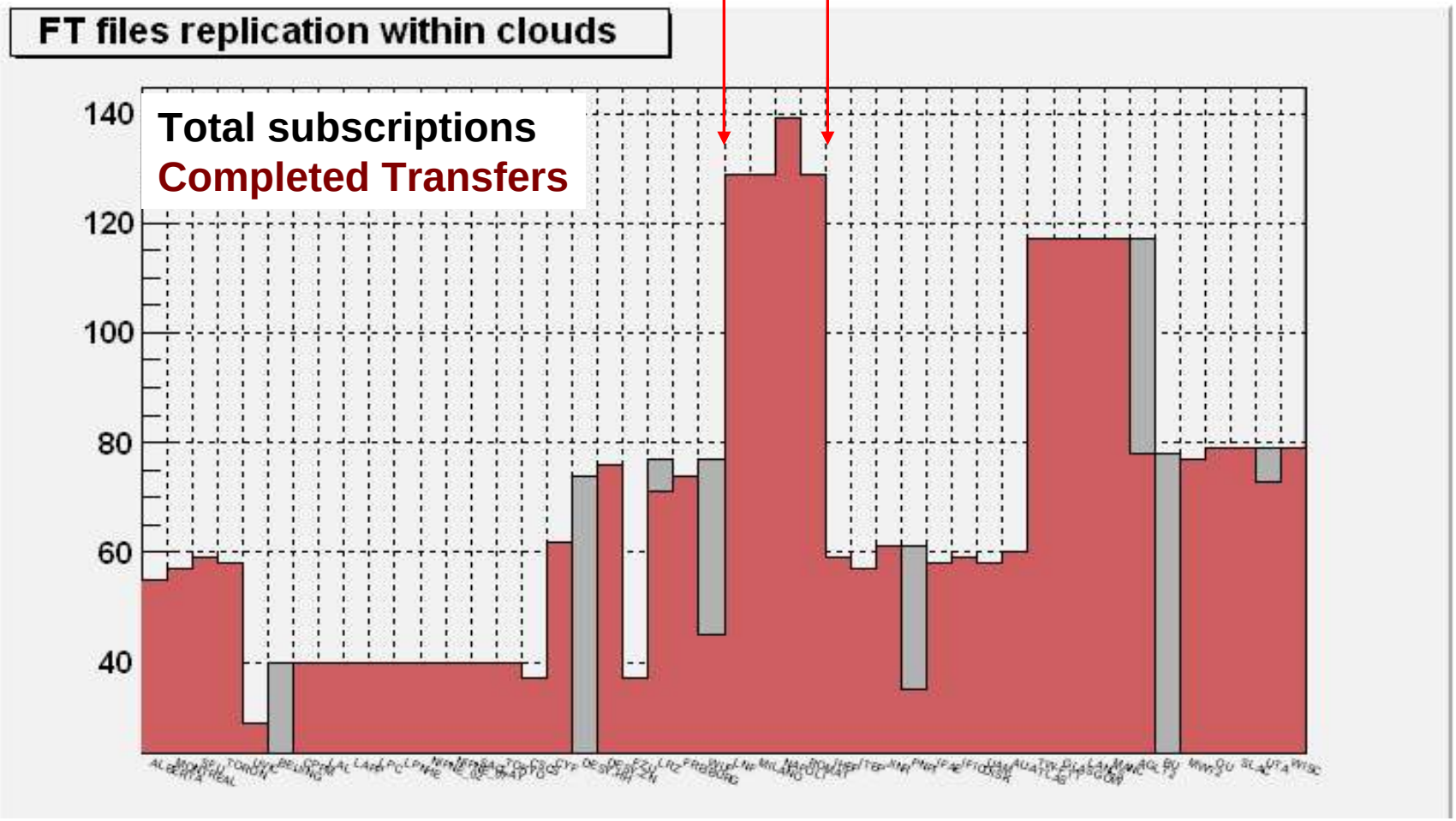
DDM Functional Test Oct 2007. Data Transfer to Tier-1s



Functionl Test – Oct 2007

DDM Functional Test Oct 2007. Data Transfer to Tier-2s

T2 italiani – efficienza 100%

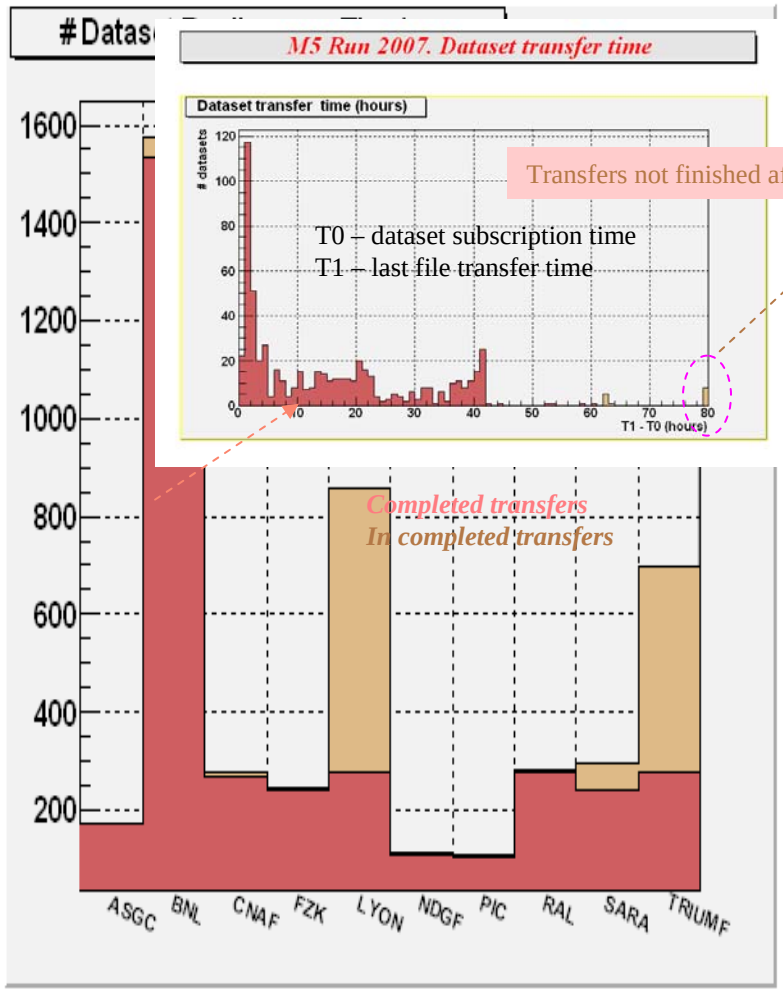


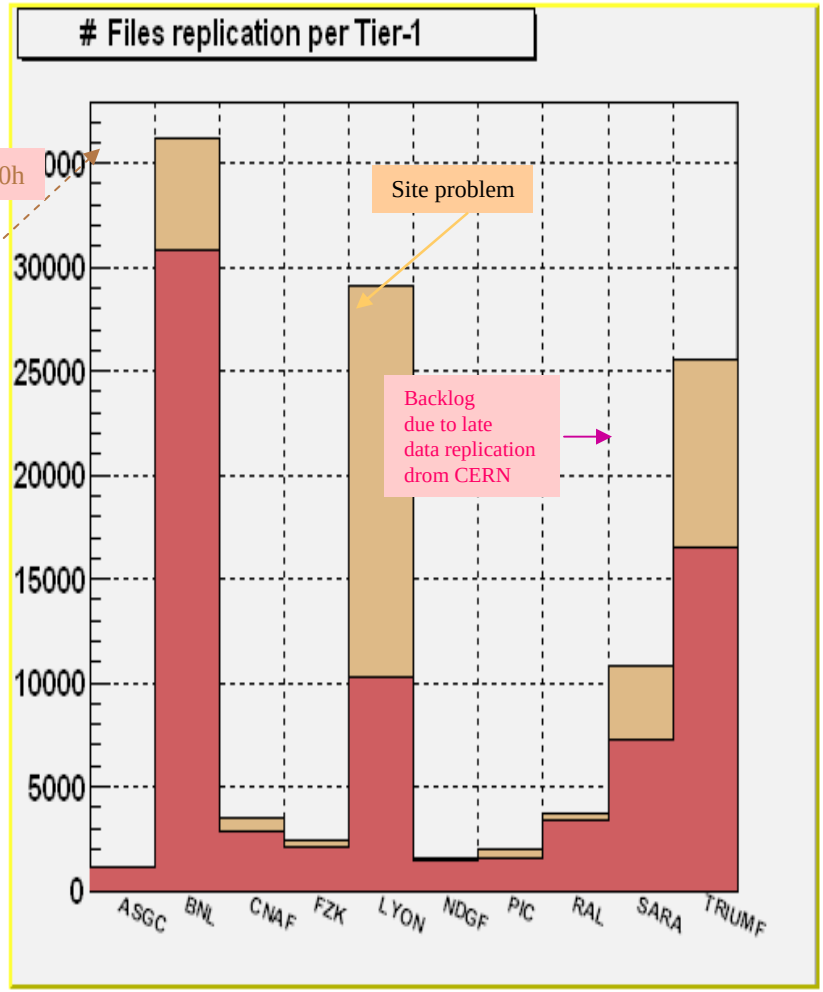
M5 Cosmic Run – Oct/Nov 2007

- 2 settimane di Run
  - Detector integration la prima settimana e reale Data Taking nella seconda
- Data sample totale (analizzabile): ~ 90 TB
- RAW Data su Tape e ESD su Disco
- No AOD o DPD
  - analisi effettuata sui RAW Data
- Raw Data e ESD distribuiti in Dataset a tutti i Tier-1 secondo lo share previsto dal CM
  - Copia intera ad alcuni Tier-1 che ne hanno fatto richiesta (BNL, Lyon, Triumf)
- Utilizzo degli end-point srm di produzione
  - Storm al CNAF

- Non è stato un test completo di computing
  - trasferimenti a basso throughput e non molto stabili
  - Solo RAW data per l'analisi
- Tuttavia un buon test sull'efficienza del sistema di trasfermiento

