ISTITUTO NAZIONALE DI FISICA NUCLEARE

Sezione di Pisa

INFN/CCR-09/06
23 Dicenbre 2009

CCR-33/2009/P

CALCOLO SCIENTIFICO: PRIME METODOLOGIE QUANTITATIVE PER UN AMBIENTE DI PRODUZIONE

Alberto Ciampa, Enrico Mazzoni

INFN-Sezione di Pisa, Largo B. Pontecorvo, 3, I-56127 Pisa, Italy

Abstract

Vengono definiti il contesto e l'attività di Calcolo Scientifico, prendendo ad esempio la Sezione di Pisa, intesa come produzione, pensando ad un approccio di tipo industriale. Si propone una prima metodologia per la valutazione quantitativa dei livelli di produzione, dell'efficienza nell'utilizzo degli impianti e della distribuzione dei costi.

La valutazione dei consumi e l'attribuzione dei costi sono basate sul consumo di energia elettrica: viene descritto come questo parametro possa essere considerato solo un esempio pratico e viene accennato a quali tipi di consumi, costi, modalità possa essere estesa l'applicazione della metodologia proposta.

I risultati di un survey con applicazione alla Sezione di Pisa sono allegati in appendice.

*Published by **SIS–Pubblicazioni***
Laboratori Nazionali di Frascati

Indice

1 Introduzione

Il presente lavoro scaturisce da una richiesta del Direttore di Sezione relativa ad un survey sulle attività di calcolo scientifico presso la Sezione di Pisa. Ponendoci dal punto di vista di un ipotetico settore di Calcolo Scientifico di Sezione, abbiamo inteso questa come una attività di produzione, pensando ad un approccio di tipo industriale:

- Non è, in generale, importante il singolo pezzo, ma il flusso di produzione.
- La qualità del sistema è data dal livello di produzione e dalla sua efficienza.
- Salvo in casi importanti non entriamo nel merito di cosa si produce, ma ci interessa come si produce, intendendo come vengono usati gli impianti.
- Si può, anzi è indispensabile, misurare il costo di produzione in due differenti condizioni: in termini generali ed in modo disaggregato per le diverse "linee di produzione" (gruppi di utenti).
- In linea di principio si può, e si deve essere pronti ad accettare "commesse" sia interne (istituzionali, quindi obbligatorie) sia esterne.

Questo lavoro costituisce il primo tentativo di fissare delle metodologie (e delle approssimazioni) per la valutazione quantitativa dei livelli di produzione, di efficienza e della distribuzione dei relativi costi di ciò che viene prodotto.

2 Obiettivi

Definito il contesto "Calcolo Scientifico" ci poniamo i seguenti obiettivi:

- Misurare le potenzialità di produzione in relazione alle risorse globali:
  - Spazio (unità rack), alimentazione elettrica, raffreddamento, reti.
- Misurare il livello attuale di produzione, inteso come capacità lorda disponibile, in relazione ai sistemi di calcolo e storage.
- Misurare il livello di capacità produttiva utilizzata (*efficienza generale*) e, ove possibile, l'efficienza con la quale viene utilizzata (*efficienza specifica*, funzione della specifica produzione).
- Misurare il costo effettivo per unità prodotta (per es. "day-core"), analizzando eventuali differenze tra le diverse metodologie produttive e "linee di produzione"

Per le stime dei costi, generali e disaggregati, è stato preso in considerazione il costo della corrente elettrica consumata: si tratta, però, solo di un esempio pratico. La metodologia fondamentale, vero obiettivo del presente lavoro, consiste nel riuscire ad effettuare un accounting affidabile delle risorse di base per ciascun tipo di produzione fino alla singola unità prodotta (unità di calcolo effettuato). Con tale metodologia di accounting, qui applicata ai consumi elettrici, si punta a poter attribuire in modo affidabile tutti i tipi di costi: funzionamento in tutte le sue categorie, ammortamenti ed anche, dotandosi di una opportuna organizzazione, costi di personale. Una volta raggiunto, tale obiettivo renderà possibile:

- l'analisi dell'efficienza delle diverse "linee di produzione" e lo sviluppo di strategie di ottimizzazione;
- una corretta valutazione di eventuali "commesse" anche esterne, indispensabile per poter percorrere la strada di servizi verso terzi;
- una programmazione delle esigenze di calcolo scientifico dei vari gruppi di utenti/esperimenti.

Un generale ed affidabile accounting disaggregato per le spese di Calcolo (applicabile anche alle Reti ed ai Servizi) potrà risultare utile in sede di definizione delle politiche di finanziamento da parte dell'Istituto verso tali attività, permettendo la possibilità di scelta tra ciò che debba ricevere finanziamenti "a corpo" e ciò che debba riceverne "a consumo".

3 Contesto: il Calcolo Scientifico alla Sezione di Pisa

Dal punto di vista logistico ed infrastrutturale il Calcolo Scientifico ed i servizi di Sezione (e-mail, web, spazio storage degli account, servizi interattivi ecc.) condividono molte risorse. L'analisi che segue partirà dal tentativo di separare per quanto possibile i due ambienti. Per attività di Calcolo Scientifico e relative risorse si intendono:

- Calcolo Grid: worker node, macchine che ospitano i diversi servizi middleware Grid, server SRM.
- Farm e cluster di esperimento o di Gruppo, utilizzate in batch o in interattivo.
- Gestione dello spazio di storage utilizzato dalle due attività sopra elencate, compresi i disk server necessari.
- Gestione della rete LAN di sala, per la parte asservita ai sistemi precedentemente menzionati.

- La rete WAN dedicata al calcolo Grid (l'altro link viene considerato generale di Sezione).

3.1 Paradigmi di Utilizzo

L'attività di produzione può avvenire in diversi modi. Qui si raggrupperanno in due soli paradigmi di utilizzo, caratterizzati come segue:

- Grid:
  - solo modalità batch (gestore locale di code LSF);
  - risorse di calcolo condivise tra le VO accettate;
  - assegnazione delle risorse dinamica in proporzione al numero di richieste presenti, regolata dal meccanismo di "fair-share" che tiene conto della percentuale di finanziamento al Grid Data Center effettuata da ciascuna VO;
  - accounting delle risorse in base all'utilizzo effettivo (corretto con la percentuale generale di utilizzo, la già menzionata *efficienza generale*).
- Farm e Cluster di esperimento:
  - l'esperimento titolare delle risorse può accedervi come preferisce;
  - risorse di calcolo riservate all'esperimento che le ha acquisite;
  - accounting delle risorse in base alla allocazione statica e non all'utilizzo dinamico;

3.2 Comunità di Utenti

Politiche verso gli utenti:

- Grid:
  - in generale si apre l'accesso a tutte le VO riconosciute da INFN-GRID;
  - a chi non ha partecipato al finanziamento del Grid Data Center viene riconosciuto un valore di fair-share "di ingresso" di pochi punti percentuali;
  - a chi lo richieda, ed in generale agli utenti maggiori o con importanti problemi, si fornisce assistenza per sottomettere job che vadano a buon fine e si cura, ove possibile, il raggiungimento di una buona efficienza di utilizzo di CPU.
- Farm e Cluster di esperimento:

– richieste di housing ben motivate per una farm o un cluster: tali richieste sono state accettate, previa verifica della sussistenza di validi motivi che impedissero l'utilizzo di Grid;

– accounting delle risorse, inclusi tutti i servizi necessari all'operatività della farm/cluster (unità rack, reti, eventuali reti veloci, corrente, etc.);

– assistenza, solo in caso di evidenti e specifici problemi;

– hosting (in via di perfezionamento) a vari livelli di servizio di farm/cluster dedicati a commesse "esterne".

4 Tipologie di risorse

4.1 Risorse Globali

- *Spazio.* L'organizzazione dei rack di sala è fissata, così come è fissata in buona misura la loro destinazione d'uso. È quindi noto lo spazio disponibile per le risorse di Calcolo Scientifico, misurato in Rack Unit. Nel presente lavoro non ne verrà tenuto conto nell'accounting (ma viene calcolato per commesse esterne), ma è facile risalire al suo utilizzo dai consumi di risorse hardware di base (vedi paragrafo 4.2).

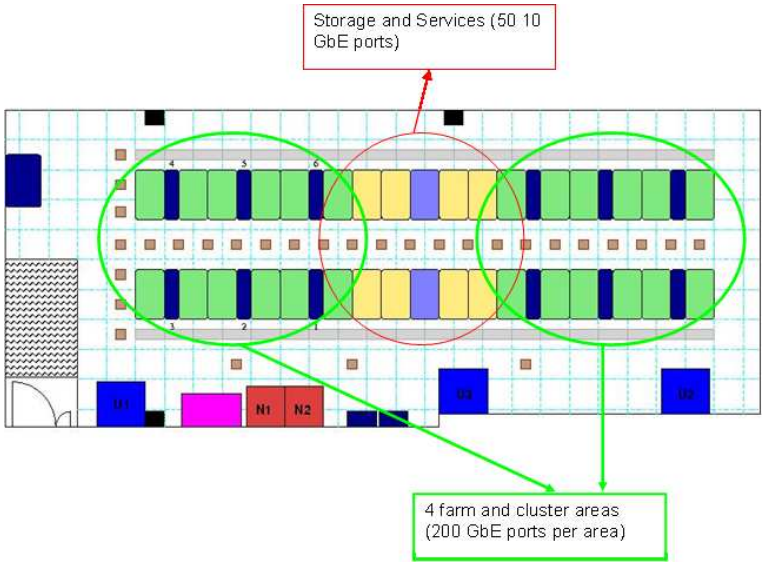


Figura 1: Layout sala calcolo INFN-Pisa

- *Alimentazione elettrica.* La disponibilità di corrente elettrica viene considerata una delle caratteristiche principali che determinano il livello potenziale di produzione.

