GARR

The Italian Academic & Research Network

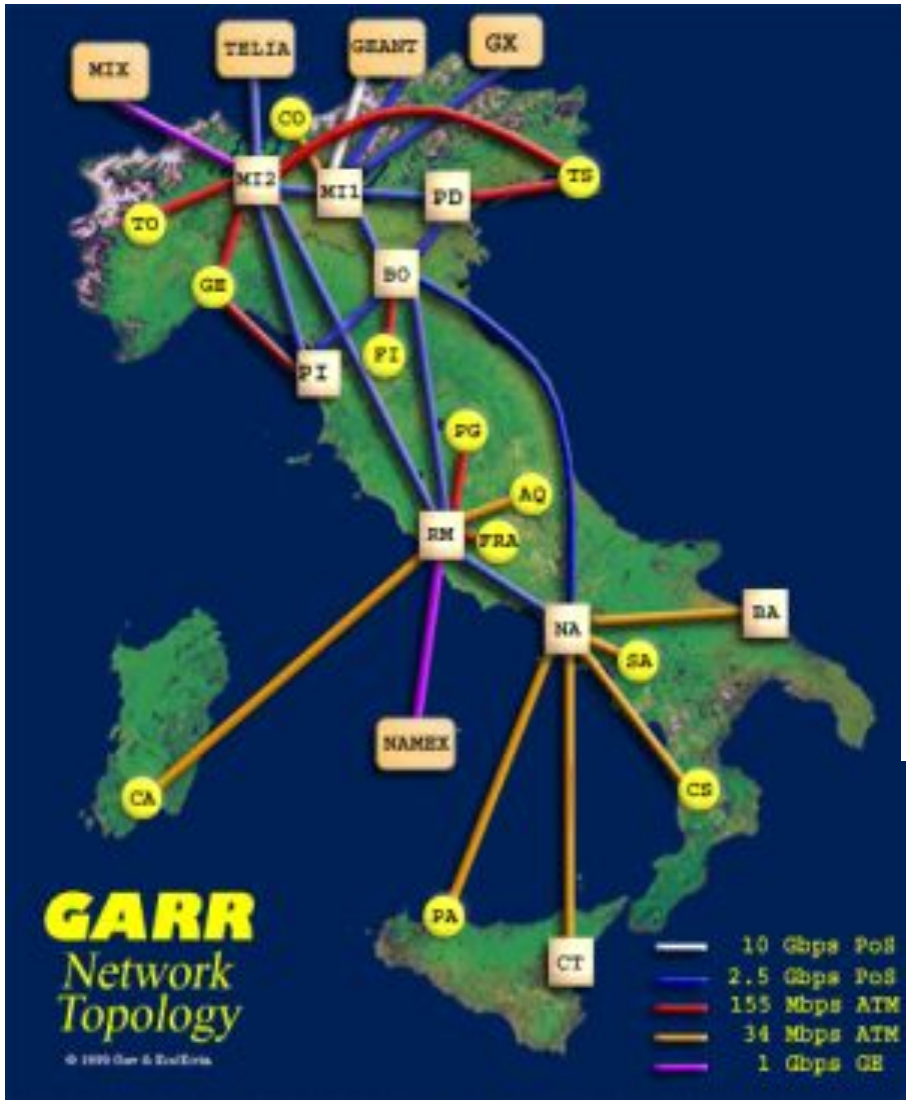
Ente per le Nuove tecnologie, l'Energia e l'Ambiente

www.garr.it

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

GARR - II Borsisti Day

Infrastruttura ENEA-GRID & rete GARR



www.garr.it

-12 Centri di ricerca in Italia
- 6 Computer Centres: Casaccia, Frascati, Bologna, Trisaia, Portici e Brindisi

Obiettivi della Borsa di Studio

- Studio delle criticità inerenti l'utilizzo di file system distribuiti su WAN nel contesto ENEAGRID
- Sviluppo di interfacce orientate all'amministratore o all'utente, focalizzate all'analisi dei problemi specifici dei file system su WAN
- Utilizzo delle informazioni raccolte per implementare soluzioni ottimizzate per la configurazione e la gestione dei file system in questione

Fasi del lavoro svolto

- Ricerca ed analisi delle potenzialità di un open-software di monitoring: la scelta di Zabbix
- Implementazione nel centro di Frascati
- Messa in produzione in diversi C.R. ENEA
- Passaggio al Distributed Monitoring
- Creazione sonde specifiche su AFS in ambiente WAN
- Dimostrazione dell'efficacia del sistema di monitoring tramite l'individuazione di problematiche specifiche per AFS su WAN
- Lo studio del file system GPFS non è stato possibile causa ritardi di implementazione su WAN nell'infrastruttura ENEAGRID

Zabbix Monitoring System

ZABBIX
MONITORING SYSTEM

• **Zabbix** è un'applicazione **web-based Open Source** multi-piattaforma di monitoraggio e analisi per sistemi informatici su reti locali e distribuite. Adattabile perfettamente alle caratteristiche richieste dall'ambientazione **GRID-Computing su WAN** (distributed monitoring).

