Stato del Calcolo e "produzioni" MC Italiane

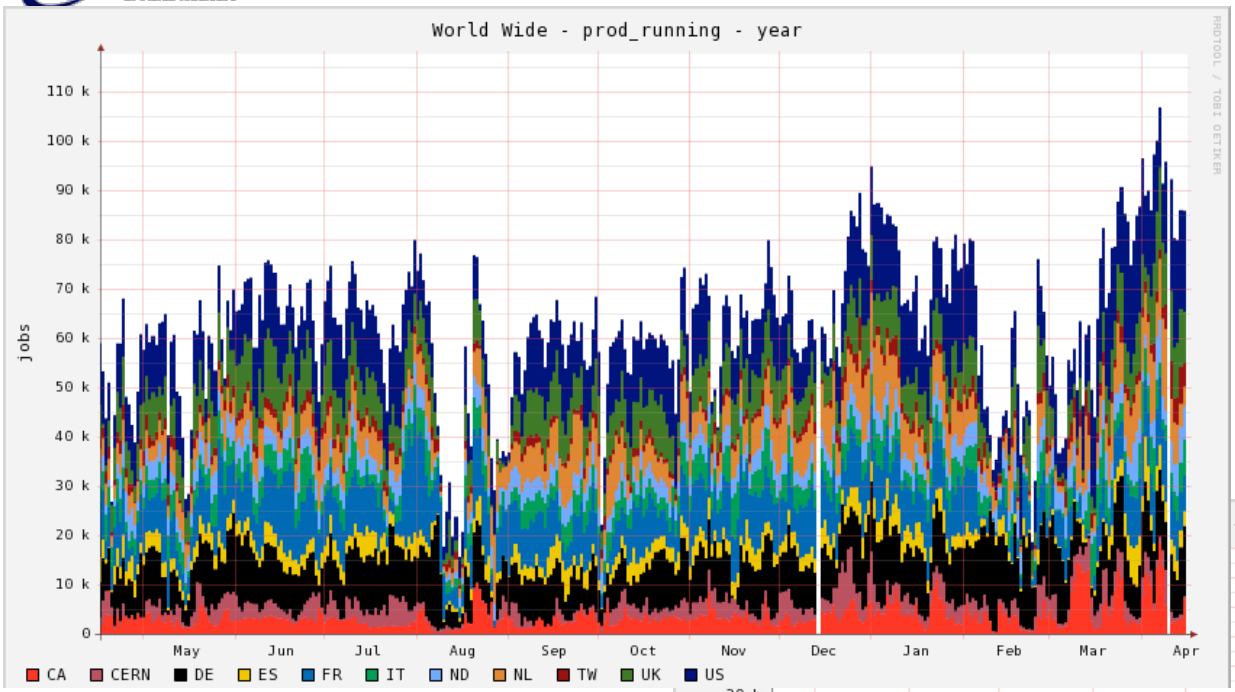
Gianpaolo Carlino

INFN Napoli

- Attività di Computing in Atlas
- La distribuzione dei dati
- L'accesso ai dati
- Utilizzo risorse Italiane

Roma Tor Vergata, 17 Aprile 2012

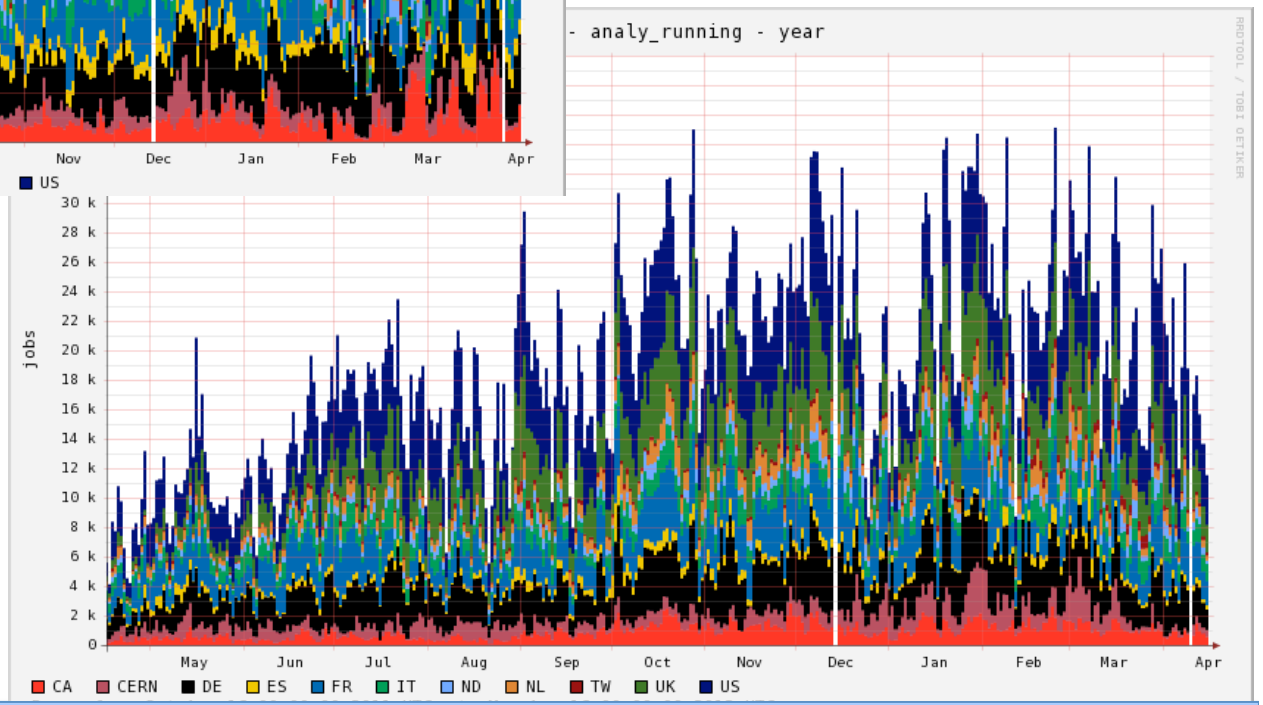
Attività di Computing in ATLAS



Numero di job simultanei nell'ultimo anno

Produzione:

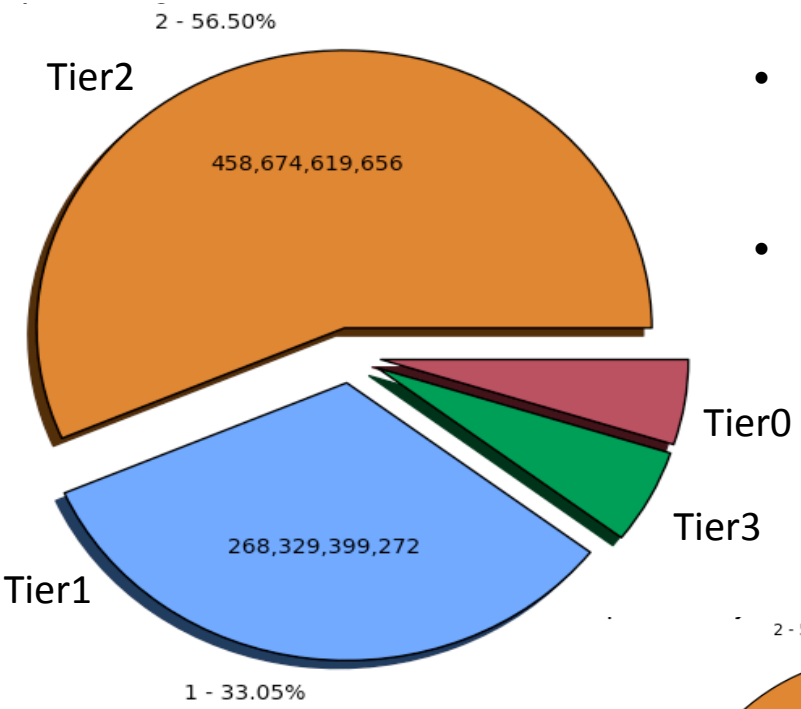
- > 60k job, constante
- incremento inizio 2012 per reco MC11



Analisi:

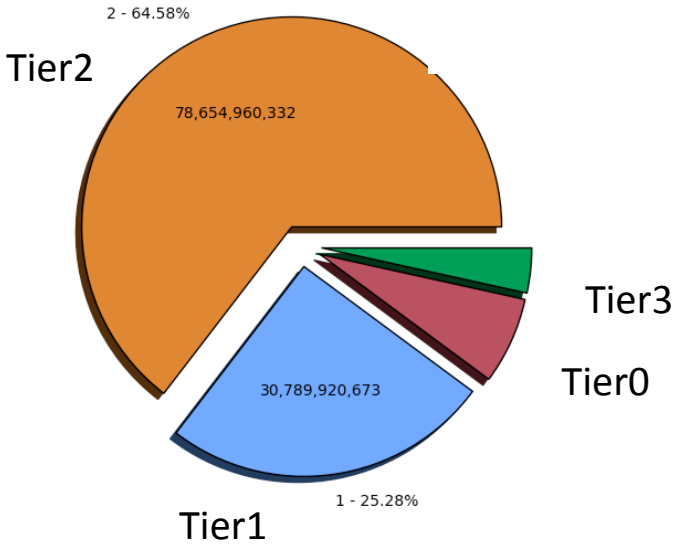
- crescita lineare nel 2011 fino a > 20k job
- Alta attività nel 2012 per le conferenze invernali
- In attesa di statistica per riprendere l'attività seria

Attività nei Tier

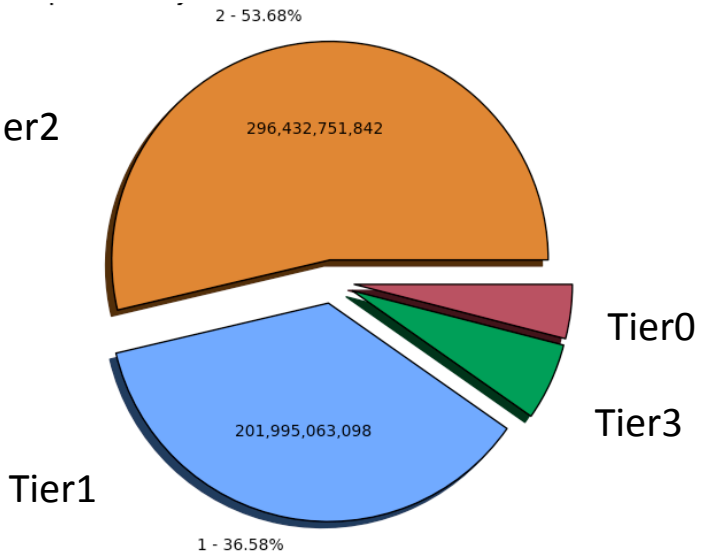


- I Tier2 forniscono la maggioranza delle risorse
- Contributo dei Tier3 non trascurabile