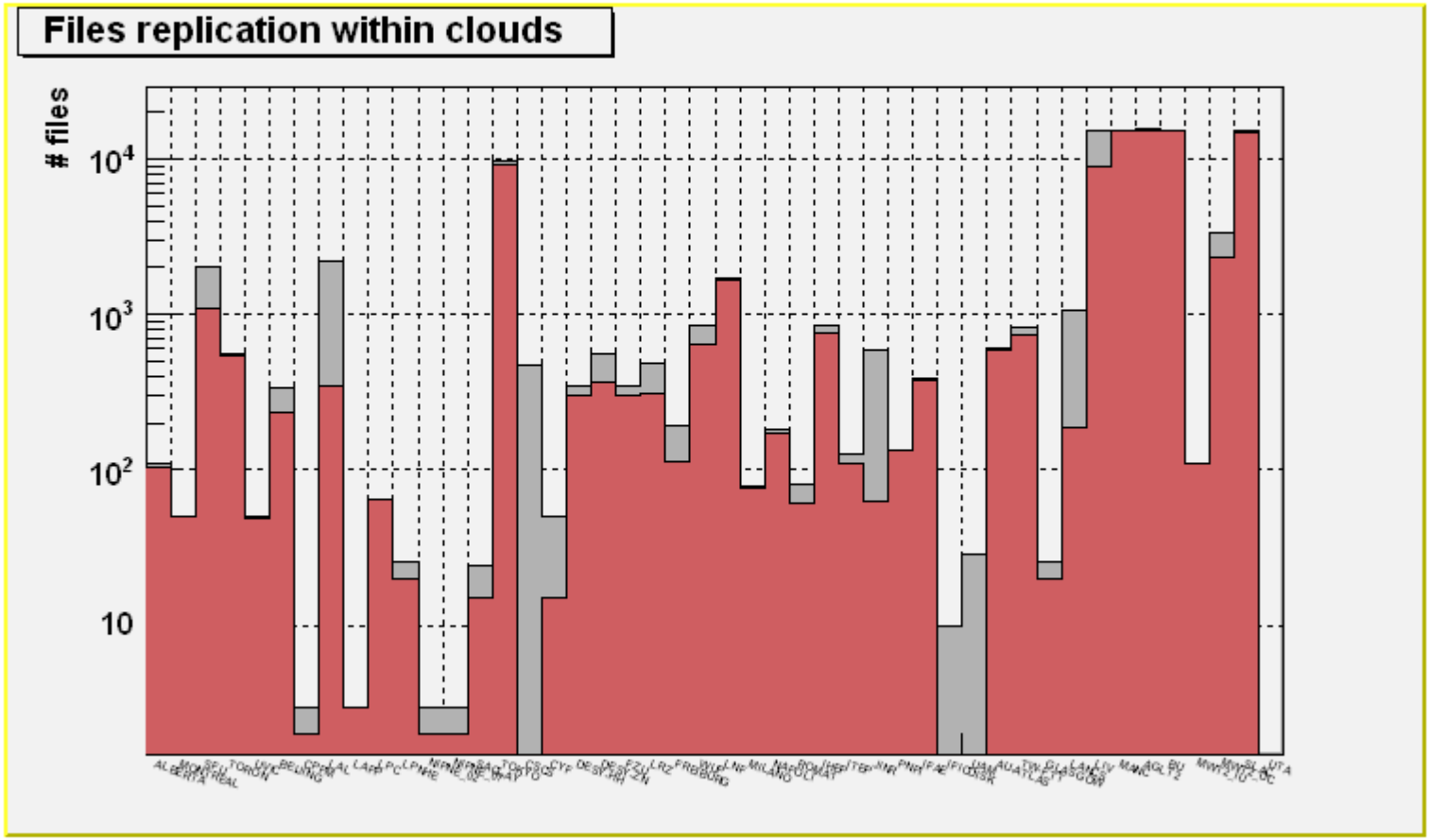
M5 Cosmic Run – Oct/Nov 2007





M5 Cosmic Run – Oct/Nov 2007

M5 Cosmic Run Oct-Nov 2007. Data Transfer to Tier-2s



T0 Export Exercise – Oct 2007

Obiettivi:

- Throughput al 100% MoU
  - Come se la macchina operasse 24h/day ~ 1 GB/sec
  - MoU prevede 720 MB/sec
- Operazioni completamente automatizzate senza intervento
- Corretto share tra dati da inviare su tape (tipo RAW) e su disco (tipo AOD e ESD)

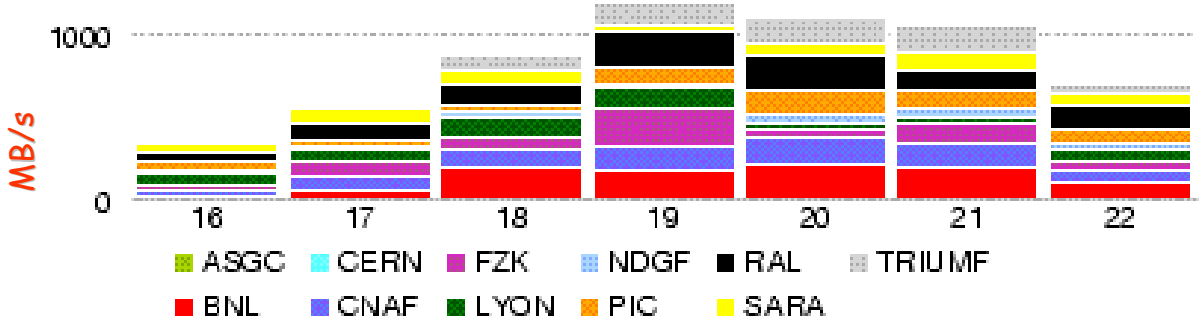
In Italia

- Thoughput da sostenere con continuità 100 MB/s
- Test del nuovo srm end-point per il disco (T0D1): STORM
  - Sviluppato interamente dall'INFN

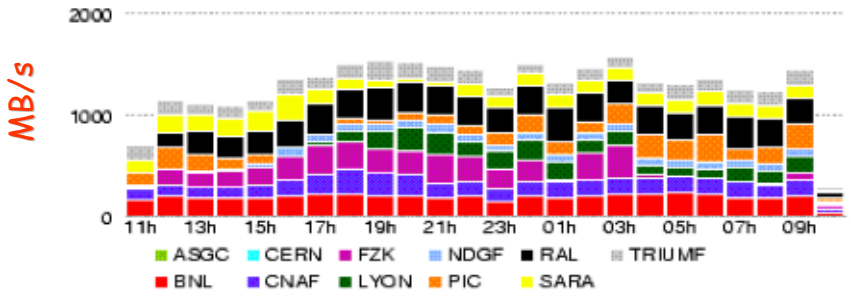
T0 Export Exercise – Oct 2007

Obiettivo raggiunto !

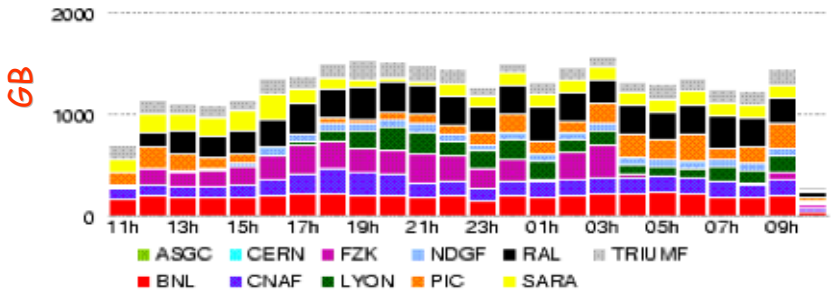
Rate di ~ 1.2 GB/sec per un periodo prolungato con un set incompleto di Tier-1



Throughput MB/s

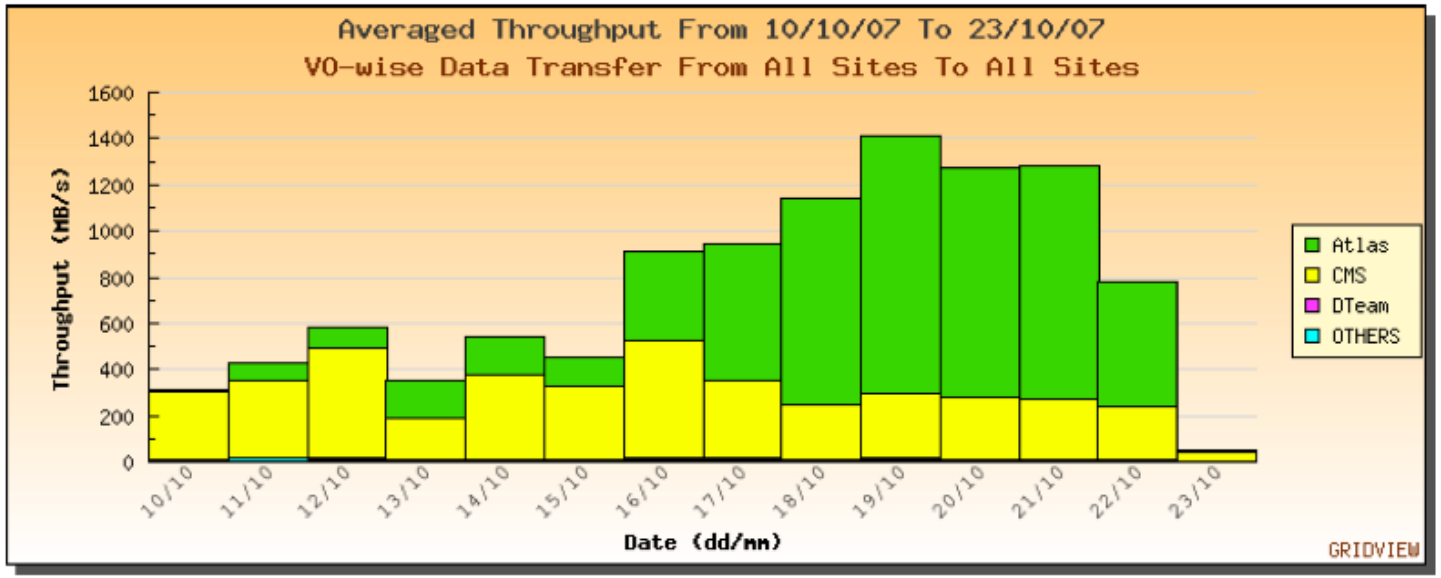


Data transferred GB



T0 Export Exercise – Oct 2007

- Il Tier-0 e i Tier-1 multi esperimento hanno dimostrato di poter supportare l'attività contemporanea di due esperimenti: ATLAS e CMS
- Test di attività contemporanea tra i 4 esperimenti LHC nel 2008: CCRC08



