I Requisiti INFN per la nuova rete GARR-X

CCR Garr-x working group report

Workshop su Calcolo e Reti dell'INFN

Rimini, Maggio 7-11 2007

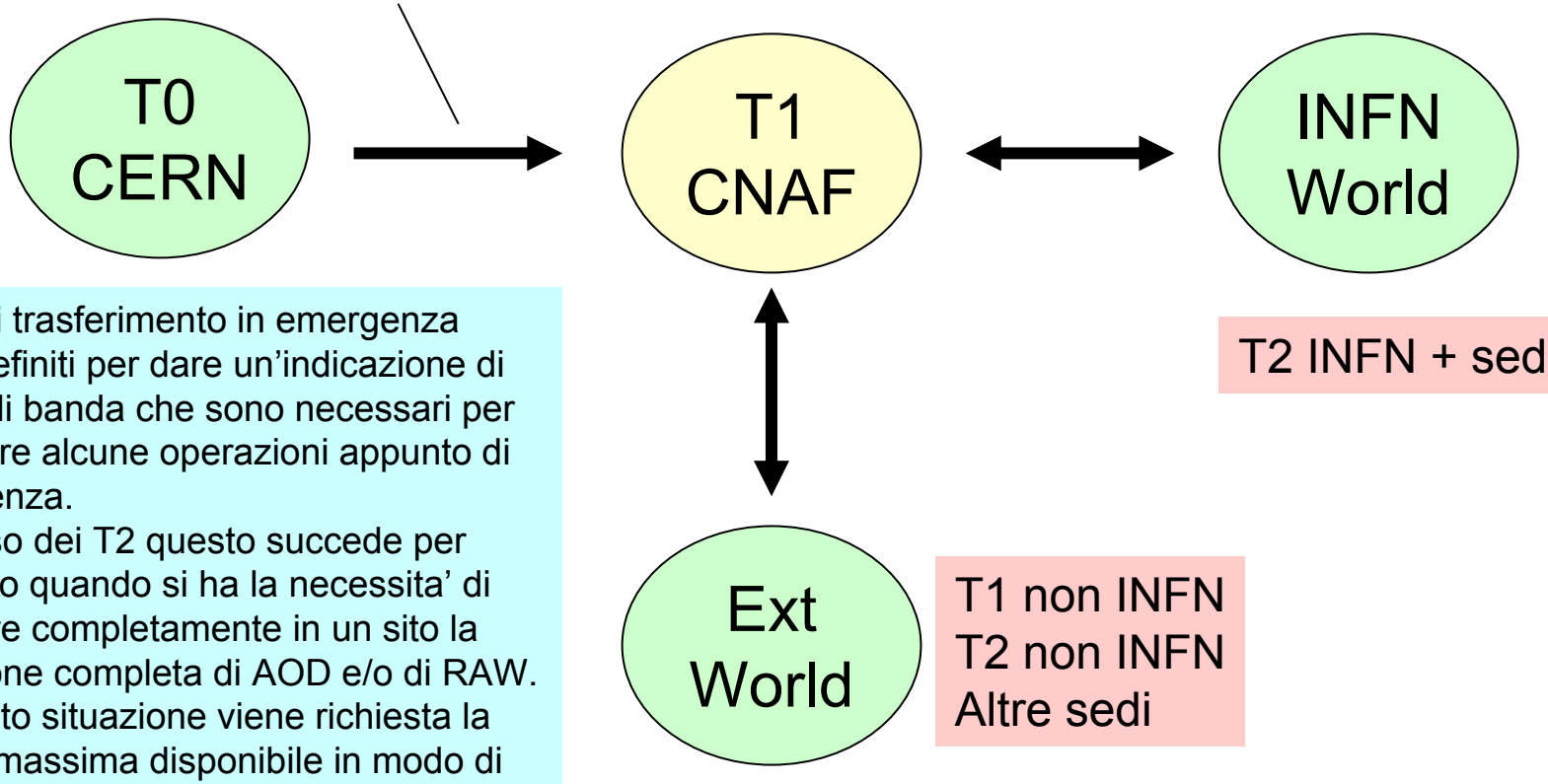
Gianpaolo Carlino (INFN - Sez. di Napoli)
Roberto Gomezel (INFN - Sez. di Trieste)
Gaetano Maron (INFN- LNL)
Alberto Masoni (INFN - Sez. di Cagliari)
Mauro Morandin (INFN - Sez. di Padova)
Davide Salomoni (INFN - CNAF)
Stefano Zani (INFN - CNAF)

Goals del working group INFN su GARR-X

- Analisi dell'uso corrente della rete GARR
- Requisiti di rete dell'INFN fino al 2011
- Definizione dei parametri di rete per applicazioni particolari come applicazioni multimedia (audio/video conferencing, VoIP, streaming, ecc.) o applicazioni che prevedono un controllo remoto real time.
- Definizione di Service Level Agreement per le tratte di rete piu cruciali per GARR-X
- Prime esplorazioni su nuovi modelli di calcolo e nuove applicazioni su rete geografica introdotte dalle proprieta' di GARR-X (Lambda Switching, L2,L3, VPN, etc.).