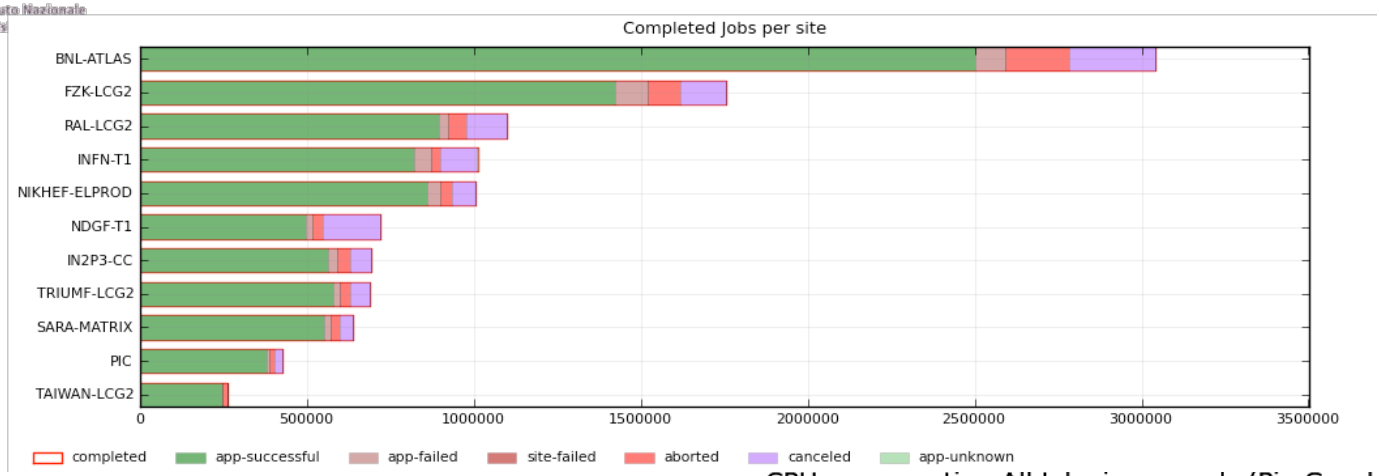
Analisi

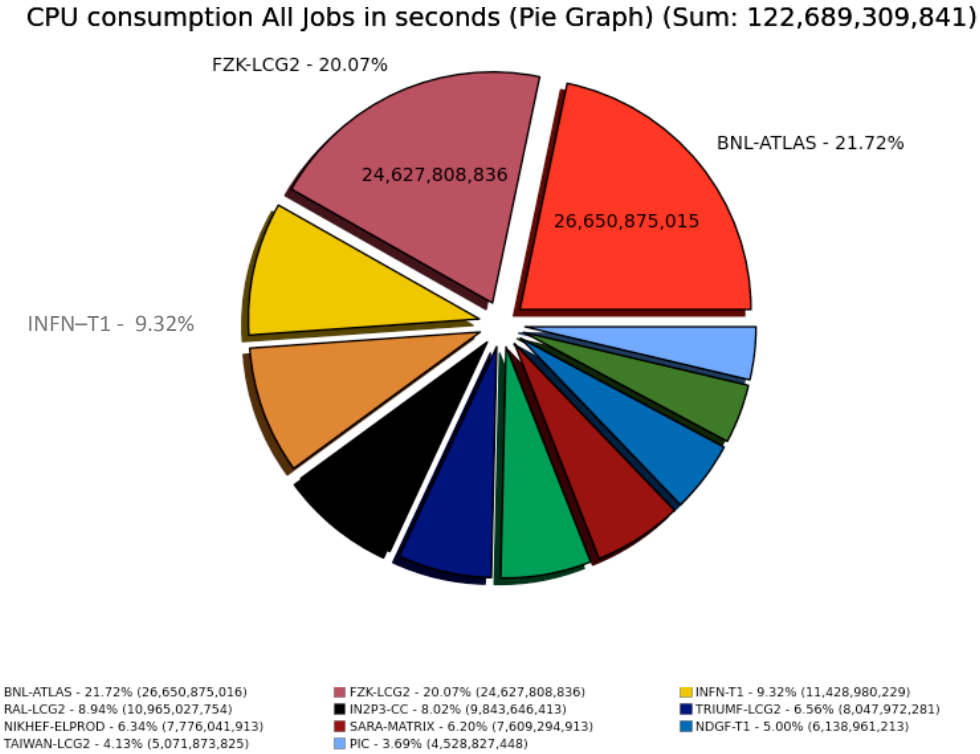


Produzione



Utilizzo risorse al CNAF

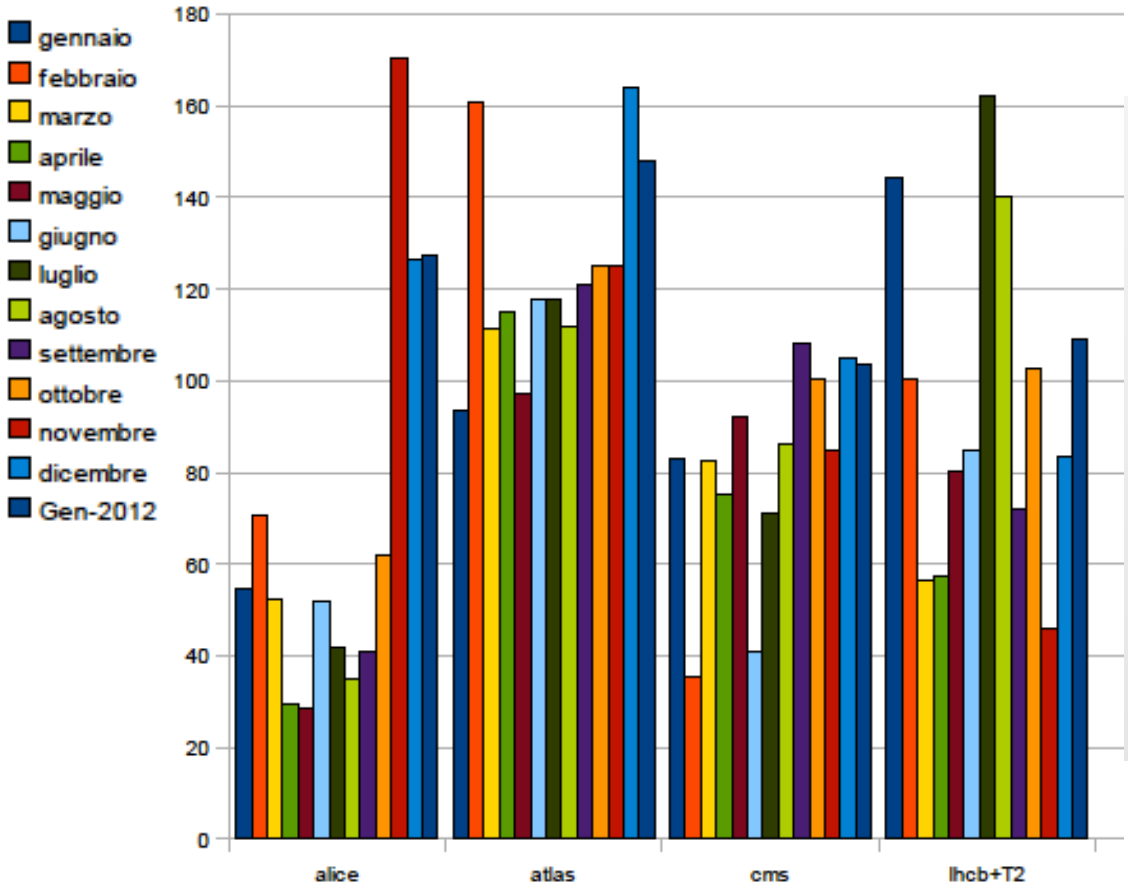


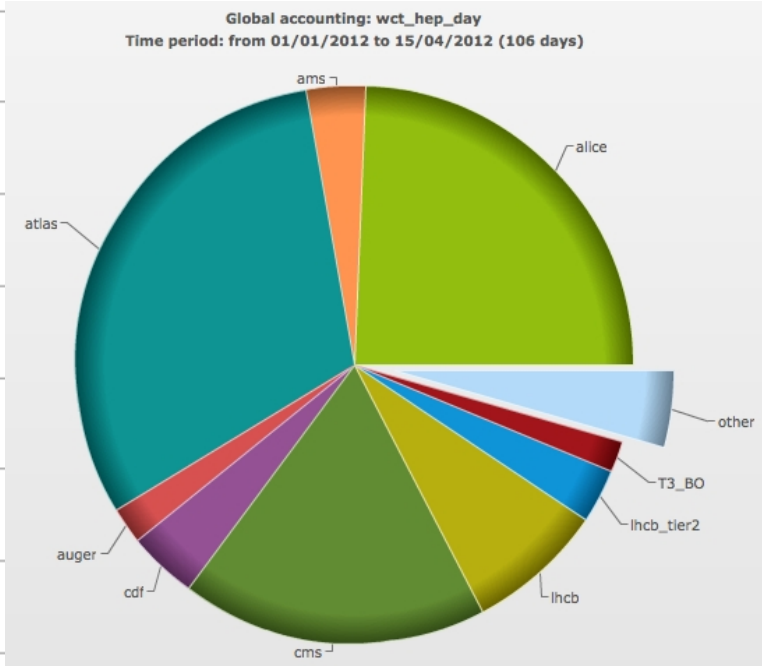


Il CNAF è tra i migliori Tier1 di ATLAS e lotta per la terza posizione

Utilizzo risorse al CNAF

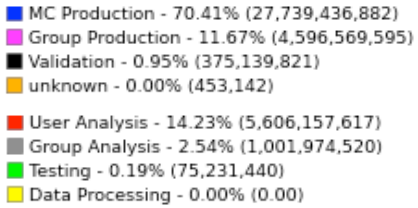
WCT HEP-SPEC day medi / Pledged 2011