T0 Export Exercise – Oct 2007

Al CNAF:

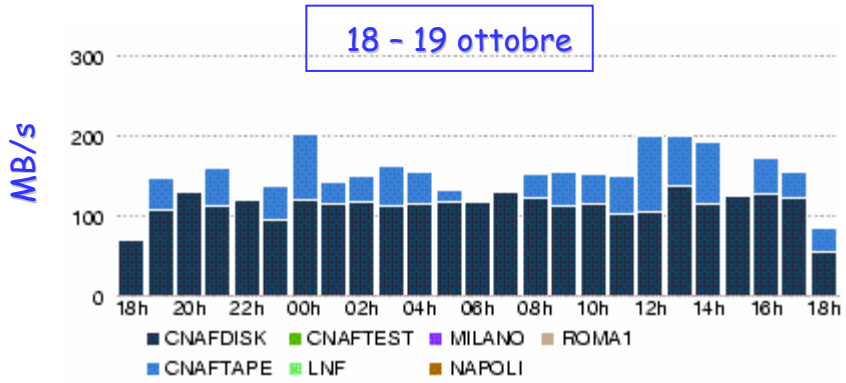
- Utilizzo del nuovo srm come disco T0D1: STORM
- Castor come tape endpoint T1D0
- Descrizione del Cluster GPFS (server, disco, rete) da LUCA

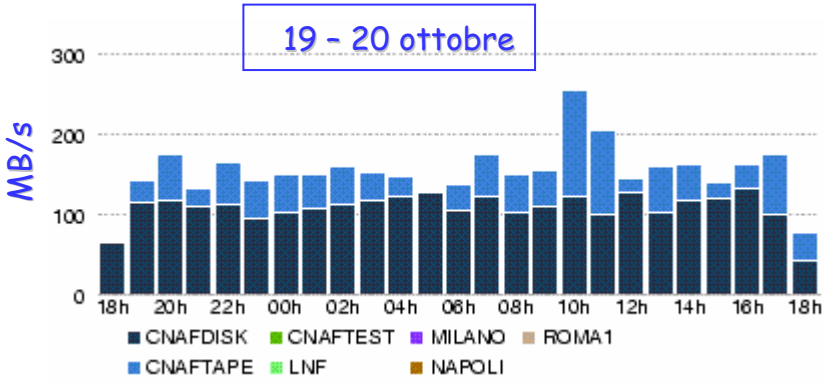
T0 Export Exercise – Oct 2007

Al CNAF:

- Nel periodo 18-21 ottobre si è superato, con continuità, il throughput previsto di 100 MB/s di ~ 50%
- Efficienze medie superiori al 90%
- Si è deciso di utilizzare STORM come srm definitivo a partire dal run di cosmici M5
- Buoni risultati di Castor Tape ma efficienza ancora da migliorare

Cloud	Efficiency	Throughput
ASGC	0%	0 MB/s
BNL	89%	185 MB/s
CERN	0%	0 MB/s
CNAF	90%	147 MB/s
CNAFDISK	94%	116 MB/s
CNAFTAPE	73%	30 MB/s
CNAFTEST	0%	0 MB/s
LNF	0%	0 MB/s
MILANO	0%	0 MB/s
NAPOLI	0%	0 MB/s
ROMA1	0%	0 MB/s
FZK	47%	130 MB/s
LYON	64%	75 MB/s
NDGF	72%	36 MB/s
PIC	74%	109 MB/s
RAL	96%	188 MB/s
SARA	15%	73 MB/s
TRIUMF	97%	149 MB/s





Computing Operations – fine 2007

- I test della seconda parte del 2007 mostrano un deciso miglioramento delle performance del sistema di distribuzione e autorizzano ad essere fiduciosi sulla reperibilità dei dati per l'analisi nel Tier-1 e nei Tier-2 nel 2008.

- Ovviamente bisogna dimostrare che questi risultati possono essere ottenuti con continuità e in presenza di molte attività concorrenti

- Il risultato del T0 throughput test e del M5 cosmic run ha mostrato una buona affidabilità del nuovo srm STORM. Si è quindi deciso di metterlo definitivamente in produzione.
  - primo caso di srm 2.2 in produzione in Atlas

Sistema di Distribuzione dei Dati (DDM)

…… Fondamentale la collaborazione con gli utenti !!

Ma gli utenti devono essere rassicurati che il sistema di trasferimento dei dati può funzionare con efficienza e velocità e non è necessario reperire i dati con mezzi alternativi

I test di questo autunno mostrano un miglioramento delle performance del sistema di distribuzione e autorizzano ad essere fiduciosi sulla reperibilità dei dati per l'analisi nei Tier-2

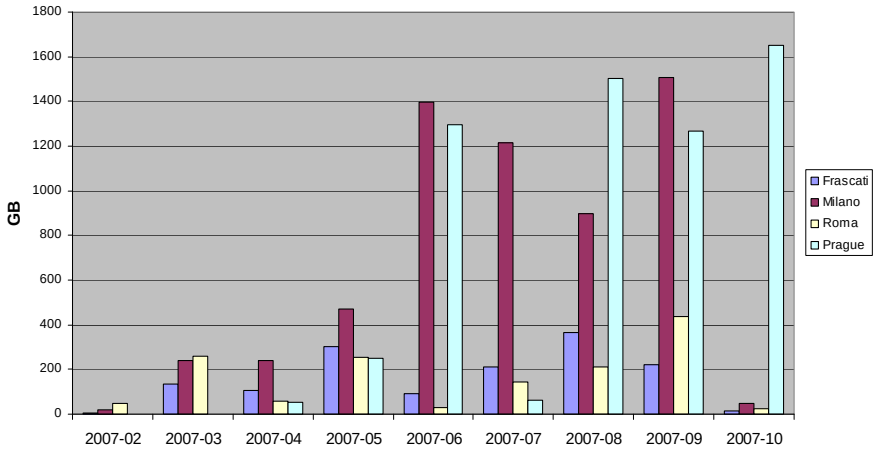
Sarà necessario

- interagire con i gruppi e il coordinatore della fisica per definire la distribuzione degli AOD nei Tier-2 in base alle attività
- l'aiuto di chi fa analisi per il monitoraggio dei trasferimenti

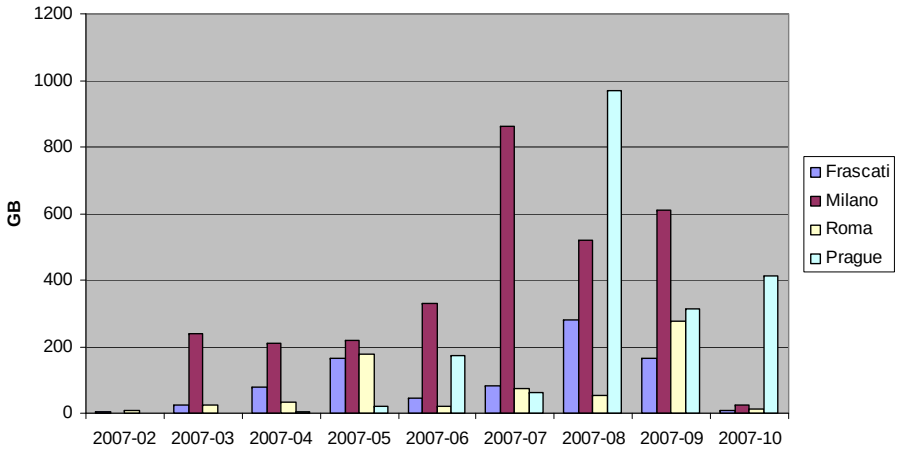
Maggiore interazione tra le comunità di computing e di fisica perché nei nostri siti siano disponibili i dati necessari per l'analisi e vengano utilizzati

Utilizzo effettivo degli AOD

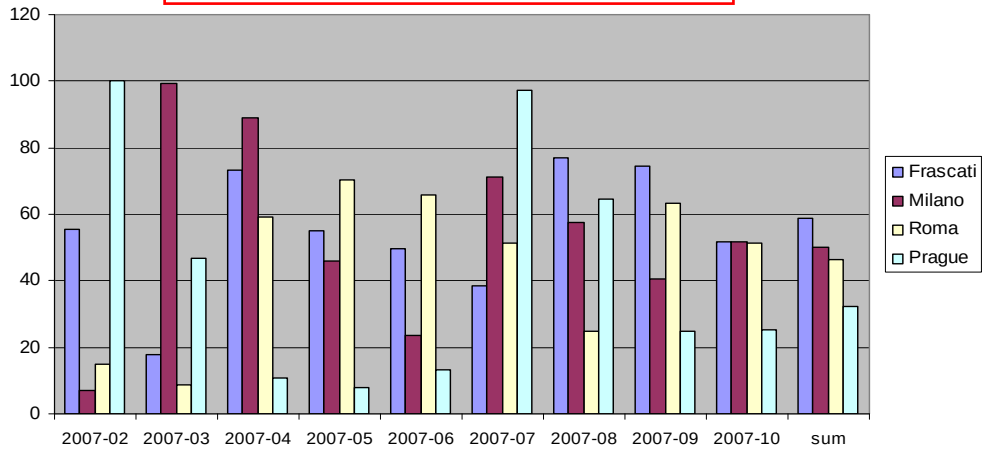
AOD nei Tier2



AOD utilizzati (GB)



Percentuale di AOD utilizzati



~ 50–60 % - AOD utilizzati
78% - DQ2 replications !

L'utilizzo è per repliche in altri siti e non per l'analisi

Esperienza con STORM

Migrazione da CASTOR (disk) a STORM

- 78 TB trasferiti per un totale di circa 1.5 M files
  - esclusi file corrotti o da archiviare su tape
- Effettuata dal 16 al 23 Novembre
  - 5 giorni (3,5 giorni effettivi) , 1 giorno per controlli e preparazione del nuovo storage per l'entrata in produzione, 1 giorno per l'aggiornamento del catalogo
- Trasferimenti effettuati utilizzando 12 diskservers con carico distribuito.
- Throughput 300 MBps. Efficienza 100%.

