GARR

The Italian Academic & Research Network

ENEA

Ente per le Nuove tecnologie, l'Energia e l'Ambiente

www.garr.it

Problematiche di rete nella sperimentazione di file-system distribuiti su WAN per applicazioni di GRID-Computing

Rapporto attività di studio marzo/12 - dicembre/12

Andrea Petricca

Due anni fa..

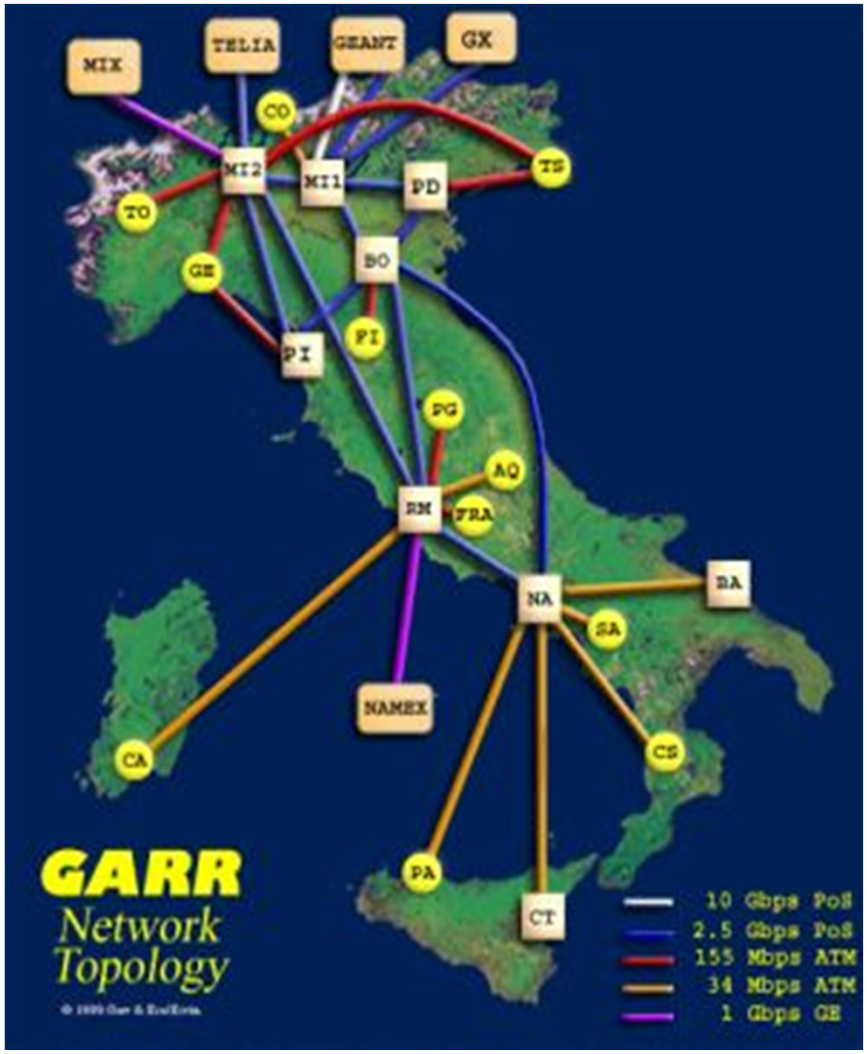
La rete è in continua evoluzione, vi è quindi la necessità di garantire in ambito GRID-Computing la qualità dei servizi offerti all'utenza.