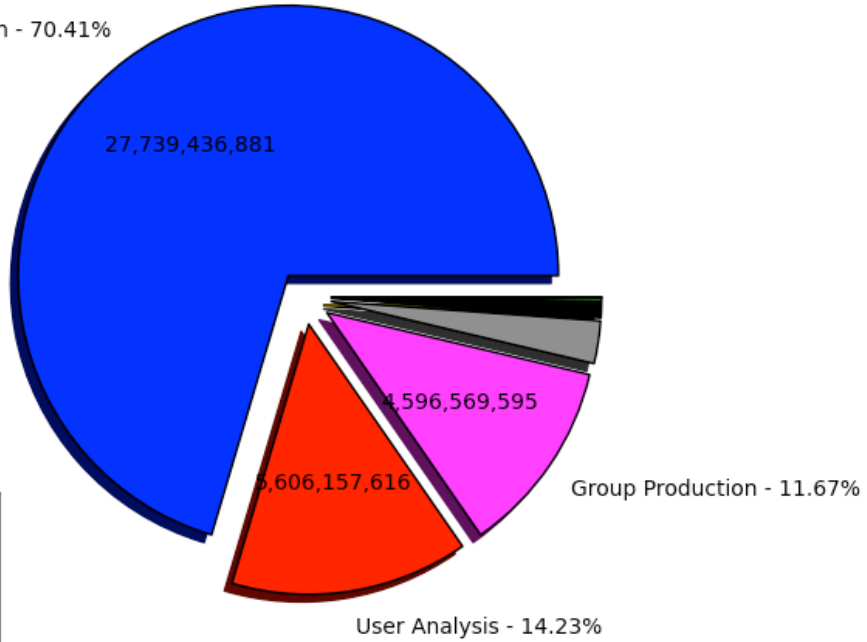




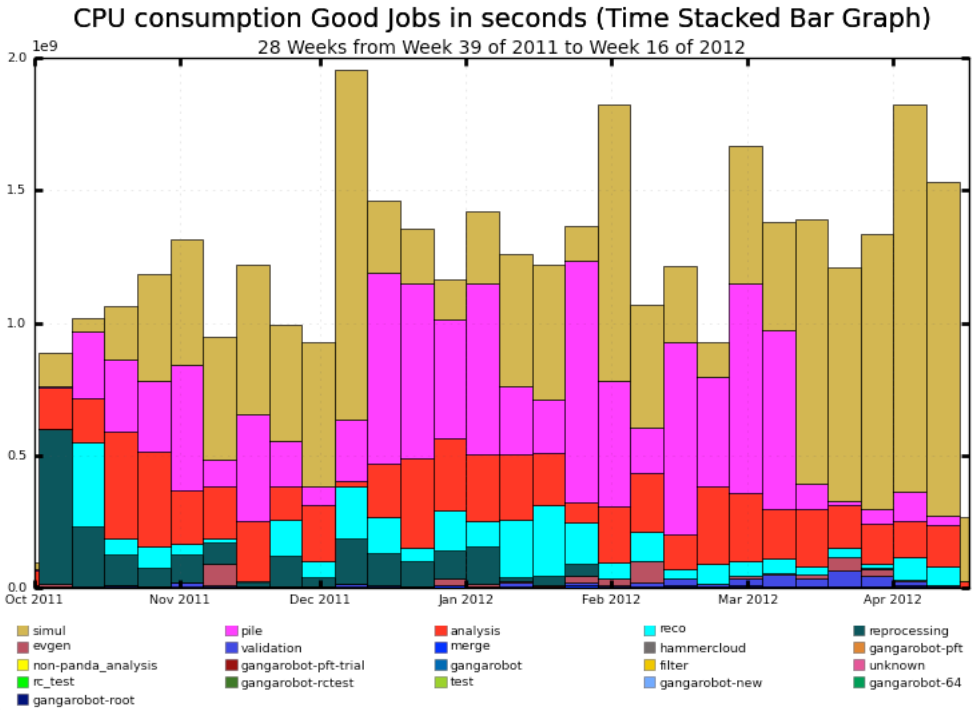
Attività al CNAF

Uso delle risorse suddiviso per attività





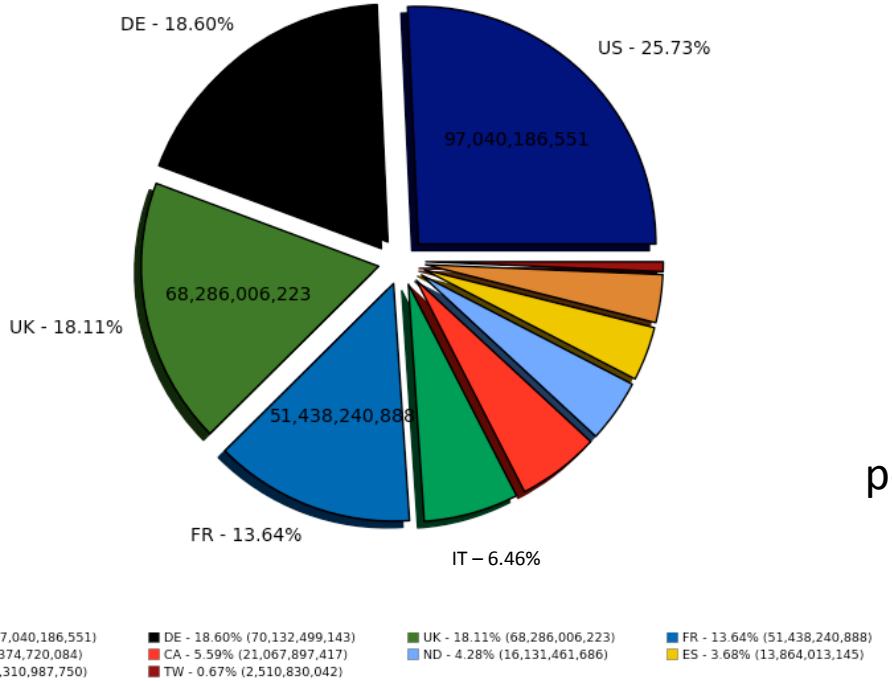
simile agli altri Tier1



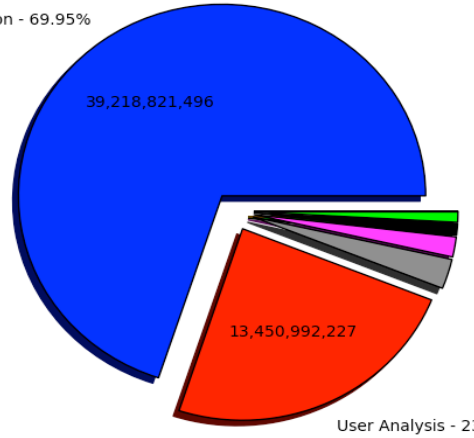
Corretto rapporto prod/analisi

Uso risorse nei Tier2

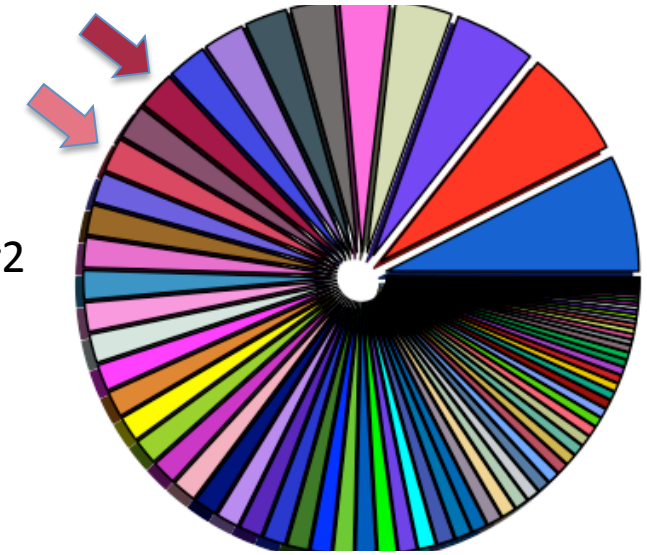
Uso risorse per "*Processing Cloud*"