Esigenze per i prossimi anni

Per ogni tratta sono stati stimati rate di trasferimento e necessita' di accesso :
- medio (average)
- massimo (max)
- in emergenza (emergency)

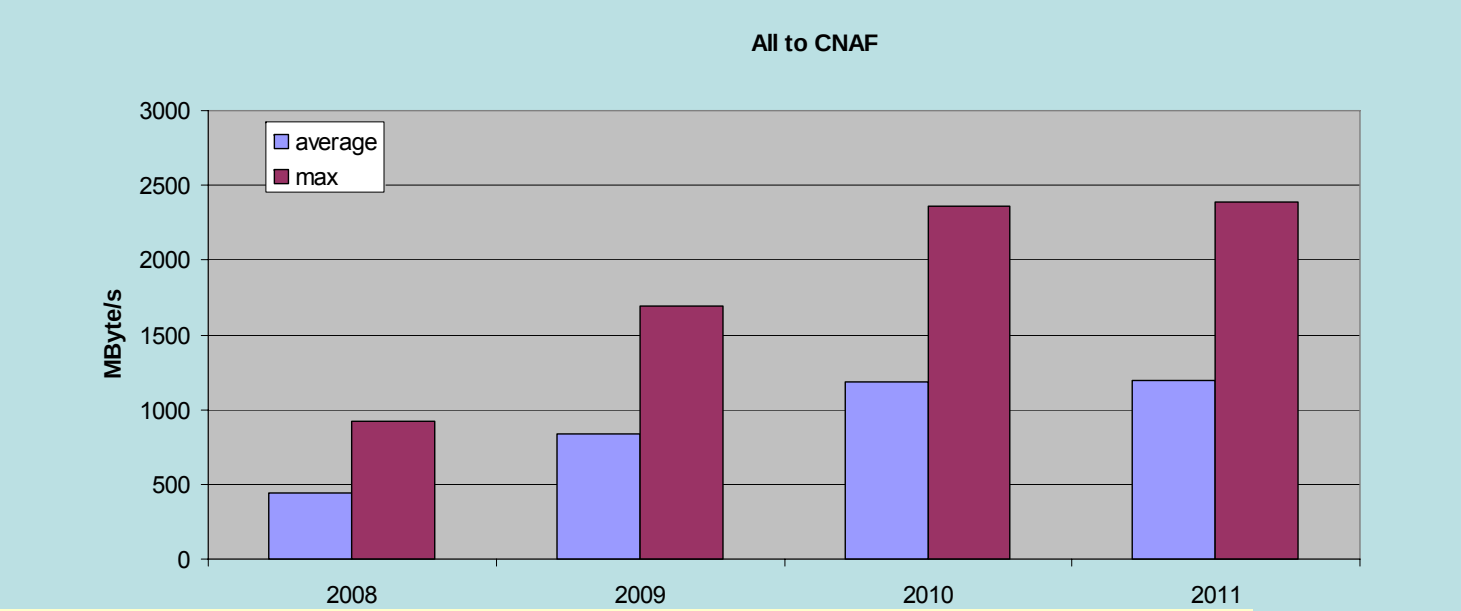


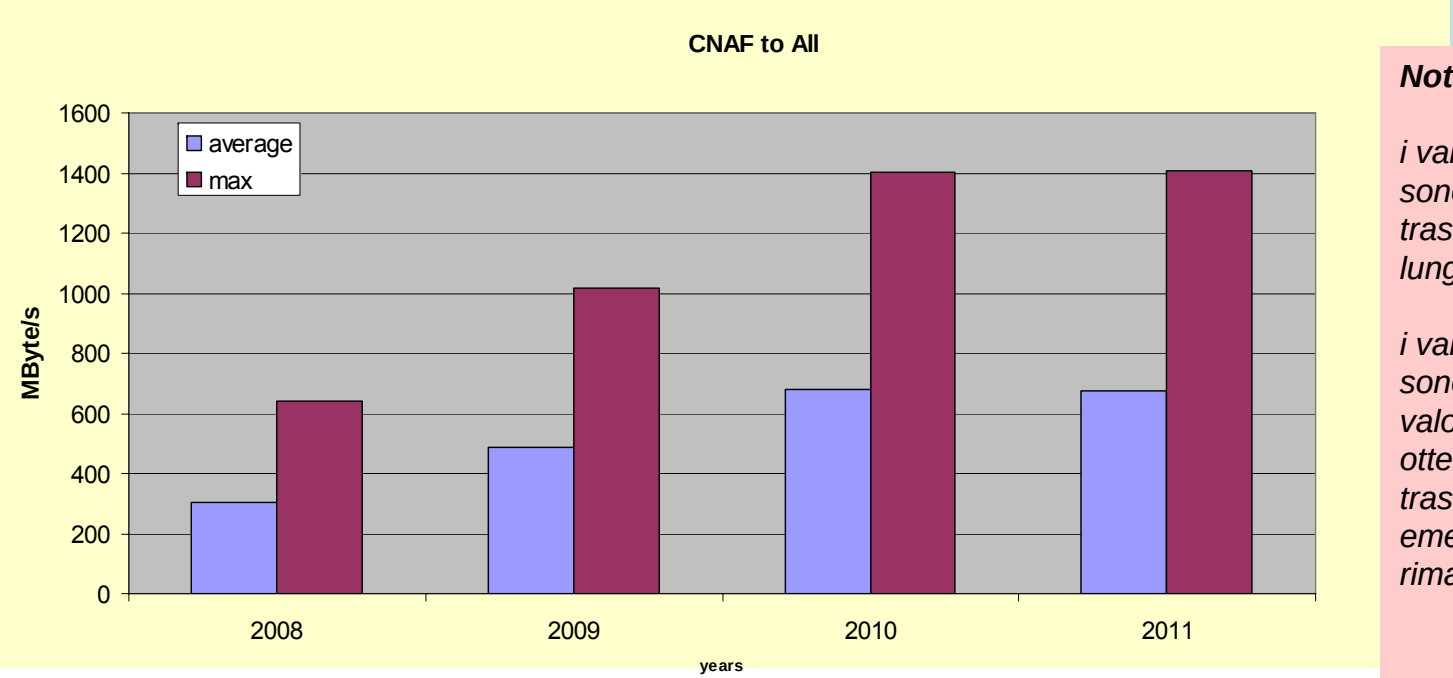
T2 INFN + sedi INFN

T1 non INFN
T2 non INFN
Altre sedi

I rate di trasferimento in emergenza sono definiti per dare un'indicazione di picchi di banda che sono necessari per compiere alcune operazioni appunto di emergenza.
Nel caso dei T2 questo succede per esempio quando si ha la necessita' di ricopiare completamente in un sito la collezione completa di AOD e/o di RAW. In questo situazione viene richiesta la banda massima disponibile in modo di ridurre i tempi di copia.