Viene inclusa nell'accounting di produzione, con una serie di criteri che permettono l'assegnazione di consumo per ciascuna produzione (gruppo di utenti). Non viene fatta distinzione tra corrente UPS e non-UPS, poichè tutti i worker node sono in alimentazione "non UPS", avendo riservato l'alimentazione UPS ai soli servizi essenziali e allo storage. Nel presente lavoro viene usata come "case study" per dimostrare l'applicabilità della generale metodologia di accounting proposta.

- *Raffreddamento.* La capacità di raffreddamento (effettiva e potenziale, entrambe note alla Sezione di Pisa) è l'altra fondamentale grandezza che caratterizza il possibile livello di produzione. Non viene esplicitamente inclusa nell'accounting di produzione, ma vi rientra sotto forma di consumo elettrico della percentuale di raffreddamento usata (determinata a sua volta dal consumo delle macchine di produzione). In questo modo risulterà facile includere in futuro nell'accounting anche spese di ammortamento impianti, manutenzione, etc.

- *Reti*

  - *LAN.* Sia per le farm di calcolo (è noto il numero di server di calcolo per ciascuna) sia per Grid (è noto il numero di core e di CPU, quindi di server "logici" utilizzati) è facile effettuare un accounting delle risorse utilizzate. Con la nuova organizzazione di rete la disponibilità di link di caratteristiche opportune non viene considerata, per ora, una limitazione e non viene tenuto in conto nel presente lavoro un accounting analitico delle risorse LAN utilizzate. In caso di grosse installazioni si richiede la fornitura di quanto necessario per i collegamenti, così come per richieste particolari (ad es. reti veloci). Il consumo elettrico (ed in generale le risorse necessarie per la rete) entrano in accounting come percentuale utilizzata dalla specifica produzione rispetto al totale.
  - *WAN.* Per Grid abbiamo un link dedicato, mentre il resto transita sul link generale di Sezione. Non si considera per ora un accounting sul traffico di rete WAN. È comunque in progetto uno studio per l'accounting disaggregato del traffico su rete WAN per gruppo di utenti.

- *Servizi.* Si intendono qui i servizi asserviti al Calcolo Scientifico: servizi Grid, SRM, disk server. Non vengono specificatamente inclusi in accounting, anche perché risulta molto difficile quantificarne la percentuale di utilizzo. Anche i servizi entrano nell'accounting partecipando al consumo di corrente, calcolato sulla base del livello di produzione rispetto al totale. In ugual modo si possono considerare

gli altri costi legati ai servizi asserviti al “Calcolo Scientifico” e quindi assegnarli ai diversi gruppi di utenti/esperimenti (produzioni).

- *Altri Servizi.* Ci riferiamo ai servizi che sono necessariamente in comune tra le attività di Calcolo Scientifico ed i Servizi di Sezione (storage misto, infrastrutture di autenticazione ed autorizzazione, DNS, DHCP, etc.). Nel presente lavoro i costi di tali servizi vengono imputati alla Sezione per la loro estrema esiguità in termini di consumi elettrici (si tratta di poche unità di server e relativamente piccolo spazio storage). Diverso sarebbe il caso nel quale si volessero prendere in considerazione i costi umani (FTE), che comunque potrebbero entrare in accounting una volta stabilita una specifica metrica di consumo (per esempio: percentuale di indirizzi di rete di un gruppo di utenti rispetto al totale per i costi di gestione del DNS e DHCP, percentuale di utenti registrati rispetto al totale per la loro gestione, etc.).

4.2 Risorse Hardware

Sono qui considerate risorse hardware di due sole tipologie (tutto il resto essendo stato analizzato prima):

- *Server di calcolo.* Intesi come i server che fanno produzione, esclusi tutti i servizi asserviti. Entrano in accounting come consumo di corrente. Tutti gli altri tipi di costi, come già accennato, possono essere attribuiti analiticamente seguendo lo stesso criterio; con alcune naturali eccezioni (come ad esempio la distinzione tra diversi modelli di server e relativi costi per farm di esperimento che hanno allocazione statica sugli utenti).

- *Storage.* Essendo lo spazio storage dedicato, viene fornito a pagamento, tenendo conto del tipo di sistema su cui risiede. Non viene considerato nessun costo di consumo (ma sarebbe facile attribuire i costi di manutenzione in base allo spazio): non va calcolato l’ammortamento perché la risorsa è stata acquisita dall’utente, non viene calcolato il consumo di corrente (per ora) essendo infatti una frazione piccola del consumo totale che viene aggiunto ai consumi dei servizi di Sezione. Naturalmente se venisse installato un importante sistema dedicato, con consumi non più irrisori, sarebbe possibile conteggiarlo.

5 Dati disponbili: Ganglia

Di seguito vengono riassunte le informazioni riguardanti i dati disponibili sulle risorse dedicate al Calcolo Scientifico. La base dati è costituita da RRDtool, l’acquisizione di questi dati e l’accesso avvengono attraverso la suite Ganglia.

Il sistema è costituito da un demone, installato su ciascun nodo della farm, che raccoglie le informazioni relative al nodo stesso ad intervalli regolari (15s) e le condivide con gli altri nodi (pertanto ognuno di essi ha una tabella completa delle informazioni relative all'intera farm), e da un collettore esterno alla farm che si occupa di fare il polling dei dati contattando uno dei nodi della farm in questione per poi procedere all'archiviazione delle informazioni in una serie di database di tipo RRD (*Round Robin Database*, database a numero massimo di record fissato) dai quali vengono successivamente generati i grafici accessibili via web.

Per mantenere limitata la dimensione del DB, partendo da questi dati primari, si procede ad una riduzione attraverso opportune "funzioni di consolidamento" creando sub-set, sempre di dimensioni fissate, di dati a granularità meno elevata. Le funzioni di consolidamento disponibili in RRDtool per questo scopo sono 4: MIN, MAX, LAST, AVERAGE.

Il collettore, oltre a raccogliere le informazioni e a provvedere al riempimento e al consolidamento dei DB si occupa anche di aggregare le informazioni per l'intera farm in modo da fornire una visione d'insieme e non solo il dettaglio del singolo nodo.

5.1 Formato RRD e consolidamento

Ogni RRD è organizzato in vari Round Robin Archive (RRA). Ciascun RRA è descritto dalla funzione di consolidamento che lo crea, dalla frazione massima di dati primari "non definiti" che è tollerata, dal numero di dati primari da utilizzare nelle operazioni di consolidamento e dal numero di punti da conservare.

Nel caso di Ganglia ogni RRD è composto da 5 RRA così fatti:

- RRA[0]=RRA:AVERAGE:0.5:1:244: una misura ogni 15s (la media è fatta su un solo punto), si conservano 244 punti, il che significa che questo RRA copre un intervallo temporale di $15s \cdot 244 = 3660s = 1.02h$;

- RRA[1]=RRA:AVERAGE:0.5:24:244: la media di 24 dati primari raccolti ogni 15s forma uno dei 244 dati conservati, il che significa che questo RRA copre un intervallo temporale di $15s \cdot 24 \cdot 244 = 87840s = 24.4h \sim 1d$ e ciascun punto rappresenta la media dei dati primari su 360s;

- RRA[2]=RRA:AVERAGE:0.5:168:244: la media di 168 dati primari, raccolti ogni 15s, forma uno dei 244 dati conservati, il che significa che questo RRA copre un intervallo temporale di $15s \cdot 168 \cdot 244 = 614880s = 7.12d \sim 1w$ e ciascun punto rappresenta la media dei dati primari su 0.7h;

- RRA[3]=RRA:AVERAGE:0.5:672:244: la media di 672 dati primari, raccolti ogni 15s, forma uno dei 244 dati conservati, il che significa che questo RRA copre un intervallo temporale di $15s \cdot 672 \cdot 244 = 2459520s = 28.47d \sim 1m$ e ciascun punto rappresenta la media dei dati primari su 2.8h;

- RRA[4]=RRA:AVERAGE:0.5:5760:374: la media di 5760 dati primari, raccolti ogni 15s, forma uno dei 374 dati conservati, il che significa che questo RRA copre un intervallo temporale di $15s \cdot 5760 \cdot 374 = 32313600s = 374d \sim 1y$ e ciascun punto rappresenta la media dei dati primari su 24h.

Nei grafici in figura 2 sono riportati alcuni esempi di dati mostrati da Ganglia per una farm riferiti a vari intervalli temporali.

