ISTITUTO NAZIONALE DI FISICA NUCLEARE

Sezione di Pisa

INFN/CCR-10/02
4 Agosto 2010

CCR-38/2010/P

PROGETTO CLUSTER GRID CSN4: LA PROPOSTA DI PISA

Alberto Ciampa[1], Ettore Vicari[2]

[1)]INFN - Sezione di Pisa, Largo Pontecorvo, 3, I-56127 Pisa, Italy
[2)]Dipartimento di Fisica Università di Pisa, Largo Pontecorvo, 3, I-56127 Pisa, Italy

Abstract

Nell'ottobre 2009 la Commissione Scientifica Nazionale 4 pubblicò, internamente all'INFN, una "Call" per la realizzazione di un cluster nazionale atto a rispondere alle esigenze di calcolo, anche parallelo, della Commissione stessa in ambito GRID.
Il presente lavoro riporta la proposta progettuale della Sezione di Pisa, presentata nel dicembre 2009, elaborata secondo i requisiti stabiliti dalla Commissione stessa con il supporto di un gruppo di esperti espressi dalla Commissione Calcolo e Reti. Tali requisiti sono riportati in allegato.
La proposta della Sezione di Pisa è risultata quella selezionata dalla CSN4, nel febbraio 2010.

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

1 RESPONSABILITÀ

La presente proposta, la sua eventuale realizzazione e la conduzione in produzione del sistema risultante ricadono sotto le seguenti responsabilità:

Responsabile Tecnico: Alberto Ciampa
Responsabile Scientifico: Ettore Vicari

2 UBICAZIONE: IL GRID DATA CENTER DI PISA

Il Cluster GRID CSN4 verrà installato nella sala che ospita il GRID Data Center di INFN Pisa ed integrato in esso, condividendo i servizi GRID, oltre alle infrastrutture. Attualmente il GRID Data Center di INFN Pisa ha in produzione circa 1700 core con 300 TB di spazio disco. La VO Theophys ha il 40% di Fair Share, pari a 680 core equivalenti.

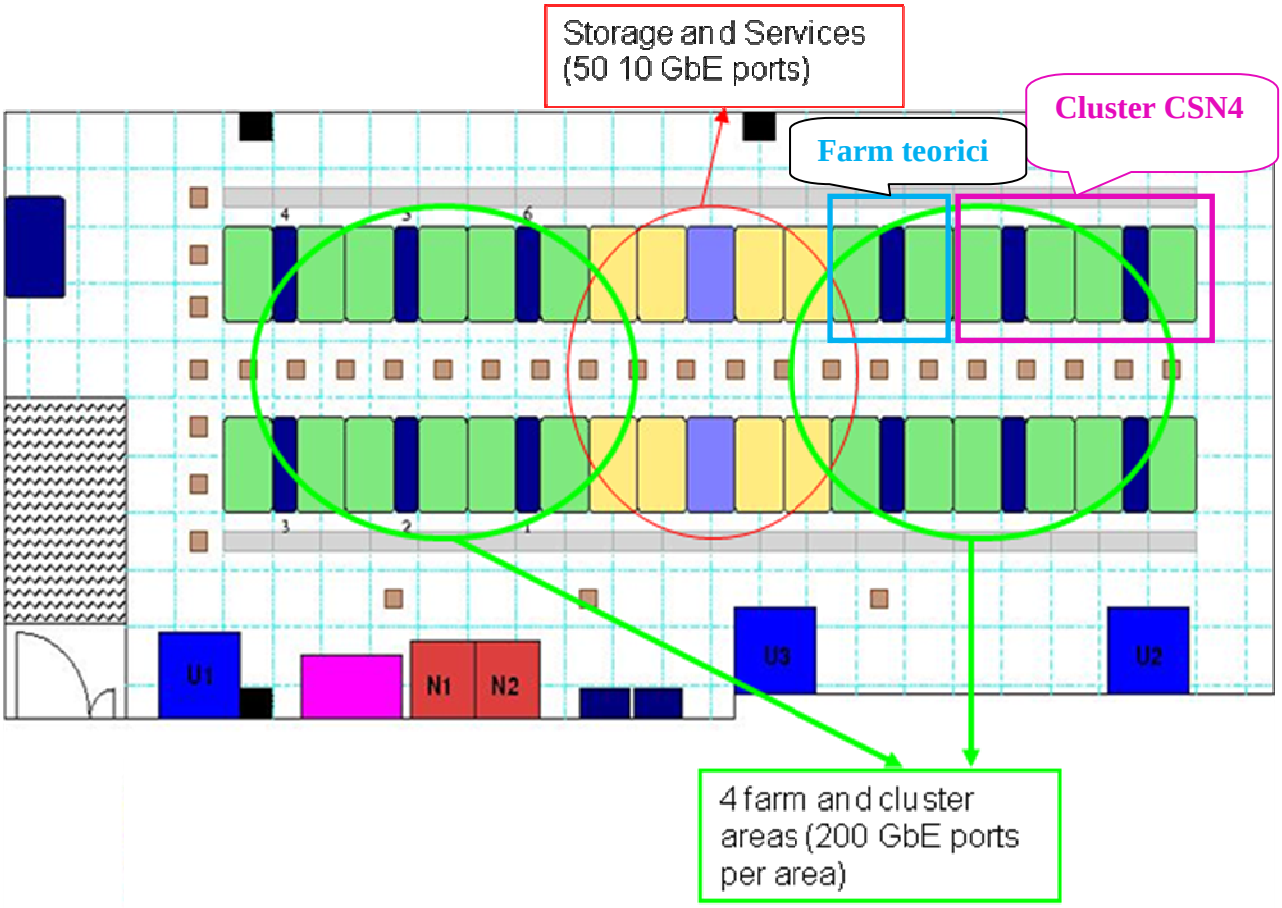

FIG. 1: La sala che ospita il GRID Data Center di Pisa.