Accesso globale al CNAF



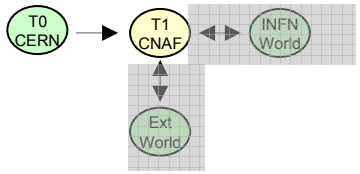


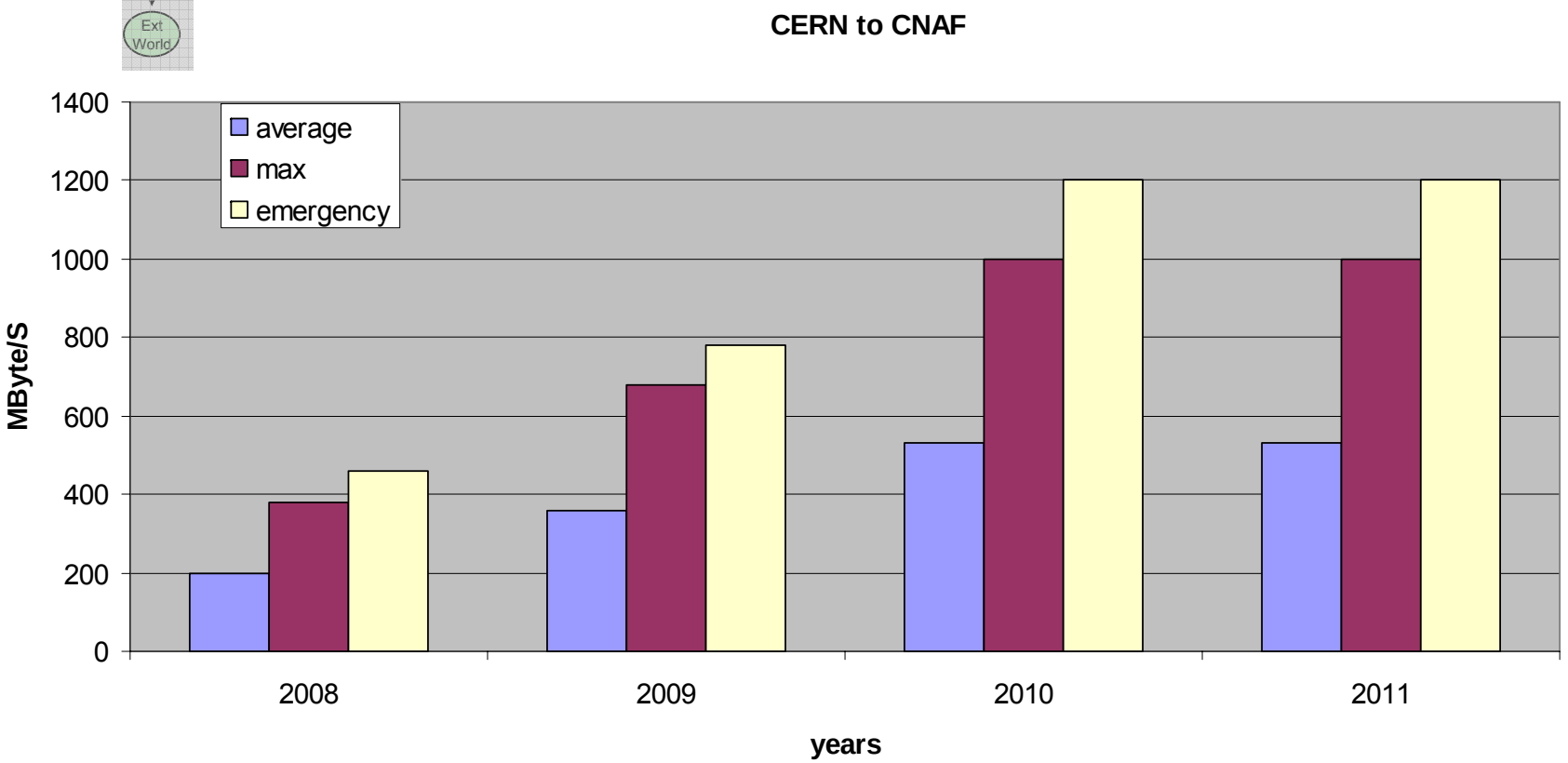
Nota ai plot presentati:

i valori di accesso medio e max sono ottenuti come somma dei trasferimenti pesati con la lunghezza del periodo di attitvità

