Monitoring di rete per le Grid

Alfredo Pagano, Mario Reale - GARR

Consortium GARR

Indice

- Perche' il monitoring di rete per le Grid
  - La Grid vista dalla rete e viceversa
- Potenzialita' della rete in fibra per le Grid
  - GARR-X
    - Vantaggi per la comunita' Grid
- Ruolo di GARR nella comunita' Grid
- Tools sviluppati in ambito EGEE
  - PerfSONAR-Lite_TSS
  - Grid Monitoring Jobs
- Conclusioni

Scopo del Monitoring di rete per siti grid, grid operations e middleware

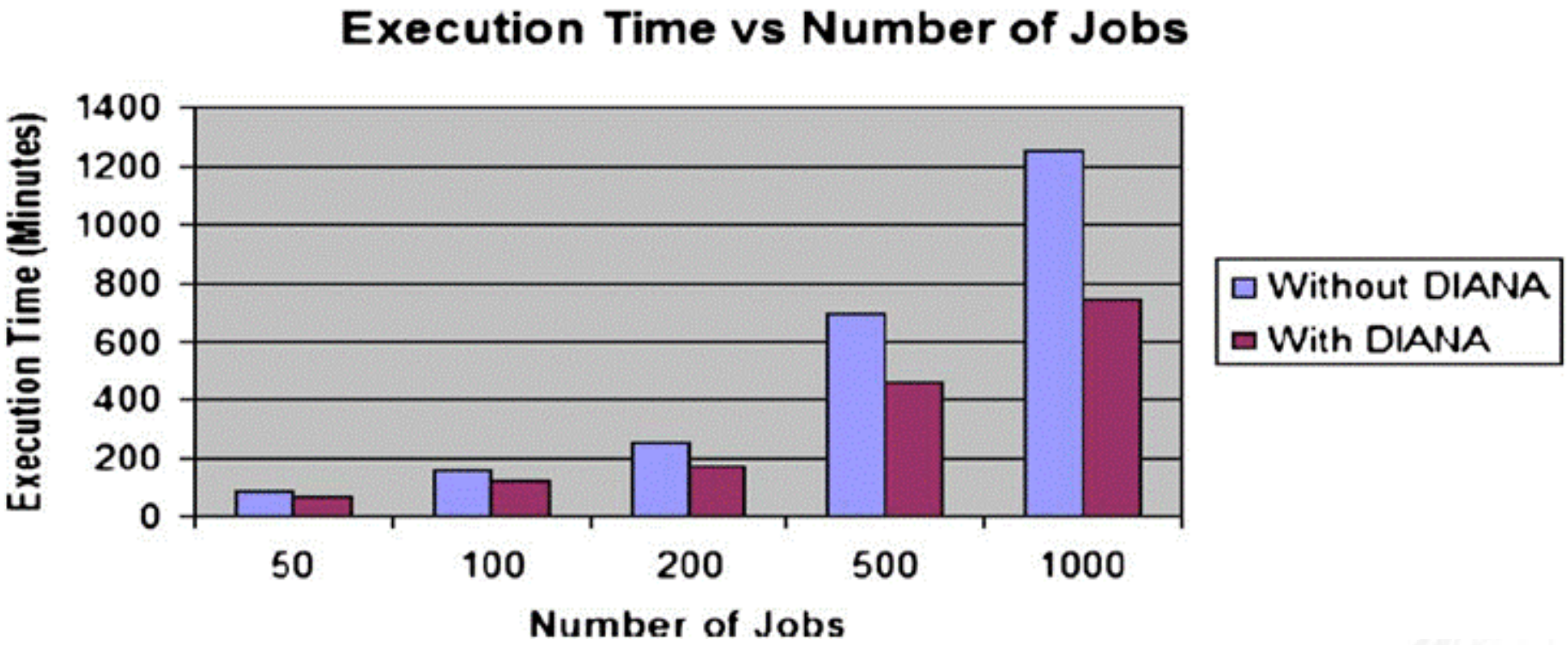
- Aiuta a diagnosticare i problemi di performance **tra siti**
  - Si osserva un data transfer molto lento. Cos'e' che non va ?
    - la rete, il server, il middleware…
  - Scomparsa di una risorsa o di un sito
    - e' caduta la rete o e' la macchina o il servizio ?
  - Le performance della mia applicazione variano con l'ora della giornata
    - c'e' un network bottleneck?
- Aiuta a diagnosticare i problemi **all'interno dei siti**
  - La maggior parte dei problemi di rete, specialmente quelli di deterioramento delle performance, non sono legati al backbone ma al "last mile"
- Permette di pianificare e prendere decisioni sul provisioning:
  - Lo SLA che ho sottoscritto coi miei provider viene rispettato ?
- Middleware performance

La Grid vista dalla rete:

- un **imponente insieme distribuito di utenti** che
  - Desidera spostare/accedere a una quantita' enorme di dati
    - LCG: ~ 12 PB/anno distribuiti sui 11 Tier1 e molteplici (>150) Tier2
    - Dal T0 dovrebbero essere "sparati fuori" circa 500-600 MB/s (~ 4 Gb/s) ai vari T1
  - Deve sottomettere jobs e trasferire dati *dovunque*
  - I cui requirements di rete sono di carattere eterogeneo
    - Middleware:
      - Accessibilita' a servers/porte nei siti Grid
      - Porte, inbound/outbound connectivity
    - Applicazioni: Requirements molto legati al tipo (HEP, Medicina, BioInformatica, Earth Observation)
      - Banda (Throughput effettivo)
      - QoS/SLA (delay RTT, OWD , jitter, path MTU…)
  - Spesso vuole doversi occupare di rete il meno possibile ☺

La Rete vista dalla Grid: ottimizzazione

- Ottimizzazione dell'accesso alle risorse e del loro



- Data Intensive And Network Aware - grid scheduling

Potenzialita' di GARR-X per la comunita' Grid

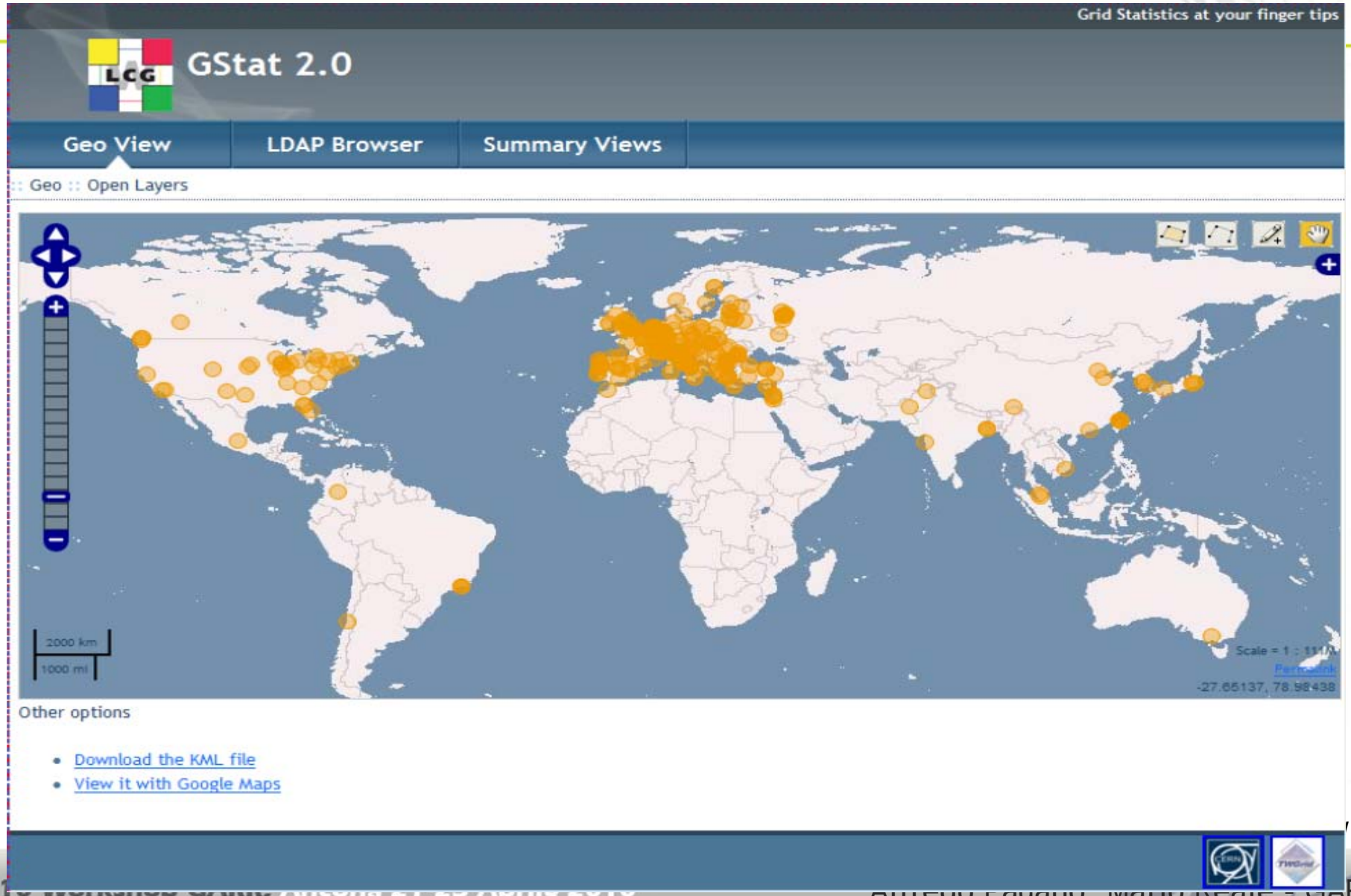
- **Accesso alla rete ad alte prestazioni**
  - per integrare risorse per il calcolo parallelo e distribuito e sistemi di storage distribuiti – per gli utenti collegati in fibra
- **Reti private**
  - per la separazione logica e fisica sia del singolo traffico utente che di comunità di utenti
  - per la creazione di reti tematiche
    - ad esempio, reti di patologia
    - con requirements specifici (per es. criptazione/sicurezza)
- **Circuiti fisici nazionali ed internazionali dedicati**
  - per il collegamento end-to-end di specifici siti che condividono dati e risorse (per es.della stessa VO)