Problemi riscontrati dall'entrata in produzione di StoRM

di interesse comune perché STORM è il primo srm versione 2 in produzione e ha fatto esperienza di tutti i problemi connessi all'interazione con gli altri sistemi

- Dal 23 Novembre STORM in produzione
- fallimenti durante il trasferimento di files da altri siti verso StoRM dovuti ad incompatibilita' tra client e server:
  1. FTS non crea la struttura di directories che avveniva a livello srm per srm1. Risolto 19-12
  2. Problemi nei trasferimenti con siti che hanno dCache come srm. File di default "volatili" per cui scomparivano dopo una breve lifetime (40h). Fix per dCache sarò disponibile a fine gennaio. Per il momento allungata la lifetime (4000h) dei file. Persi 10 kfiles.
  3. Problemi con Ganga nell'accesso ai file in corso di risoluzione

Test dei canali FTS

Test dei canali di trasfermiento FTS: T1 → T2 e T2 → T1

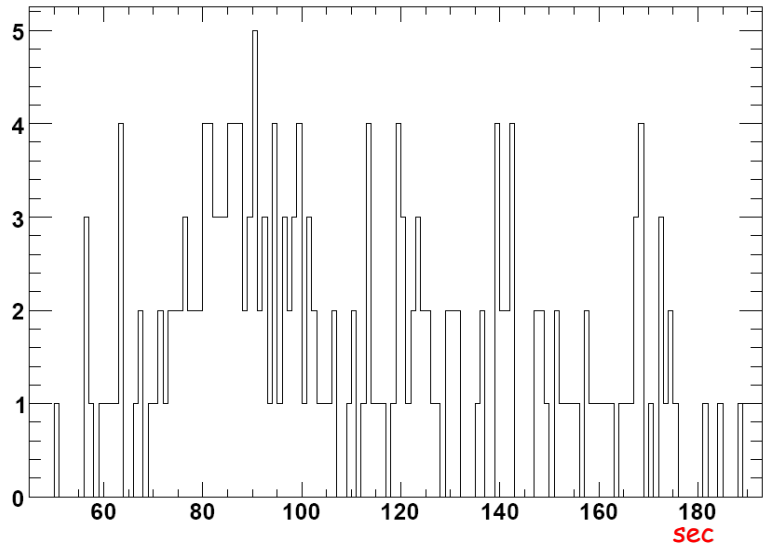
Obiettivo: verificare che i canali sono configurati in modo da garantire il throughput di trasferimento previsto dal CM

- Trasferimenti T2 → T1: File MC (RDO e HITS) prodotti nei Tier-2 e trasferiti nei Tier-1 per la ricostruzione e l'archivio
  - file da ~ 2 GB (jumbo files)
  - throughput previsto 10 / 20 MBps (normale, picco)
- Trasferimenti T1 → T2: File AOD,TAG e DPD per l'analisi
  - file da ~ 1 GB
  - throughput previsto 15 / 30 MBps (normale, picco)

Test in corso, risultati preliminari

- Trasferimento NA → CNAF:
  - throughput aggregato 77 ± 8 MBps
  - normali condizioni di operazione del sito
- Valore che soddisfa le nostre esigenze
- Non è necessario modificare i parametri del canale: numero di file trasferiti contemporaneamente (10) e numero di stream per file (5)
- Ripetere con statistica maggiore e negli altri siti

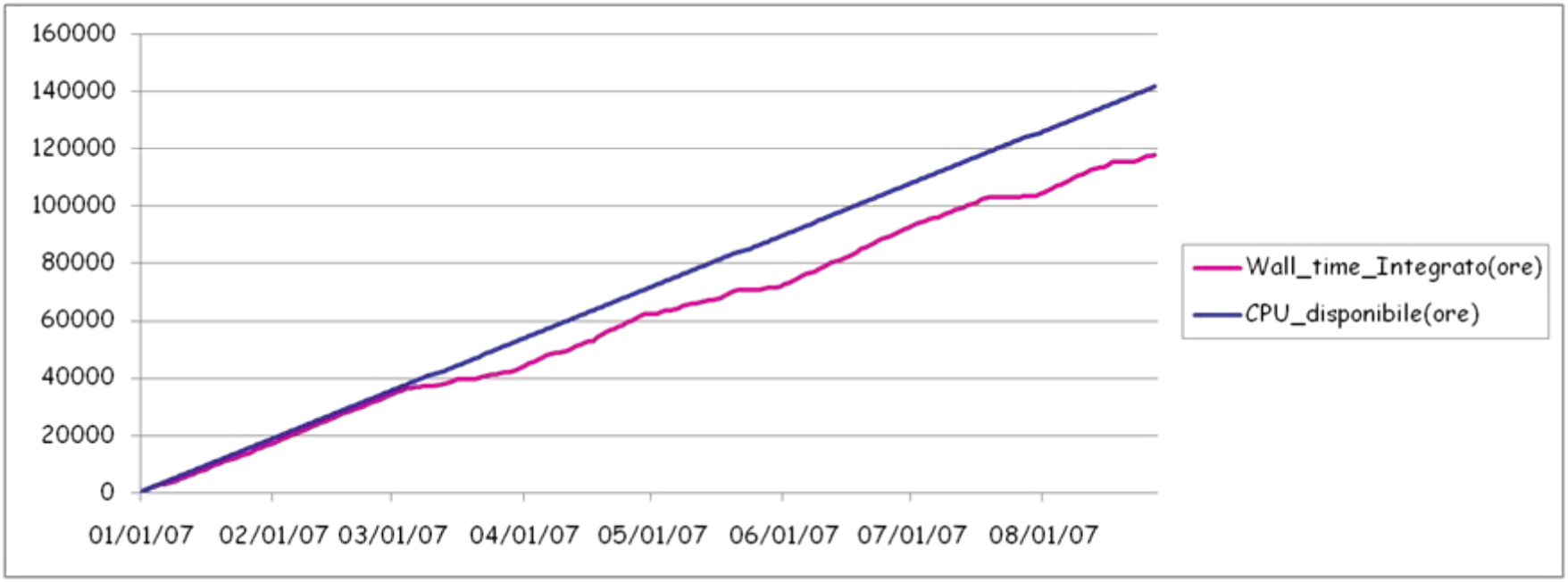
Tempo di trasferimento dei file



Uso delle risorse nei Tier-2 italiani

ProtoTier2 – Frascati

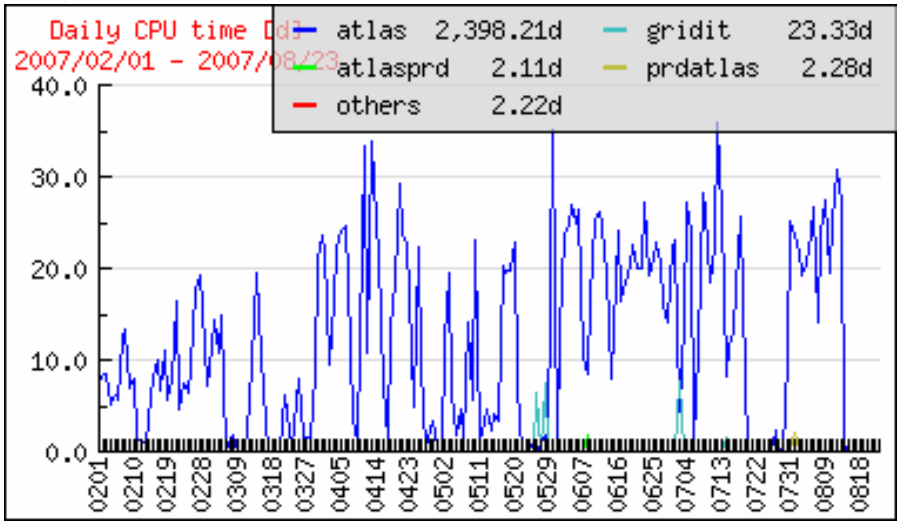
Utilizzo Risorse 10/06 - 08/07

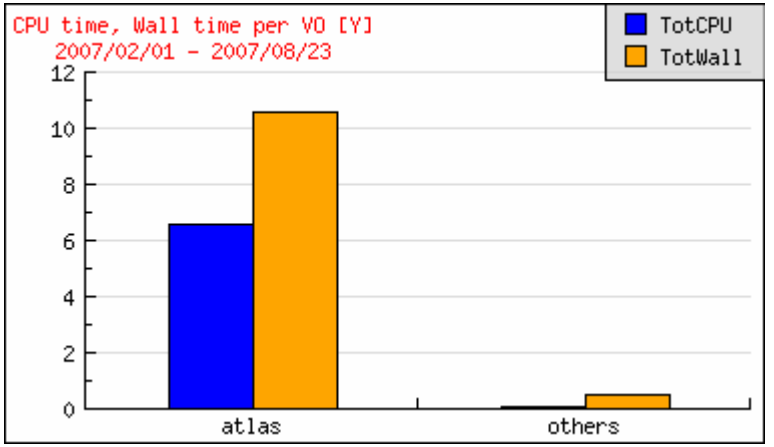


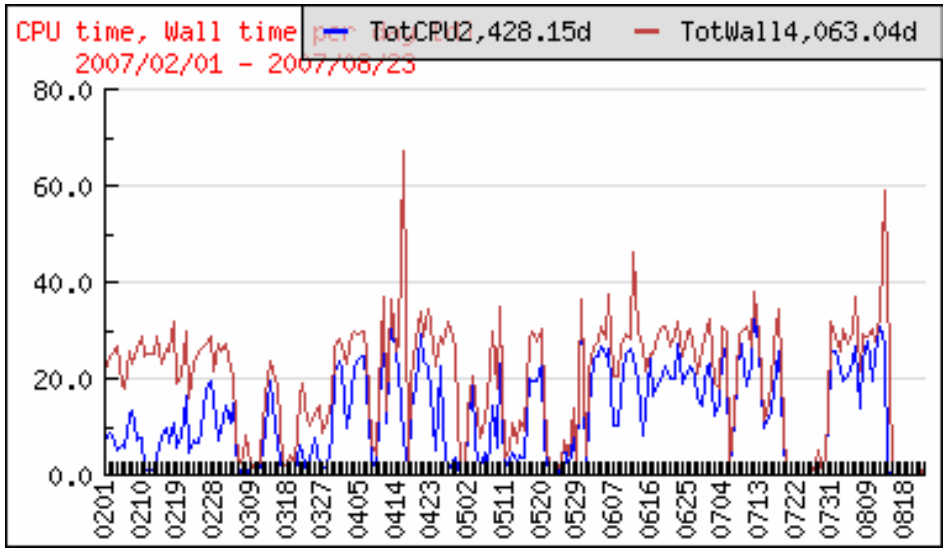
Efficienza 91% (88% per Atlas) tranne negli ultimi mesi in cui si è ridotta a causa di numerosi upgrade del middleware GRID contenenti bachi)