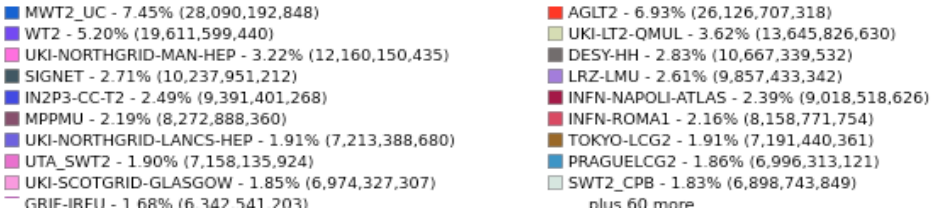


per attività



per Tier2

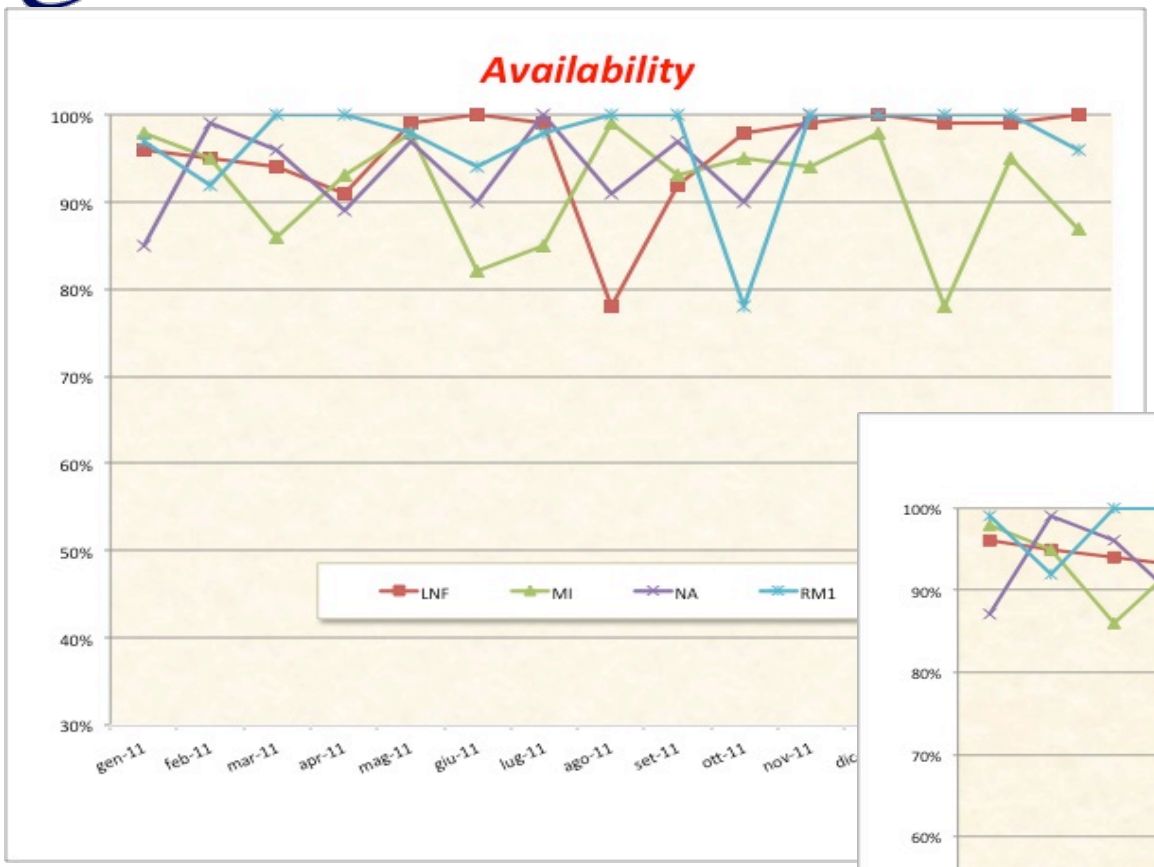




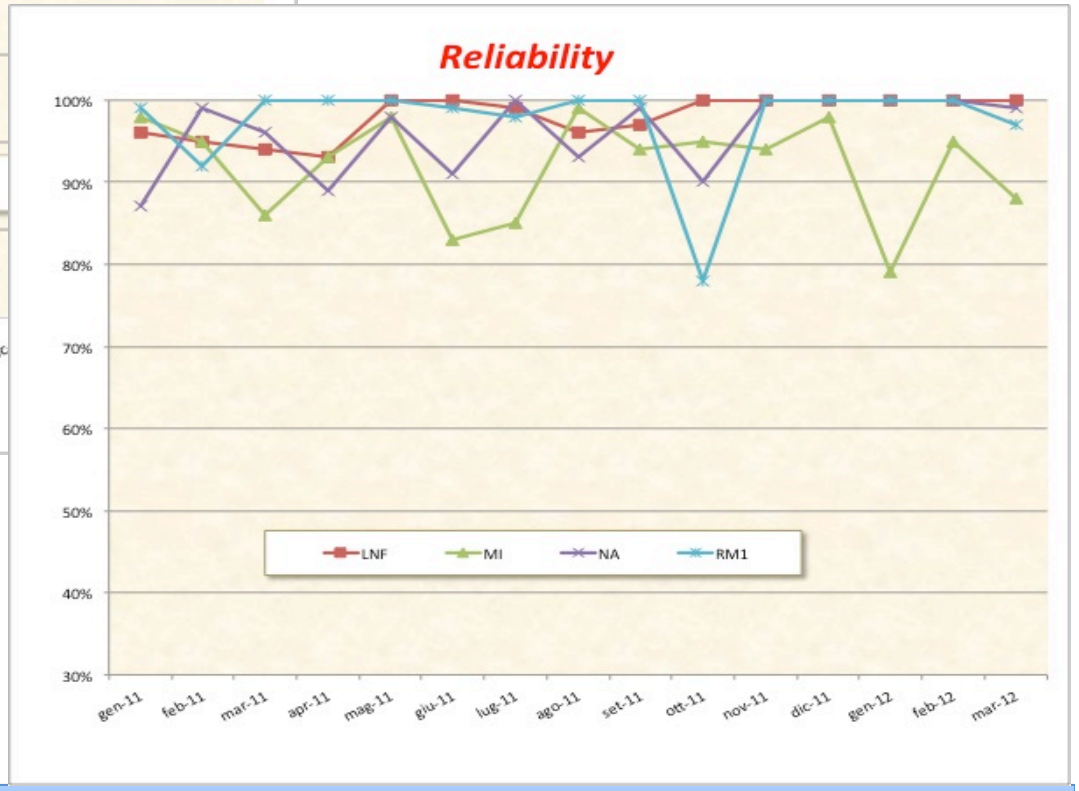
Quinta cloud

La percentuale può essere molto diversa dai pledges dichiarati a causa delle risorse a disposizione nelle varie per le attività nazionali (anche in IT)

Reliability & Availability 2011-12



Valori medi 2011-12			
Frascati		Milano	
rel	ava	rel	ava
98%	96%	92%	92%
Napoli		Roma	
rel	ava	rel	ava
96%	95%	98%	97%



Availability =
time_site_is_available/total_time

Reliability =
time_site_is_available/
(total_time-time_site_is_sched_down)

Classificazione dei Tier2

- Necessità di individuare i siti più affidabili per l'analisi cui inviare la maggior parte dei dati.
- Classificazione in base alle performance (stabilità)

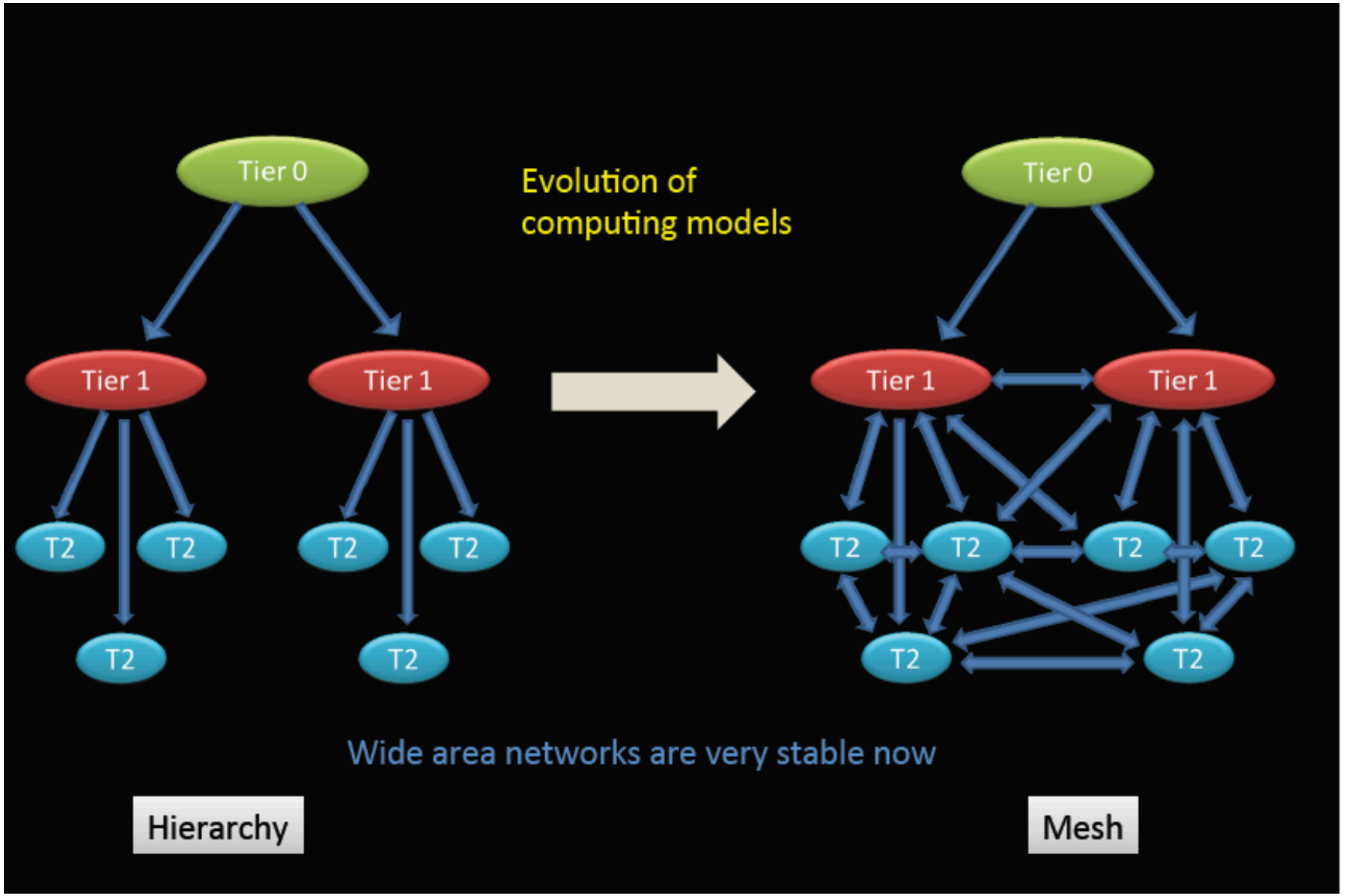
4 Gruppi

- Alpha: (60% share): T2D con rel > 90%
- Bravo: (30% share): non T2D con rel> 90%
- Charlie: (10% share): 80% < rel < 90%
- Delta: (0% share): rel <80%