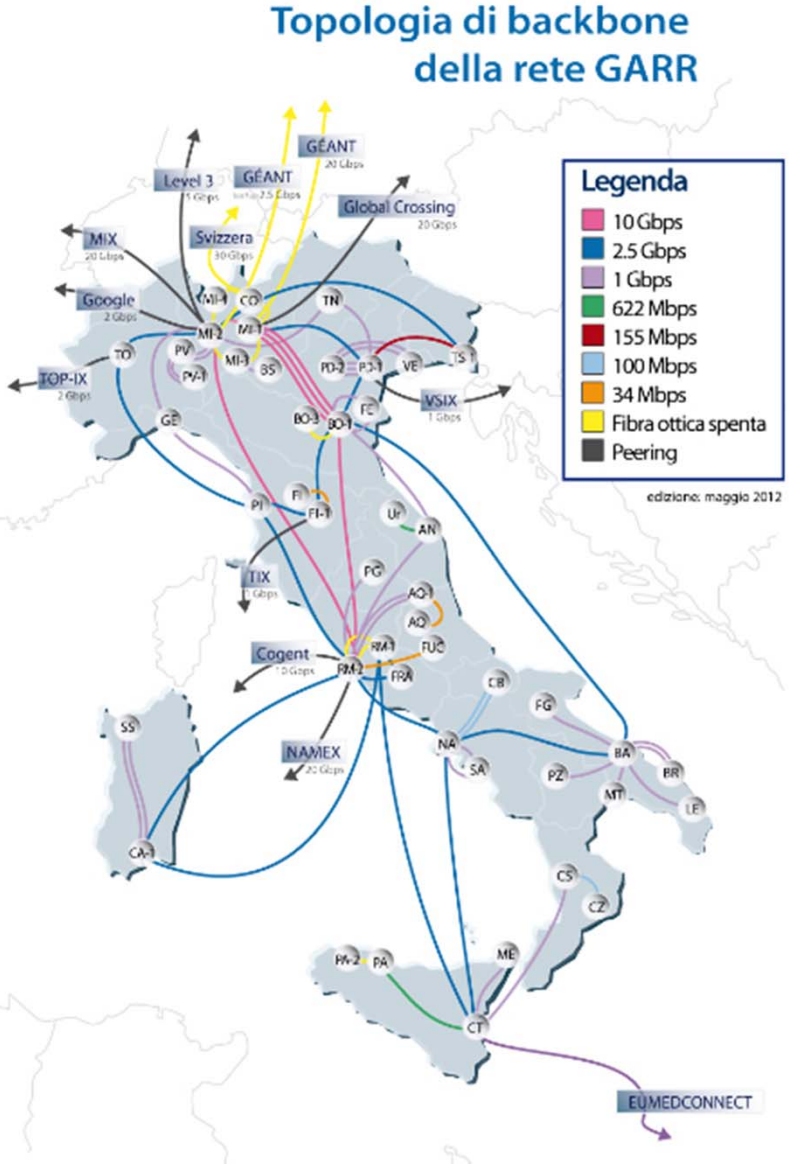
Il monitoring degli apparati diventa fondamentale per avere sia il controllo generale dell'infrastruttura sia lo strumento per individuare parametri critici con la finalità di migliorare le prestazioni e le funzionalità del sistema.

Il servizio reso dal monitoring dedicato deve quindi sia controllare la normale attività della rete, dei nodi di calcolo dell'infrastruttura e dei sistemi di storage utilizzati sia verificare e valutare le prestazioni del completo sistema di GRID-Computing.

Questo semplifica la risoluzione dei problemi e le operazioni di gestione della rete.

Due anni fa…





www.garr.it

Sistema di monitoring distribuito: Zabbix (1)

ZABBIX
MONITORING SYSTEM

Monitoring da un unico punto di accesso.

Ambiente server multilivello (Master & Child)

- Da semplice Client-Server a sistema distribuito composto da vari livelli di clienti e serventi
- Master sufficientemente potente da permettere la gestione dei dati inviati dai Server-Child