• Consente di monitorare in tempo reale servizi e apparati di rete, attraverso:

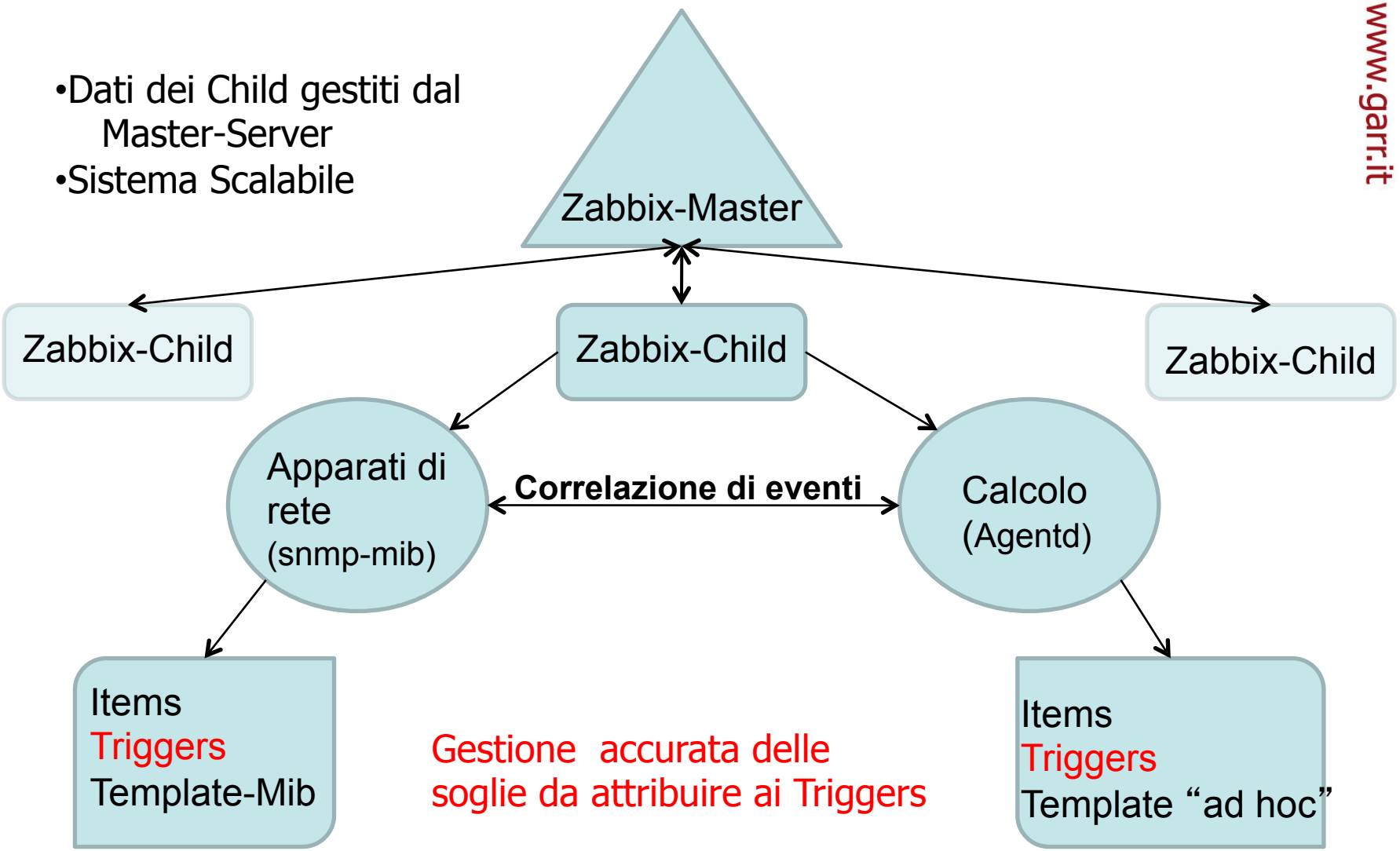
Ssh, Snmp, Telnet, IPMI, agent Zabbix (multipiattaforma), **external scripts**.

• I **dati** collezionati **real-time**, definiti tramite l'interfaccia web, vengono **memorizzati** su di un **Database** (MySQL, PostgreSQL, Oracle), per poi essere elaborati.

• Ogni **workflow** può essere astratto creando delle **template** personalizzate in aggiunta alle molte templates già disponibili.

Zabbix-Server Master&Child

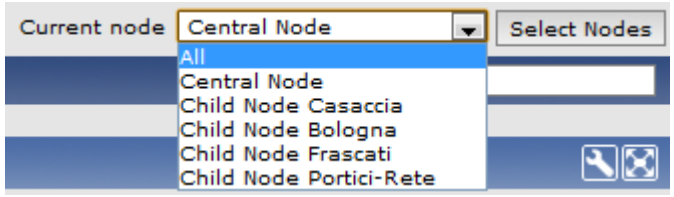
- Dati dei Child gestiti dal Master-Server
- Sistema Scalabile



Gestione accurata delle soglie da attribuire ai Triggers

Cosa è stato possibile fare con il Monitoring Distribuito

- Unico accesso sul Server Master di Portici:
  - Possibilità di switchare tramite un semplice clic da un Server ad un altro.



- Possibilità di creare screens e maps interattivi, con la caratteristica di entrare nello specifico fino alla localizzazione dell' errore rilevato dai triggers.

- Troubleshooting delle problematiche di connettività, dei disservizi e delle prestazioni.

- Possibilità di mettere in produzione scripts che agiscano parallelamente su più centri ENEA, in particolare sull' analisi del FileSystem AFS.