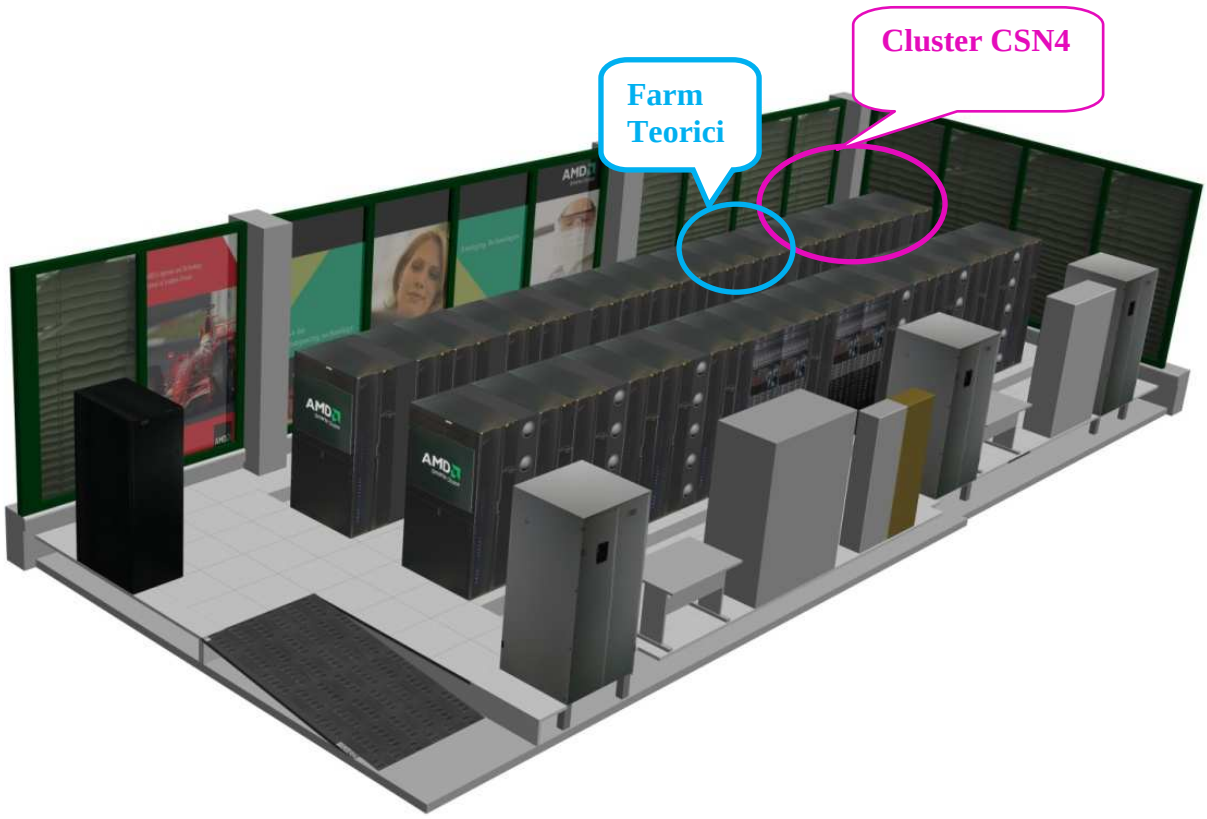

FIG. 2: Rendering tridimensionale della Sala Calcolo.

3 IMPIANTI ELETTRICO E DI CONDIZIONAMENTO

La sala è dotata di impianti di distribuzione elettrica e di condizionamento così caratterizzati:

TAB. 1: Dati caratteristici degli impianti di sala.

Potenza massima disponibile in sala	317 kVA
Potenza disponibile sotto UPS (potenza nominale UPS)	80 kVA
Carico attuale utilizzato in sala	160 kVA
Carico attuale utilizzato sotto UPS	38 kVA
Potenza nominale del gruppo elettrogeno	400 kVA (tutto il polo Fibonacci)
Capacità frigorifera disponibile in sala	235 kW
Capacità frigorifera attualmente impegnata in sala	165 kW
Stima della potenza necessaria per il nuovo cluster (in sala)	41 kVA