30 CPU dedicate per ATLAS (26 fino a marzo)

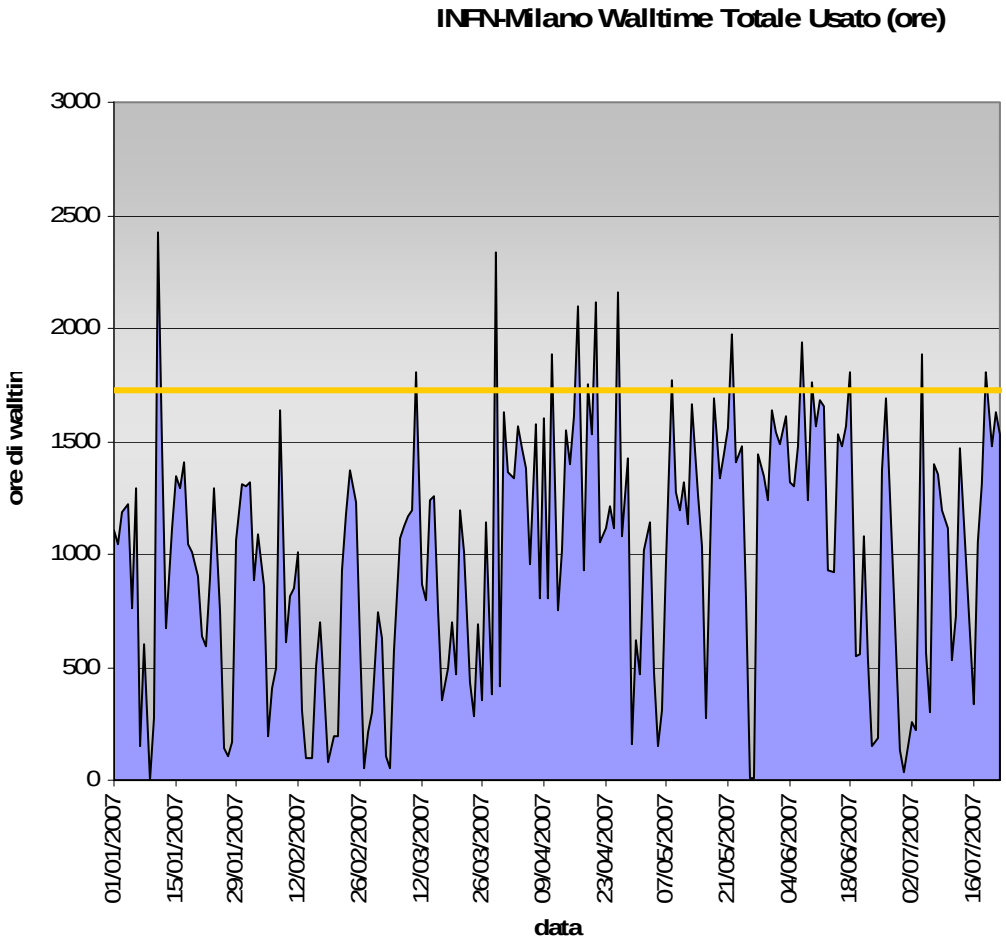
ProtoTier2 - Frascati



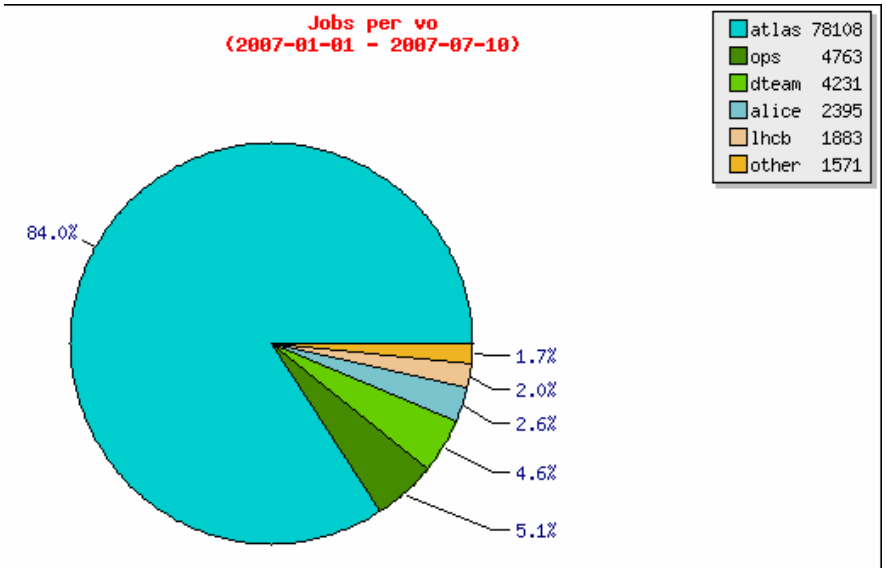




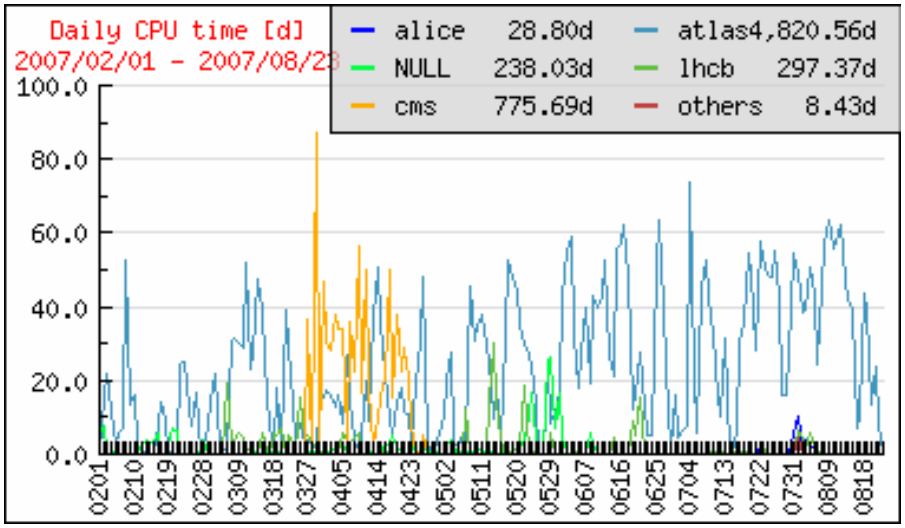
Tier2 Milano

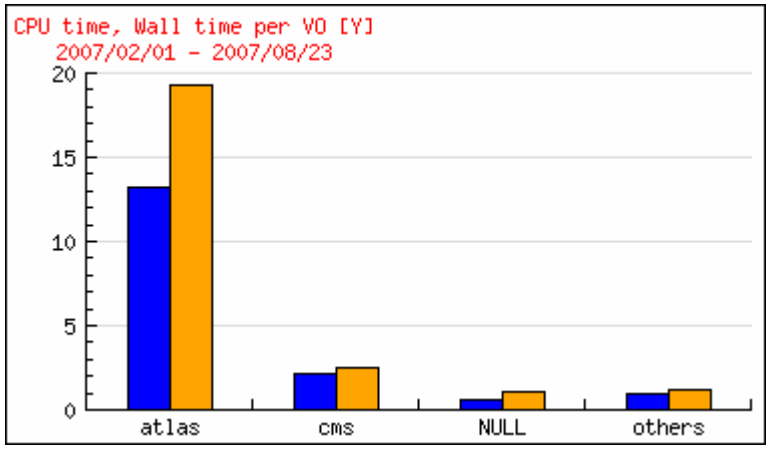


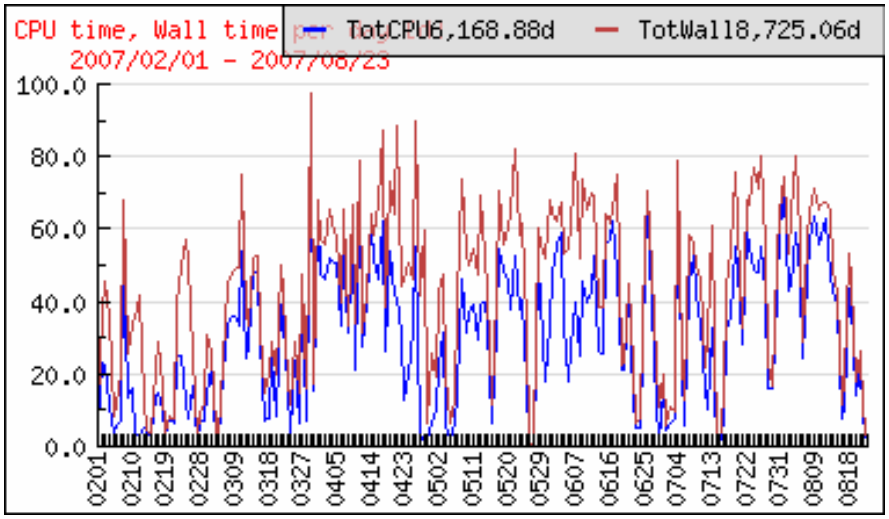
Utilizzo Risorse
10/06 - 07/07