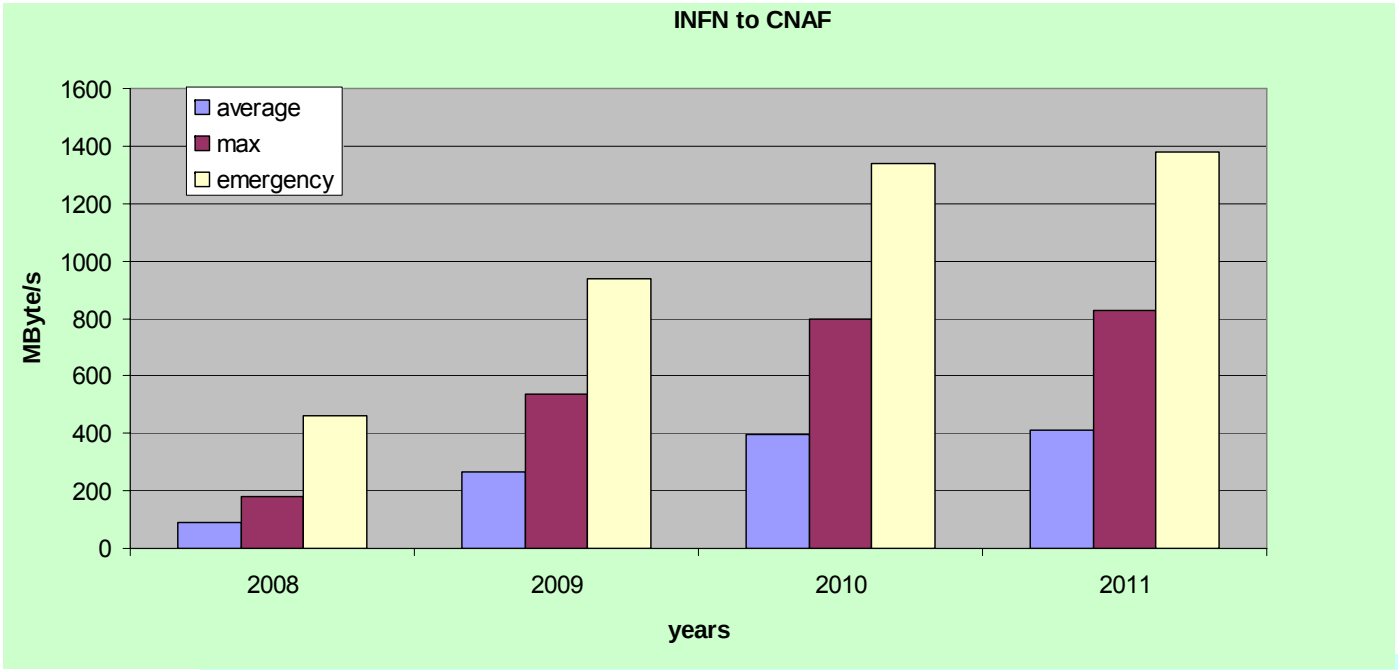
i valori di accesso in emergenza sono ricavati determinando il valore massimo dell'insieme ottenuto sommando, per ciascun trasferimento, al relativo rate in emergenza i valori max dei rimanenti trasferimenti