Potenzialita' di GARR-X per la comunita' Grid

- **Riduzione del provisioning time per il set up di circuiti end-to-end**
  - Gli utenti avranno un'unico point-of-contact
    - In casi fortunati in cui l'hardware fosse gia' online un nuovo ciruito potrebbe essere attivato in poche ore
    - Possibilita' di erogazione on-demand

GARR nella comunita' Grid

- GARR nel suo ruolo di NREN (National Research and education Network) contribuisce per il supporto rete (e non solo)
  - Progetti generali e tematici/specifici (e-Health)
- In particolare GARR partecipa a:
  - EGEE (SA2)
  - EGI (O-E-12: task di coordinamento Supporto Rete)
  - IGI (Supporto Rete)
  - DECIDE (Management e Supporto Rete)
  - EUMEDGRID-Support (Technical Coordinator)
  - EUIndiaGrid2 (Supporto Rete)
  - diagnoSIS (VPH proposal) (Management e Supporto Rete)


Grid Statistics at your finger tips
GStat 2.0
Geo View
LDAP Browser
Summary Views
:: Geo :: Open Layers
2000 km
1000 mi
Other options
Download the KML file
View it with Google Maps

EGI : European Grid Initiative

- EGI e' l'iniziativa europea permanente di Grid
  - High Throughput Computing
- L'idea e' svincolare l'infrastruttura Europea di Griglia dalla dipendenza dall'approvazione di progetti EU biennali
  - Ufficializando un impegno europeo permanente nel campo delle Grid
  - Fornendo una e-Infrastructure permanente per l'e-Science in Europa
  - Il modello e' federale:
    - ogni nazione/federazione ha una sua NGI (National Grid Initiative)
    - Esiste poi un body di coordinamento Europeo: EGI.eu
- EGI.eu E' stata fondata ufficialmente lunedi' 8 febbraio 2010 (~2 mesi fa) ad Amsterdam
  - http://www.egi.eu/cms/about/news/
  - "La DANTE delle GRID"

Struttura della Grid in Europa

L'Infrastruttura permanente di Grid in Europa al servizio della ricerca
ERA
European Research Area

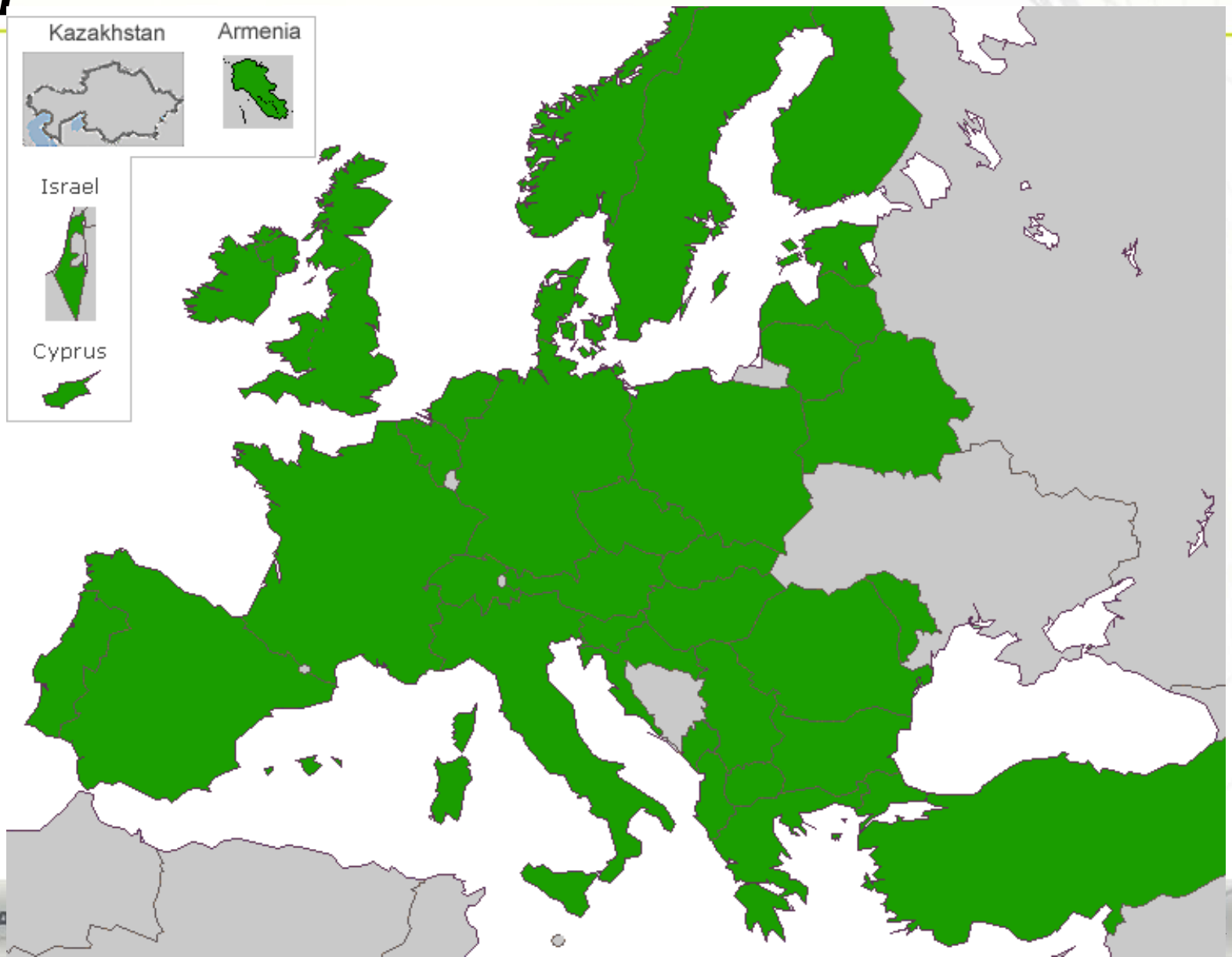
Le Iniziative Grid Nazionali nei vari Paesi Europei
National Grid Initiatives
In Italia: ***IGI***

EGI = EGI.eu + NGIs

L' organizzazione centrale (core) di coordinamento delle Grid nazionali per dare corpo ed anima ad EGI basato ad Amsterdam

Il progetto che da corpo a questo modello per EGI, definendo i compiti di EGI.eu e quelli delle NGI si chiama EGI-Inspire

Paesi che hanno firmato il MoU di EGI



GARR e EGI

- GARR (come partner di IGI) e' responsabile del coordinamento di supporto rete per l'infrastruttura EGI
  (Task Internazionale di EGI O-E-12)
  - L' attivita' e' in fase di pianificazione
    - Manpower limitato ➔ Poche funzionalita' condivise ed utili
      - Troubleshooting
      - Monitoring e2e per un sottoinsieme di siti rilevanti
- L'idea e' di creare un coordinamento permanente per EGI tra le NREN e DANTE a livello GEANT3 (analogamente a LHCOPN, e-VLBI…)