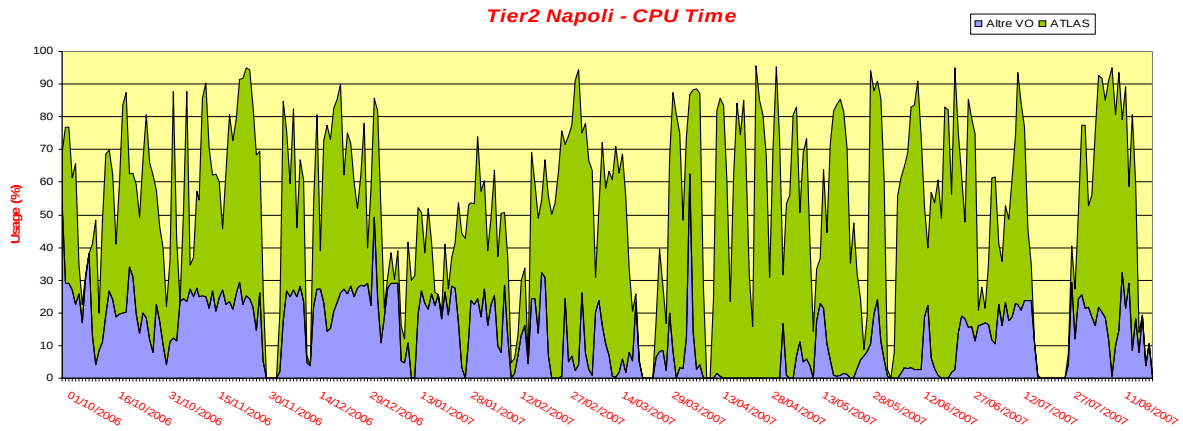
Tier2 sj- Milano





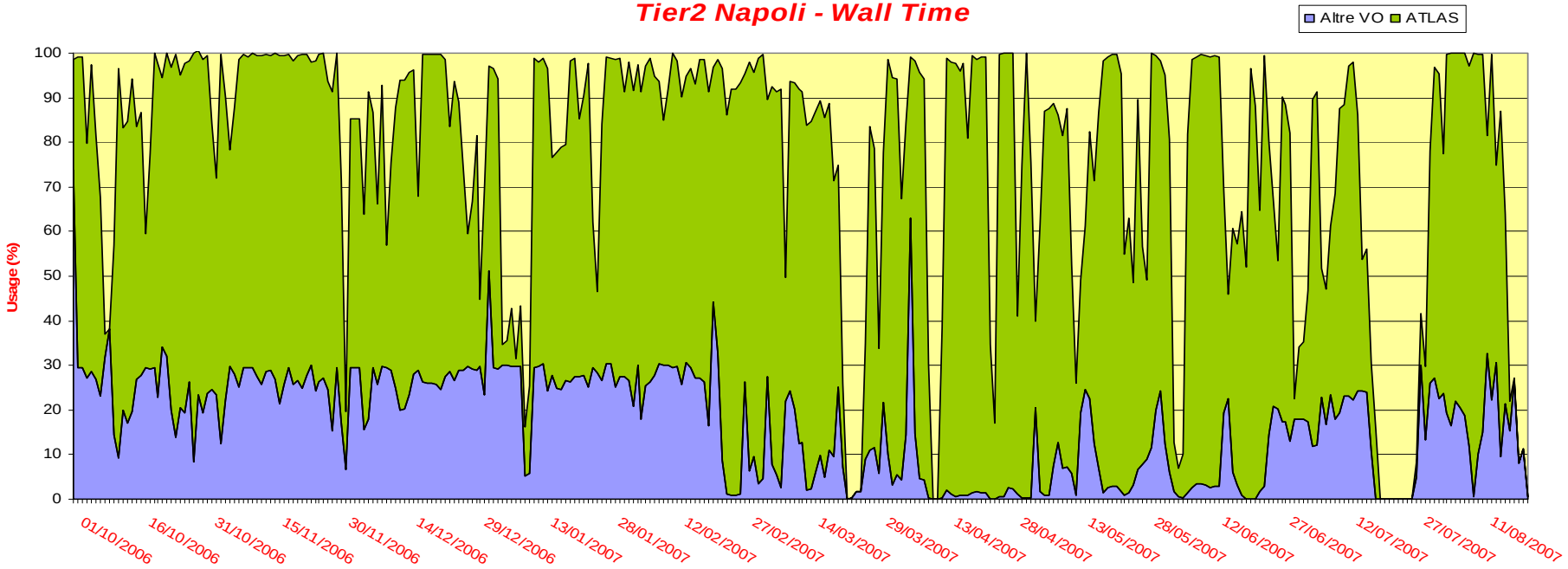


Tier2 - Napoli

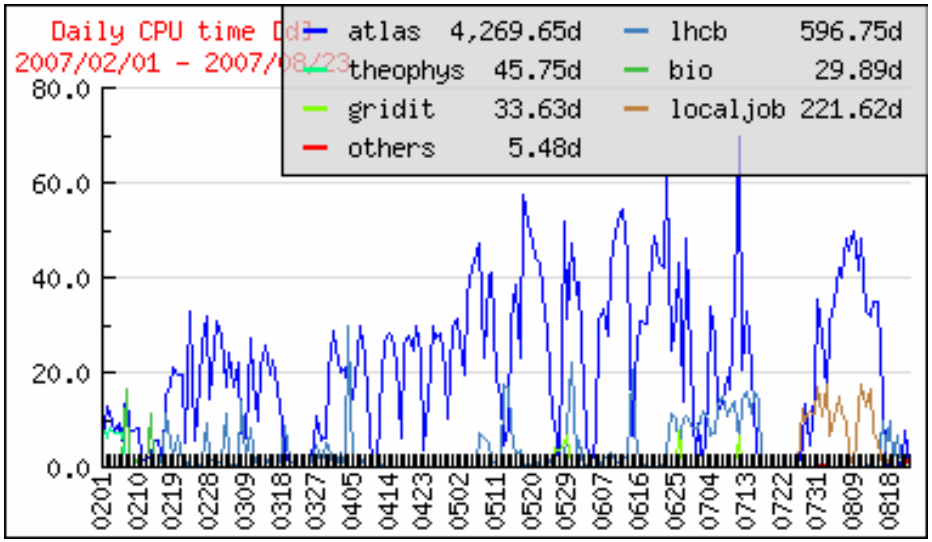


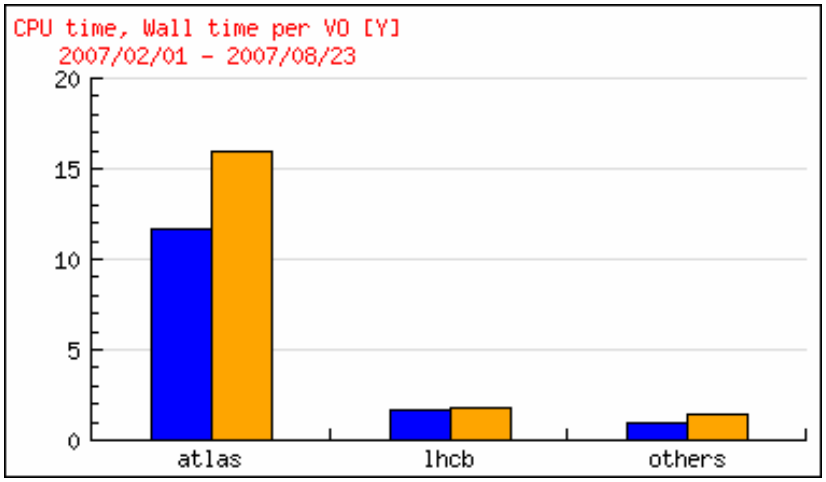
Utilizzo Risorse
10/06 - 08/07

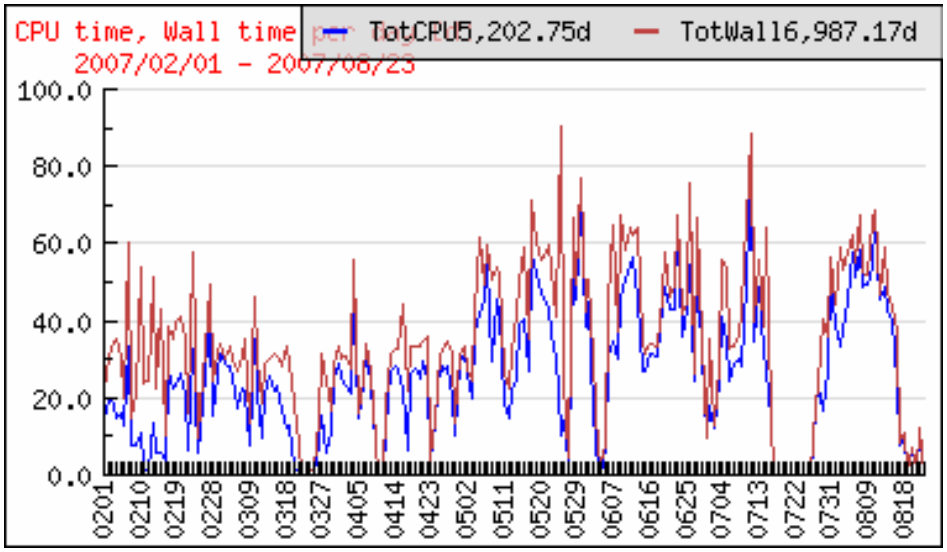
- 62 core (34 fino a aprile)
- Wall time eff. = 80%