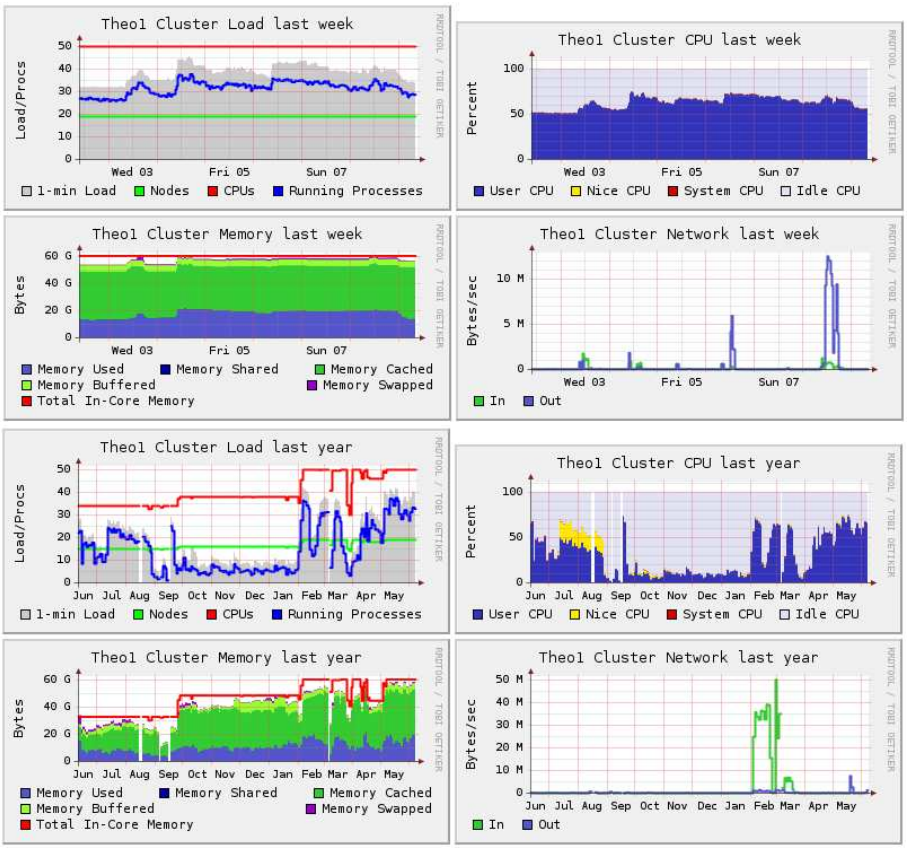


Figura 2: Esempio di schermate Ganglia con diversi intervalli temporali

5.2 Aggregazione dei dati

Come detto prima i dati dei singoli host vengono aggregati sia a livello di farm sia a livello di grid per presentare dei “summary”. Tale aggregazione è realizzata calcolando, per ogni

step temporale, la somma dei dati relativi a ciascun host acceso in quel momento. Inoltre nei DB viene aggiunto il relativo numero di host accesi. Nella figura 3 sono riportati alcuni esempi di dati “summary” mostrati da Ganglia per una farm riferiti a vari intervalli temporali.

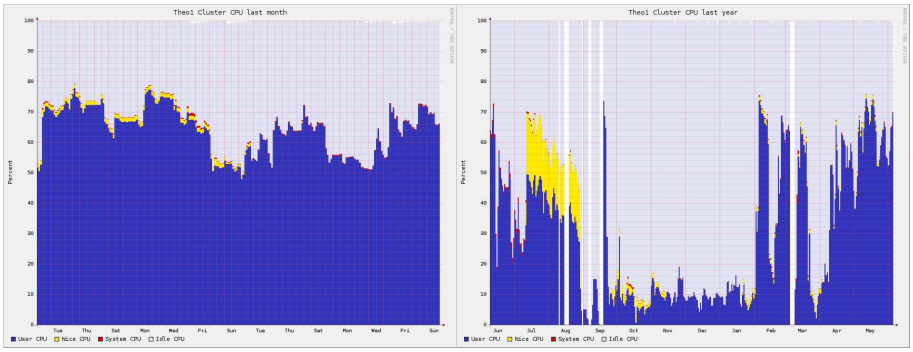


Figura 3: Esempio di dati aggregati per una farm

5.3 Dati disponibili

Per quanto riguarda l’utilizzo dei core, per ciascun nodo, viene misurata la percentuale di utilizzo nei vari stati: idle, system, nice, user e I/O wait (se supportato dal kernel) riferito al totale dei core presenti nel nodo. Quindi sono disponibili i seguenti DB:

- cpu_idle.rrd
- cpu_nice.rrd
- cpu_system.rrd
- cpu_user.rrd
- cpu_wio.rrd
- cpu_num.rrd

con ovvio riferimento al contenuto.

6 Elaborazioni per il Calcolo Scientifico

Ricordiamo come premessa che per ciascun server il consumo elettrico è determinato dal numero di CPU, mentre il livello di produzione dal numero di core.

Per quantificare il livello di utilizzo (produzione) di ciascuna farm/cluster interessa valutare il numero di core della farm/cluster occupati da ciascun tipo di carico e poi calcolarne l'integrale nell'arco di tempo di nostro interesse. A questo scopo l'idea è di partire dai dati "summary" e capire se e come sia possibile avere una buona stima di ciò che ci serve. Si considerano in dettaglio quali siano i dati disponibili, quali siano le grandezze che possiamo agevolmente calcolare e quindi confrontare con le grandezze, utili ai nostri scopi, che vorremmo calcolare.

6.1 Dati disponibili: dati primari

Per ogni istante t_i e per ciascun host h vengono misurate le seguenti grandezze:

- C_i^h numero di core presenti nell'host h all'istante t_i
- PI_i^h percentuale di core in stato idle nell'host h all'istante t_i
- PN_i^h percentuale di core in stato nice nell'host h all'istante t_i
- PS_i^h percentuale di core in stato system nell'host h all'istante t_i
- PU_i^h percentuale di core in stato user nell'host h all'istante t_i
- PW_i^h percentuale di core in stato I/O wait nell'host h all'istante t_i, questo parametro potrebbe non essere presente se il kernel del sistema non fornisce l'informazione.

Se il numero di host presenti nel cluster è k nei "summary" troveremo le seguenti grandezze:

- $\sum_{h=1}^{k} C_i^h$
- $\sum_{h=1}^{k} PI_i^h$ e analogamente per le altre percentuali.

Dato che uno degli scopi della survey è valutare l'efficienza di utilizzo di una farm, invece di utilizzare le percentuali relative ai vari stati (idle, I/O wait, nice e user) sono state definite due sole percentuali: una relativa alla "produzione",è stata chiamata "working" e indicata con Pw, e una relativa alla "non-produzione", denominata "sleeping" e indicheremo con Ps che sono così definite

$$Pw_i^h = PN_i^h + PU_i^h \tag{1}$$

$$Ps_i^h = PI_i^h + PW_i^h + PS_i^h \tag{2}$$

nel seguito si farà riferimento solo a queste percentuali.

6.2 Dati disponibili: dati consolidati

Per ciascun tipo di consolidamento si fa una media su n dati primari, con n che dipende dal consolidamento considerato. Quindi per ogni host viene salvato:

- $\frac{1}{n}\sum_{i=1}^{n} C_i^h$
- $\frac{1}{n}\sum_{i=1}^{n} Pw_i^h$ e analogamente per Ps

Quindi se il numero di host presenti nel cluster è k nel "summary" si troveranno le seguenti grandezze:

- $\frac{1}{n}\sum_{h=1}^{k}\sum_{i=1}^{n} C_i^h$
- $\frac{1}{n}\sum_{h=1}^{k}\sum_{i=1}^{n} Pw_i^h$ e analogamente per Ps

6.3 Dati calcolati: dati primari

Per ciascun host è possibile calcolare il prodotto fra il numero dei core e la percentuale di utilizzo per ciascuna tipologia di impiego, ovvero si è in grado di conoscere l'utilizzo globale del singolo host:

$$C_i^h \cdot Pw_i^h \tag{3}$$

analogamente per Ps. Utilizzando i dati del "summary" si riesce a calcolare per ciascuna farm:

$$\sum_{h=1}^{k} C_i^h \cdot \sum_{h=1}^{k} Pw_i^h \tag{4}$$

mentre si vorrebbe calcolare:

$$\sum_{h=1}^{k} C_i^h \cdot Pw_i^h \tag{5}$$

dato in grado di fornirci l'utilizzo globale della farm.

6.4 Dati calcolati: dati consolidati

Visti i paragrafi precedenti per i dati consolidati quello che si riesce a calcolare per ciascun host è:

$$\frac{1}{n}\sum_{i=1}^{n} C_i^h \cdot \frac{1}{n}\sum_{i=1}^{n} Pw_i^h = \frac{1}{n^2}\sum_{i=1}^{n} C_i^h \cdot \sum_{i=1}^{n} Pw_i^h \tag{6}$$

analogamente per Ps. Utilizzando i dati del "summary" si riesce quindi a calcolare:

$$\sum_{h=1}^{k} \frac{1}{n} \sum_{i=1}^{n} C_i^h \cdot \sum_{h=1}^{k} \frac{1}{n} \sum_{i=1}^{n} Pw_i^h = \frac{1}{n^2} \sum_{h=1}^{k} \sum_{i=1}^{n} C_i^h \cdot \sum_{h=1}^{k} \sum_{i=1}^{n} Pw_i^h \tag{7}$$

mentre si vorrebbe calcolare:

$$\frac{1}{n} \sum_{h=1}^{k} \sum_{i=1}^{n} C_i^h \cdot Pw_i^h \tag{8}$$

6.5 Discussione dati calcolati

A questo punto si pone il problema di capire l'approssimazione e gli errori fra quello che si desidera valutare e quello che si riesce a calcolare con i dati disponibili. Il dato di partenza che si può calcolare (3) è esattamente quello ciò che si voleva, resta però da capire quali differenze ci siano fra le relazioni 4 e 5, così come fra le relazioni 7 e 8.