EGEE SA2 *(Network Support Activity)*

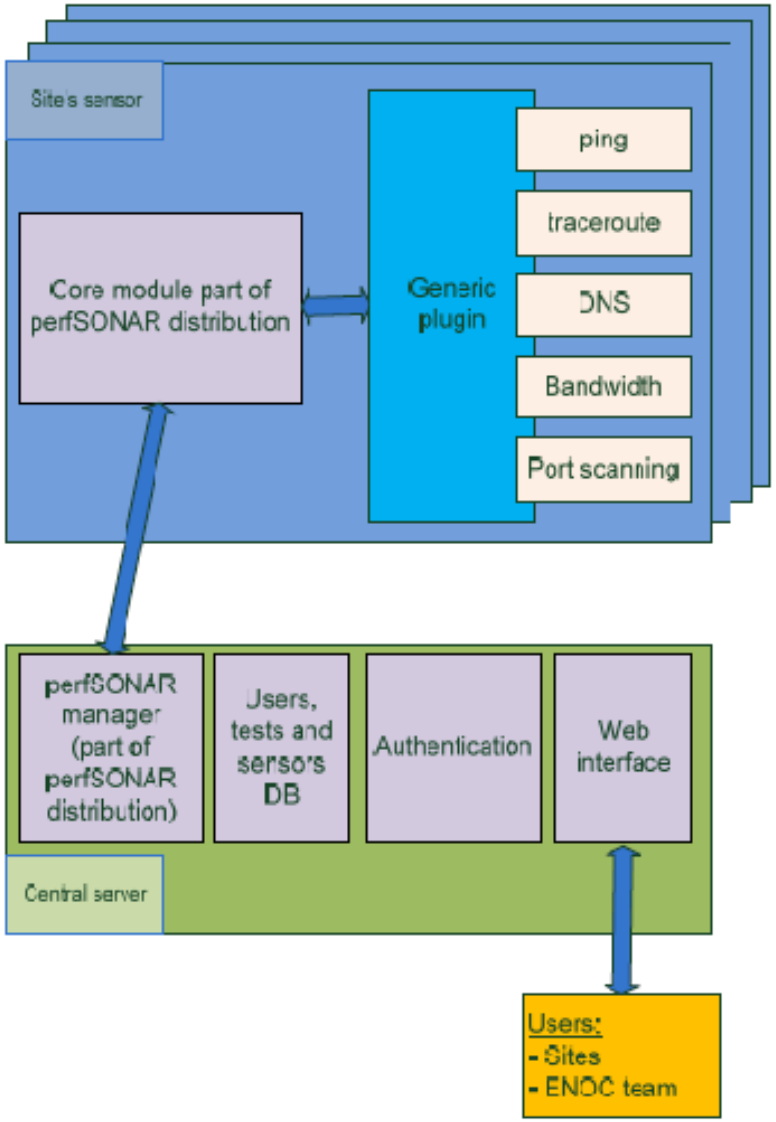
- gLite IPv6 compliance
  - Esame dettagliato del codice gLite
    - Code checker statico e dinamico
    - Bug Submiting
    - Testing di componenti middleware in IPv6
    - Tutorials su IPv6 programming
- Monitoring di rete
  - PerfSONAR-Lite_TSS
  - Job based monitoring

PerfSONAR-Lite_TSS

- Idea: fornire un tool **leggero** e **on demand** di troubleshooting di rete per i siti Grid
  - Basato sul protocollo web service PerfSONAR
- Enfasi e' su poche analisi – on demand - che impattano sulla funzionalita' del middleware:
  - Ping
  - Traceroue
  - Reverse DNS lookup
  - Port Scan
  - BWCTL (IPERF) bandwidth test
- Sistema sviluppato da DFN/RRZE e in validazione a GARR, RENATER, NDGF

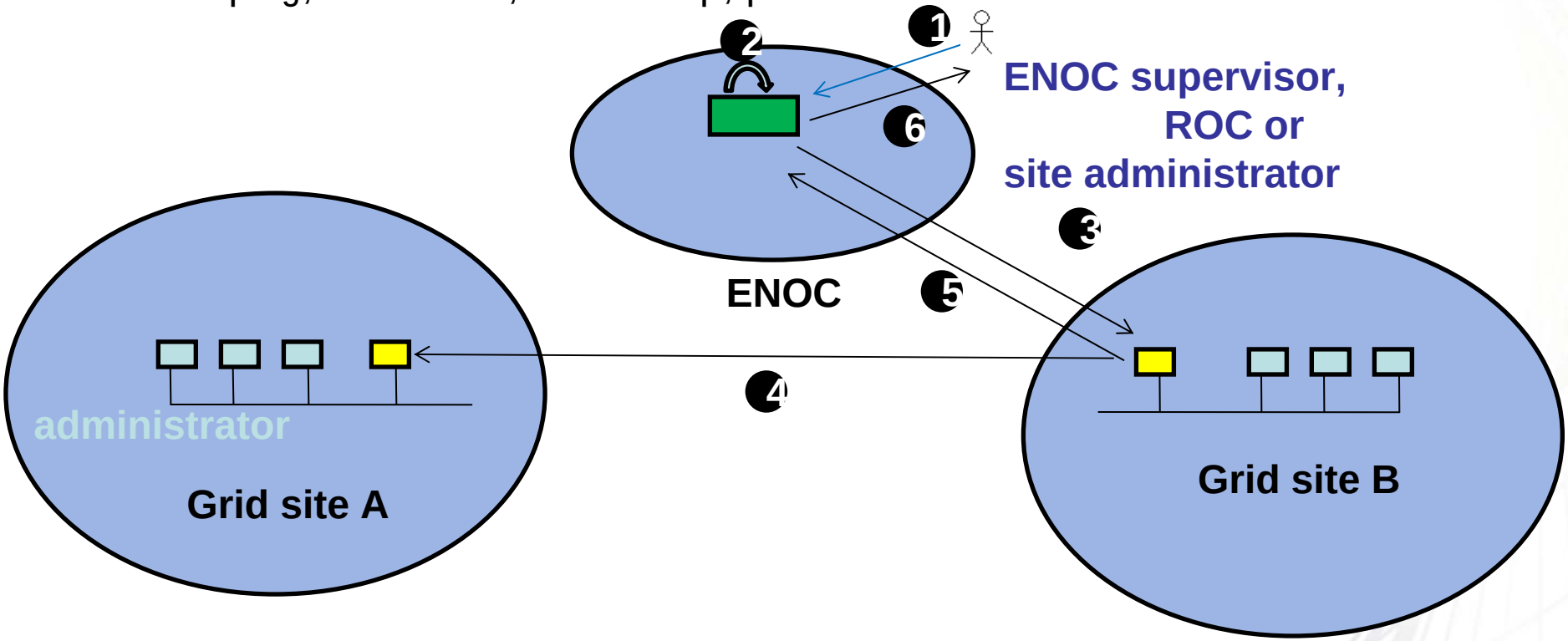
Architettura PerfSONAR-Lite_TSS

- web server centrale per accesso alle misure
- un client light-weight client in ogni sito
- funzioni di base fornite da plugin compatibili con pS
- Gli utenti sono:
  - Siti
  - Coordination team