T1	tier	%	T2 Group	tiers status
CA	AUSTRALIA-ATLAS_DATADISK	1.2 %	bravo	registered: 6 ready: 6
	CA-ALBERTA-WESTGRID-T2_DATADISK	1.2 %	bravo	
	CA-MCGILL-CLUMEQ-T2_DATADISK	1.2 %	bravo	
	CA-SCINET-T2_DATADISK	2.4 %	alpha	
	CA-VICTORIA-WESTGRID-T2_DATADISK	2.4 %	alpha	
	SFU-LCG2_DATADISK	2.4 %	alpha	
DE	CSCS-LCG2_DATADISK	2.4 %	alpha	registered: 11 ready: 11
	CYFRONET-LCG2_DATADISK	0.6 %	charlie	
	DESY-HH_DATADISK	2.4 %	alpha	
	DESY-ZN_DATADISK	2.4 %	alpha	
	GOEGRID_DATADISK	2.4 %	alpha	
	HEPHY-UIBK_DATADISK	1.2 %	bravo	
	LRZ-LMU_DATADISK	2.4 %	alpha	
	MPPMU_DATADISK	2.4 %	alpha	
	PRAGUELCG2_DATADISK	1.2 %	bravo	
	UNI-FREIBURG_DATADISK	2.4 %	alpha	
	WUPPERTALPROD_DATADISK	2.4 %	alpha	
ES	IFAE_DATADISK	2.4 %	alpha	registered: 6 ready: 6
	IFIC-LCG2_DATADISK	2.4 %	alpha	
	LIP-COIMBRA_DATADISK	0 %	delta	
	LIP-LISBON_DATADISK	2.4 %	alpha	
	NCG-INGRID-PT_DATADISK	2.4 %	alpha	
	UAM-LCG2_DATADISK	2.4 %	alpha	
FR	BEIJING-LCG2_DATADISK	2.4 %	alpha	registered: 11 ready: 11
	GRIF-IRFU_DATADISK	0 %	delta	
	GRIF-LAL_DATADISK	0.6 %	charlie	
	GRIF-LPNHE_DATADISK	2.4 %	alpha	
	IN2P3-CPPM_DATADISK	1.2 %	bravo	
	IN2P3-LAPP_DATADISK	2.4 %	alpha	
	IN2P3-LPC_DATADISK	0 %	delta	
	IN2P3-LPSC_DATADISK	2.4 %	alpha	
	RO-02-NIPNE_DATADISK	0 %	delta	
	RO-07-NIPNE_DATADISK	0 %	delta	
	TOKYO-LCG2_DATADISK	1.2 %	bravo	
IT	INFN-FRASCATI_DATADISK	1.2 %	bravo	registered: 4 ready: 4
	INFN-MILANO-ATLASC_DATADISK	0.6 %	alpha	
	INFN-NAPOLI-ATLAS_DATADISK	2.4 %	alpha	
	INFN-ROMA1_DATADISK	2.4 %	alpha	
NL	RRC-KI_DATADISK			registered: 9 ready: 8
	CSTCDIE_DATADISK	0 %	delta	
	IL-TAU-HEP_DATADISK	1.2 %	bravo	
	JINR-LCG2_DATADISK	1.2 %	bravo	
	RU-PNPI_DATADISK	0 %	delta	
	RU-PROTVINO-IHEP_DATADISK	1.2 %	bravo	
	TECHNION-HEP_DATADISK	0 %	delta	
	TR-10-ULAKBIM_DATADISK	0.6 %	charlie	
	WEIZMANN-LCG2_DATADISK	1.2 %	bravo	
UK	UKI-LT2-QMUL_DATADISK	2.4 %	alpha	registered: 13 ready: 13
	UKI-LT2-RHUL_DATADISK	1.2 %	bravo	
	UKI-LT2-UCL-HEP_DATADISK	0 %	delta	
	UKI-NORTHGRID-LANCS-HEP_DATADISK	2.4 %	alpha	
	UKI-NORTHGRID-LIV-HEP_DATADISK	1.2 %	bravo	
	UKI-NORTHGRID-MAN-HEP_DATADISK	2.4 %	alpha	
	UKI-NORTHGRID-SHEF-HEP_DATADISK	1.2 %	bravo	
	UKI-SCOTGRID-ECDF_DATADISK	2.4 %	alpha	
	UKI-SCOTGRID-GLASGOW_DATADISK	2.4 %	alpha	
	UKI-SOUTHGRID-BHAM-HEP_DATADISK	1.2 %	bravo	
	UKI-SOUTHGRID-CAM-HEP_DATADISK	1.2 %	bravo	
	UKI-SOUTHGRID-OX-HEP_DATADISK	2.4 %	alpha	
	UKI-SOUTHGRID-RALPP_DATADISK	1.2 %	bravo	
US	AGLT2_DATADISK	0.6 %	charlie	registered: 7 ready: 7
	ILLINOISHEP_DATADISK	0 %	delta	
	MWT2_DATADISK	2.4 %	alpha	
	NET2_DATADISK	2.4 %	alpha	
	OU_OCHEP_SWT2_DATADISK	0 %	delta	
	SLACXRD_DATADISK	2.4 %	alpha	
	SWT2_CPB_DATADISK	2.4 %	alpha	

Distribuzione dei dati

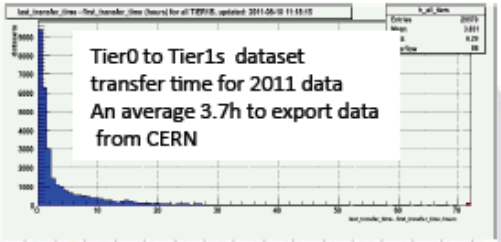
Evoluzione del Computing Model

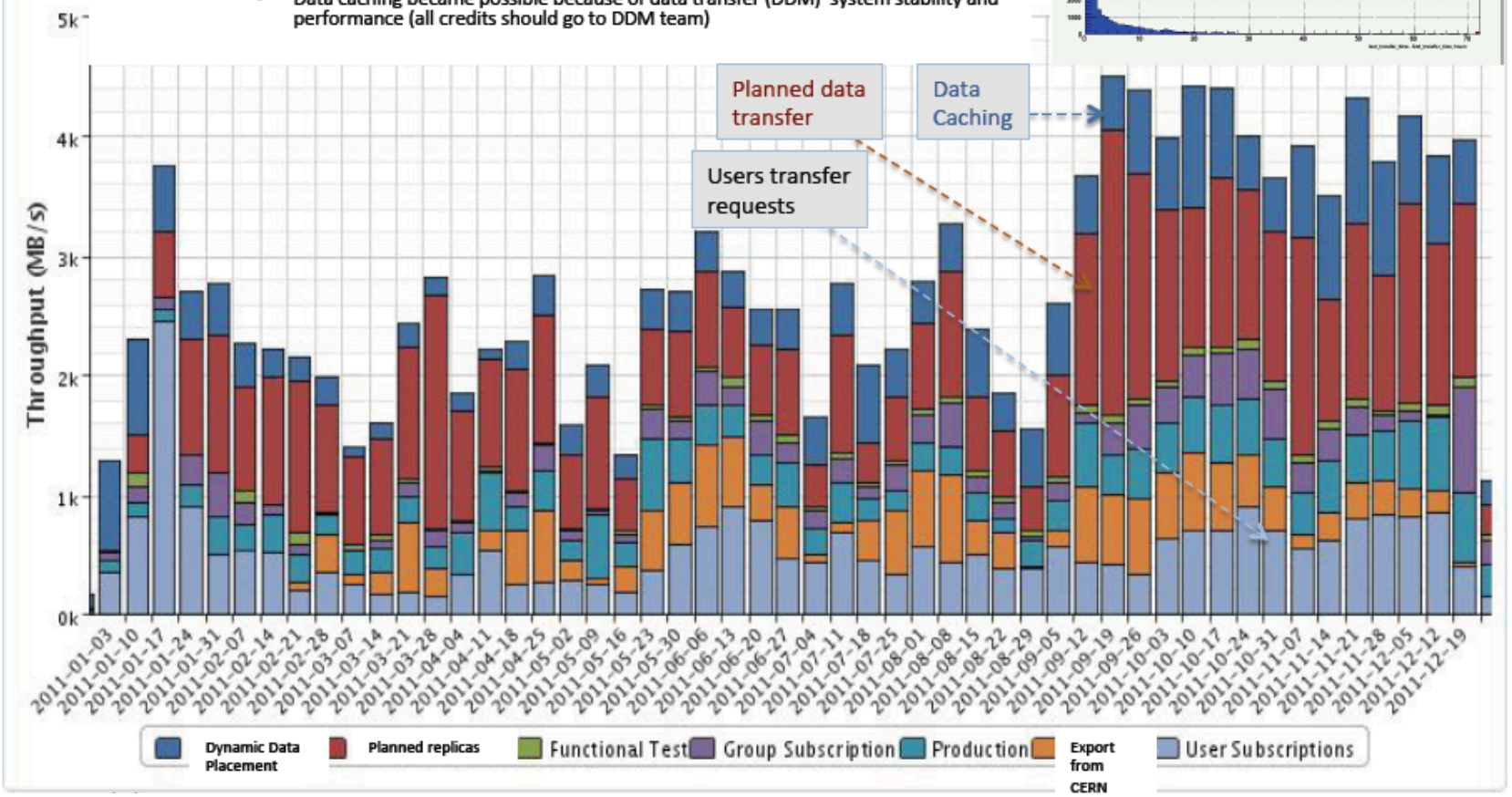


Distribuzione dei dati

- Statica (planned) = distrubuzione predefinita secondo share fissati
- Dinamica (data caching) = distribuzione in base alla popolarità dei dati