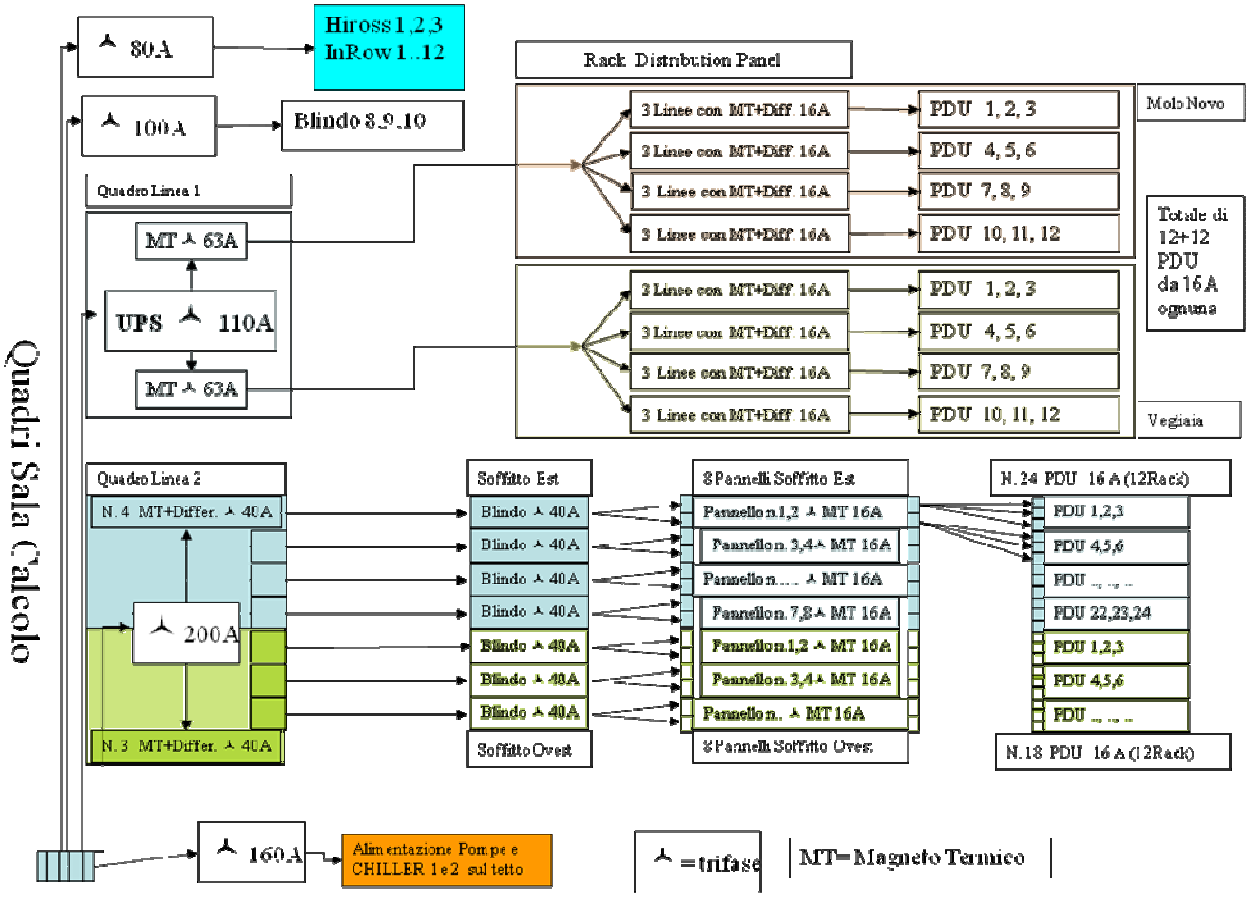

FIG. 3: Schema della distribuzione elettrica di sala.

Il sistema di condizionamento si compone di due sottosistemi che lavorano concorrentemente:

- Sottosistema InRow APC:
    - 12 APC InRow a circolazione di acqua refrigerata, con una capacità totale di raffreddamento potenziale massima di 240 kVA;
    - 2 chiller per la produzione di acqua refrigerata (posizionati in copertura ed alimentati con un circuito dedicato) con una capacità di raffreddamento di 160 kVA. L'impianto è già predisposto per l'aggiunta di un terzo chiller con capacità di raffreddamento di 80 kVA.
- Sottosistema Hiross:
    - Tre unità ad espansione diretta poste in sala, con capacità di raffreddamento totale di 75 kVA.

4 REALIZZAZIONE DEL CLUSTER GRID: INTERVENTI NECESSARI

4.1 Adeguamento degli Impianti

Gli impianti di distribuzione elettrica e di condizionamento non necessitano di alcun intervento per il loro potenziamento legato alla realizzazione del progetto.

E' previsto, e già parzialmente finanziato dalla Sezione, un intervento di parziale ristrutturazione e svecchiamento della parte interna di distribuzione elettrica (a valle dei quadri di sala) per permettere un aumento della percentuale di utilizzo della corrente disponibile. L'intervento includerà anche il completamento del sistema di monitoring ed accounting della corrente erogata da ciascuna delle linee di alimentazione asservite alla Sala Calcolo e impianti relativi. Costo 18.000 Euro IVA inclusa, coperto dalla Sezione per il 50%.

L'impianto elettrico che alimenta la sala di Pisa è dotato di un gruppo di continuità che serve, per scelta architetturale, solo una delle linee di alimentazione. A tale linea sono collegati i server centrali, i server essenziali per il funzionamento di GRID, lo storage e gli switch di rete di sala. I worker nodes non sono alimentati sotto continuità. Un secondo gruppo di continuità è in corso di installazione: è il gruppo da 1150 kVA dismesso dal CNAF che verrà trasportato a Pisa ed installato entro il primo quarto del 2010.

4.2 Acquisizione Server, Storage e Rete

I server barebone sono già presenti a Pisa: sono 128 server Acer Altos 2p 1U. Il costo di acquisizione dei server barebone è stato di 4.800 Euro IVA inclusa, coperto dalla Sezione e da rimborsare.

Sono inoltre già presenti i processori, 256 AMD Opteron 2356 (Quad core, 2.3 GHz) e i dischi, uno per server, da 80 GB a tecnologia SATA.