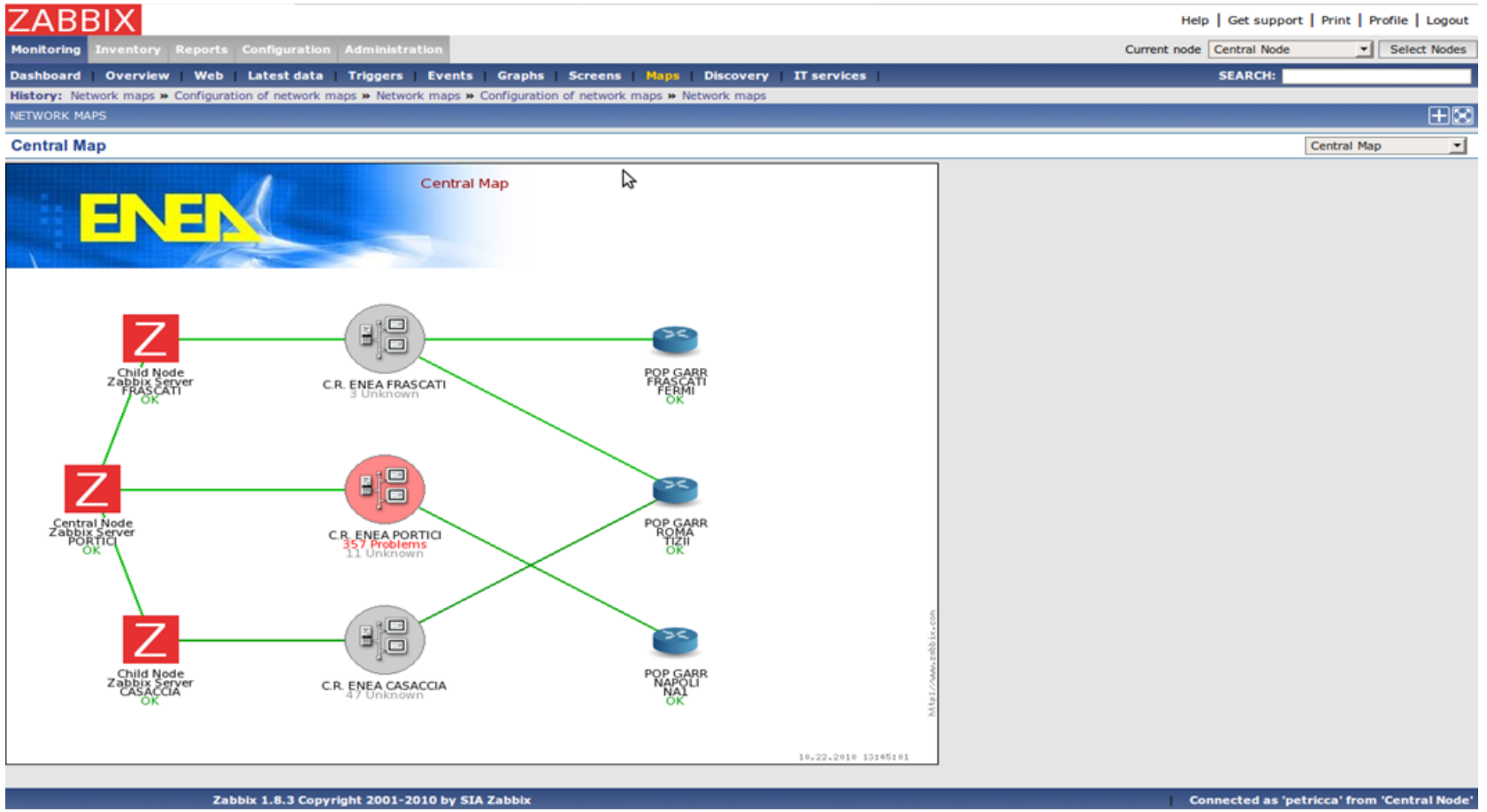
Quali problematiche ha dato il Monitoring Distribuito

- I figli costano!!!
  - Ottimizzazione della gestione del database centrale
- Replicazione delle informazioni
  - A vantaggio della centralizzazione delle informazioni
  - Customizzazione più complessa

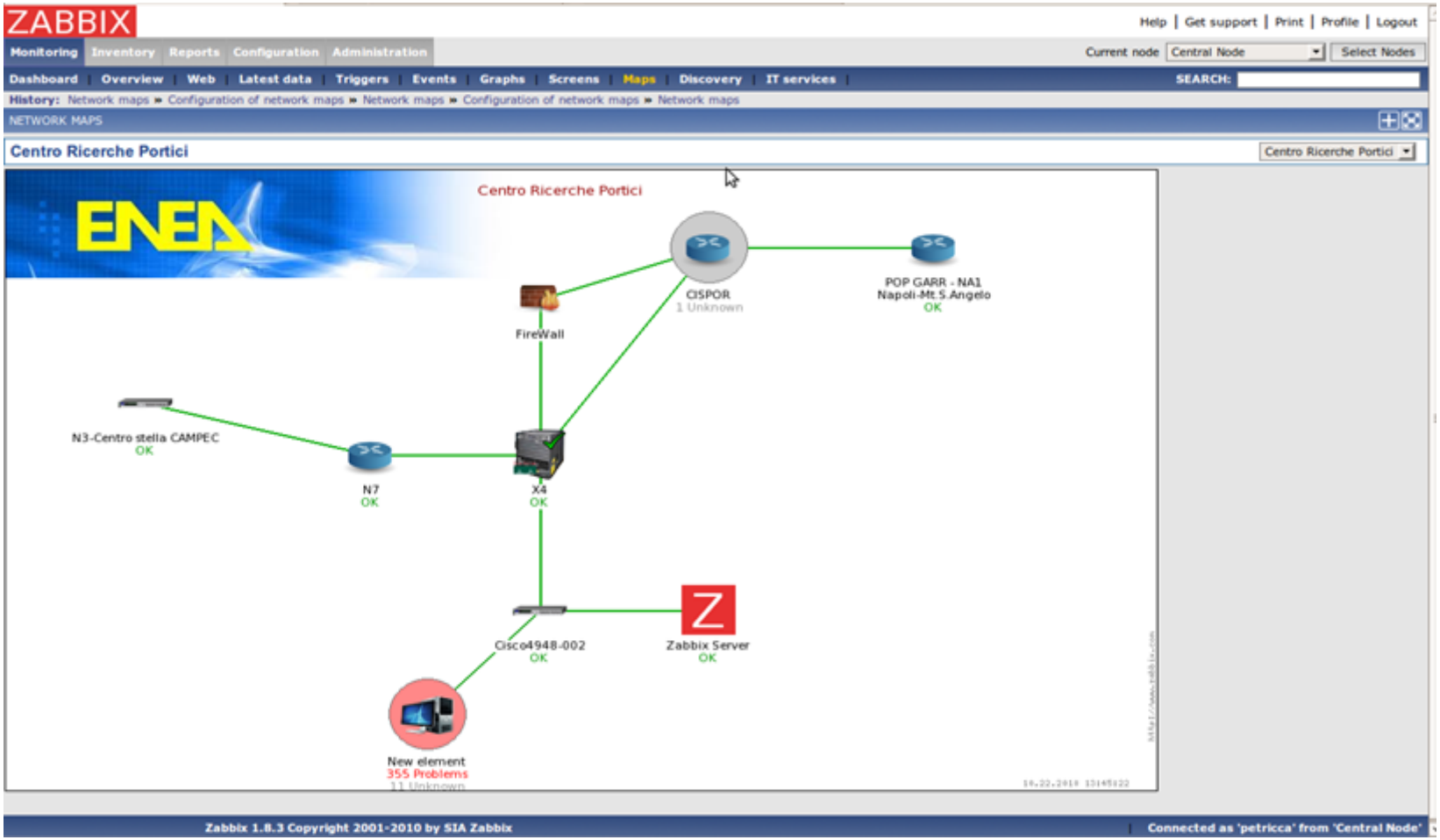
Modalità di segnalazione di errore o criticità