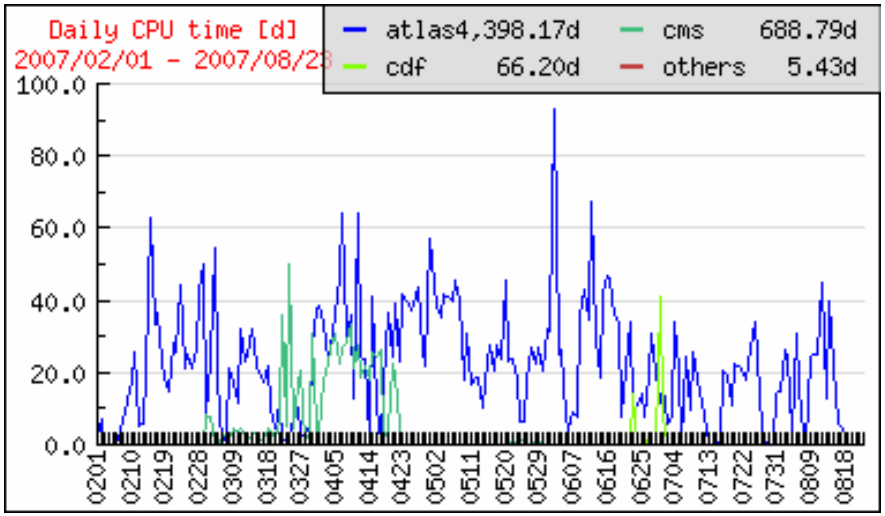
Tier2 - Napoli

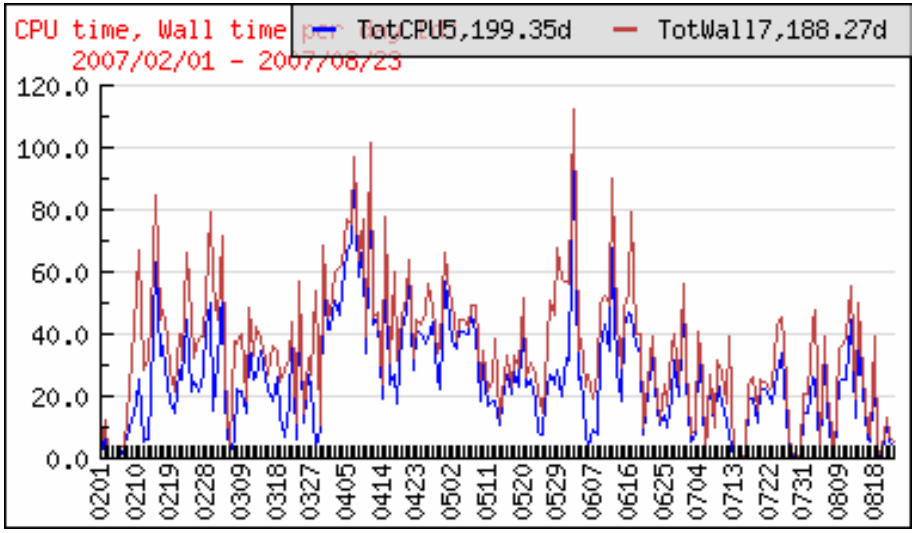


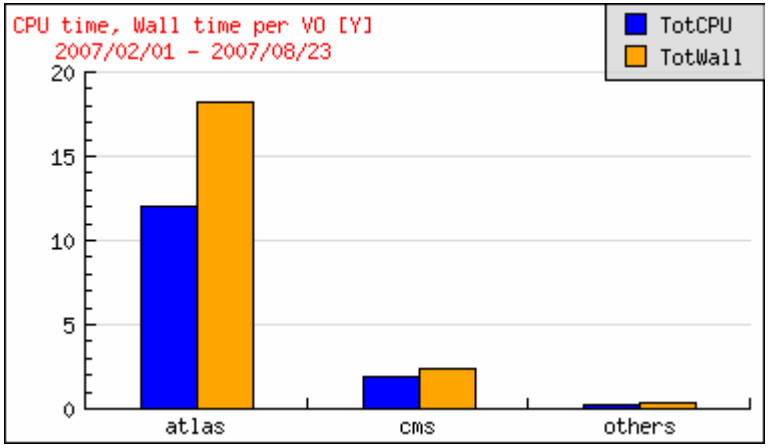




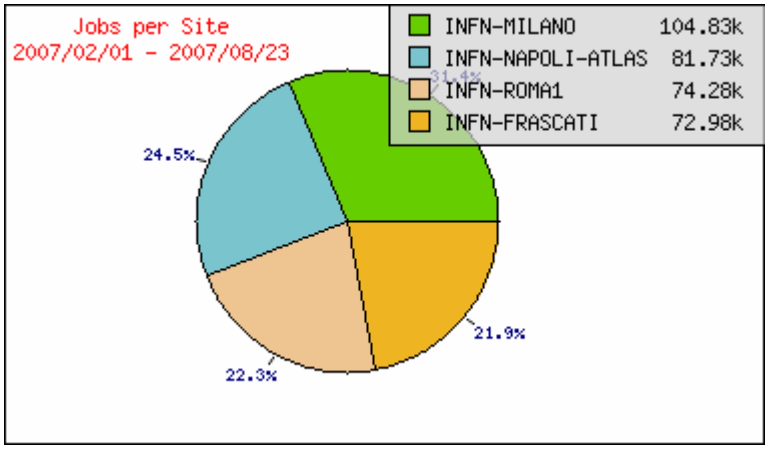
Tier2 - Roma I

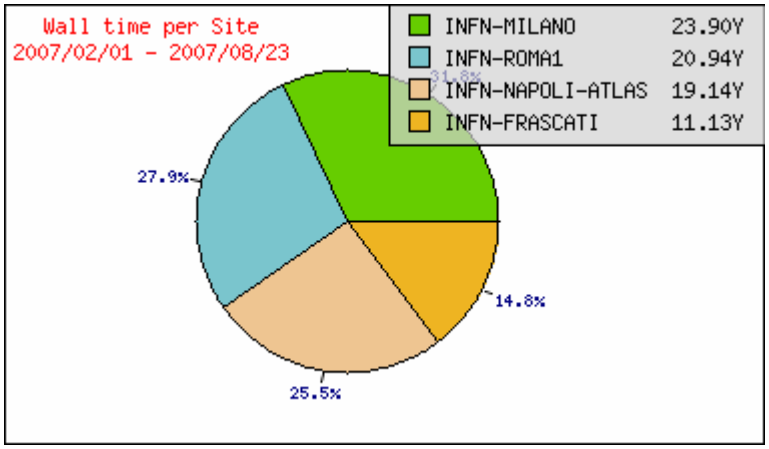

Daily CPU time [d]
2007/02/01 - 2007/08/23
atlas4,398.17d
cms 688.79d
cdf 66.20d
others 5.43d
100.0
80.0
60.0
40.0
20.0
0.0
0201
0210
0219
0228
0309
0318
0327
0405
0414
0423
0502
0511
0520
0529
0607
0616
0625
0704
0713
0722
0731
0809
0818

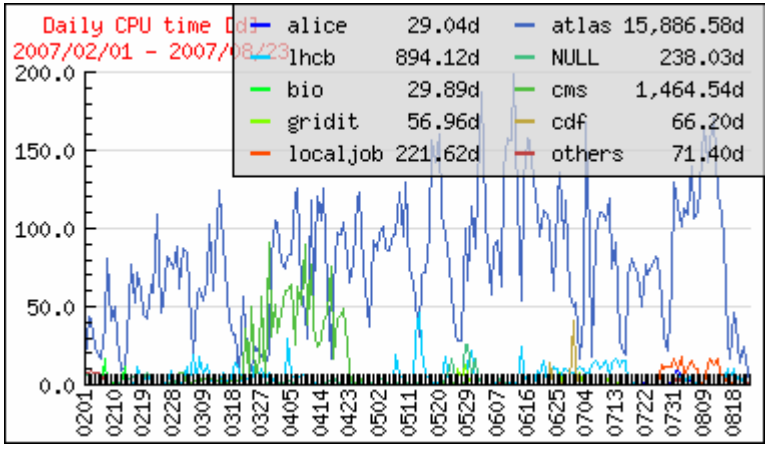

CPU time, Wall time
2007/02/01 - 2007/08/23
TotCPU5,199.35d
TotWall7,188.27d
120.0
100.0
80.0
60.0
40.0
20.0
0.0
0201
0210
0219
0228
0309
0318
0327
0405
0414
0423
0502
0511
0520
0529
0607
0616
0625
0704
0713
0722
0731
0809
0818

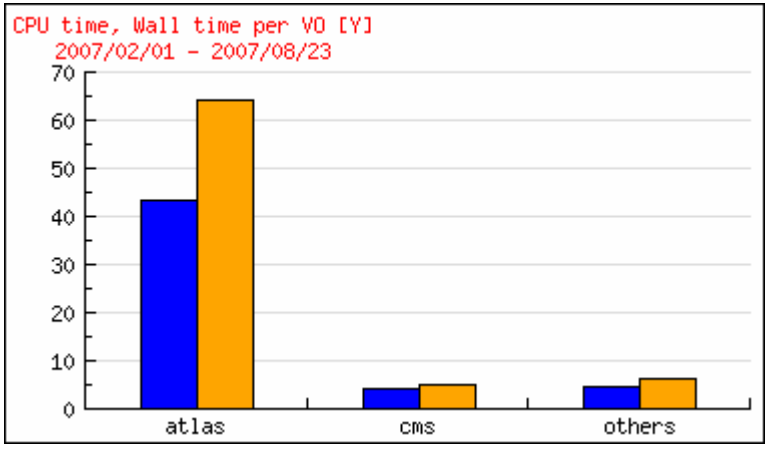

CPU time, Wall time per VO [Y]
2007/02/01 - 2007/08/23
TotCPU
TotWall
20
15
10
5
0
atlas
cms
others