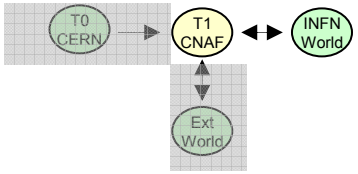
Accesso al CNAF dal CERN (T0)

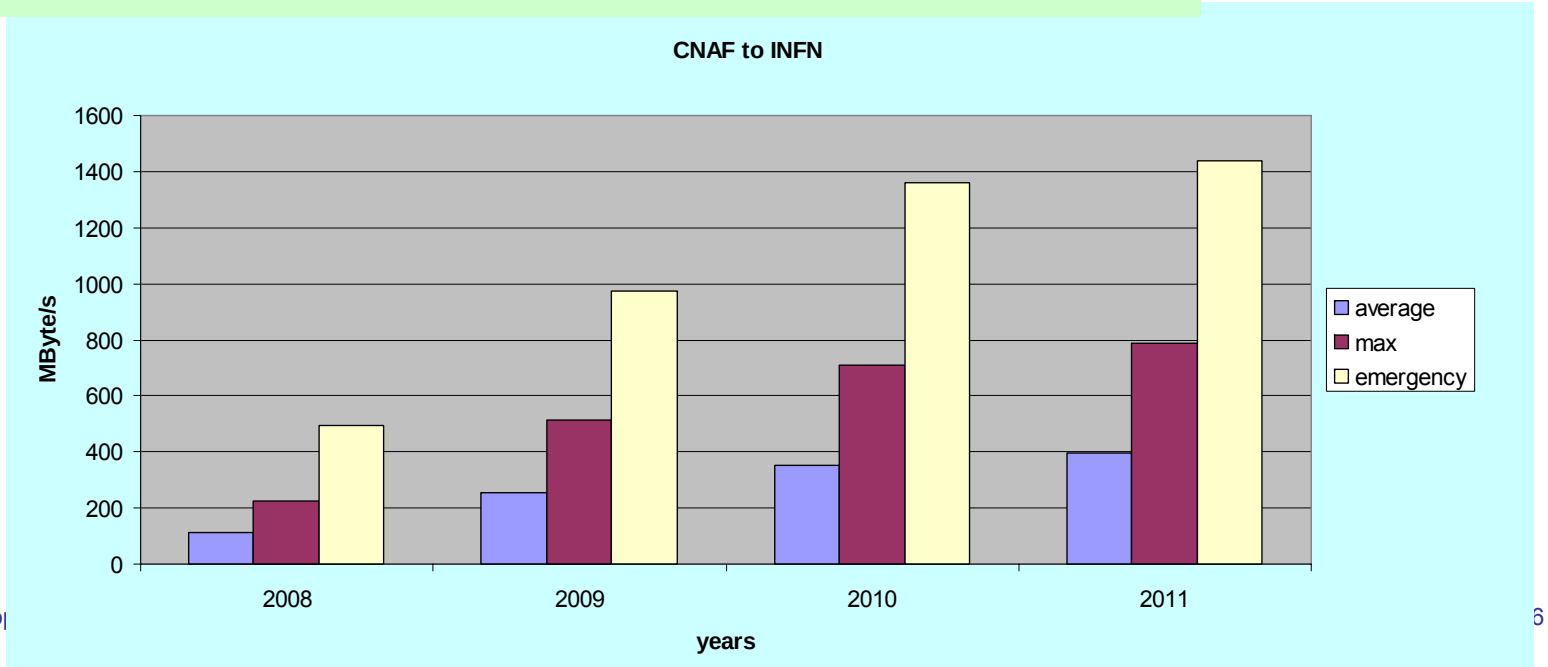