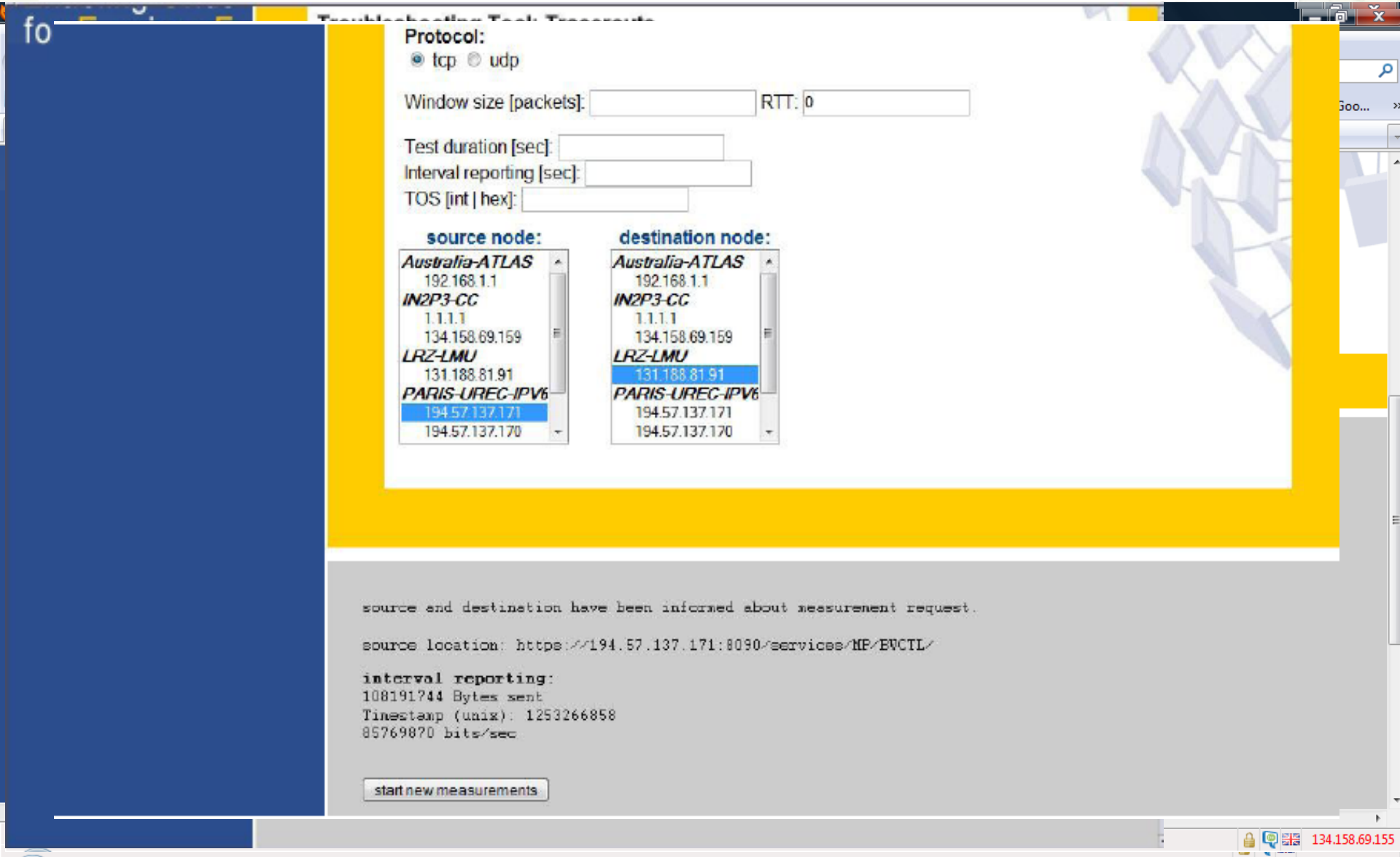


Architettura di PerfSONAR-Lite TSS

- Network troubleshooting tool
  - Esegue tests on demand da un sito Grid in maniera gestita dal team centrale:
    - ping, traceroute, DNS lookup, port scan e bandwith



BWctl Test

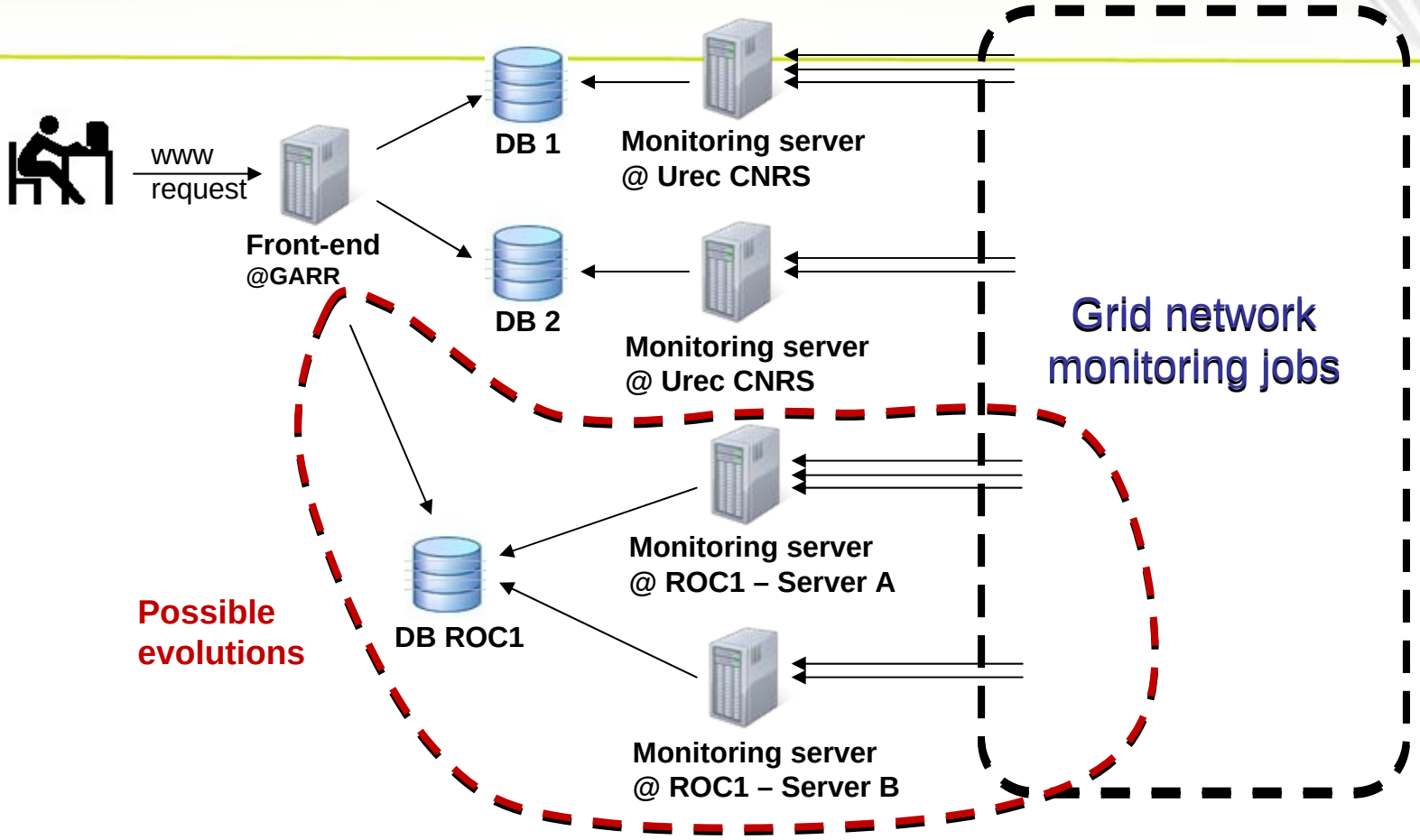

Protocol:
tcp
udp
Window size [packets]:
RTT: 0
Test duration [sec]:
Interval reporting [sec]:
TOS [int | hex]:
source node:
destination node:
Australia-ATLAS
192.168.1.1
IN2P3-CC
1.1.1.1
134.158.69.159
LRZ-LMU
131.188.81.91
PARIS-UREC-IPV6
194.57.137.171
194.57.137.170
source end destination have been informed about measurement request.
source location: https://194.57.137.171:8090/services/MP/BWCTL/
interval reporting:
108191744 Bytes sent
Timestamp (unix): 1253266858
85769870 bits/sec
start new measurements
134.158.69.155

Job based monitoring

(EGEE SA2 in partnership UREC/CNRS)

- **Paradigma**: controllare la Grid tramite la Grid stessa
  - Senza il bisogno di installazioni presso i siti Grid
  - Utilizzando l'infrastruttura di AuthN/AuthZ della Grid stessa
- **Come implementarlo**: attraverso Jobs di Grid
  - Pilot jobs in esecuzione presso i siti effettuano misure di monitoring
  - un server centrale le pubblica via Web