- Mail/sms
- Dashboard: visualizzazione della tipologia di errore e relativa gravità
- Overview: visualizzazione valore puntuale real-time
- Screens grafici

Fasi di individuazione errore (1)



Fasi di individuazione errore (2)



Fasi di individuazione errore (3)

Nome host

Time	Host	Description	Status	Severity
2011.Feb.14 07:26:31	virgilio.brindisi.enea.it	Test-trigger Down calls on 7000	OK	High
2011.Feb.14 07:26:30	linuxfs.trisaia.enea.it	Test-trigger Down calls on 7000	OK	High
2011.Feb.14 05:26:31	virgilio.brindisi.enea.it	Test-trigger Down calls on 7000	PROBLEM	High
2011.Feb.14 05:26:30	linuxfs.trisaia.enea.it	Test-trigger Down calls on 7000	PROBLEM	High
2011.Feb.14 02:26:29	linafs1.frascati.enea.it	Test-trigger calls waiting on 7000 max for 10 count >5	OK	Warning

Gravità errore

EVENTS

Group ENEA AFS FileServers

Displaying **1** to **4** of **4** found

Filter

Time	Description	Status	Severity
2011.Feb.16 11:58:15	Test-trigger Down calls on 7000	UNKNOWN	High
2011.Feb.14 07:26:30	Test-trigger Down calls on 7000	OK	High
2011.Feb.14 05:26:30	Test-trigger Down calls on 7000	PROBLEM	High
2011.Feb.09 16:10:29	Test-trigger Down calls on 7000	OK	High

Integrazione tra GINS e Zabbix (GARR  ENEAGRID)