High Performance Monitoring

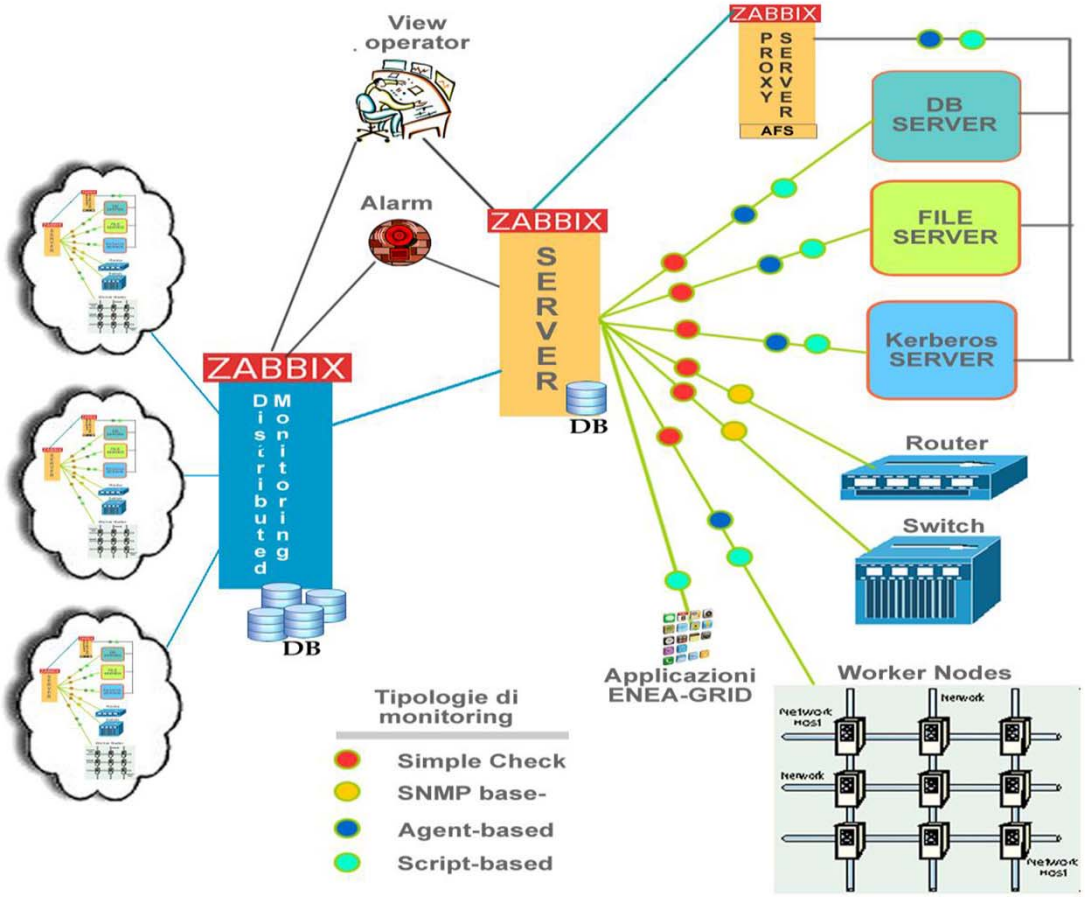
- Riduzione del carico dei singoli Zabbix-Server
- Possibilità di utilizzare macchine server virtuali

Troubleshooting delle problematiche di connettività, dei disservizi e delle prestazioni

Sistema di monitoring distribuito: Zabbix (2)

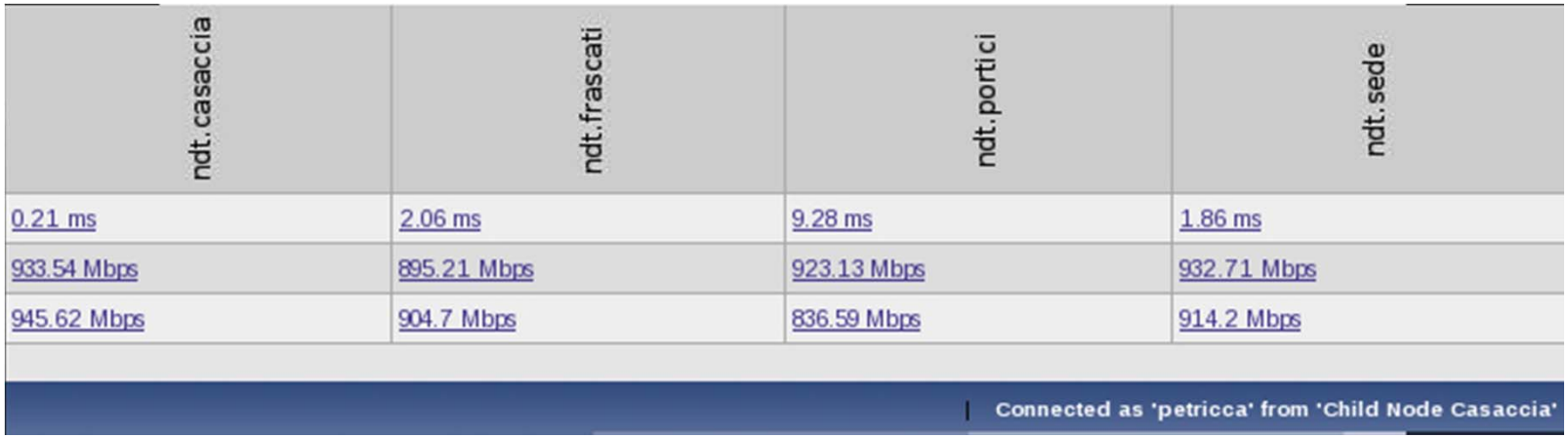
Situazione attuale server installati e configurati nei centri ENEA.