L'unica componente da acquisire è la memoria RAM. Il costo stimato è di 41.400 Euro (IVA compresa) per dotare l'intero sistema di 1 GB di RAM per core (512 moduli da 2 GB di RAM DDR2-667 Registered pari ad 1 TB complessivo). I server possono essere dotati fino a 32 (o 64 utilizzando moduli da 4GB) GB di RAM ciascuno (4 GB per core) scalando linearmente il costo su indicato.

L'assemblaggio dei server sarà a cura del Settore di Calcolo Scientifico della Sezione di Pisa ed eseguito utilizzando personale esterno, con un costo previsto di 5.000 Euro IVA inclusa (il costo include l'installazione del sistema operativo e la cablatura della rete).

Per il sistema di storage verrà utilizzato il sistema SAN DDN S2A 9900 già in produzione, espandendone la capacità secondo le richieste. La capacità massima di espansione del singolo sistema è di 600 TB (utilizzando dischi SATA da 1 TB); attualmente il sistema è dotato di 120 TB di disco raw.

Il costo previsto per aggiungere 10 TB raw è di 8.400 Euro IVA inclusa.

Per il collegamento Ethernet (1 link 1Gb per server) sono possibili due scenari:

- Scenario A (integrazione nell'architettura di rete esistente): connessione allo switch centrale GRID modello Force10 E1200i. Sono necessarie due schede da 90 porte 1 Gb ciascuna: Euro 50.000 IVA compresa e il materiale per la cablatura (trunk cables, patch panels e cavi RJ45) per un costo di 10.000 Euro IVA compresa.

- Scenario B (low cost): 4 switch da 48 porte a 1 Gb da utilizzare uno per rack, con uplink a 1 Gb, costo (incluso materiale per il cablaggio): 5.000 Euro IVA compresa.

I due scenari presentano importanti differenze:

- Prestazioni: lo scenario B connette i server a 1 Gb, ma con uplink di rack (32 sistemi) a 1 Gb, mentre lo scenario A connette ciascun server con un link dedicato a 1 Gb allo switch centrale;

- Affidabilità e gestione: l'integrazione del cluster nella infrastruttura generale di rete del GRID Data Center offre evidenti maggiori possibilità di gestione, mentre la soluzione "low cost" comporterebbe limitazioni nella flessibilità di configurazione. La soluzione integrata avrebbe una affidabilità molto maggiore, utilizzando una apparecchiatura ridondata e di classe superiore, oltre ad una più facile gestione che si tradurrebbe nella possibilità di offrire un migliore supporto agli utenti.

In entrambi gli scenari l'installazione e la cablatura dei server sarà a cura del Settore di Calcolo Scientifico della Sezione di Pisa.

Il cluster verrà realizzato mediante rete veloce InfiniBand DDR (20 Gb/sec), secondo quanto descritto nel successivo paragrafo dedicato alla rete veloce.

4.3 Posizionamento e Alimentazione

Il cluster verrà alloggiato in 4 rack (32 server per rack).

Per il posizionamento in sala e l'alimentazione del cluster sono necessari i seguenti materiali:

- N.4 rack APC 42U (p/n AR3100): Euro 5000 IVA inclusa.

- N. 4 PDU (Power Distribution Unit) APC da 16A trifase con 33 prese: Euro 3400 IVA inclusa.

4.4 LAN e WAN

Il cluster sarà integrato nella infrastruttura di rete Ethernet del GRID Data Center esistente. Si riporta di seguito le schema della infrastruttura di rete, riferendoci al solo scenario di integrazione nello switch centrale Force10 già presente (scenario A su descritto).