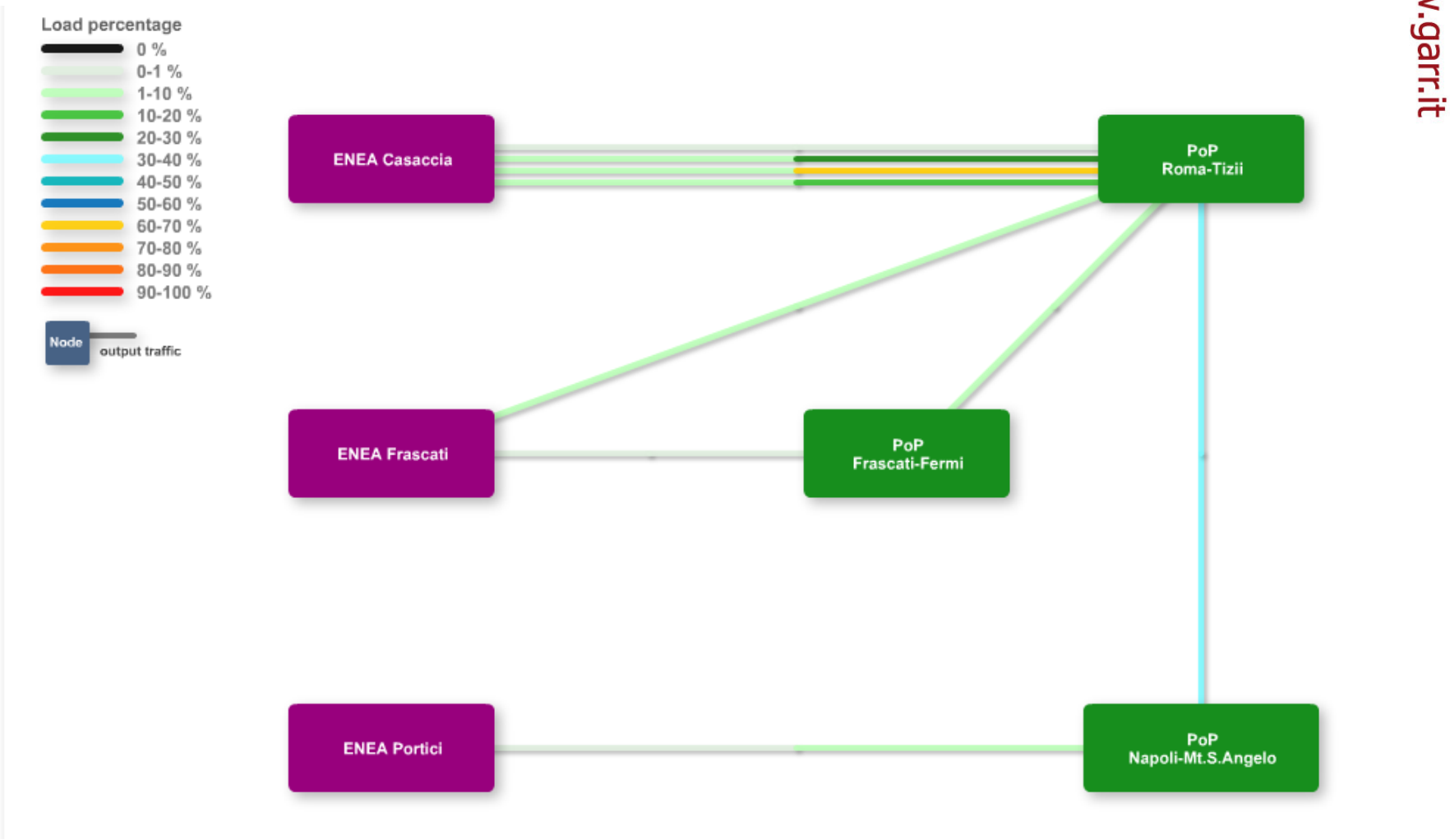
GARR Integrated Networking Suite

Owned by: sw.dev@garr.it

ZABBIX
MONITORING SYSTEM

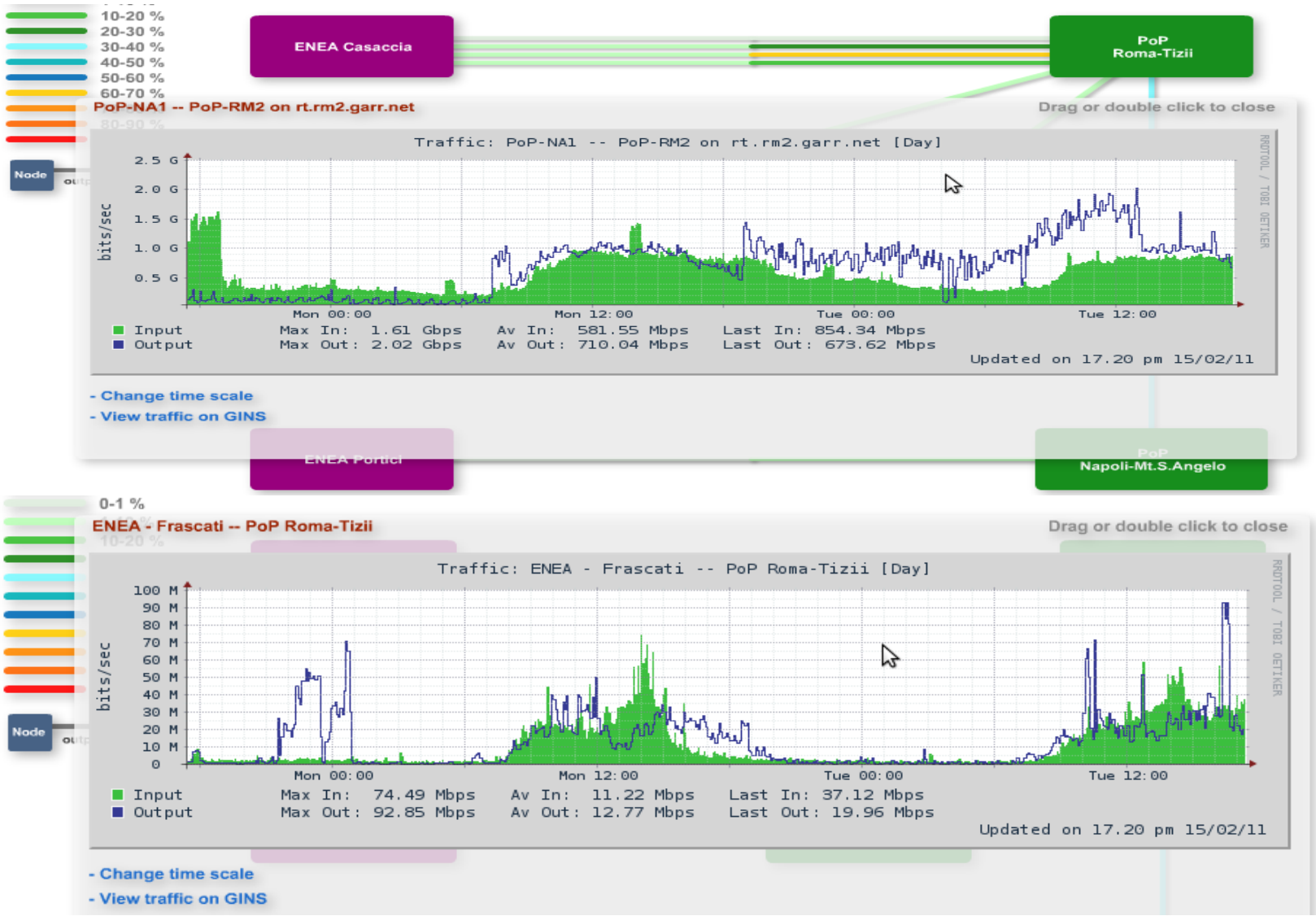
- Interattività tra Zabbix e GINS.
  Possibilità di integrazione dei due sistemi di Monitoring
- Analisi dello stato del C.R. completo:
  - ✓ Rete GARR (relativa al C.R.).
  - ✓ Rete interna.
  - ✓ Servizi di rete.
  - ✓ Cluster GRID (external scripts AFS, GPFS).