Supponendo che il numero di core sia costante, sia nel tempo sia fra gli host del cluster, ossia $C_i^h = \alpha$, si vede che le relazioni sono praticamente uguali; infatti, la 5 diviene:

$$\alpha \cdot \sum_{h=1}^{k} Pw_i^h \tag{9}$$

mentre la 4 diviene:

$$k\alpha \cdot \sum_{h=1}^{k} Pw_i^h \tag{10}$$

Le equazioni (9) e (10) sono uguali a meno di un fattore k (numero di host del cluster) che però è disponibile nei dati misurati e permette quindi di correggere il calcolo.

Sempre nell'ipotesi di costanza del numero di core la 8 diviene:

$$\frac{\alpha}{n} \cdot \sum_{h=1}^{k} \sum_{i=1}^{n} Pw_i^h \tag{11}$$

mentre la 7 diviene:

$$\frac{k\alpha}{n} \cdot \sum_{h=1}^{k} \sum_{i=1}^{n} Pw_i^h \tag{12}$$

Di nuovo le equazioni (11) e (12) sono uguali a meno di un fattore k che si può correggere utilizzando i dati disponibili.

In conclusione si riesce a calcolare esattamente ciò che ci interessa approssimando un numero di core costante, resta quindi da discutere quanto questa approssimazione sia accettabile. I motivi per cui il numero di core può cambiare nel tempo sono due:

1. uno dei core della CPU si spegne per qualche motivo;

2. uno dei nodi del cluster si spegne.

Si può tranquillamente escludere il primo caso poichè l'esperienza di questi anni ha mostrato che un problema ad uno dei core provoca il crash del nodo riconducendoci di fatto al secondo caso; inoltre i kernel attualmente in uso non permettono di spegnere il singolo core a "run time" per ridurre i consumi.

Resta quindi il solo caso 2 ma, poichè Ganglia considera solo i dati provenienti dagli host accesi, i contributi ai dati presenti nei DB e le relative medie tengono già conto dell'eventuale down di alcuni host.

Un diverso discorso va fatto per quanto riguarda la costanza del numero di core fra i vari host, condizione che in termini più formali può essere espressa assumendo che il fattore α non dipenda dall'indice h. Questa ipotesi è verificata nella maggioranza delle farm di nostro interesse.

Alla luce di quanto discusso si può concludere che il calcolo eseguito con i dati disponibili è una buona approssimazione di ciò a cui siamo interessati, anche se in alcune farm l'ipotesi di costanza assoluta del numero di core non è strettamente verificata.

È un limite di questa analisi il fatto che il parametro di reale interesse, ossia la quantità di core che si trovino in un dato stato di utilizzo, non sia disponibile come dato primario ma debba essere ricavato dai dati disponibili. In questa operazione si evidenzia la perdita di informazioni dovuta, da un lato, alla procedura di riassunto dei dati nel passaggio dal singolo nodo al cluster e, dall'altro, alla procedura di consolidamento nel tempo dei dati stessi attraverso le operazioni di media.

Per poter superare questi limiti vi sono varie possibilità:

- per quanto riguarda il consolidamento dei dati e la relativa perdita di informazioni dovuta alle procedure di calcolo della media, è possibile ampliare la dimensione dei database RRD in modo da conservare per un intero anno sia i dati primari (quelli misurati ogni 15 secondi) sia i dati consolidati (quelli risultanti dalle medie). In questo modo si può scegliere la granularità necessaria per eseguire i calcoli. Chiaramente per poter fare questo si deve modificare il codice sorgente di Ganglia e quindi crearne una versione customizzata.

- per quanto riguarda la possibilità di avere disponibile direttamente la quantità di core in un certo stato di utilizzo si può sfruttare la capacità di Ganglia di definire delle

metriche personalizzate da misurare e mettere nei database RRD. Sfruttando questa capacità prevista dal software si potrebbe configurare il servizio di ciascun nodo di una farm affinchè sia misurato direttamente il parametro $C_i^h Pw_i^h$ e analogamente per le altre. In questo caso, fortunatamente, non è necessario modificare il codice sorgente, semplicemente si procede ad una configurazione opportuna.

Chiaramente si potrebbero implementare entrambe le migliorie se si valutasse questo strumento utile per le necessità di survey delle risorse di calcolo scientifico.

Attualmente è stata modificata la parte di rappresentazione per il web dei dati di Ganglia in modo da graficare il numero totale di core C_i, la parte working $C_i Pw_i$ e quella sleeping $C_i Ps_i$ di ciascuna farm. Nella figura 4 è riportato un esempio di questi grafici per vari intervalli temporali.

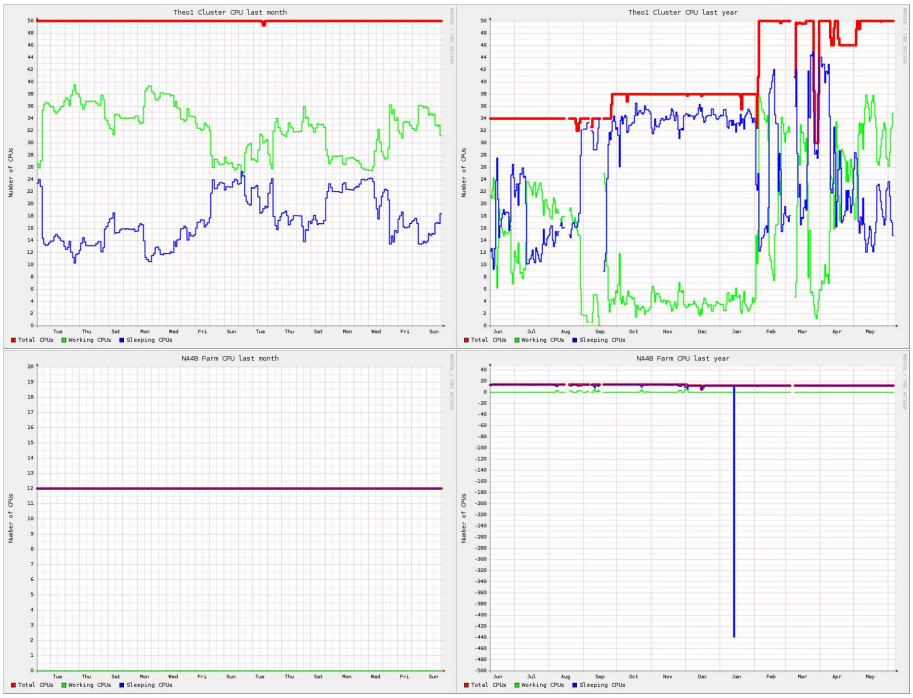


Figura 4: Esempio di grafico modificato con numero totale di core, working e sleeping

6.6 Valutazione integrali

Visto quanto riportato in precedenza, si è proceduto all'estrazione dai DB dei dati riguardanti il numero di core (in funzione del tempo e relativamente all'ultimo anno) per ciascuna delle farm non Grid al fine di valutare:

- l'integrale del numero di core disponibili nella farm nell'intervallo di tempo considerato. Da questo integrale verrà calcolato il consumo lordo totale della farm (passando dai dati di core a quelli di CPU);
- la relazione 7 e quindi farne l'integrale per avere una stima dell'efficienza di utilizzo della farm valutando quanto del consumo lordo totale di cui sopra sia stato utilizzato per fare "produzione" e quanto sia stato utilizzato semplicemente per tenere accesi i sistemi.

In questa operazione sono state eseguite le correzioni ai dati di seguito descritte:

1. eventuali intervalli di tempo in cui mancavano dati per errori di lettura sono stati riempiti facendo una interpolazione lineare fra gli estremi disponibili;
2. se la mancanza dei dati coincideva con la parte iniziale o terminale dell'intervallo di nostro interesse si è estrapolato per costanza rispetto al primo o all'ultimo dato disponibile;
3. si è verificato ed eventualmente imposto che la somma dei core working e sleeping in ogni istante fosse uguale al numero totale di core disponibili.

In figura 5 sono riportati i dati di due farm dopo le correzioni precedentemente descritte (confrontare questi dati con quelli di figura 4 riportanti i soliti dati prima delle correzioni).