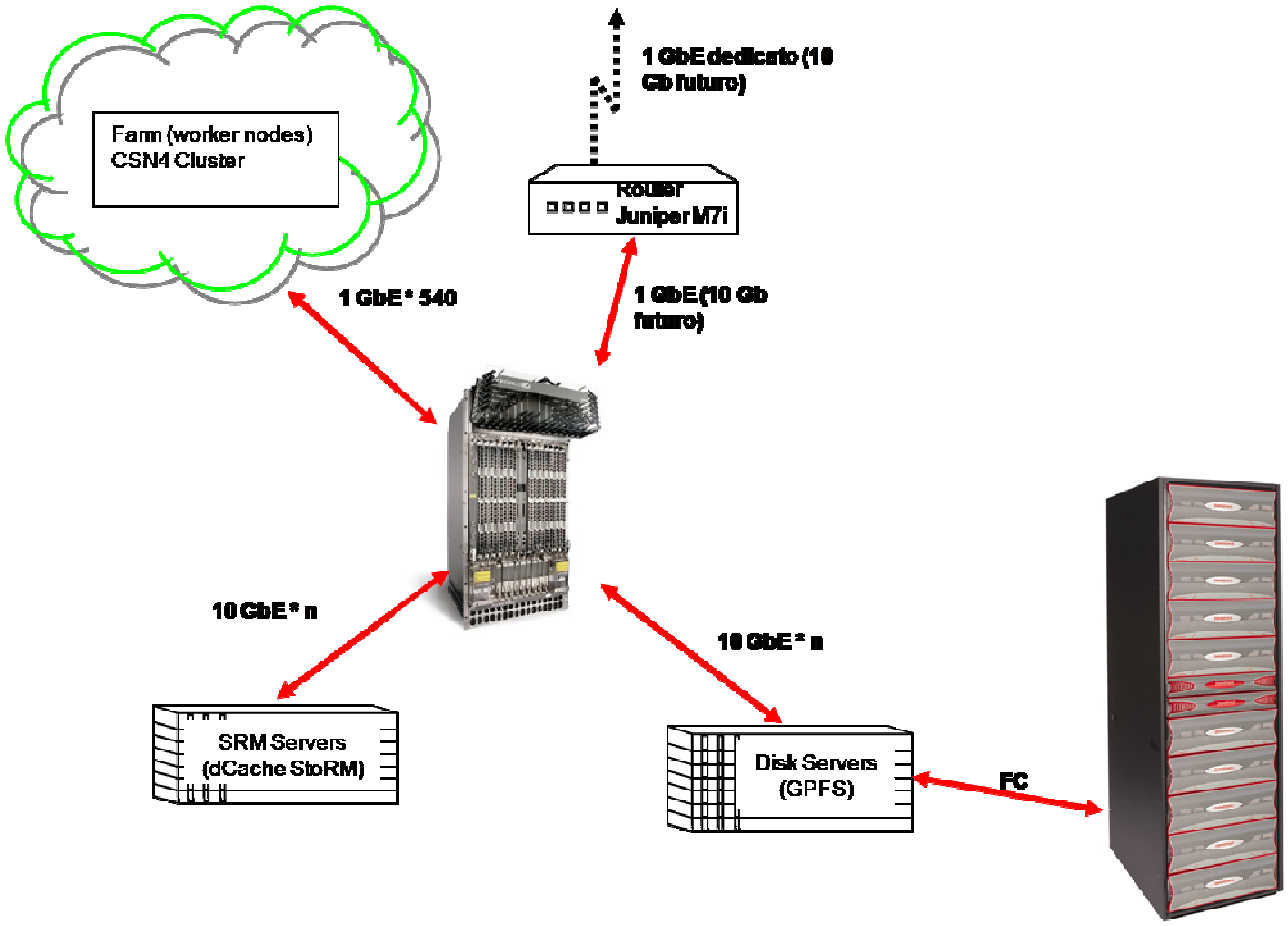

FIG. 4: Schema della infrastruttura di rete

L'accesso alla rete geografica è realizzato tramite un link 1 GbE dedicato a GRID, con possibilità di backup tramite il link GbE di Sezione.

4.5 Rete Veloce

La rete veloce sarà InfiniBand DDR, con infrastruttura realizzata mediante un solo switch a 144 porte al quale verranno collegati tutti i 128 server. Per la realizzazione della rete veloce verrà utilizzato il materiale sotto indicato, tutto già presente:

- Switch IB DDR Cisco SFS 7024 da 144 porte in rame
- 128 schede IB DDR Cisco SFS HCA320 A1
- Cavi in rame CX4

L'installazione e cablatura della rete veloce sarà a cura del Settore di Calcolo Scientifico della Sezione di Pisa.

4.6 Installazione, Configurazione e Accensione

L'installazione e configurazione sarà cura del Settore di Calcolo Scientifico della Sezione di Pisa, che ha in gestione l'infrastruttura di GRID-Pisa.

Il tempo necessario per la messa in produzione del cluster è di quattro mesi, a partire dalla emissione degli ordini per il materiale necessario.

4.7 Messa in produzione, Gestione e Monitoring

Il GRID Data Center di Pisa ha supportato la VO Theophys dalla sua entrata in produzione, attualmente è in produzione una farm da 1700 core utilizzata anche da Theophys secondo il meccanismo di fairshare, determinando la percentuale di fairshare in base al numero di core presenti finanziati dal gruppo IV (attualmente 40%). I job vengono sottomessi tramite il gestore di code LSF.

Per il GRID Data Center la Sezione di Pisa utilizza una licenza LSF site (con numero illimitato di core) di tipo demo, a titolo gratuito, per cui non vengono indicati costi previsti per l'integrazione del cluster in oggetto.