Tier2 Italiani

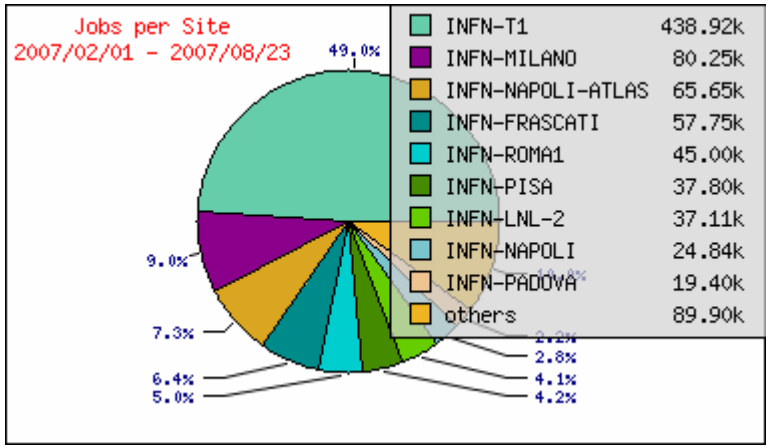


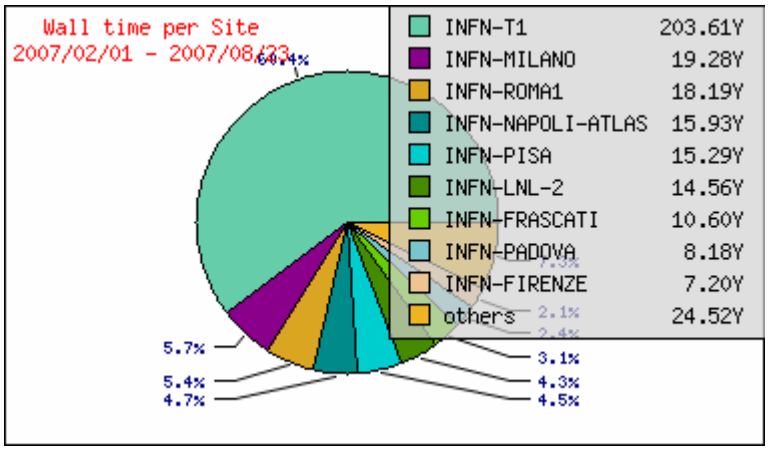


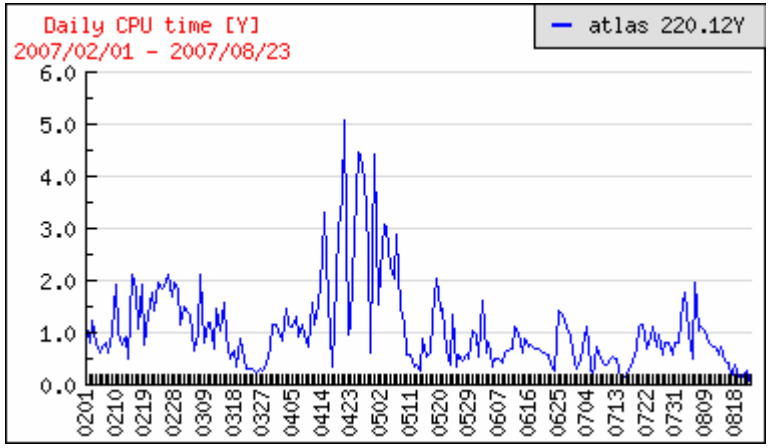


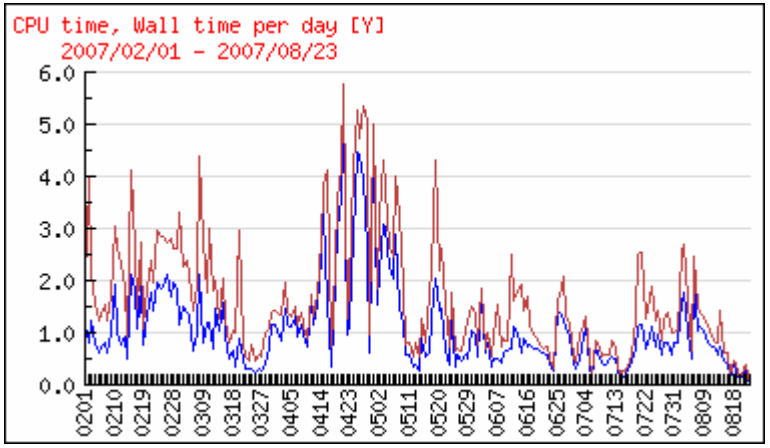


Atlas @Tier2 Italiani

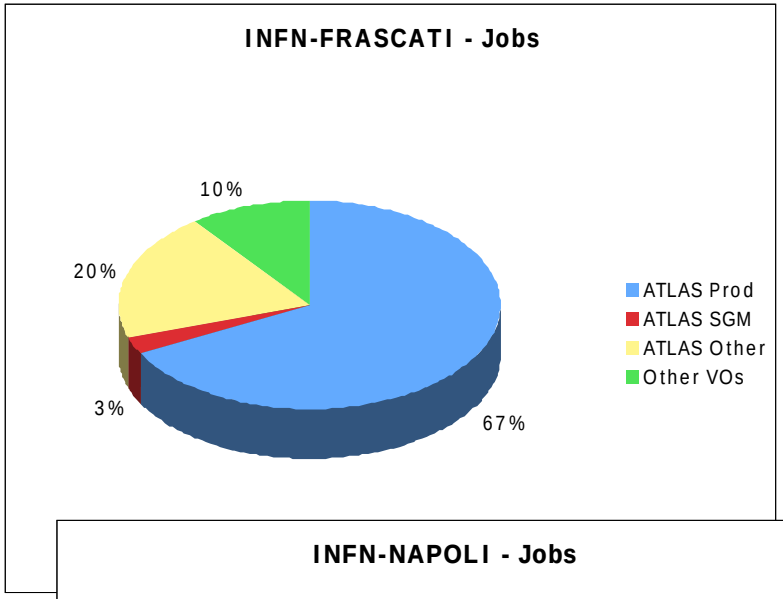


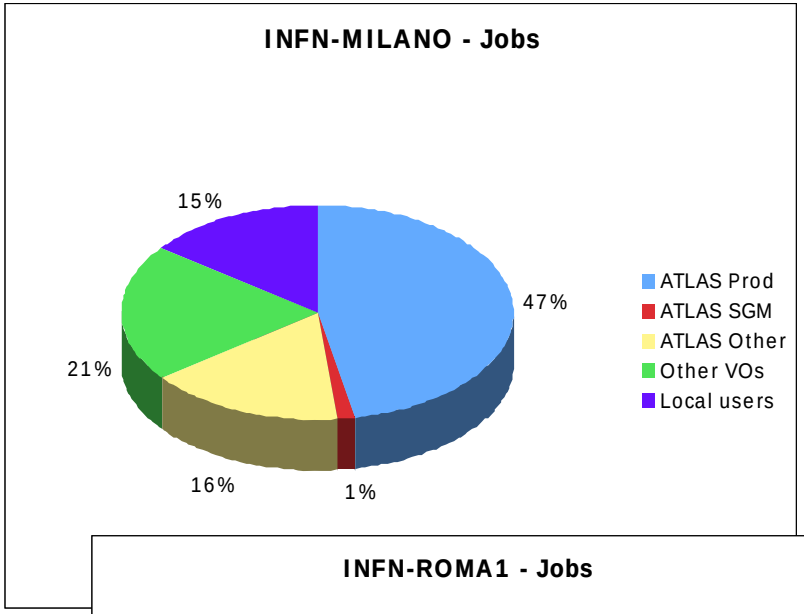


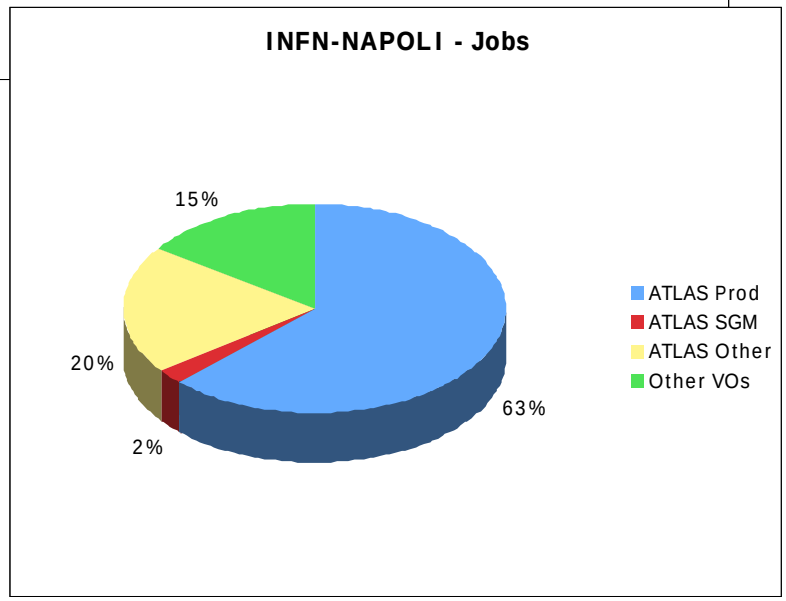


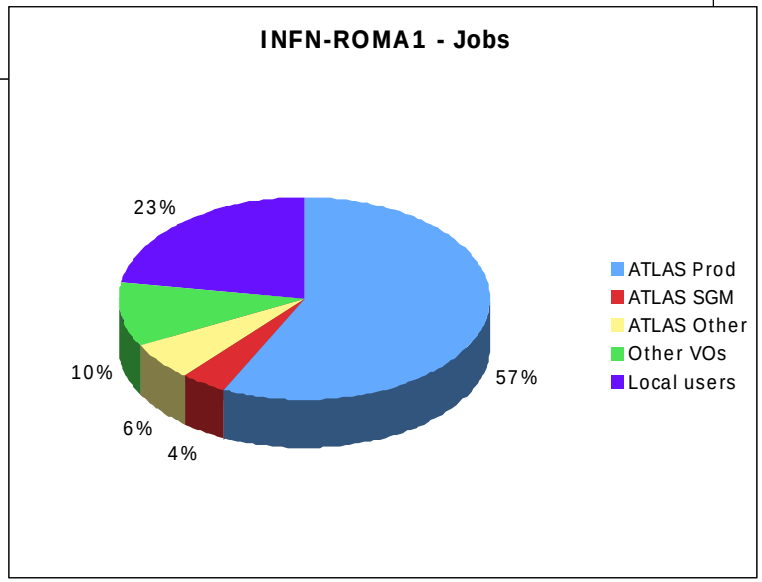


Job Sottomessi ai Tier2









-Talk 4b -

I Sistemi di storage dei Tier-2 italiani

1. Descrizione dei sistemi di storage e di rete dei Tier-2 italiani
2. Definizione delle necessità di throughput per l'analisi previste per il 2008
3. Descrizione dei test effettuati nei Tier-2 italiani
4. Presentazioni di test svolti su sistemi equivalenti in altri siti di ATLAS
   - Test di scalabilità effettuati a Glasgow
   - Il Tier-2 di Tokyo
5. Confronto tra le diverse soluzioni
   - Hardware: DAS vs SAN
   - Middleware: DPM vs STORM/GPFS
6. Strategia di ATLAS
   - Soluzioni previste per il 2008
   - Pianificazione di attività e test

Strategia di ATLAS