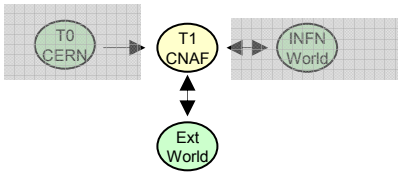
Accesso a/da INFN World dal/al CNAF

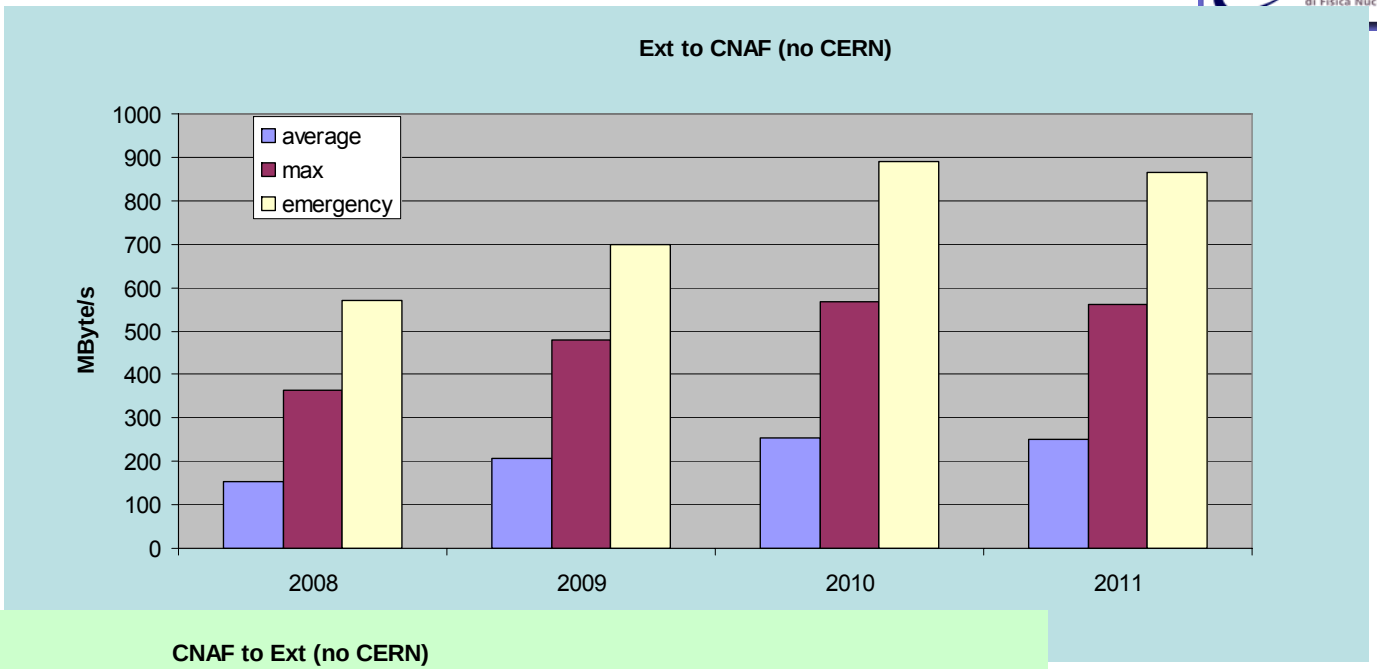


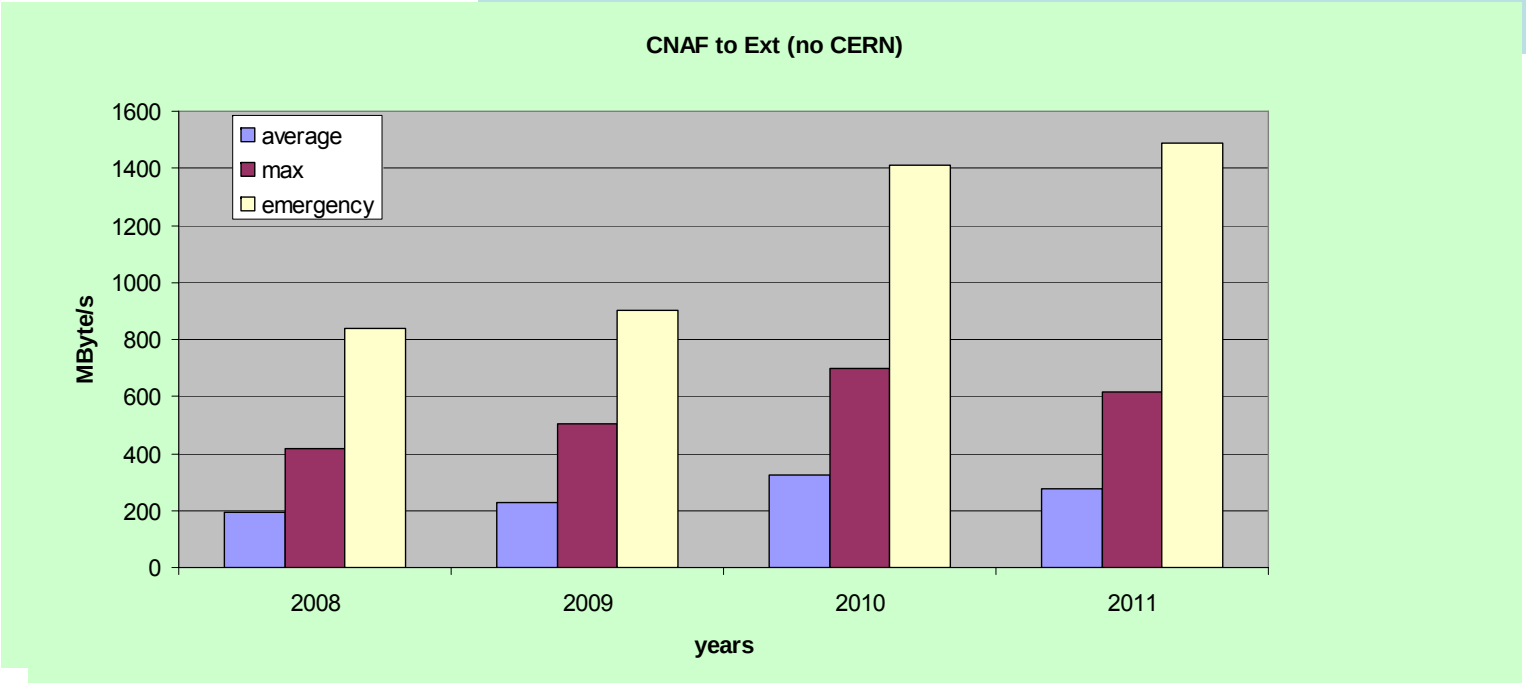




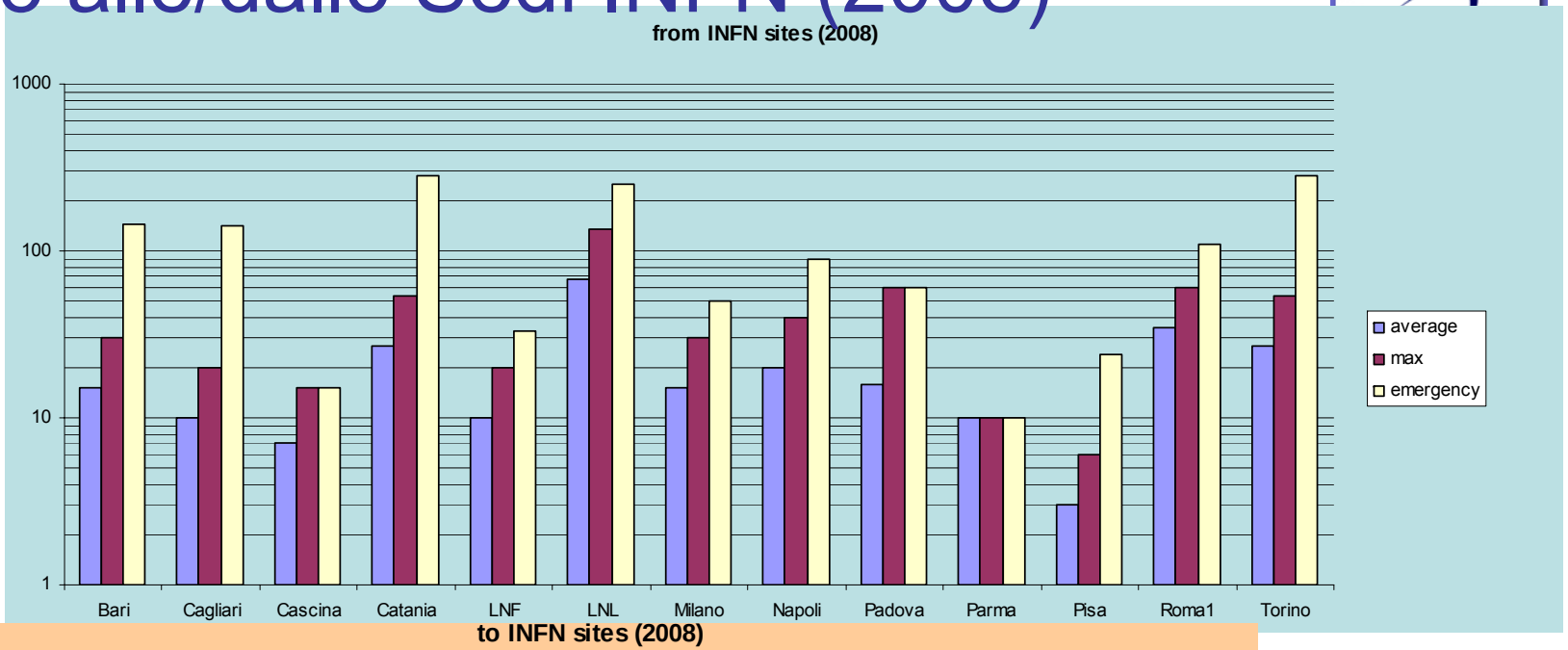