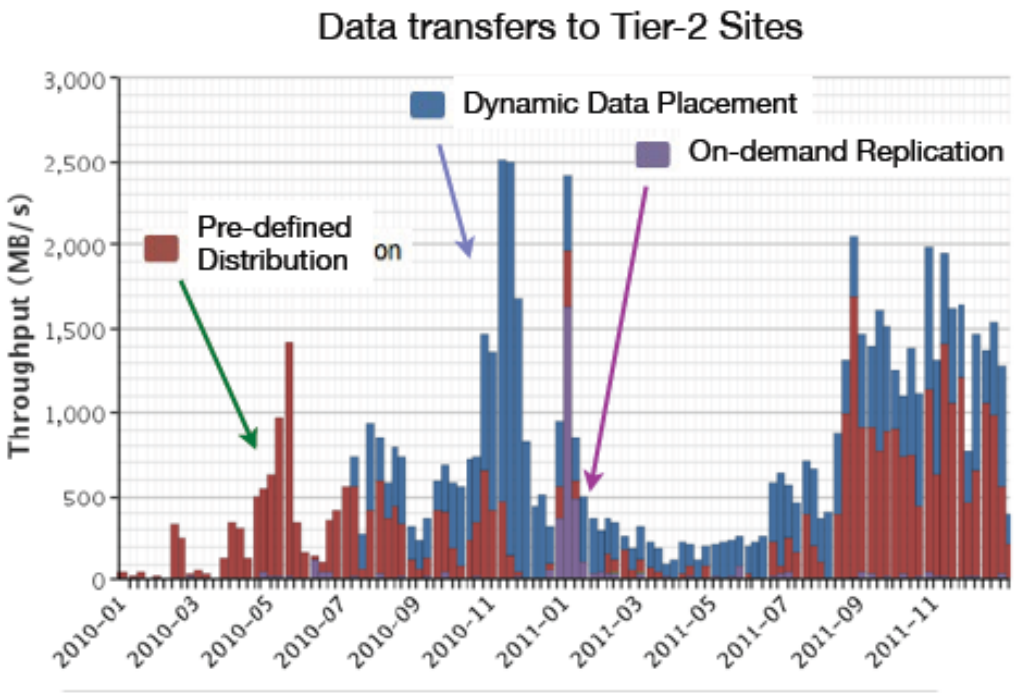
- 2011 Weekly Average Data Throughput per week in MB/s
  - Balance between planned replicas, dynamic placement and user requests
  - Excellent data transfer efficiency
    - Success rate ~100%
      - With automatic retry policy
    - Data caching became possible because of data transfer (DDM) system stability and performance (all credits should go to DDM team)





Distribuzione dei dati

- Nel 2011 si è cercato il giusto rapporto tra il volume di dati trasferiti dinamicamente e staticamente per permettere che una frazione significativa dell'analisi fosse svolta nei Tier2 insieme ad un uso saggio del disco
- Le comunità di utenti fanno capo soprattutto ai Tier2 (cpu e disco dedicati, accesso diretto), era dannoso concentrare l'analisi solo ai Tier1 (inizi 2011)



Replica dei dati prestabilita:

- Tier1, replica per ridondanza (consolidamento), copia primaria
- Tier2, replica per l'analisi, primaria
- Tier2, replica extra per l'analisi, copia secondaria

Determinata dallo share della cloud (Tier1) e dalla classificazione (Tier2)

Replica dinamica dei dati

- Tier1 e Tier2, basata sulla popolarità dei dati, copia secondaria

Determinata dall'utilizzo dei siti

Dati disponibili in Italia

- DATADISK
  - Area di storage centrale (pledged)
  - contiene sia i dati preplaced prodotti centralmente che quelli prodotti dal PD2P, circa il 10% del totale. I secondari vengono cancellati quando non utilizzati
- GROUPDISK
  - Aree gestite dai gruppi (pledged), dati sempre più utilizzati per l'analisi
  - CNAF: SUSY e TOP
  - Milano: MS e EGAMMA
  - Napoli: HIGGS e TRIGGER
  - Roma1: HIGGS e MUONI
- LOCALGROUPDISK
  - Area di storage (quasi) permanente per gli utenti italiani
  - Uso da massimizzare sfruttando l'accesso diretto dai *Tier3*, root o Proof on Demand, o con job di griglia (prun)

- Sottoscrizione automatica dell'output dell'analisi sul proprio LOCALGROUP quando si esegue l'analisi in griglia, in qualsiasi sito (solo per i dati da conservare)
  - option "--destSE=INFN-NAPOLI-ATLAS_LOCALGROUPDISK "
- Da pochissimo disponibile un tool, basato su dq2, che manda una mail quando un dataset arriva nel sito specificato (grazie a Manoj)
  - utilissimo per minimizzare la latenza dell'analisi

Accesso ai dati

Accesso ai dati

- Accesso in GRID
  - Pathena, Prun, Proof on Demand
  - Analisi diretta o skimming su AOD, ESD, D3PD, NTUP
  - Mette a disposizione un numero molto maggiore di risorse sul singolo sito o, parallelizzando l'attività, su più centri di calcolo distribuiti a livello geografico
  - Necessario per tutte le attività su grandi sample di dati o urgenti
- Accesso diretto
  - Root, Proof, athena per sviluppo e test di codice
  - Analisi finale
  - Accesso diretto dalle CPU *à la Tier3* ai dati nello storage del Tier2. Nativo nei siti Storm (MI) o, grazie al protollo XROOTD adottato alla fine dell'anno scorso, nei siti DMP (LNF, NA, RM1)
  - Evita la duplicazione delle NTUP per l'analisi finale localmente nel Tier3

Tipica analisi in 2 step:

1. Skimming in GRID su dati di gruppo con produzione di NTUP light salvate in LOCALGROUP
2. Analisi interattiva con ROOT/PROOF o in GRIGLIA con PoD sui dati

vedi esempio analisi H→ZZ di Valerio

Proof on Demand (PoD)

Elisabetta Vilucchi e
Roberto De Nardo (LNF)

- **Proof** – tool che parallelizza l'analisi con ROOT su diversi core della stessa macchina (Proof Light) o su un cluster di nodi
- **Demand** – possibilità di usare, a richiesta, i nodi di una farm destinata prevalentemente ad altri scopi (per esempio Tier2/3 in GRID)

- L' utente lancia il master PoD sulla propria macchina che apre un canale verso il cluster e consente all'utente di richiedere l'allocazione di un certo numero di job slot
- A questo punto l'utente invia la richiesta che si traduce in un numero di job in coda pari al numero di job slot richieste
- L'utente puo' controllare l'allocazione delle risorse che gli appariranno disponibili non appena i job vanno in stato di running. A questo punto l'utente puo' sottomettere l'analisi con Proof sui nodi di calcolo in cui compare un job PoD running
- Proof, in base ai nodi a disposizione, suddivide l'analisi in un certo numero di job paralleli

Proof on Demand (PoD)

Elisabetta Vilucchi e
Roberto De Nardo (LNF)

Proof on Demand su un cluster Grid: Tier2 o Tier3

- E' stato sviluppato un plugin di PoD per gLite, che da la possibilità agli utenti di attivare un cluster Proof “on demand” su una farm in Grid con middleware gLite
- Gli utenti, connettendosi ad una UI, possono lanciare PoD e riservare un certo numero di nodi sulla farm di un Tier2/3
- La gestione delle risorse e' simile a quella del cluster locale e il codice per il setup di PoD viene fatto direttamente da cvmfs, disponibile ormai nella maggior parte dei siti di ATLAS
- Da ultimare, prevista per un prossimo sviluppo, un plugin di PoD per Panda, il sistema ufficialmente riconosciuto dalla collaborazione per fare analisi, in modo tale che sia possibile visualizzare nell'accounting dell'esperimento anche i job di analisi fatti con questo plugin

Proof on Demand (PoD)

Elisabetta Vilucchi e
Roberto De Nardo (LNF)