GINS: Weathermap per ENEA(1)



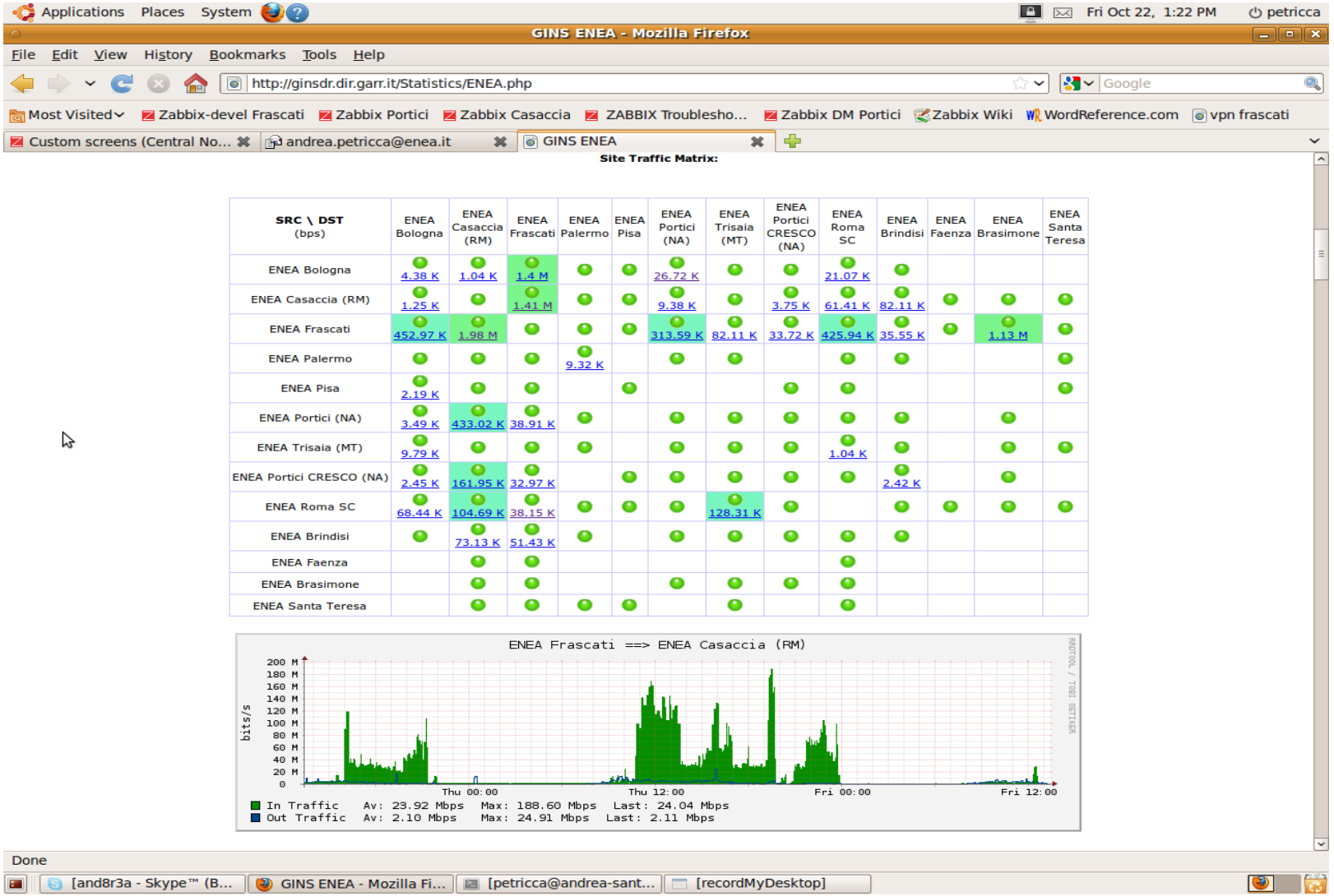
www.garr.it

GINS: Weathermaps per ENEA(2)



www.garr.it

GINS: Site Traffic Matrix per ENEA



SRC \ DST (bps)	ENEA Bologna	ENEA Casaccia (RM)	ENEA Frascati	ENEA Palermo	ENEA Pisa	ENEA Portici (NA)	ENEA Trisaia (MT)	ENEA Portici CRESCO (NA)	ENEA Roma SC	ENEA Brindisi	ENEA Faenza	ENEA Brasimone	ENEA Santa Teresa
ENEA Bologna	4.38 K	1.04 K	1.4 M			26.72 K			21.07 K				
ENEA Casaccia (RM)	1.25 K		1.41 M			9.38 K		3.75 K	61.41 K	82.11 K			
ENEA Frascati	452.97 K	1.98 M				313.59 K	82.11 K	33.72 K	425.94 K	35.55 K		1.13 M	
ENEA Palermo				9.32 K									
ENEA Pisa	2.19 K												
ENEA Portici (NA)	3.49 K	433.02 K	38.91 K										
ENEA Trisaia (MT)	9.79 K								1.04 K				
ENEA Portici CRESCO (NA)	2.45 K	161.95 K	32.97 K							2.42 K			
ENEA Roma SC	68.44 K	104.69 K	38.15 K				128.31 K						
ENEA Brindisi		73.13 K	51.43 K										
ENEA Faenza													
ENEA Brasimone													
ENEA Santa Teresa													

Vantaggi offerti dall'integrazione di GINS & Zabbix

- Tramite l'**interazione Gins-Zabbix**, è possibile localizzare **in maniera esaustiva** le problematiche esistenti.

- **Si eliminano le condizioni di ambiguità** tra i possibili fattori di errore o rallentamento del sistema di Grid-Computing.

- **Ogni aspetto** è controllato da un **unico punto di accesso** al sistema di monitoring.