- ✓ Frascati
- ✓ Casaccia
- ✓ Portici
- ✓ Bologna
- ✓ Brindisi
- ✓ Trisaia



Analisi di banda: Server NDT

- Nelle sedi di Casaccia, Frascati, Portici, Roma Sede, Trisaia e Brindisi si è deciso di installare dei server NDT per monitorare costantemente la banda disponibile tra le diverse sedi per ottimizzare i collegamenti ed individuare i probabili colli di bottiglia.
- I server, tutti operanti sulla porta 80 e non 7123, sono identificati con ndt.'nomesede'.enea.it
- Ciò permette a tutti gli amministratori di rete di controllare costantemente la qualità dei collegamenti tra le sedi ENEA.
- Importante è stata la possibilità di automatizzare le operazioni di analisi su Zabbix. È il software stesso che tramite degli script esterni, esegue i test NDT, storicizzando i risultati e mettendo in allarme in caso di problematiche derivanti la rete.



ndt.casaccia	ndt.frascati	ndt.portici	ndt.sede
0.21 ms	2.06 ms	9.28 ms	1.86 ms
933.54 Mbps	895.21 Mbps	923.13 Mbps	932.71 Mbps
945.62 Mbps	904.7 Mbps	836.59 Mbps	914.2 Mbps

Connected as 'petricca' from 'Child Node Casaccia'

GPFS su WAN

- L'attività svolta di maggiore interesse è stata la messa in produzione di un laboratorio dove poter testare il file-system distribuito GPFS su WAN.
- La possibilità di avere a disposizione delle macchine e dei dischi dedicati ha permesso la creazione di un cluster GPFS, che è stato poi adattato per i vari test che si volevano svolgere.
- Le configurazioni sono state sostanzialmente due: in single-cluster e in multi-cluster.
- Sono stati fatti dei test che mettessero a confronto le velocità di trasferimento dati tra le sedi ENEA, e con metodi differenti.

GPFS su WAN: Single-Cluster (1)

- Il primo passo è stato quello di creare macchine sostanzialmente identiche dal punto di vista software e il più possibile nell'hardware.
- Sono state configurate 5 macchine, trasferite poi fisicamente nelle sedi di Frascati (2), Portici (1), Brindisi (1) e Trisaia (1).
- Dopo alcune difficoltà di configurazione, siamo riusciti a strutturare il cluster GPFS su WAN, verificando poi la consistenza dei dati, l'affidabilità delle repliche e le performance di trasferimento tra le sedi dove era presente la macchina appartenente al cluster GPFS.
- Abbiamo confrontato poi le velocità di trasferimento dati con il più comune comando SCP, verificando notevoli vantaggi nell'uso del file-system GPFS.

GPFS su WAN: Single-Cluster (2)

- Le macchine utilizzate, dal punto di vista di rete, sono state tutte inserite, ove possibile, nella LAN di calcolo Enea.
- Le macchine di Frascati e di Portici sono collegate su porte ad 1Gb/s allo switch dedicato al Calcolo, che a sua volta è collegato in fibra ottica al router centrale della sede, quest'ultimo connesso al Pop GARR della sede Enea.
- La macchina di Trisaia utilizza un doppio collegamento in load balancing a 200Mb/s con la dorsale del GARR.
- La macchina di Brindisi utilizza un collegamento ad 1Gb/s condiviso con il polo universitario di Brindisi, per cui nel peggiore dei casi ha una linea dedicata di 250Mb/s.