Funzionamento del sistema

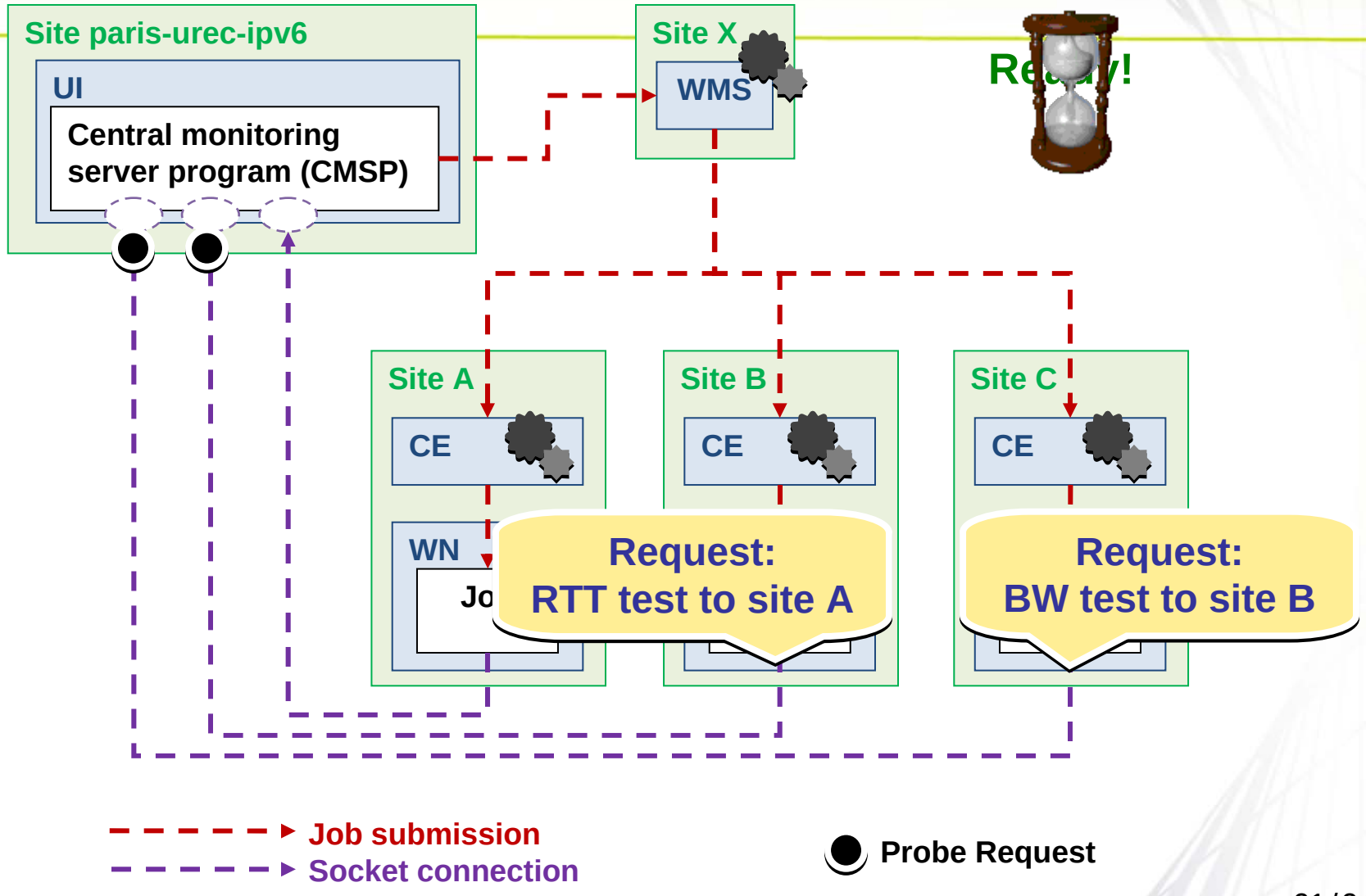


Frontend: Apache Tomcat, Ajax, Google Web Toolkit (GWT)

Backend: PostgreSQL

Linguaggi implementativi: Python, bash script

Initialization of grid jobs



Metriche attuali

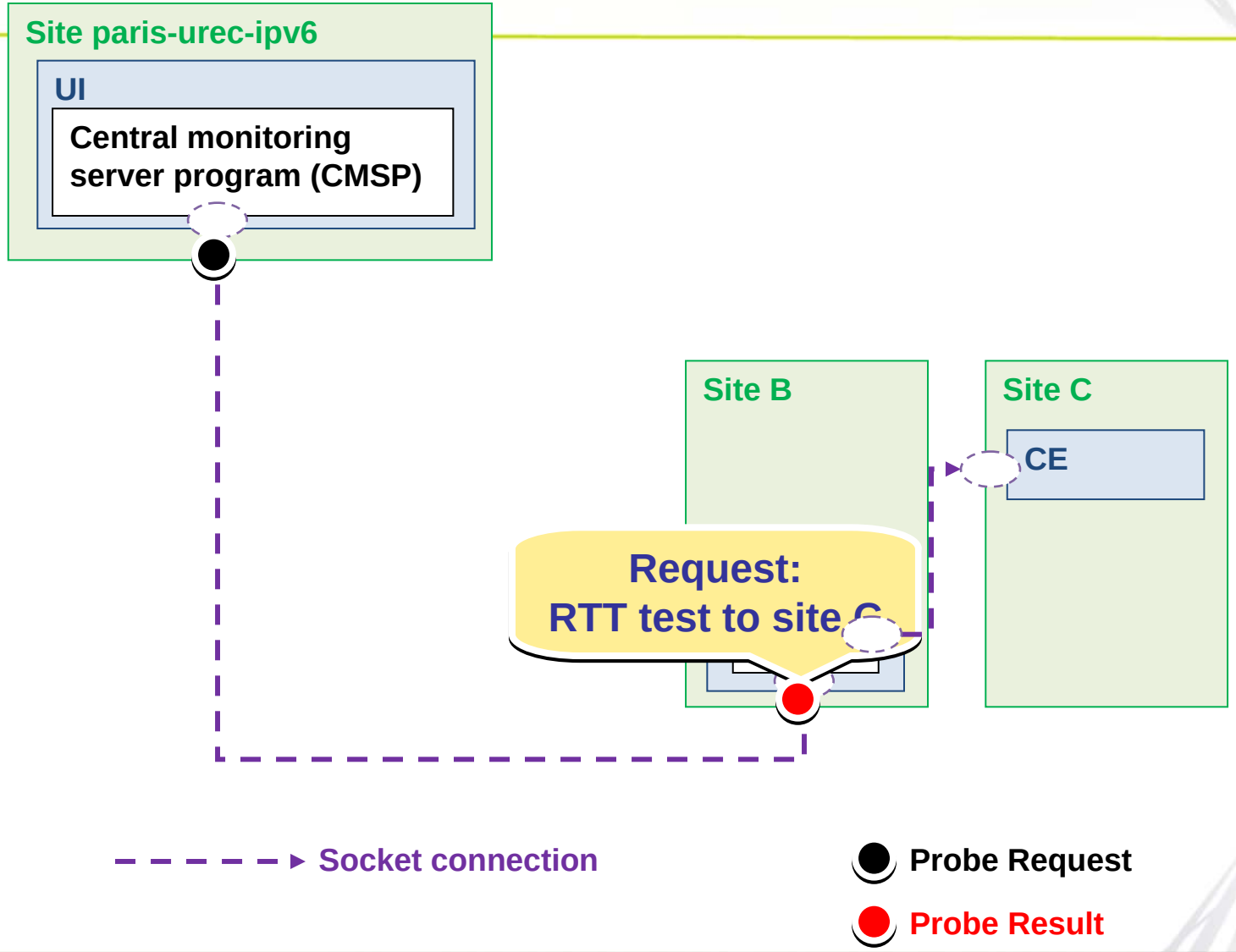
- Latency test
  - TCP RTT
  - Every 10 minutes
- Hop count
  - Iterative connect() test
  - Every 10 minutes
- MTU size
  - Socket (IP_MTU socket option)
  - Every 10 minutes
- Achievable Bandwidth
  - TCP throughput transfer via GridFTP transfer between 2 Storage Elements
  - Every 8h

Per limitare il numero di connessioni aperte questi 3 tests sono eseguiti simultaneamente

Osservazioni

- L'architettura adottata associa ad 1 job molte misure (1 job::molte probes)
  - tiene conto del ritardo legato allo start job
  - minimizza il rischio di job abort in fase di start
- La connessione TCP connection e' iniziata dal job
  - Non necessita porte aperte sul WN → piu' sicuro
- C'e' comunque un meccanismo di autenticazione tra il job ed il server
- Alta scalabilita'
- La durata del job e' non puo' eccedere il valore del parametro *GlueCEPolicyMaxWallClockTime*
  - Per questo motivo ci sono 2 jobs in esecuzione in ogni sito
    - Uno principale in esecuzione
    - Uno di supporto in attesa
      - Che entra in esecuzione qualora termini il principale