- Il grado di servizio dell'**intero sistema** di GRID geo-distribuito viene **monitorato nel suo complesso.**

Caratteristiche di rete del file system AFS

- L'architettura di AFS
  - Architettura Client-Server: client, file-server e db-server
  - Caching aggressivo su disco locale: performance e scalabilità
  - Livello di trasporti: Rx Protocol su UDP
  - Windows size dinamica con limite a 32
- Anche con banda elevata: Throughput limitato unicamente dal RTT, problema della latenza.
- In fase di studio il protocollo RxTcp

RTT tra centri ENEA – Limiti Rx Protocol

DA / A	Frascati	Portici	Casaccia	Bologna	Trisaia	Brindisi
Frascati	0.224 ms	6.036 ms	16.110 ms	11.463 ms	24.517 ms	17.668 ms
Portici	5.756 ms	0.138 ms	5.817 ms	12.242 ms	20.568 ms	8.239 ms
Casaccia	2.271 ms	5.340 ms	0.137 ms	34.783 ms	25.679 ms	19.293 ms
Bologna	24 ms	50 ms	30 ms	0. ms	53 ms	43 ms
Trisaia	23 ms	14 ms	25 ms	51 ms	0. ms	11 ms
Brindisi	17.539 ms	8.662 ms	19.098 ms	32.309 ms	13.699 ms	0.394 ms

Esempio MaxDataRate:
rtt ≈ 24 ms (tra Frascati e Trisaia)
Max. Datarate = windowsize * packetsize / rtt = 32 * 1440 / .024 ≈ 1.9 MB/s

rtt ≈ 14 ms (tra Trisaia e Portici)
Max. Datarate = windowsize * packetsize / rtt = 32 * 1440 / .014 ≈ 3.3 MB/s

rtt ≈ 6.036 ms (tra Frascati e Portici)
Max. Datarate = windowsize * packetsize / rtt = 32 * 1440 / .006 ≈ 7.7 MB/s

AFS – script di monitoring

AFS mette a disposizione una serie di strumenti, parametri e contatori che permettono l'analisi dello stato di carico e di congestione dei servizi offerti.

Afsmonitor (client vs. file server):

- Callback AFS – Resend & Reread
  - Trigger warning if last.value(5#)>2%

Rxdebug (file server):

- Calls waiting for a thread
  - Trigger warning if last.value(5#)>5
  - Trigger high if last.value < 0

Udebug (db server):

- Sync-Site AFS – Sincronia e Consistenza dbserver
  - Trigger high if last.value = 0

Caso pratico: utente X(1)

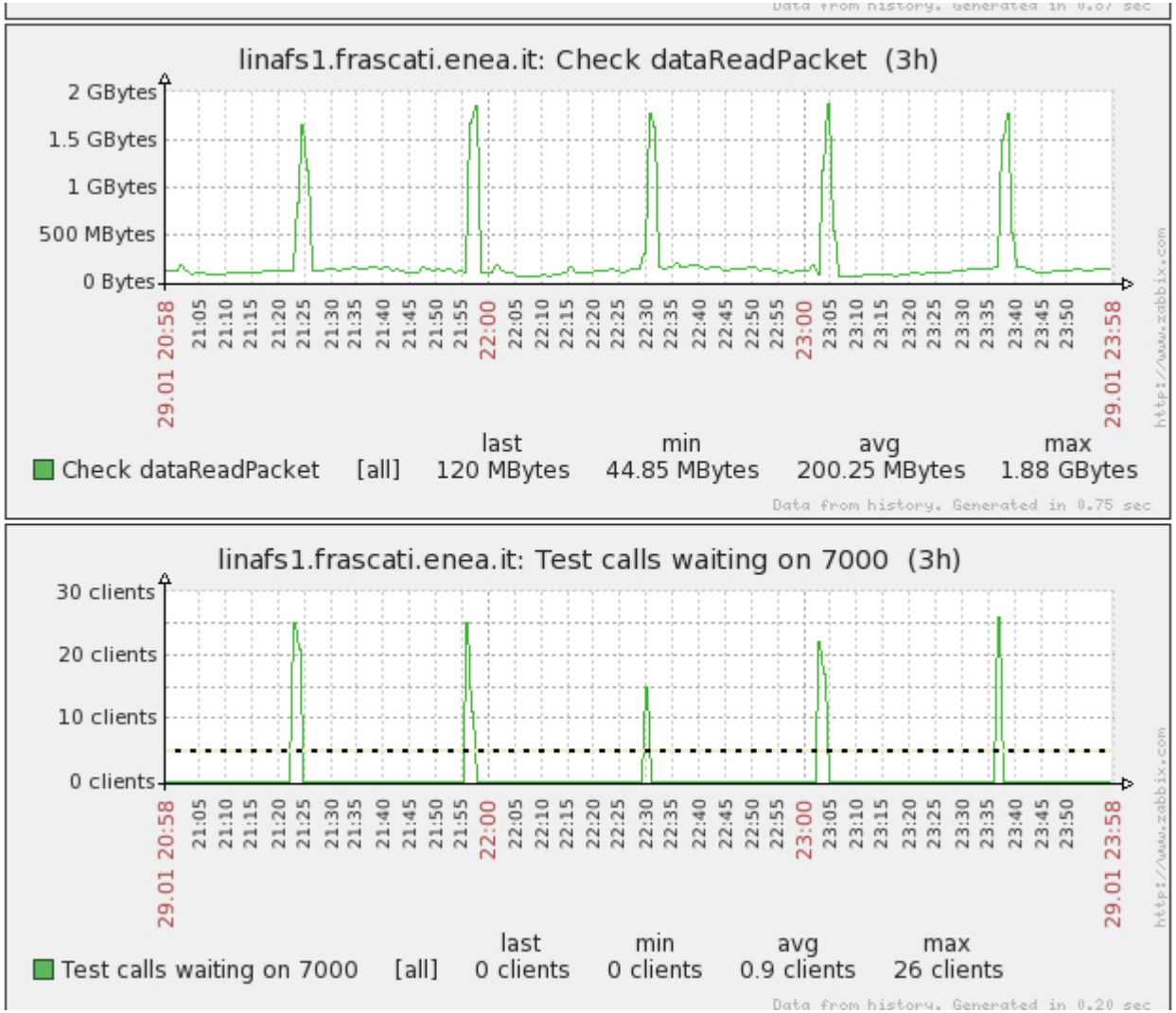
Il caso utente X rientra in quella casistica di eventi nei quali i jobs fatti girare dall'utente non rispettano le regole di "buon uso" della griglia computazionale di ENEA.