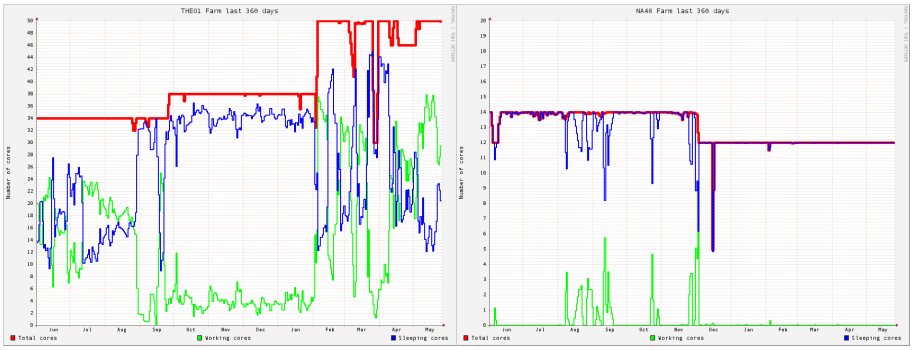


Figura 5: Grafici dell'uso dei core dopo le correzioni

Fatte queste correzioni ai dati si è proceduto al calcolo dei carichi working e sleeping secondo la relazione 7 opportunamente corretta per il numero di host da cui poi si sono ottenuti grafici come quelli riportati in figura 5. Su questa serie di dati si è quindi proceduto alla valutazione numerica degli integrali:

$$\int C(t)dt \backsimeq \sum_{j=1}^{m} C_j \cdot \Delta t$$

$$\int C(t) \cdot Pw(t)dt \backsimeq \sum_{j=1}^{m} C_j \cdot Pw_j \cdot \Delta t$$

$$\int C(t) \cdot Ps(t)dt \backsimeq \sum_{j=1}^{m} C_j \cdot Ps_j \cdot \Delta t$$

I valori C_j, Pw_j e Ps_j sono i dati disponibili nel RRD per il consolidamento che meglio si accordano con l'intervallo temporale oggetto della nostra analisi, ossia un anno. Quindi questi valori sono frutto delle operazioni di misura e media descritte nei paragrafi precedenti per cui soggetti alle approssimazioni sopra discusse.

6.7 Valutazione integrali: un affinamento

Riprendendo quanto discusso nei paragrafi 6.4, 6.5 e 6.6 è possibile valutare un modo alternativo per il calcolo degli integrali di nostro interesse.

L'obiettivo è calcolare:

$$\frac{1}{n} \sum_{h=1}^{k} \sum_{i=1}^{n} C_i^h Pw_i^h$$

mentre con i dati a disposizione e seguendo lo schema con il quale vengono consolidati ("summary") si calcola:

$$\sum_{h=1}^{k} \frac{1}{n} \sum_{i=1}^{n} C_i^h \sum_{h=1}^{k} \frac{1}{n} \sum_{i=1}^{n} Pw_i^h$$

Poichè si possono effettuare i prodotti $C_i^h Pw_i^h$ non solo a livello della farm totale ma anche a livello dei singoli host calcolando le percentuali working e sleeping sulla somma dei core dei singoli host, si ha la possibilità di integrare sommando sul tempo ed infine sugli host dell'intera farm ottenendo così la formula desiderata. È necessario però considerare che C_i^h e Pw_i^h sono medie ottenute attraverso la procedura di consolidamento e che sono valori non per singolo core ma per host. Effettuare il calcolo in questo modo è sicuramente più preciso, ma comporta delle complicazioni nei passaggi di correzione dei dati mancanti o non corretti (paragrafo 6.6).

Nel paragrafo precedente infatti sono state enunciate le correzioni che vengono eseguite sui dati per compensare eventuali buchi dovuti a problemi di lettura. Rivediamoli alla luce della possibilità di eseguire i calcoli su dati "summary" o per singolo host. Il caso di dati non corretti (i cosiddetti "picchi") non presenta problemi. I dati possono mancare in tre situazioni diverse:

1. i dati mancano all'inizio dell'intervallo di integrazione;

2. i dati mancano nel mezzo dell'intervallo;

3. i dati mancano alla fine dell'intervallo.

Utilizzando i dati "summary" si prensentano, quindi, tre nuovi scenari corrispondenti:

1. poichè la farm esisteva nel periodo di cui mancano i dati (altrimenti avremmo integrato partendo da un istante successivo) si può concludere che la farm fosse spenta o non raggiungibile. Il dato mancante si può correggere con 0. La correzione fa perdere elementi di accounting perchè il sistema di monitoring non ha funzionato per un dato periodo;

2. nel secondo caso si è autorizzati ad affermare che il sistema di monitoring non abbia preso dati ma la farm fosse in funzione e quindi i dati disponibili vengono interpolati linearmente. Se questo tipo di intervallo coincidesse con periodi previsti di down della farm si sarebbe semplicemente diviso in due parti l'intervallo di integrazione eliminado quindi il buco;

3. nel terzo caso, essendo noi a fare l'integrale, si suppone che si sappia se e quando è stata dismessa o spenta una farm dunque l'intervallo di integrazione non comprenderà un periodo di cui non si possano avere dati se non per uno degli altri motivi. Quindi si deve concludere che la farm fosse accesa ed il sistema di monitoring non abbia raccolto dati. Siamo quindi autorizzati ad approssimare estrapolando per costanza con l'ultimo dato disponibile.

Nel caso dei singoli host lo scenario si complica. I "buchi" di dati non riguardano più la totalità di una farm ma singoli host. Sappiamo benissimo che una farm non è costante nel tempo nel numero di host: i nodi, infatti, vengono riavviati in seguito a problemi oppure se ne aggiungono di nuovi. Questo fa sì che non sia ovvio associare la presenza di un "buco" ad un errore di lettura del sistema (se un host è spento non manda letture) e quindi applicare un qualsiasi algoritmo di correzione. In particolare per i punti 1 e 3 può davvero essere accaduto che l'host sia stato aggiunto alla farm solo a partire da una certa data, oppure che sia stato rimosso dalla farm da un certo giorno o sia semplicemente fermo, in ogni caso questa informazione non è disponbile nei dati di Ganglia e quindi non è possibile fare una correzione ragionevole di questo tipo di buchi. Solo i buchi di tipo 2 sono ancora facilmente correggibili con una approssimazione lineare come nel caso dei dati "summary" (ma imponendo così una approssimazione forte: si esclude il momentaneo down di un singolo host). Si potrebbe anche decidere di correggere tutti i buchi presenti sui singoli host utilizzando in questi intervalli temporali le informazioni disponibili nel "summary" distribuite sui singoli host. In questo modo però si creerebbe

una correlazione fra i due metodi, perdendo in principio i vantaggi rispetto al conteggio effettuato sui dati “summary”.

Carenze inevitabili nella acquisizione dei dati possono inficiare o addirittura peggiorare la maggiore precisione del metodo di calcolo basato sui singoli host.

Nel caso di farm in cui il numero di host è piuttosto costante nel tempo o in cui al limite aumenta si possono utilizzare le stesse regole di estrapolazione sia per il conto fatto con il “summary” sia per quello fatto con i singoli host, senza utilizzare le informazioni provenienti da uno per correggere la mancanza di informazioni dell’altro. In questo modo si calcolano gli stessi integrali seguendo i due metodi e quindi si valuta la bontà del metodo basato sui “summary”, che presenta minori incertezze in presenza di dati lacunosi.

Date queste premesse è stato deciso di procedere nel modo seguente:

1. selezionare un campione ristretto di farm che avessero mantenuto la loro composizione in termini di numero e qualità degli host costanti nel tempo, o almeno per un lasso di tempo di qualche mese;

2. individuare un lasso temporale in cui le farm scelte non presentassero problemi nei dati particolarmente gravi e che quindi potessero essere corretti ragionevolmente bene con approssimazioni lineari;

3. eseguire il calcolo dell’integrale per le farm del punto 1 sull’intervallo di tempo del punto 2 sia utilizzando i dati dei singoli host sia quelli “summary” secondo le metodologie fin qui discusse;

4. confrontare i risultati dei due integrali calcolati al punto 3. Questo ha fornito un’indicazione sull’errore introdotto dall’utilizzo dei dati “summary”.

Farm	$\int C$	$\int CPw$	$\int CPs$	$\int CPw / \int C$	$\int CPs / \int C$
fluent	142128	76132.421	66073.338	0.536	0.465
theo1	53040	26717.449	26322.107	0.504	0.496
gruppo5	14112	1569.918	12542.046	0.111	0.889

Tabella 1: Integrale core-ora calcolato con i dati “summary”

Come si può vedere dal confronto dei numeri riportati nelle tabelle 1 e 2 i due metodi danno risultati paragonabili in termini assoluti e praticamente uguali se si valuta il rapporto fra i core working/sleeping ed il totale dei core. Questo permette di concludere che il calcolo basato sui dati “summary”, corretto dividendo per il numero totale di host presenti in una farm, è una buona approssimazione dei dati a cui siamo interessati.

Farm	$\int C$	$\int CPw$	$\int CPs$	$\int CPw / \int C$	$\int CPs / \int C$
fluent	144000	78378.188	65622.016	0.544	0.456
theo1	53040	26760.830	26278.619	0.504	0.495
gruppo5	14112	1569.918	12542.032	0.111	0.889

Tabella 2: Integrale core-ora calcolato con i dati dei singoli host

6.8 I Dati e la Metodologia di Calcolo

È opportuna una riflessione alla fine di questi due lunghi capitoli. Perchè tutto questo lavoro per calcolare le percentuali di efficienza delle farm, un paradigma di calcolo già oggi non maggioritario ed in via di diminuzione rispetto a Grid? Perchè il nostro obiettivo è la determinazione di una metodologia che sia la più generalmente applicabile possibile e per far ciò occorre sapere quali dati sono necessari (e sufficienti).

Confrontiamo i dati disponibili nel caso del paradigma Grid rispetto alle farm.