GPFS su WAN: Prestazioni single-cluster

Viene riportata di seguito la tabella relativa ai test eseguiti tra le macchine del Cluster GPFS.

La scrittura e la lettura dei file è stata svolta utilizzando il comando lmdd, parametrizzato per eseguire trasferimenti di file della dimensione di 150MB.

I dati sono stati trasferiti dal volume locale montato sul file-system GPFS verso i volumi montati sullo stesso file-system localizzati nelle altre sedi.

	/gwan_FRA1 (MB/s)		/gwan_FRA2 (MB/s)		/gwan_POR1 (MB/s)		/gwan_TRI1 (MB/s)		/gwan_BRI1 (MB/s)	
Franscati 1	65	75	90	90	45	45	10	10	28	10
Frascati 2	63	63	100	100	45	45	10	10	45	14
Portici 1	60	70	70	90	50	50	10	10	30	30
Trisaia 1	10	10	10	10	10	10	70	78	10	10
Brindisi 1	8	60	6	60	28	60	8	11	50	75

Write file (of)
Read file (if)

GPFS LMDD OF IF
150 MB dati

GPFS su WAN: Confronto con SCP (1)

Viene riportata di seguito la tabella relativa ai test eseguiti tra le macchine utilizzando il comando SCP. La scelta del file da utilizzare è stata quella di riutilizzare lo stesso file creato per il test con il file-system GPFS. Dato che il comando SCP contempla la possibilità di comprimere i dati che si intendono inviare, ci siamo assicurati che il file fosse completamente randomico e non costituito da tutti zeri, facilmente comprimibili.

	/gwan_FRA1 (MB/s)			/gwan_FRA2 (MB/s)			/gwan_POR1 (MB/s)			/gwan_TRI1 (MB/s)			/gwan_BRI1 (MB/s)		
Franscati 1	28	7,5	35	33	8	47	28	7,1	35	7,9	7,1	6,4	15,8	7,9	14,9
Frascati 2	28	7,5	47	23	7,9	28	11	7,1	10	1,2	3,8	1,4	1,3	5,3	3,5
Portici 1	28	7,9	35	20	7,5	35	17	7,9	35	1,3	4,6	1,2	9,5	5,7	9,5
Trisaia 1	5,6	7,1	5,9	6,7	7,1	5	5	7,1	7,1	20	25	33,3	11	13	11
Brindisi 1	2,8	6,1	2,7	2	6,8	2	2,5	6,8	1,9	10,9	7,5	10,9	23,6	7,9	28,4

Write file (compressione: NO, cifrario: 3DES)
Write file (compressione: YES, cifrario: BLOWFISH)
Write file (compressione: NO, cifrario: BLOWFISH)

SCP
150MB dati

GPFS su WAN: Confronto con SCP (2)

In questo test in scrittura, sono stati eseguiti delle variazioni sui set del comando scp in funzione della tipologia di cifratura dei dati e della possibilità di compressione.

Vanno fatte delle considerazioni in funzione dei risultati ottenuti.

Il cifrario BLOWFISH, scelto come test alternativo al cifrario 3DES (default del comando scp) è risultato sostanzialmente migliore, sia nei collegamenti di rete ad 1Gb/s delle sedi di Frascati e Portici, sia nei collegamenti a 250Mb/s di Trisaia e Brindisi.

Va sottolineato come la compressione dei dati possa non influire e tanto più peggiorare le prestazioni di trasferimento nel caso di collegamenti più veloci (1Gb/s), mentre risulta vantaggiosa nei collegamenti limitati da una banda inferiore (200-250Mb/s).

GPFS su WAN: Confronto con SCP (3)

- Come è possibile verificare dalla tabella seguente, vi è un notevole vantaggio nell'utilizzare trasferimenti dati tramite il file-system GPFS, qualsiasi sia la tipologia di connessione di rete tra le due macchine.