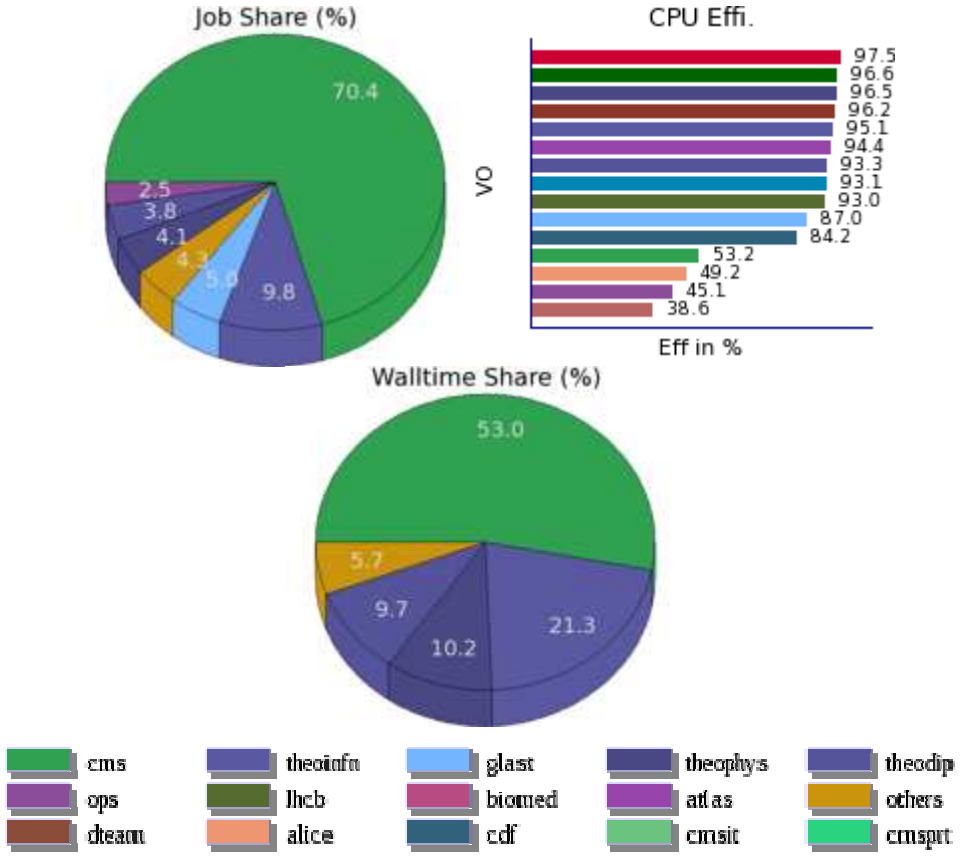

FIG. 5: Attività del GRID Data Center di Pisa dell'ultimo anno per VO.

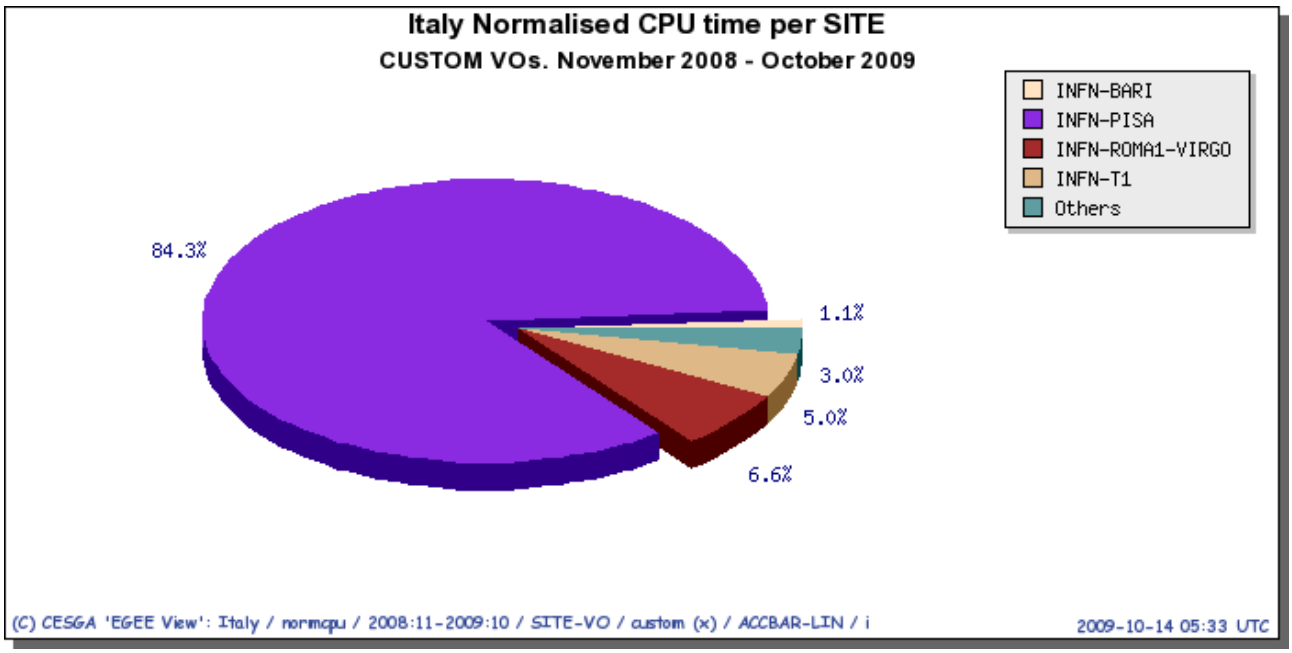

FIG. 6: Spaccato dell'attività GRID della VO Theophys in Italia nell'ultimo anno.

La VO Theophys è organizzata in gruppi: TheoINFN, TheoDip e Theolong. Il gruppo Theolong è associato ad una coda specifica (che utilizza una specifica partizione di GRID-Pisa contenente 80 core) che non ha limiti di CPU, per servire una specifica richiesta espressa dal gruppo.

La messa in produzione e la gestione del cluster prevedono la sua integrazione con la struttura GRID esistente. Sarà naturalmente possibile definire partizioni separate (utilizzando lo stesso metodo impiegato per Theolong) ed eventualmente gruppi e code separate per esigenze specifiche (per esempio utilizzo in cluster di tutto o parte/parti del sistema).

L'accesso allo storage avverrà tramite lo SRM StoRM (già in produzione a Pisa) che si avvale della infrastruttura centrale di storage in uso a GRID (SAN centrale mediante GPFS). Lo spazio disco verrà reso disponibile, mediante acquisto di disco da installare nella SAN, su richiesta.

Nello spazio disco dedicato al cluster CSN4 sarà allocata la home condivisa e lo spazio accessibile via SRM per l'esportazione ed importazione di file dati in ambito GRID.

L'integrazione del cluster in oggetto in una struttura GRID già esistente e nella quale la VO Theophys già svolge attività di calcolo da tempo rende trasparente l'ingresso del nuovo cluster per gli amministratori del GRID Data Center. Gli utenti della VO Theophys hanno, peraltro, già accesso alle strutture di supporto previste da INFN-GRID.