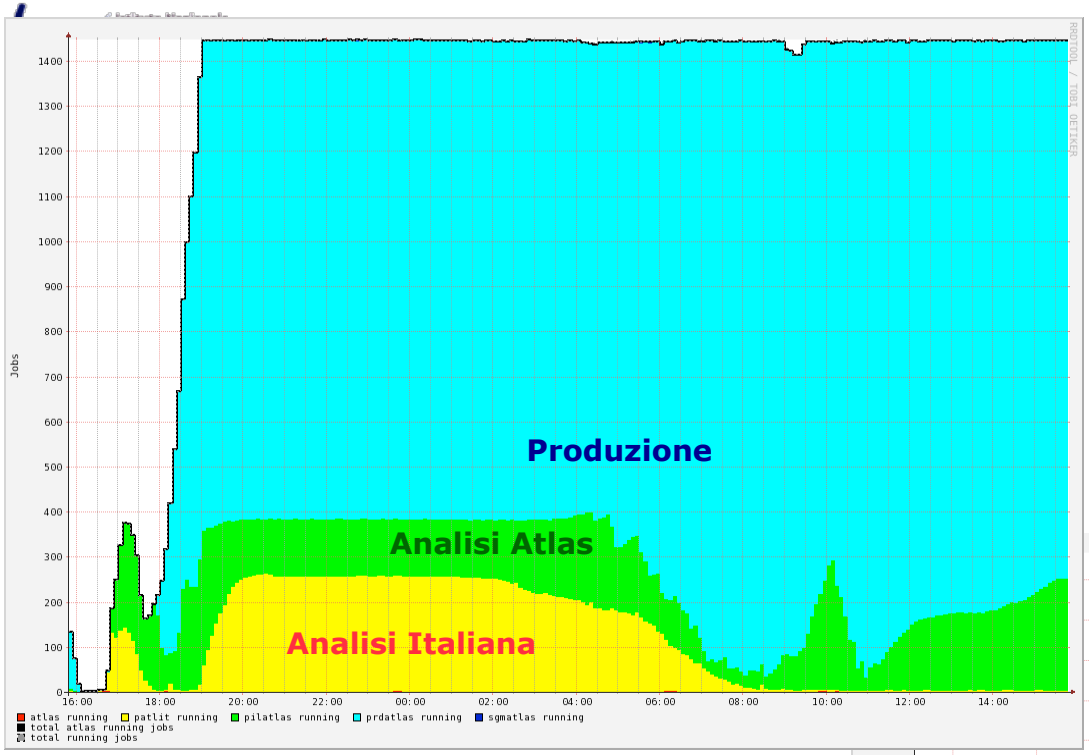
- Da domani (18-19/04) si terrà un tutorial ai Laboratori Nazionali di Frascati per illustrare l'utilizzo di Proof, PoD e del nuovo plugin per gLite: http://agenda.infn.it/conferenceDisplay.py?confId=4933
- Per partecipare è sufficiente possedere un certificato digitale, essere iscritti alla VO ATLAS ed avere un account al Cern
- Molto utile per tutti coloro che fanno analisi

Utilizzo risorse Italiane

Utilizzo efficiente delle risorse: priorità degli utenti italiani nei siti Grid di ATLAS Italia

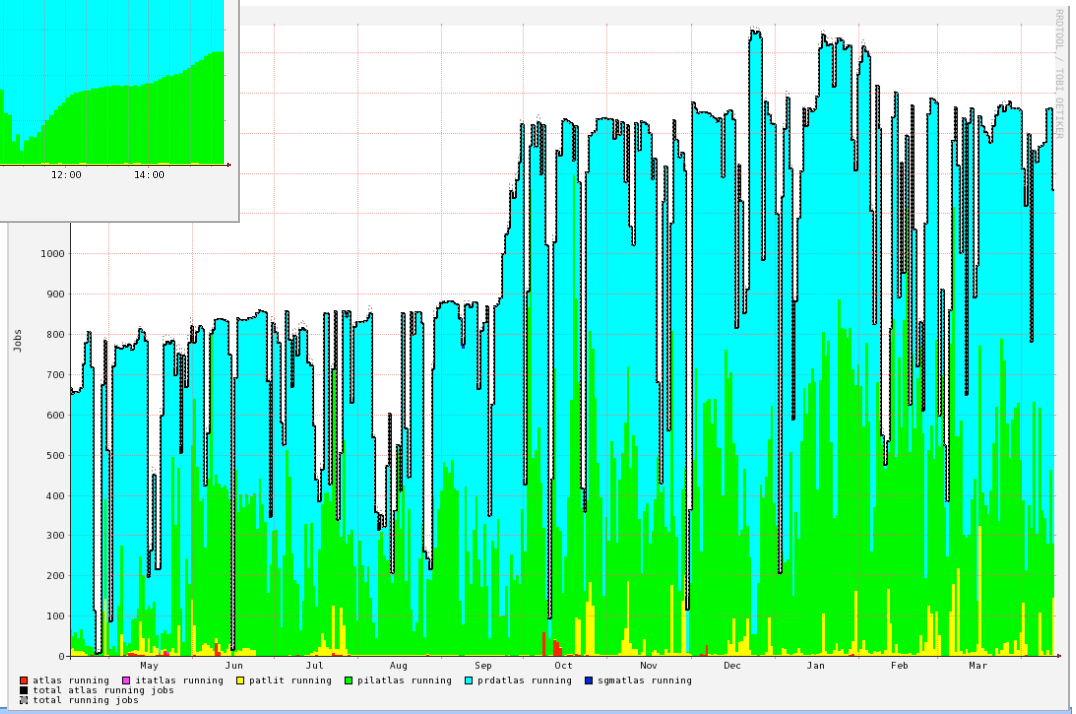
- **Gli utenti italiani possono utilizzare uno share dedicato dei siti**
  - **Macchine fisicamente riservate o slot ad alta priorità**
  - **Gli utenti italiani, identificati automaticamente da Panda tramite la presenza del gruppo VOMS /atlas/it nel loro proxy accedono agli share dedicati, senza alcuna operazione da parte loro**
  - **I job sottomessi ad un sito italiano da un utente italiano che ha il gruppo /atlas/it partiranno prima degli altri job di analisi, anche se accodati in momenti successivi**
  - **Per assicurarsi di avere l'accesso allo share di /atlas/it è sufficiente controllare nel momento della registrazione in VOMS di richiedere il gruppo /atlas/it, anche se il certificato non proviene dall'INFN, a patto che si faccia parte della comunità italiana**
  - **Nel caso in cui non si possegga il gruppo si può farne richiesta anche dopo essersi registrati**

Risorse dedicate attività IT



Job running in un Tier2:

← Ultima settimana
↓ Ultimo anno



- L'analisi viene svolta efficacemente nei Tier2 Italiani
- Le risorse dedicate (dedicabili) sono significative
- Permettono ai job italiani di andare in run più velocemente senza essere accodati agli altri
- Non solo analisi, anche prod MC

Produzione Monte Carlo

ATLAS accetta le produzioni Monte Carlo private (G4, digi e reco) su risorse *unpledged*

- ***Additional Production***
  - Produzione di gruppo in aggiunta a quelle già concordate
  - Deve essere approvata dal Physics Coordinator specificando nella richiesta che si tratta di produzione Addizionale
    - Sanity check contro richieste poco motivate
  - Eseguita centralmente (Prod sys) ma i dati non sono distribuiti dal DDM
    - Solo il PD2P può spostare questi dati
    - Il team di simulazione invia al sito specificato che offre le risorse

- ***Private Production***
  - User production diverse da quelle ufficiali di ATLAS. Non vista di buon occhio
  - Eseguita individualmente
  - Richiesta la validazione confrontandola con un sample di riferimento eseguito centralmente

Produzione Monte Carlo

Qualche numero …….

- *Simulation Time = 3300 HepSpec06 sec*
- *Reconstruction Time = 830 HepSpec06 sec*
  - valori a 8 TeV ottenuti dalla media dei tempi a 7 TeV sull'intero set di canali (e # totale di eventi) scalato a 8 TeV

Potenza di calcolo a disposizione:

- Valor medio di un singolo core = 10 HepSpec06
  - 1 evento richiede ~ 7 minuti per sim e reco
- Potenza di calcolo totale nei Tier2 (estate 2012) = ~52 kHS06
- Stima potenza di calcolo per attività italiane = ~10 kHS06
  - Dopo aver sottratto: WLCG Pledges = 26.6 kHS06 e varie = ~2.5 kHS06 e considerato: efficienza = 85%, risorse non disponibili = 5%
  - da usare per analisi e produzione MC con percentuali da decidere

10^5 eventi al giorno e $3x10^6$ al mese (nell'ipotesi share 50%)

Produzione Monte Carlo

10^5 eventi al giorno e $3x10^6$ al mese (nell'ipotesi share 50%)

Attenzione, è una novità, quindi è da testare e verificare

- Verificare se i gruppi di fisica sono d'accordo, soprattutto nel caso di produzioni private, e le procedure da seguire in tal caso
- Dal punto di vista dei siti, richiede la pubblicazione delle CPU totali e pledged in Panda e il test che la procedura di assegnazione dei task e delle risorse dedicate funzioni correttamente

- Siamo pronti a collaborare con i gruppi italiani e i responsabili dell'analisi per definire la lista di produzioni MC e il corretto peso rispetto all'analisi e collaborare con ATLAS per i test