RTT, MTU, hop count test

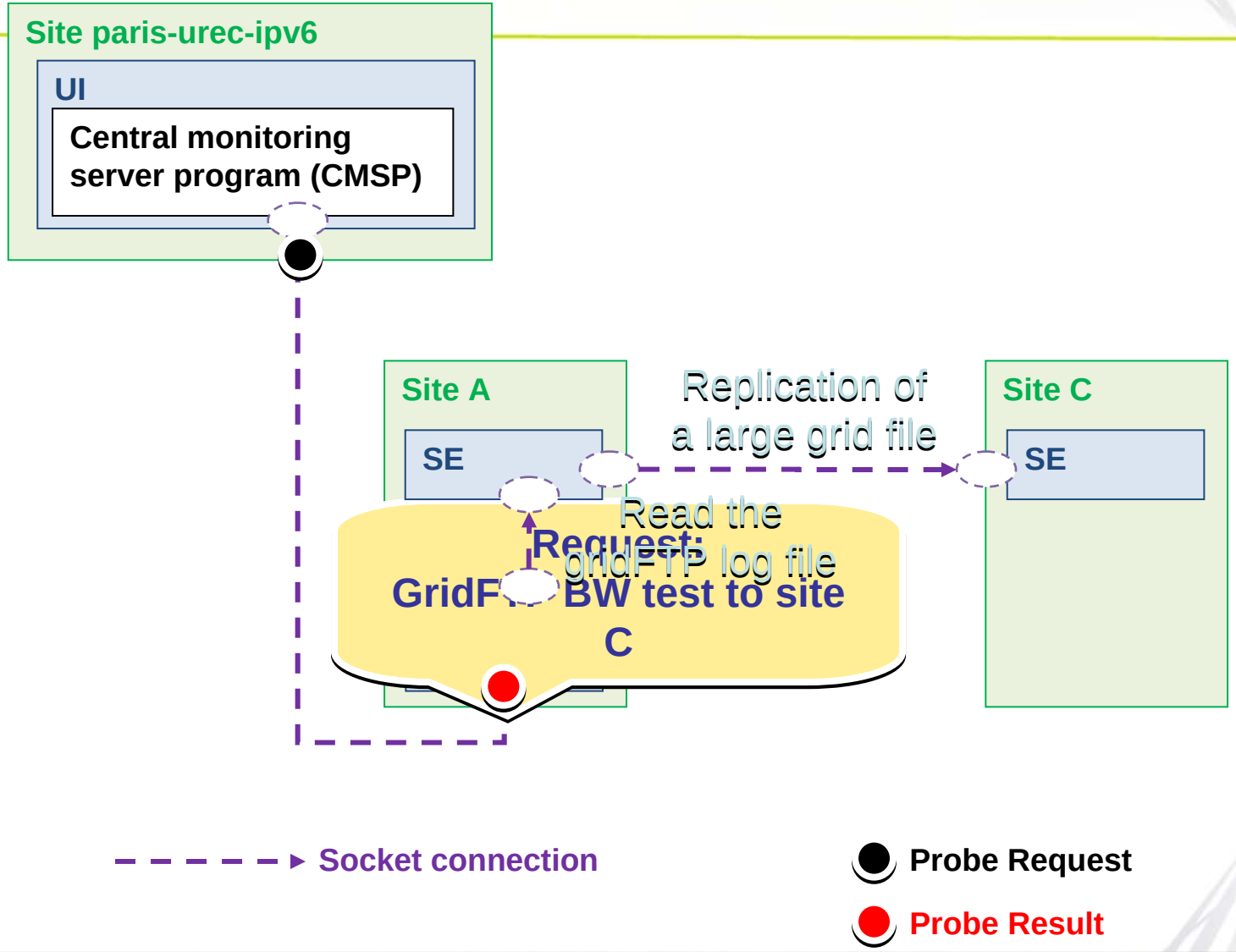


RTT, MTU, hop count test

- L' RTT misurato e' il tempo necessario per una chiamata a TCP 'connect()'
  - Dato che una connect() implica il round-trip di pacchetti specifici:
    - SYN ->
    - SYN-ACK <-
    - ACK ->
  - I risultati sono comparabili a quelli del 'ping'
- L'MTU e' dato dal' opzione socket IP_MTU
- Il numero di hops viene calcolato in maniera iterativa
- Tutte queste misure richiedono:
  - di connettersi su una porta accessibile (*) di una macchina nel sito remoto
  - Di chiudere successivamente la connessione (non c'e' invio dati)

(1): Usiamo la porta del gatekeeper sul CE (2119)

GridFTP BW test



Vincoli Tecnici

- Quando esegue un job, l'utente Grid viene mappato localmente sul Worker Node (WN):
  - Non dispone di **privilegi di root** sul WN
  - ➢ Alcune operazioni low-level non sono possibili (per es. **aprire un socket ICMP in ascolto non e' possibile)**
- Diversita' di ambiente sui Worker Nodes (OS diversi, 32/64 bits…)
  - Es.:far scaricare ed eseguire al job un tool esterno puo' diventare complesso (anche se scritto in un linguaggio portabile/bytecode)
- Il sistema deve convivere con gli overhead legati alla Grid (latenza nello start di un job…)

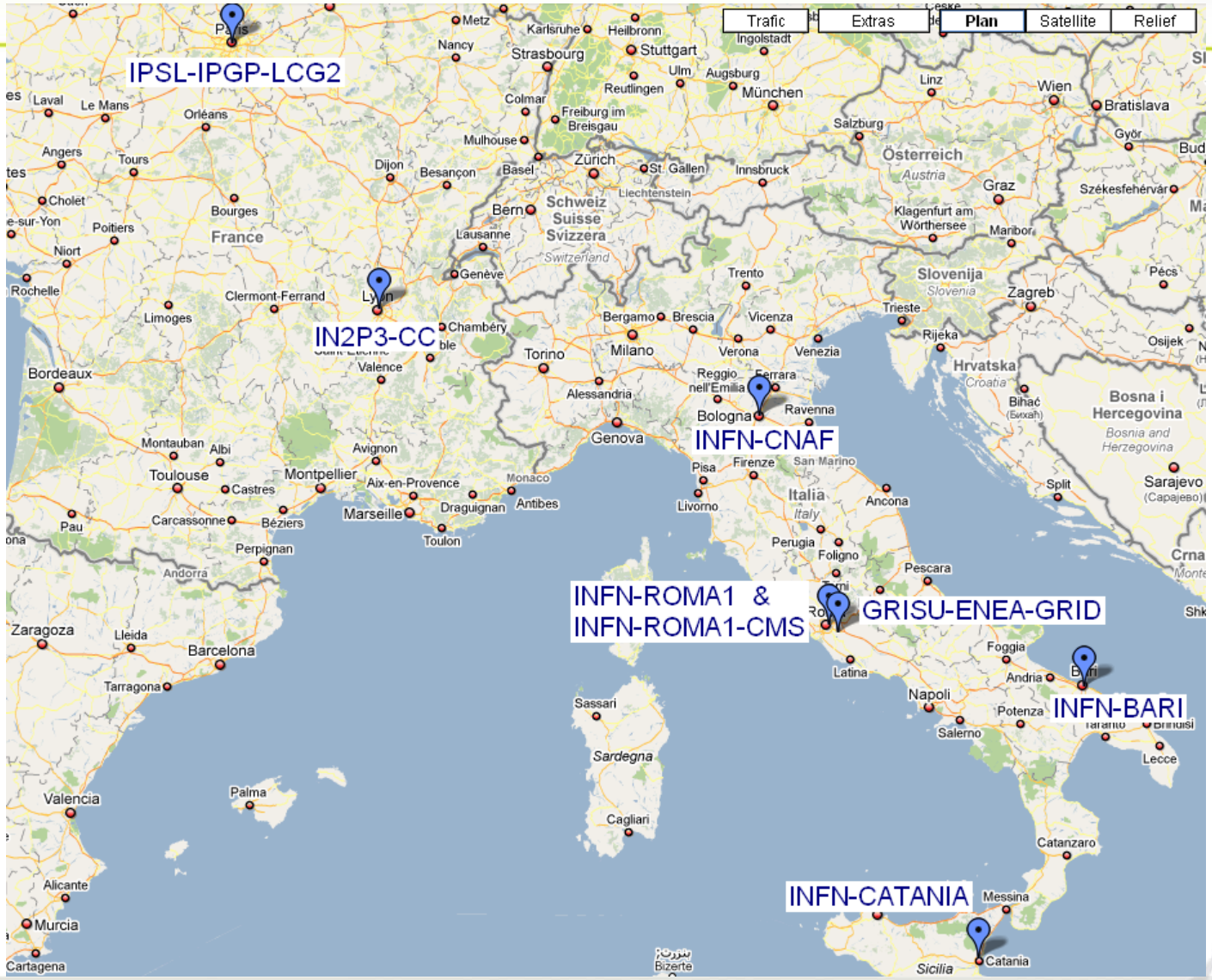
Choice of network paths, scheduling

- Monitorare tutti i possibili path e2e e' troppo oneroso:

  N x (N-1) con N ~ 300 siti

- Dobbiamo necessariamente considerare un sottoinsieme
  - Di rilevanza per le VO, o gli esperimenti, o i path piu' "gettonati"
  - Per maggiori informazioni: https://edms.cern.ch/document/1001777
- Il sistema e' completamente configurabile per quanto riguarda la scelta dei path e2e e lo scheduling delle misure
  - L'amministratore specifica una lista di tests fornendo:
    - Il sito sorgente
    - Il sito destinazione
    - Il tipo di test
    - La frequenza

Sperimentazione: 8 Siti



La GUI


A GRID Network Monitor based on GRID Jobs
Select DB:
Global
Mario Reale (user)
Measurements
Last 30 days
Current Week
Today
Chart
Zoom: 1' 5' 1h 1d 5d 1m 3m 6m 1y Max
12:20 April 16, 2010
rtt 0.031
0.175
0.150
0.125
0.100
0.075
0.050
8 pm Tue Apr 13 8 am 12 pm 4 pm 8 pm Wed Apr 14 8 am 12 pm 4 pm 8 pm Thu Apr 15 8 am 12 pm 4 pm 8 pm Fri Apr 16 8 am
Fri Apr 16 2010 10:00:06 GMT+0200 rtt infn-roma1 in2p3-cc 0.0292458534241
Fri Apr 16 2010 09:50:06 GMT+0200 rtt infn-roma1 in2p3-cc 0.0314381122589
Fri Apr 16 2010 09:40:06 GMT+0200 rtt infn-roma1 in2p3-cc 0.0313789844513
Page 1 of 5
Showing measurements 1 - 100
Jobs gLite ID
Logout

Prossimi Sviluppi

- Sistema di Trigger per mettere in piedi degli allarmi per I network administrators
- Ulteriore miglioramento della GUI per rendere piu' correlabile l'informazione
- Aggiunta di misure on-demand
  - Non solo schedulate

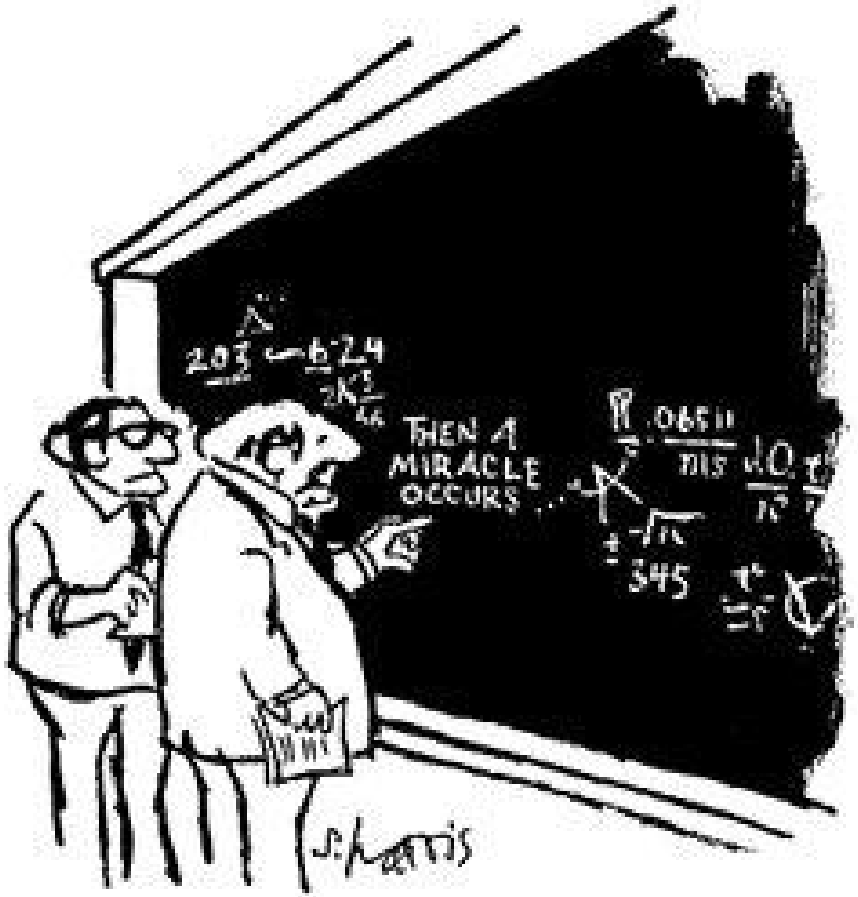
Conclusioni

- Le Grid possono contare su una rete che sostanzialmente funziona bene
- Questo non diminuisce l'importanza del monitoring di rete per le Grid
  - Enfasi principale e' sull' on-demand e sul troubleshooting
  - Un ruolo importante per il monitoring e' legato alle attivita' di tipo PERT per identificare I bottleneck in caso di performance sotto le aspettative
- GARR ha un ruolo importante in molte iniziative Grid
  - Prime fra tutte IGI ed EGI (coordina il supporto rete a livello Europeo)
  - In molti altri progetti (e-Health, beni culturali, regionali (mediterraneo, india..))
- Abbiamo visto 2 esempi di tools in sviluppo che hanno buone potenzialita' d'utilizzo:
  - PerfSONAR-Lite_TSS ed i NetMon Grid Jobs
- Anche IPv6 giochera' un ruolo importante nei prossimi anni
  - Per la compatibilita' del middleware
  - Occorrera' disporre di strumenti di monitoring adeguati

Referenze

- EGEE
  - http://www.eu-egee.eu
- EGI
  - http://www.egi.eu
- IGI
  - http://www.italiangrid.org
- EUMEDGRID-Support
  - http://www.eumedgrid.eu
- EUIndiaGrid2
  - http://www.euindiagrid.eu
- DIANA
  - http://www.springerlink.com/content/x7l1k413g5600g74/
- PerfSONAR-Lite_TSS
  - https://enoc-troubleshooting.gridops.org
- Job based monitoring
  - http://twiki.cern.ch/bin/view/EGEESA2

Grazie!

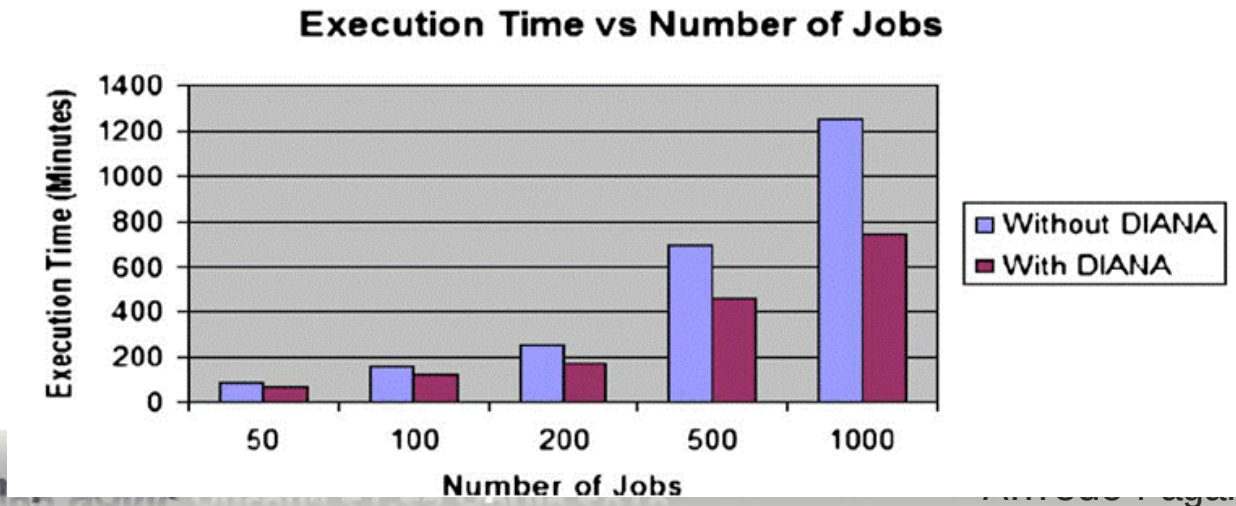


"I think you should be more explicit here in step two."