- Grid (dati LSFMON):
  - numero di core, numero di job running, numero di job queued *disponibili solo come media del periodo*;
  - numero di job eseguiti, walltime, CPUtime $\forall$VO *solo come somma sul periodo*.
- Farm (dati GANGLIA):
  - numero di core, percentuale di utilizzo per i vari modi (system, idle, wait, user, nice) $\forall$farm e $\forall$host *come medie del periodo e vari sottoperiodi* (schema RRD).

Il lavoro per lo sviluppo ed il test della tecnica usata per le farm è stato effettuato con l'obiettivo di verificare se un insieme di dati analogo a quelli forniti da GANGLIA fosse sufficiente per una affidabile valutazione degli utilizzi e delle efficienze di utilizzo delle risorse.

A questo punto l'uso di tale tecnica può essere esteso al paradigma Grid, sapendo quali dati far estrarre da LSFMON (o analogo sistema di monitoring delle code batch). Più in generale, tale tecnica potrà essere utilizzata in tutti quei casi nei quali sia applicabile una metrica lineare con la quantità di utilizzo (ancora più in generale: una qualsiasi metrica lineare opportuna), con diversi capitoli di costi e soprattutto su diverse tipologie di apparati (per esempio reti con metrica il traffico, storage con metrica lo spazio o servizi con metrica il numero di utenti o il numero di erogazioni per servizi di tipo discreto).

7 Prime Metodologie Quantitative

7.1 Dati Globali

In Figura 6 è riportato un esempio di schermata del quadro sinottico di sala.

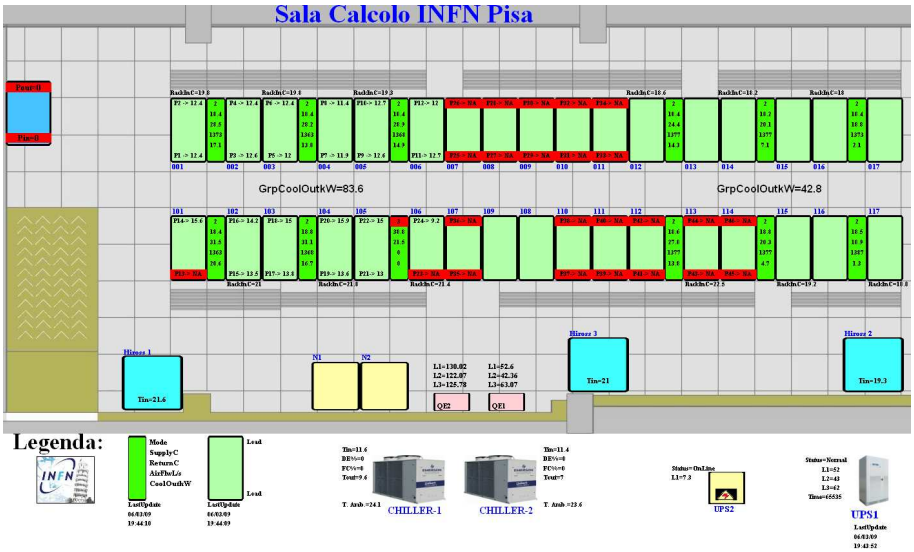


Figura 6: Sinottico sala calcolo

Come menzionato le risorse globali sono lo spazio, la corrente elettrica, il condizionamento e la rete.

Per ciascuna di queste risorse vengono date:

- la quantità disponibile
- la parte (percentuale) utilizzata
- la potenzialità di espansione (ove possibile)

7.2 Dati Caratteristici della Produzione

Le risorse considerate sono di due tipi: server e storage. Viene indicata la quantità disponibile e viene data una indicazione delle potenzialità di espansione. La parte utilizzata rientra nell'analisi dei processi di produzione. In particolare vengono indicati:

- server di produzione
- server di servizio: intesi come servizi asserviti alla produzione (Grid, SRM, disk server ecc.)
- spazio di storage utilizzato per la produzione

7.3 Dati Consuntivi della Produzione dell'Ultimo Anno

Per le misurazioni d'ora in poi ci avvarremo dei normali strumenti di monitoring e accounting disponibili: LSFMON (monitor del gestore di code Grid, sviluppato internamente ma disponibile per gli altri siti) consultabile in `http://farmsmon.pi.infn.it/lsfmon/`, Ganglia (monitor delle farm) consultabile in `http://farmsmon.pi.infn.it/`. Potremo utilizzare solo i dati storici che questi tool conservano per il loro funzionamento standard, che sono insufficienti per misurazioni accurate. Questa situazione potrà essere ovviata in futuro predisponendo la memorizzazione dei dati ritenuti utili per le prossime analisi, delle quali la presente costituisce il primo prototipo. Nella presente situazione è possibile analizzare solo i dati consuntivi dell'ultimo anno, e non periodi frazionari.

Si dovranno inoltre fare una serie di assunzioni ed approssimazioni che verranno di volta in volta indicate (per una analisi della metodologia e degli errori si veda il paragrafo 6).

La prima grandezza da stimare, che sarà utile per molte delle analisi successive, è il:

- *"consumo cpu lordo"*. Ovvero la quantità di corrente consumata da una cpu (non un singolo core) per l'attività di produzione. Con le seguenti note:

  - viene considerato un uguale consumo per tutte le cpu, indipendentemente dal numero di core;
  - le prestazioni di un core sono considerate costanti (1.5 KSI2k) e le capacità di produzione di una cpu lineari con il numero di core di cui è dotata (1, 2, 4);
  - non è considerato imputabile all'attività di Calcolo Scientifico il consumo di server, rete e storage dedicato ai Servizi generali di Sezione;
  - al Calcolo Scientifico è imputato non solo il consumo delle cpu che effettuano produzione, ma anche quello delle cpu dedicate ai servizi asserviti alla produzione. Analogamente viene imputata al Calcolo Scientifico la parte percentuale (in base alle cpu) di consumi di rete e storage. Questi consumi vengono conteggiati distribuendoli sulle cpu di produzione (in questo senso il consumo si intende "lordo").

  Si ha: *"consumo cpu lordo" = (corrente utilizzata in sala + copertura)/($\sum$ [cpu produzione + cpu servizi generali])*.

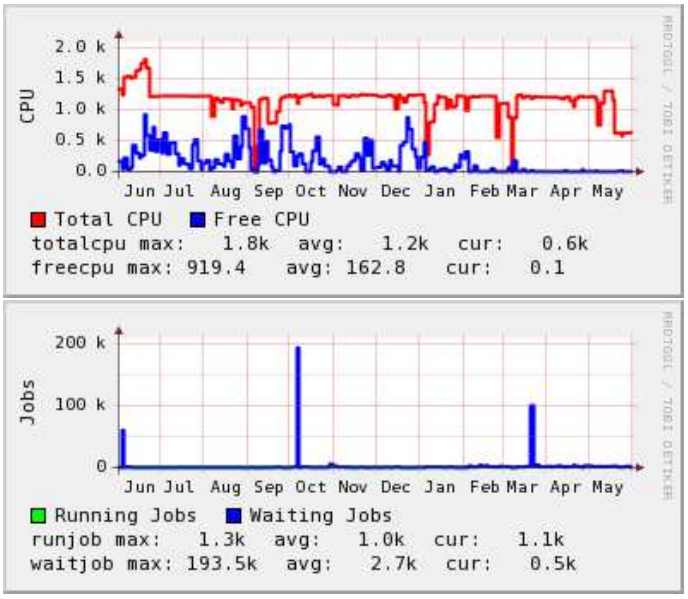


Figura 7: Grafici GStat relativi all'utilizzo del Grid Data Center di Pisa: CPU nella parte superiore, job slot nella parte inferiore

7.3.1 GRID

Un esempio dei grafici di GStat relativi al Grid Data Center di Pisa è riportato in figura 7.

Vengono riportate le seguenti grandezze:

- consuntivo di produzione nell'ultimo anno espresso in sec-core (o day-core, year-core):
  - vengono conteggiati solo i secondi durante i quali un core è stato allocato ad una VO (walltime);
  - il tempo di core viene conteggiato indipendentemente dall'esito del job;
  - non viene conteggiato il tempo durante il quale il core non aveva in esecuzione un job (jobslot vuoto, core free);
  - Viene conteggiato il tempo durante il quale il core è rimasto inattivo in attesa di I/O (viene così conteggiato il tempo utilizzato al lordo dell'efficienza di utilizzo o efficienza specifica).
- percentuale di utilizzo della capacità di produzione:
  - si dovrà calcolare: $100 \cdot (1 - \int coreliberi / \int core)$, ma non sono presenti questi dati storici

  - è nota la media dei core liberi e di quelli disponibili, quindi $100 \cdot (1 - \overline{coreliberi}/\overline{core})$

- corrente utilizzata:

  - si dovrebbe calcolare: $consumolordocpu \cdot \int cpuproduzioneGRID$ ma non si hanno gli storici
  - si può approssimare con: $consumolordocpu \cdot \overline{cpuproduzioneGRID}$. In realtà non si conosce nemmeno la $\overline{cpuproduzioneGRID}$, bensì la $\overline{coreGRID}$, dalla quale è possibile stimare la $\overline{cpuproduzioneGRID}$ estrapolando linearmente il rapporto core/CPU di produzione Grid della situazione attuale. Questa stima è accurata non essendo variato il rapporto tra macchine in produzione con CPU dual e quad core (stiamo assumendo che quando non è disponibile un certo numero di CPU la loro distribuzione tra dual e quad core si mantenga quella presente sull'insieme delle CPU).