L'utenza, non rendendosi conto della localizzazione effettiva dei dati che venivano elaborati, lanciava un job parallelo su molti nodi di Portici che contemporaneamente andavano a scrivere dati temporanei su di un file server di Frascati

Il job, andava quindi ad occupare tutti i threads disponibili sul file server, creando un numero elevato di utenze in "call waiting" per circa 5 minuti, sovraccaricando oltretutto inutilmente la rete.

Problema individuato tramite la correlazione dei dati rilevati dagli external scripts creati.

Caso pratico: utente X (2)



Stato attuale in numeri: Zabbix-Centri ENEA

- Casaccia: 100 Agent monitorati – 4 Template “ad hoc”
- Frascati: 40 Agent monitorati – 6 Template-Mib “ad hoc”
- Portici: 640 Agent monitorati – 15 Template “ad hoc”
- Bologna: 10 Agent monitorati – 2 Template “ad hoc”

Nei centri di Portici e Frascati vengono monitorati costantemente gli 8 File Server e 8 Db Server che compongono la struttura di ENEAGRID

Zabbix monitora tramite SNMP circa 700 apparati di rete dislocati su tutta l’ infrastruttura ENEA.

Risultati ottenuti e Conclusioni (1)

- Implementazione di un' infrastruttura di monitoring
  - Frascati, Portici, Casaccia e Bologna
- Sistema completo di analisi delle prestazioni del file system distribuito AFS su WAN
  - ISP GARR, apparati di rete, risorse e servizi orientati al calcolo scientifico
  - Scripts esterni relativi al funzionamento di AFS (client, file server e dbserver)
- Metodologia per la risoluzione di problematiche comuni nel contesto GRID-Computing

Risultati ottenuti e Conclusioni (2)

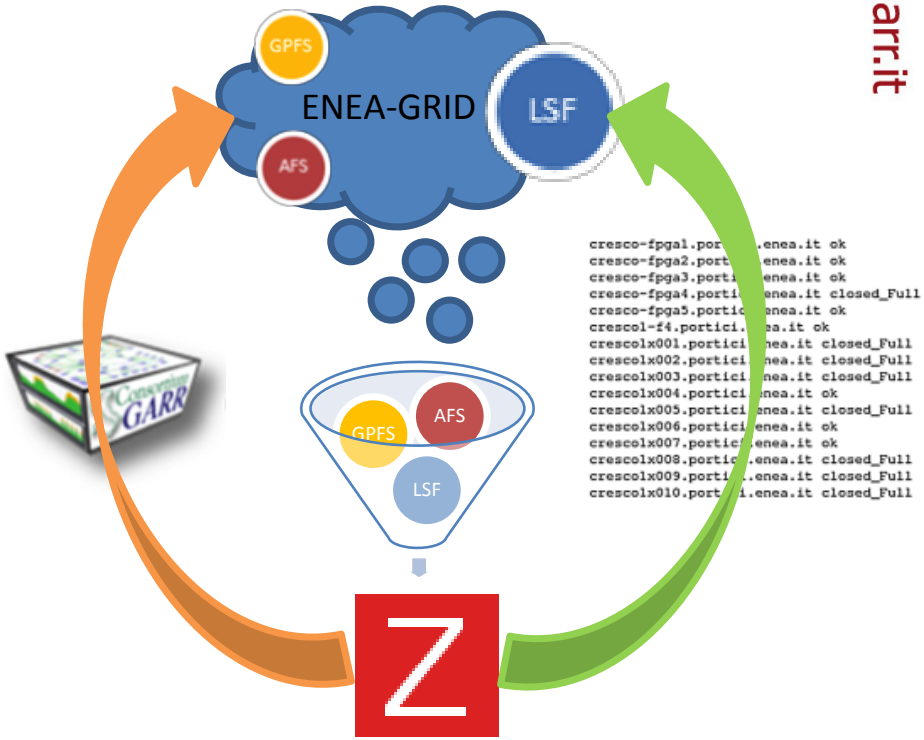
- Individuazione di scenari ottimali di esecuzione job seriali e paralleli secondo le richieste dell' utenza
- La criticità di tali sistemi si riflette nella specificità dei controlli effettuati e nella frequenza del loro utilizzo per garantire alla griglia ENEA il massimo standard possibile di disponibilità delle risorse per il calcolo scientifico.
- Il lavoro svolto, fino al mese di ottobre, è stato oggetto di una presentazione alla Conferenza GARR 2010 – Welcome to the Future Internet! 26-28 Ottobre 2010, Tornino ed è stato selezionato dal Comitato di Programma come "full paper" per la pubblicazione a cura del GARR negli atti ufficiali della conferenza stessa

Proposte prolungamento attività(1)

- Estendere a tutta ENEAGRID il software di monitoring Zabbix nella sua specificità per il controllo dei file system distribuiti
- Customizzazione di Zabbix in funzione dei diversi livelli di autorizzazione di accesso in base alle specificità dell' utenza
- Analisi del file system distribuito GPFS su WAN
- Integrazione delle informazioni relative alle prestazioni su rete dei file system distribuiti con i componenti di ENEAGRID

Proposte prolungamento attività(2)

- Monitoring ATTIVO di ENEAGRID
- Dati Zabbix utilizzati come input per ENEAGRID
- Meccanismi di priorità di accesso alle risorse di calcolo



- GRAZIE A TUTTI PER L'ATTENZIONE!!!