In aggiunta ai tool standard disponibili per il monitoring dei siti di INFN-GRID, a Pisa sono in funzione:

- Ganglia, per il monitoraggio dello stato delle farm e cluster da un punto di vista fisico (http://farmsmon.pi.infn.it/)

- Sinottico, per il monitoring dello stato di funzionamento degli impianti di sala, con dati specifici per i singoli condizionatori e singoli rack (http://farmsmon.pi.infn.it/sinottico/)

- LSFMON, per il monitor in tempo reale e storico dell'attività del gestore delle code (http://farmsmon.pi.infn.it/lsfmon/)

- JOBMON, per il monitor dell'attività GRID a livello del singolo job (link provvisorio https://gridse.sns.it/jobmon/pisa/, in futuro http://farsmon.pi.infn.it/jobmon)

Non sono previsti costi di software aggiuntivi.

4.7.1 Nota sull'utilizzo di configurazioni Cluster mediante MPI sotto GRID.

L'utilizzo di configurazioni Cluster mediante MPI sotto GRID è supportato da INFN GRID ma presenta la necessità di un periodo di test e sperimentazioni. Durante questo periodo, per facilitare queste attività minimizzando l'impatto sul sistema, il Settore di Calcolo Scientifico della Sezione metterà a disposizione gratuitamente dei gruppi di CSN4 che lo richiederanno un sistema Sun X4600 dotato di 8 processori quad-core (32 core totali) nella stessa configurazione dei server del cluster. Il sistema sarà reso accessibile sia via GRID (eventualmente su una coda separata) che mediante coda locale.

5 COSTI DI ESERCIZIO E MANUTENZIONE

Le spese di esercizio del cluster consistono nei costi di energia e di manutenzione e sono così determinati:

- Costi di energia. Il consumo del sistema è stimato in 1.4A per server, con un profilo di carico all'80% ed inclusivo del consumo per gli apparati di rete, per un totale di 179A=41KW. Il consumo per il raffreddamento è calcolato nel 35% del consumo dal sistema, cioè 14.4KW. Per un totale di 55.4KW (di cui solo 41 KW in sala).

- Costi di manutenzioni. Nello specifico:

    - Server: 0 Euro, i guasti saranno gestiti tramite spare part già acquisite (abbiamo parti equivalenti a 10 server completi);

    - Storage: l'intero sistema DDN S2A 9900 ha un costo previsto di manutenzione di 15.000 Euro IVA inclusa, tale spesa verrà suddivisa tra i gruppi di utenti in base alla percentuale di spazio sul totale. Per i 10 TB dedicati al cluster CSN4 si prevede una quota di 1.000 Euro/anno IVA inclusa.

    - Apparati di rete. Nel caso dello scenario A (integrazione nel Force 10) 3.500 Euro/anno IVA inclusa, per lo scenario B ("low cost") 0 Euro.

- Impianti di condizionamento e UPS. Si intende la percentuale rispetto al costo totale della manutenzione degli impianti calcolata in base al consumo del sistema sul totale: 5.000 Euro IVA inclusa.

6 PERSONALE RICERCATORE E TECNICO COINVOLTO IN LOCO E NON

Per l'utilizzo per cluster integrato nella struttura GRID in funzione a Pisa non è prevista la necessità di aggiunta di personale specifico. Il settore di Calcolo Scientifico della Sezione di Pisa vede coinvolti 3 tecnologi e 2 tecnici: Alberto Ciampa, Silvia Arezzini, Enrico Mazzoni (GRID e Rete), Dario Fabiani (GRID) e Federico Calzolari (Scuola Normale Superiore, dedicato a GRID e Cluster). Più in dettaglio:

- L'utilizzo in configurazione "farm", con sottomissione ed esecuzione di job su singolo core, comporta la gestione di 1024 core aggiuntivi rispetto alla configurazione attuale, e questa attività non necessita di aggiunta di personale sistemistico;

- L'utilizzo in configurazione "cluster" del sistema (o di sue parti), rimanendo sempre all'interno del paradigma GRID, non presenta una situazione che necessita di aggiunta di personale sistemistico. Diversa sarebbe la situazione nel caso in cui l'uso in cluster dovesse richiedere di uscire dal paradigma standard GRID. Il Settore di Calcolo Scientifico della Sezione di Pisa ha esperienza con la gestione di cluster ed è disponibile per fornire supporto sistemistico volto al training di una persona, a carico della comunità di utenti, per metterla in grado di erogare, a sua volta, supporto sistemistico agli utenti.

- Il supporto agli utenti rientra, per le normali richieste, all'interno di quanto il Settore di Calcolo Scientifico della Sezione già fornisce per tutti gli utenti della GRID locale; per richieste più complesse, o per escalation, esiste il servizio di supporto centrale di INFN-GRID.

7 RIEPILOGO TOTALE DEI COSTI

Tutti i costi si intendono IVA inclusa.

Per i motivi esposti nel paragrafo "Posizionamento, Alimentazione e Rete Ethernet" la soluzione integrata è fortemente consigliata; la soluzione "Low cost" viene presentata per indicare il costo del minimo occorrente per mettere in produzione il sistema.

I costi Esterni non INFN si riferiscono a attrezzature già fornite a titolo gratuito da partner tecnologici. Tali forniture sono frutto di progetti conclusi e non hanno altre ricadute o impatti sul presente progetto.

Soluzione con rete integrata nel GRID Data Center (Scenario A, raccomandato).