- consumo di corrente per day-core:

  - questo è uno dei parametri fondamentali che caratterizzano il paradigma di produzione in esame (Grid): il costo di produzione (in elettricità, per questo lavoro) della singola unit, al lordo della efficienza di utilizzo degli impianti da parte dell'utente (efficenza generale);
  - si dovrebbe calcolare:

    $$correnteGRIDutilizzata/\int numerocorediproduzioneoccupati$$

  - si approssima con:

    $$correnteGRIDutilizzata/\overline{numerocorediproduzioneoccupati}$$

    a sua volta pari alla media di job running, valendo per INFN-Pisa job-slot=numero di core;
  - in questo modo si tiene conto del livello medio di utilizzo dell'insieme delle farm Grid (efficienza generale).

7.3.2 Farm di Esperimento

Preso in generale questo paradigma di produzione differisce in modo sostanziale dal precedente (Grid).

Volendo calcolare l'efficienza globale di una produzione (gruppo di utenti) in Grid si ha:

$$e(VO) = e(CPU) \cdot e(GRID)$$

ove

- $e(VO)$ è l'efficienza globale di una Virtual Organization
- $e(CPU)$ è l'efficienza media di utilizzo della cpu per la VO (*CPU-time/wall-time*: efficienza specifica)
- $e(GRID)$ è l'efficienza di utilizzo dell'impianto di produzione (*core utilizzati/core totali*: efficienza generale)

mentre per una farm di esperimento si ha:

$$e(ESP) = e(CPU) \cdot e(Farm)$$

ove, però:

- $e(CPU) = 1$, perchè viene conteggiato come tempo di CPU solo quello effettivamente in esecuzione (CPU user e nice) e non i diversi stati di idle e wait (CPU idle, I/O wait ecc.). Questo perchè si suppone che una farm non abbia frequenti richieste di I/O non locale (cosa invece normale in Grid)
- $e(Farm)$ è una caratteristica della farm (e quindi dell'esperimento) essendo $e(Farm) = \int cpuused / \int cpu$.

Nel caso di Grid:

- ciascuna VO paga secondo la propria percentuale di utilizzo l'inefficienza di utilizzo dell'impianto di produzione;
- ciascuna VO deve gestire l'efficienza con la quale utilizza le CPU. Questa efficienza è influenzata da molti fattori: tipo e qualità del software applicativo, architettura di LAN e WAN, architettura dello storage utilizzato, etc.

In una farm di esperimento:

- ad una produzione viene accollato il costo totale del funzionamento della farm, indipendentemente dal livello di utilizzo. È a questo punto ininfluente ogni eventuale differenza nella efficienza specifica dell'utilizzo delle CPU della farm stessa.

Vengono riportate grandezze logicamente analoghe a quelle relative al paradigma Grid, per ora considerate in generale sull'insieme delle farm, nel prossimo capitolo prese disaggregate per singole farm:

- consuntivo di produzione nell'ultimo anno: espresso in sec-core (o day-core, year-core):
  - viene conteggiata solo la percentuale di core considerata "used" cioè impegnata in "user" o "nice", considerati gli stati "in produzione" per le farm di esperimento;
  - non si ha alcuna informazione su come sia stato impegnato il tempo di CPU (di core): batch o interattivo, nè da chi;
  - non viene conteggiata come produzione la percentuale di core spesa in "idle", "system" o "wait" di vario tipo.
- percentuale di utilizzo della capacità di produzione:
  - viene calcolato: $100 \cdot \int coreinproduzione / \int core$. Nel caso delle farm di esperimento, a differenza di Grid, abbiamo i dati di Ganglia che possono essere disaggregati (anche se, ovviamente, non infittiti) e quindi si possono calcolare gli integrali necessari invece delle sole medie, come descritto nel capitolo 6.
- corrente utilizzata:
  - si può stimare il *(consumo lordo farm)* come *(consumo CPU lordo)*$*\int$*(CPU delle farm)*, avendo i dati di Ganglia.
- consumo di corrente per day-core:
  - analogamente a quanto detto nel caso del paradigma Grid, è il costo di produzione (in elettricità) della singola unit, al lordo della efficienza di utilizzo degli impianti da parte degli utenti;
  - si calcola: *(consumo lordo farm)/*$\int$*(numero di core di produzione occupati)*. Anche in questo caso sappiamo calcolare l'integrale (con una approssimazione dovuta alla densità dei dati) avendo i dati storici da Ganglia.

7.4 Dati Disaggregati della Produzione dell'Ultimo Anno

In questa sezione vengono calcolate le stesse grandezze descritte nei paragrafi 7.3.1 e 7.3.2, con le stesse metodologie, ma in modo disaggregato per gruppo di utenti/VO/Esperimento. In particolare, per ciascun utente vengono dati:

- paradigma di produzione: Grid, farm di esperimento
- numero di core che compongono la farm al termine del periodo di osservazione (solo per la modalità farm)
- day-core utilizzati nel periodo
- percentuale di efficienza:
  - *(efficienza generale)*(efficienza di CPU)* nel caso Grid
  - *(day-core utilizzati)/(day-core allocati)* per le farm di esperimento
- costo (in consumo elettrico espresso in Wh) di un day-core tenuto conto della percentuale di efficienza
- eventuali rilevanti questioni aperte con l'utente

Vengono analizzati tutti gli utenti/gruppi/esperimenti che hanno una farm e tutte le VO "non sporadiche", cioè che nel periodo hanno calcolato almeno 1 day-core (> 86400 secondi).

Congiuntamente ed a completamento della survey verranno organizzati brevi incontri con tutti gli utenti coinvolti per approfondimenti sui casi specifici. Tutti i gruppi/esperimenti che hanno una farm hanno un referente locale per il Calcolo Scientifico, cosa non necessariamente vera per le VO. Per queste ultime che abbiano utilizzato il Grid Data Center esclusivamente in remoto e non abbiano un referente locale verrà interpellato via e-mail il referente nazionale. Tutte queste informazioni saranno raccolte in una tabella come quella riportata di seguito. I risultati della survey per la Sezione INFN di Pisa sono riportati in appendice.

Utente	Mod	#core	d/core	%eff	Wh/(day-core)	Major Issues

8 Ringraziamenti

I più sentiti ringraziamenti vanno al Prof. Tarcisio Del Prete che ha rivisto tutto l'articolo ed al quale ci siamo rivolti ponendo specifiche questioni relative al capitolo 6, ricevendone suggerimenti, consigli e, a sua volta, intriganti contro-domande.

A Risultati survey

Calcolo Scientifico alla Sezione di Pisa dell'INFN

Enrico Mazzoni, Alberto Ciampa

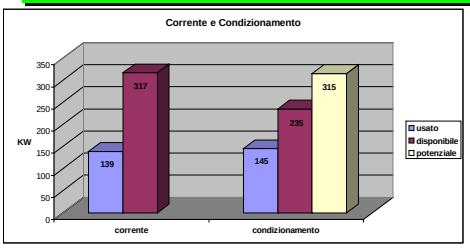
Dati salienti al giugno 2009

Sala

- **Spazio RACK**: **34 rack**, di cui 33 da 42U ed uno da 48U

- **Corrente elettrica installata**: **1380A+450A(cop.)** = **317.4KW + 103.5KW(cop.)** = 200A ⅄ +100A ⅄ +50A ⅄ (cond.)+110A ⅄ (UPS)+150A ⅄ (copert.)
- **Corrente impiegata (media)**: **602A + 228A (cop.)** = **138.5KW + 52.4KW (cop.)** = **43.6%, 50.6% (cop.)**

- **Capacità di dissipazione dei condizionamenti**: **235 KW** = 3*25 + 2*80 KW
- **Capacità di dissipazione impiegata (media)**: **145 KW** = **61.7%**
- **Capacità di dissipazione potenziale massima**: **315 KW** = 3*25 + 3*80

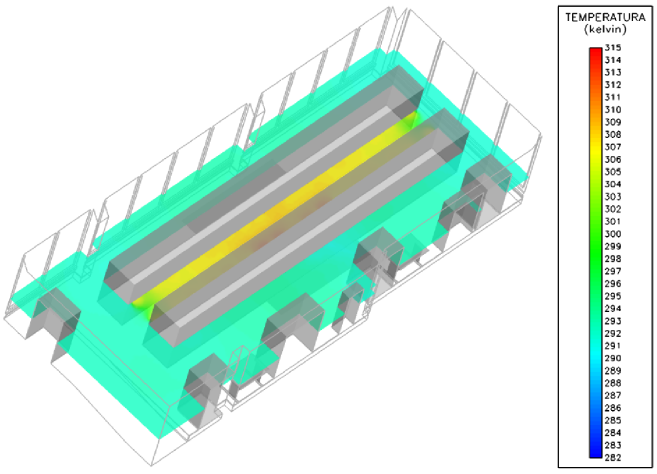


- **RETE:**
  - **LAN: 900 * 1GbE + 40 * 10 GbE = mix 14*90 1GbE e 14*10 10GbE (max 14 board)**
  - **WAN: 1 Gb/s dedicato GRID + 400 Mb/s Sezione (MAN)**

Calcolo Scientifico

- **Core di calcolo**: **1567** = 1.332 (GRID) + 235 (Farm Esperimento)
- **Potenza di calcolo**: **2.35MSI2k**
- **Core di infrastruttura per Calcolo Scientifico**: **98** (GRID+dCache+GPFS)
- **Capacità di storage**: **300TB** lordi

- **Potenzialità massima: 7.000 core + 1 PB storage.** Con attuale tecnologia (quad core, 1 TB disk), verificato con simulazione termofluidodinamica



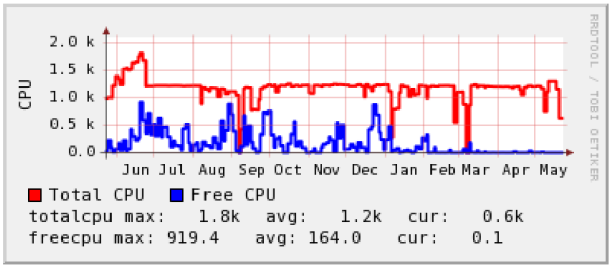
Profilo di utilizzo del periodo 1/6/2008 – 31/5/2009

Con ***Consumo CPU [Lordo]*** si intende il consumo medio di una CPU in produzione compresi i consumi di tutti server di servizio asserviti, la rete LAN, lo storage presente ed il relativo consumo per condizionamento.