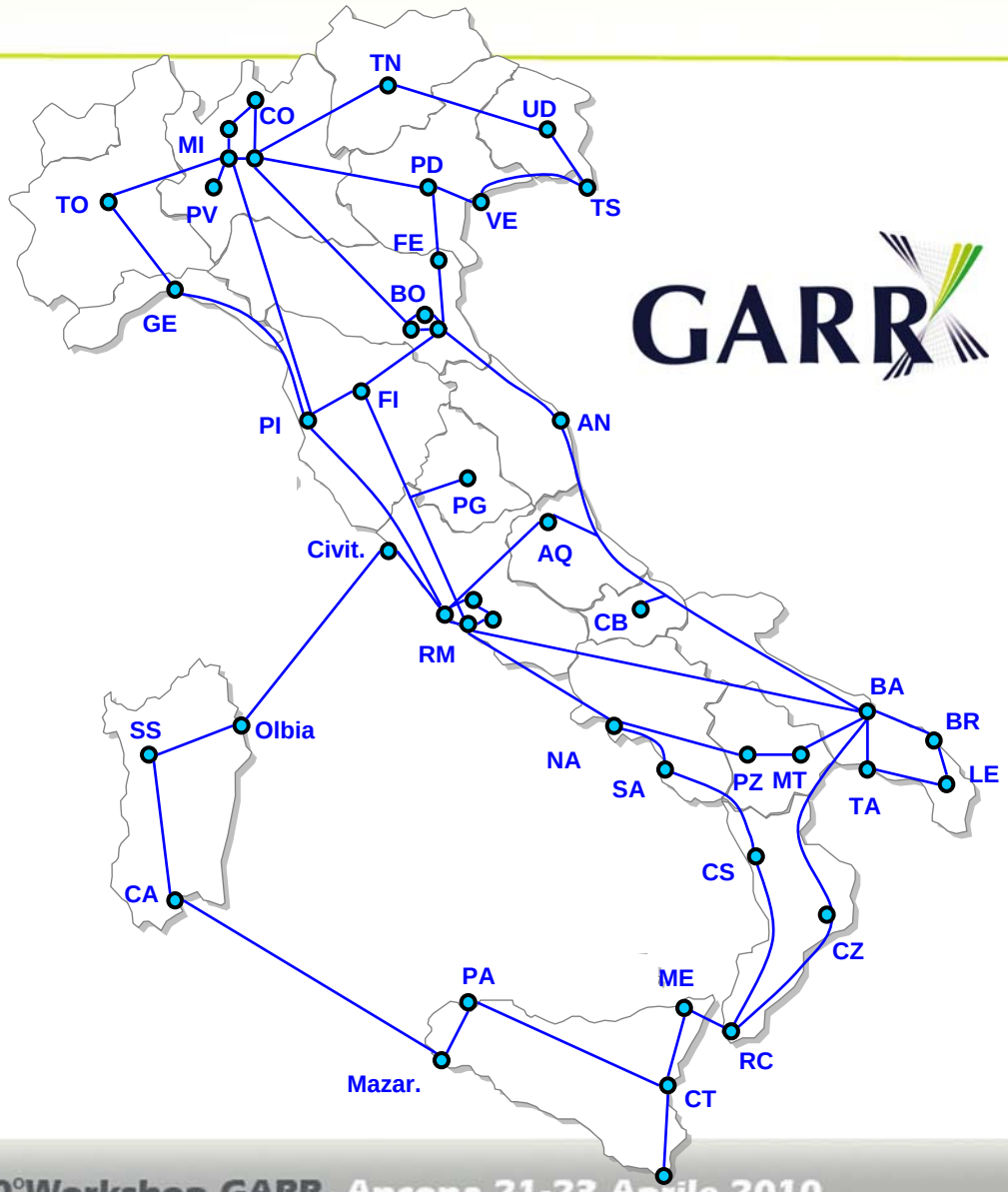
BACKUP

Scopo del Monitoring di rete per i servizi grid ed il middleware

- Ottimizzazione delle performance
  - Si desidera migliorare le prestazioni di un file transfer tra due siti
  - Si desidera conoscere qual'e' il Computing Element piu' "vicino" ai miei dati per poi sottomettergli un job
  - Definire una funzione di costo di accesso ai dati/alle risorse



Infrastruttura in fibra di dorsale di GARR-X

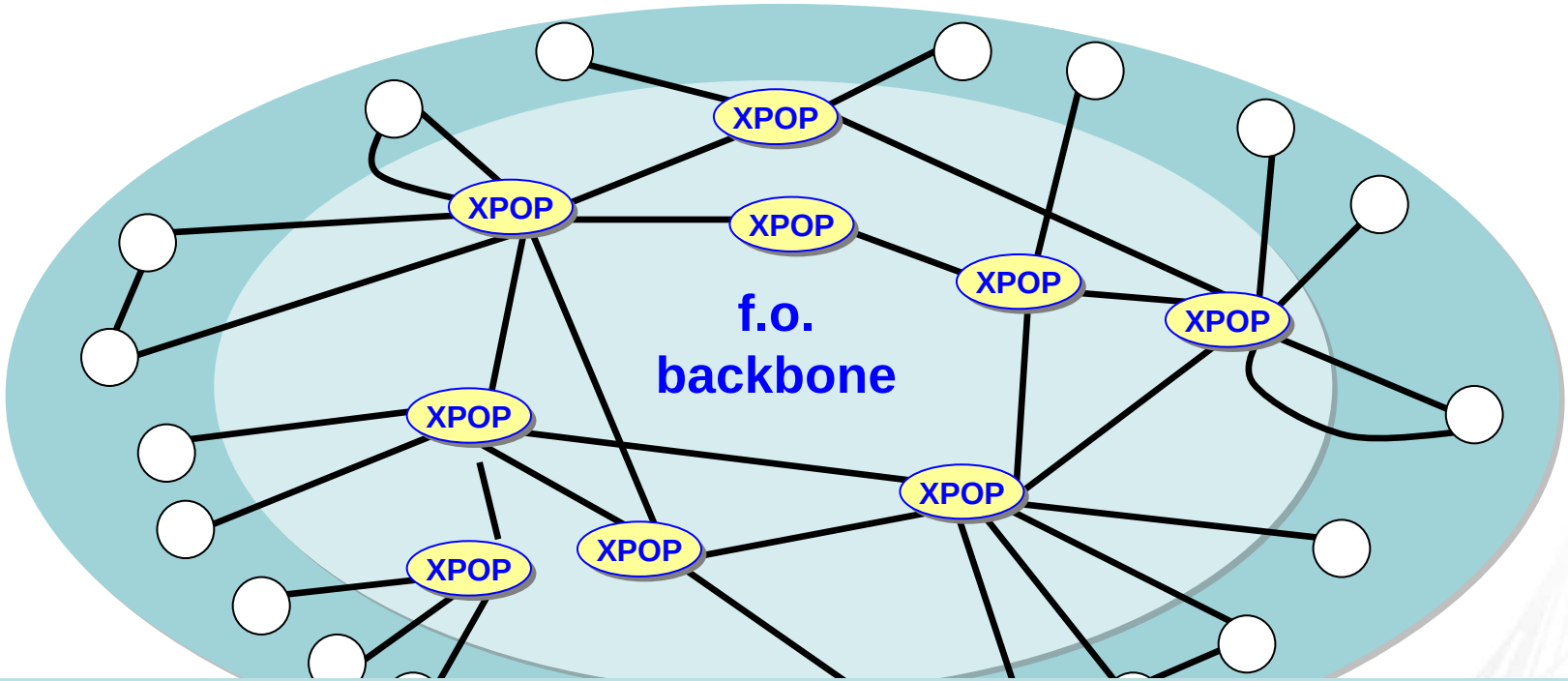


Alcune cifre di sintesi

- 45 PoP GARR principali in 35 citta' (inizialmente), estendibili in 3 anni a 55 citta'.
  - 99% co-locati con sedi utente
- 60 apparati trasmissivi
  - 150 nodi di amplificazione (uno ogni 70 km di fibra)
- 10.500 km fibre di dorsale
- 1.500 km fibre di accesso (non presenti in figura)

GARR-X

La rete ottica di GARR-X e' capillare:
Fibra ottica per il backbone
Fibra ottica per l'accesso



➔Possibilità di erogare i medesimi servizi
a tutti gli utilizzatori della rete indipendentemente
dalla loro posizione geografica