TAB. 2: Costi di realizzazione (IVA inclusa)

	Sezione	Esterni (INFN)	Esterni non INFN
Adeguamento Impianti elettrico e di condizionamento	9.000	9.000	
Acquisizione server, storage e rete Ethernet		119.600	150.000 (AMD, Acer)
Posizionamento, Alimentazione (rack, distribuzione elettrica)		8.400	
Rete veloce a bassa latenza (switch, cablaggio)			144.800 (AMD, Cisco, IBM)
Installazione, Configurazione			
Messa in produzione (costi software una tantum)			
Totale investimento	**9.000**	**137.000**	**294.800**

TAB. 3: Costi di manutenzione ed esercizio (esclusa energia) (IVA inclusa)

	Sezione	Esterni	Esterni non INFN
Costi di manutenzione annui	3.500	6.000	
Costi di licenze software annui			
Costo di manutenzione ed esercizio per i primi 3 anni	**10.500**	**18.000**	

TAB. 4: Costi energetici (stima) (IVA inclusa)

	Sezione	Esterni	Esterni non INFN
Costi energetici annui	15.000	35.000	
Costo energetico per 3 anni	**45.000**	**105.000**	

Soluzione “Low Cost” (Scenario B, di ripiego).

TAB. 5: Costi di realizzazione (IVA inclusa)

	Sezione	Esterni (INFN)	Esterni non INFN
Adeguamento Impianti elettrico e di condizionamento	9.000	9.000	
Acquisizione server, storage e rete Ethernet		64.600	150.000 (AMD, Acer)
Posizionamento, Alimentazione (rack, distribuzione elettrica)		8.400	
Rete veloce a bassa latenza (switch, cablaggio)			144.800 (AMD, Cisco, IBM)
Installazione, Configurazione			
Messa in produzione (costi software una tantum)			
Totale investimento	**9.000**	**82.000**	**294.800**

TAB. 6: Costi di manutenzione ed esercizio (eslusa enegia) (IVA inclusa)

	Sezione	Esterni	Esterni non INFN
Costi di manutenzione annui		6.000	
Costi di licenze software annui			
Costo di manutenzione ed esercizio per i primi 3 anni		**18.000**	

TAB. 7: Costi energetici (stima) (IVA inclusa)

	Sezione	Esterni	Esterni non INFN
Costi energetici annui	15.000	35.000	
Costo energetico per 3 anni	**45.000**	**105.000**	

In entrambe le soluzioni il costo per la configurazione con 2 GB/core al posto di 1 GB/core è di 41.400 Euro aggiuntivi (IVA inclusa).

Soluzione con rete integrata nel GRID Data Center (A, raccomandato).

TAB. 8: TCO (Total Cost of Ownership: investimento + ricorrenti) per i primi tre anni (IVA inclusa)

Sezione	Esterni	Esterni non INFN
64.500	260.000	294.800

TAB. 9: Finanziamento su tre anni (IVA inclusa)

2010	2011	2012
180.000	40.000	40.000

Soluzione "Low Cost" (B, di ripiego).

TAB. 10: TCO (Total Cost of Ownership: investimento + ricorrenti) per i primi tre anni (IVA inclusa)

Sezione	Esterni	Esterni non INFN
54.000	205.000	294.800

TAB. 11: Finanziamento su tre anni (IVA inclusa)

2010	2011	2012
125.000	40.000	40.000

Allegato 1

Progetto CSN4 Cluster: Requisiti

Il cluster in oggetto rappresenterà il servizio centralizzato per il calcolo parallelo degli utenti del Gruppo IV. Il servizio erogato è inteso per un triennio (2010 – 2012) , ma il cluster potrà essere utilizzato anche negli anni successivi, per i quali la CSN4 valuterà la possibilità di Upgrade o di estensioni dell'hardware.

Il progetto deve essere realizzato ipotizzando una potenza di picco dell'ordine di **5 TFlops.**

I requisiti minimi o indispensabili sono:

- **I nodi di calcolo** devono essere multicore a 64 bit (x86_64) con almeno 1 GB di memoria per core. Riportare nel progetto il costo per un eventuale incremento di 1 GB di memoria per core.

- **Interconnessione tra i nodi** con Infiniband o rete con prestazioni (banda e latenza) equivalenti in ambiente MPI.

- **Storage**: Lo spazio home per gli utenti deve essere di almeno 10 TB, condiviso tra i nodi di calcolo ed eventualmente espandibile.

- **Software**: Il sistema operativo (x86_64) e le librerie MPI-2 devono essere supportati da InfnGrid possibilmente aggiornati all'ultima major release. Il dettaglio delle altre librerie e tools verrà concordato con il sito.

- **Accessibilità**: Il cluster dovrà offrire i servizi per l'accesso attraverso Grid nella prospettiva che questo diventi l'unica modalità supportata quando si arriverà ad un piena integrazione di MPI nell'infrastruttura di InfnGrid. Ciò dovrà consentire una flessibile ripartizione delle risorse (tempo di calcolo e storage) tra i gruppi di utenti ed eventualmente utenti all'interno dei gruppi (fair share), nonché la gestione del relativo sistema di accounting. Accessi diversi da Grid potranno essere supportati temporaneamente, ma con ridotte funzionalità in termini di flessibilità di gestione e di accounting che andranno concordate localmente. Dovranno essere resi disponibili agli utenti adeguati servizi Web per la documentazione ed il monitoraggio del sistema.

Indice