	Frascat1 (MB/s)		Frascati2 (MB/s)		Portici (MB/s)		Trisaia (MB/s)		Brindisi (MB/s)	
Frascati1	65	35	90	47	45	35	10	7,9	28	15,8
Frascati2	63	47	100	28	45	11	10	3,8	45	5,3
Portici	60	35	70	35	50	35	10	4,6	30	9,5
Trisaia	10	7,1	10	7,1	10	7,1	70	33,3	10	13
Brindisi	8	6,1	6	6,8	28	6,8	8	10,9	50	28,4

LMDD su File-System GPFS distribuito su WAN

SCP tra host geodistribuiti su WAN

- Nei casi più interessanti, ossia nei collegamenti da 1 Gb/s tra le sedi, abbiamo l'aumento di almeno un fattore 2 del data-rate di trasferimento.
- Per quanto riguarda invece trasferimenti su collegamenti <o= 250 Mb/s si notano fattori altalenanti, dipendenti probabilmente dalla condizione di traffico della rete e della saturazione della banda disponibile.

GPFS su WAN: Multi-Cluster (1)

- Il test svolto ha permesso di verificare la funzionalità multi-cluster di GPFS in ambiente geo-distribuito.
- Sono stati creati due cluster diversi, messi poi in comunicazione remota tra di loro.
- Il test ha dato esito positivo, verificando l'effettiva possibilità di una migrazione da file-system GPFS utilizzato esclusivamente in configurazione locale a file-system GPFS effettivamente su WAN.
- La migrazione delle configurazione verrà eseguita entro la fine di ottobre per i server GPFS già produzione nelle principali sedi ENEA.

GPFS su WAN: Multi-Cluster (2)

- Sono stati svolti test analoghi sulle prestazioni relative al trasferimento dati tra le sedi, ottenendo gli stessi risultati del caso single-cluster.
- Attenzione particolare è stata data alla possibilità di utilizzare delle interfacce InfiniBand per la gestione dati in locale, e una interfaccia TCP/IP che gestisca la comunicazione con l'esterno. Sono stati quindi eseguiti dei test che abilitassero e verificassero entrambe le comunicazioni.
- Queste prove hanno dato esito positivo. Si è proceduto quindi alla messa in produzione del sistema testato sul sistema in produzione CRESCO3.

- Tutto il lavoro svolto su GPFS è stato possibile grazie al collega Agostino Funel, che ha seguito, supervisionato e gestito tutta l'attività di messa in produzione.

GPFS su WAN: Operazioni sul Multi-Cluster

- Passaggio del Cluster GPFS di Portici che permette di utilizzare l'interfaccia pubblica dei nodi ed abilitazione di RDMA.

  L'operazione è stata compiuta con successo e l'abilitazione di RDMA su InfiniBand DDR ha permesso di raddoppiare le prestazioni in lettura e scrittura. I risultati per le tre tipologie di storage disponibili, in scrittura/lettura da singolo nodo sono:
  DDN9550 1.4 GB/s invece di 700 MB/s
  DDN9900 1.5 GB/s invece di 900 MB/S
  ServerSTorage E4 (MINNI) 700 MB/s invece che 400 MB/s

- Abilitazione della configurazione Multicluster GPFS tra Portici e Frascati

  I cluster GPFS di Portici e Frascati sono stati configurati in modo che entrambi possono pubblicare remotamente alcuni dei file system.

- Le prestazioni di trasferimento su WAN sono circa 30-40 MB/s ma si deve fare ancora qualche lavoro di ottimizzazione. Ricordo che AFS su WAN da prestazioni inferiori di un ordine di grandezza.

Conclusioni

L'attività da me svolta tramite la collaborazione GARR-ENEA si è concentrata sempre su due obiettivi principali: Monitoring e Prestazione File-System.

Questo ha permesso di parametrizzare un numero elevato di valori che facilitano il controllo real-time del funzionamento e delle prestazioni dell'infrastruttura di calcolo di EneaGRID in funzione dei file-systems AFS e GPFS sulla quale si basa.

Il software Zabbix, suggerito a suo tempo dal GARR, si è rivelato idoneo nell'automatizzare controlli più o meno complessi, funzionale negli allarmi e nella praticità della grafica interna con una notevole possibilità di espansione verso l'esterno.