Consumo cpu [lordo] = **239W (1.04A)** = (138.5+52.4)KW/(622[GRID]+145[farm]+33[SCeR]) = 190.9/800 KW

GRID

- **GRID day-core**: **350.971** pari a **962 anni-core**.

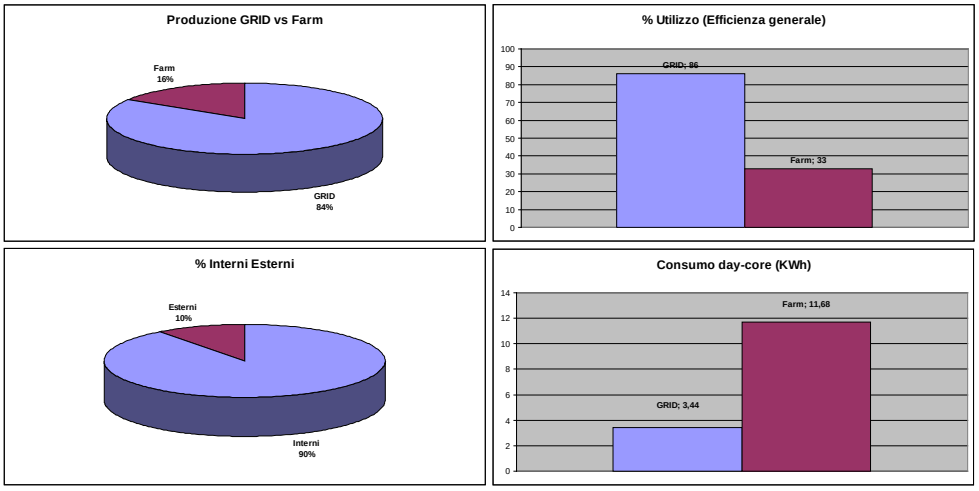


- **% di utilizzo**: **86%** = %(1200-164)/1200
- **Corrente utilizzata (lordo)**: **148.7 KW** = 239*622 W
- **Consumo di corrente per day-core (lordo)**: **3.44 KWh** = (148.7/1036)*24 KWh

Farm di esperimento

- **Farm di esperimento day-core**: **66.759**, pari a **185 anni core**.
- **% di utilizzo**: **33%**
- **Corrente utilizzata (lordo)**: **29.2 KW** = 239*122 W
- **Consumo di corrente per day-core (lordo)**: **11.68 KWh** = (29.2/60)*24 KWh

GRID vs Farm



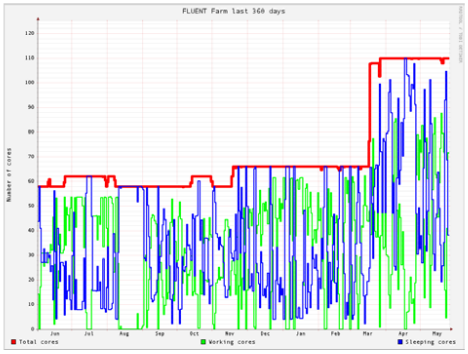
Utenti nel periodo 1/6/2008 – 31/5/2009

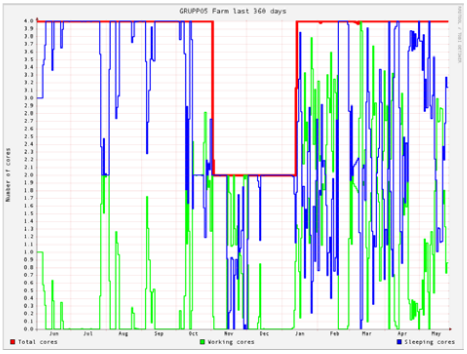
Vengono inclusi solo utenti “non sporadici”, cioè con più di 1 giorno/core (>86400 sec).

- GRID: Alice*, Argo*, ATLAS, BaBar, Biomed*, CDF, CMS, CMSit, Compchem*, ESR*, GLAST, LHCb*, OPS*, Pamela*, TheoDip, Theophys, Theoinfn, Virgo (* esterni)
- Farm di esperimento: Siena, Fluent, Theo1, CDF, ATLAS, MEG, Gruppo5, Mafalda, NA48

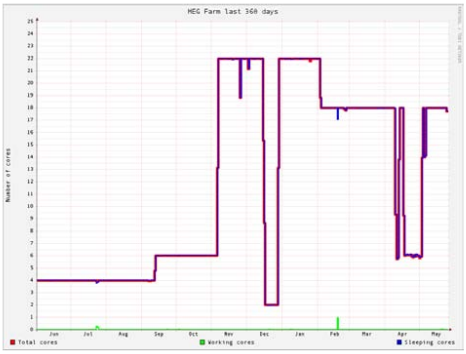
Utente	Mod.	# core	d-core	%,eff.	KWh/(day -core)	Major Issues
Alice*	GRID		7	35	9,8	
Argo*	GRID		1.821	82	4,2	
ATLAS	GRID		1.654	77	4,5	
BaBar	GRID		2.699	41	8,4	
Biomed*	GRID		2.927	56	6,1	
CDF	GRID		1.550	66	5,2	
CMS	GRID		75.834	41	8,4	Bassa eff.: LAN, Storage
CMSit	GRID		18	4	86,0	Bassa eff.
Compchem*	GRID		14.859	77	4,5	
ESR*	GRID		377	83	4,1	
GLAST	GRID		3.976	70	4,9	
LHCb*	GRID		579	74	4,6	
OPS*	GRID		29	36	9,6	
Pamela*	GRID		9.028	83	4,1	
TheoDip	GRID		38.858	77	4,5	
TheoPhys	GRID		60.949	81	4,2	
TheoINFN	GRID		49.957	82	4,2	
Virgo	GRID		47	27	12,7	Sviluppo sw
Siena	Farm	21	3.741	49	9,6	
Fluent	Farm	110	11.956	47	6,3	
Theo1	Farm	50	4.814	34	13,0	
CDF	Farm	8	39	1	202,0	Basso utilizzo
ATLAS	Farm	8	789	18	24,0	Basso utilizzo 2009
MEG	Farm	18	3	0	2.993,8	Basso utilizzo
Gruppo 5	Farm	4	322	25	23,1	
Mafalda	Farm	2	1	0	7.309,6	Basso utilizzo
NA48	Farm	12	92	2	291,1	Basso utilizzo 2009

Grafici Utilizzo Farm

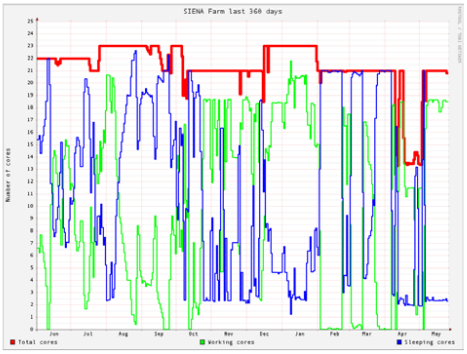

FLUENT Farm last 360 days
Number of cores
Total cores
Working cores
Sleeping cores

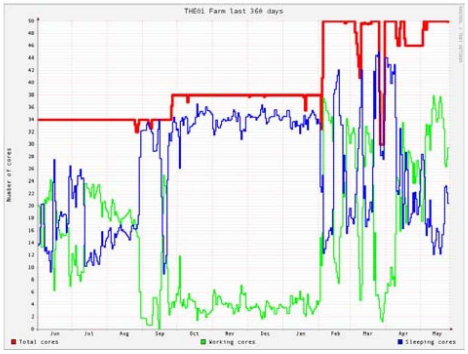

GRUPPO5 Farm last 360 days
Number of cores
Total cores
Working cores
Sleeping cores


MAFALDA Farm last 360 days
Number of cores
Total cores
Working cores
Sleeping cores


HEG Farm last 360 days
Number of cores
Total cores
Working cores
Sleeping cores


NA48 Farm last 360 days
Number of cores
Total cores
Working cores
Sleeping cores


SIENA Farm last 360 days
Number of cores
Total cores
Working cores
Sleeping cores


THE01 Farm last 360 days
Number of cores
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Jan
Feb
Mar
Apr
May
Total cores
Working cores
Sleeping cores