Accesso da/a External World al/dal CNAF

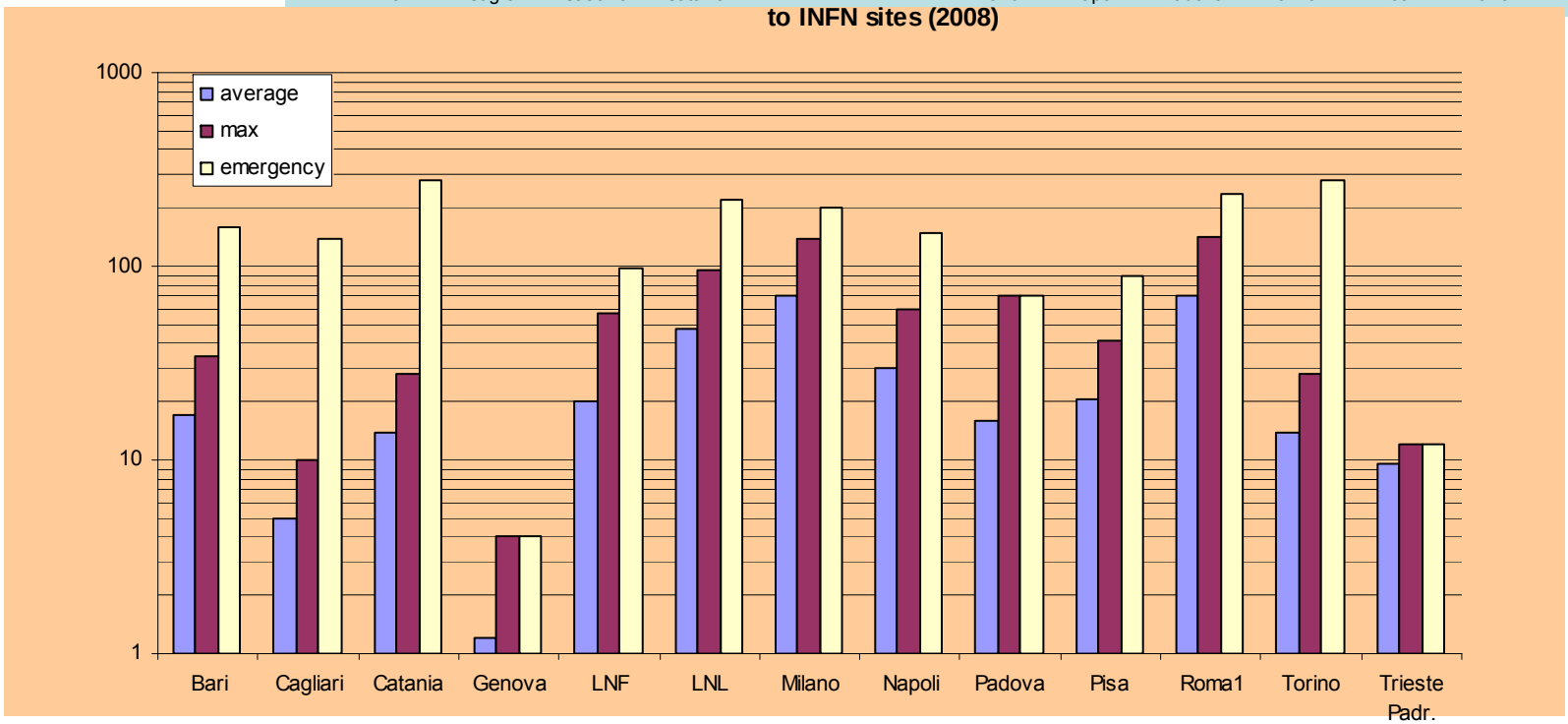




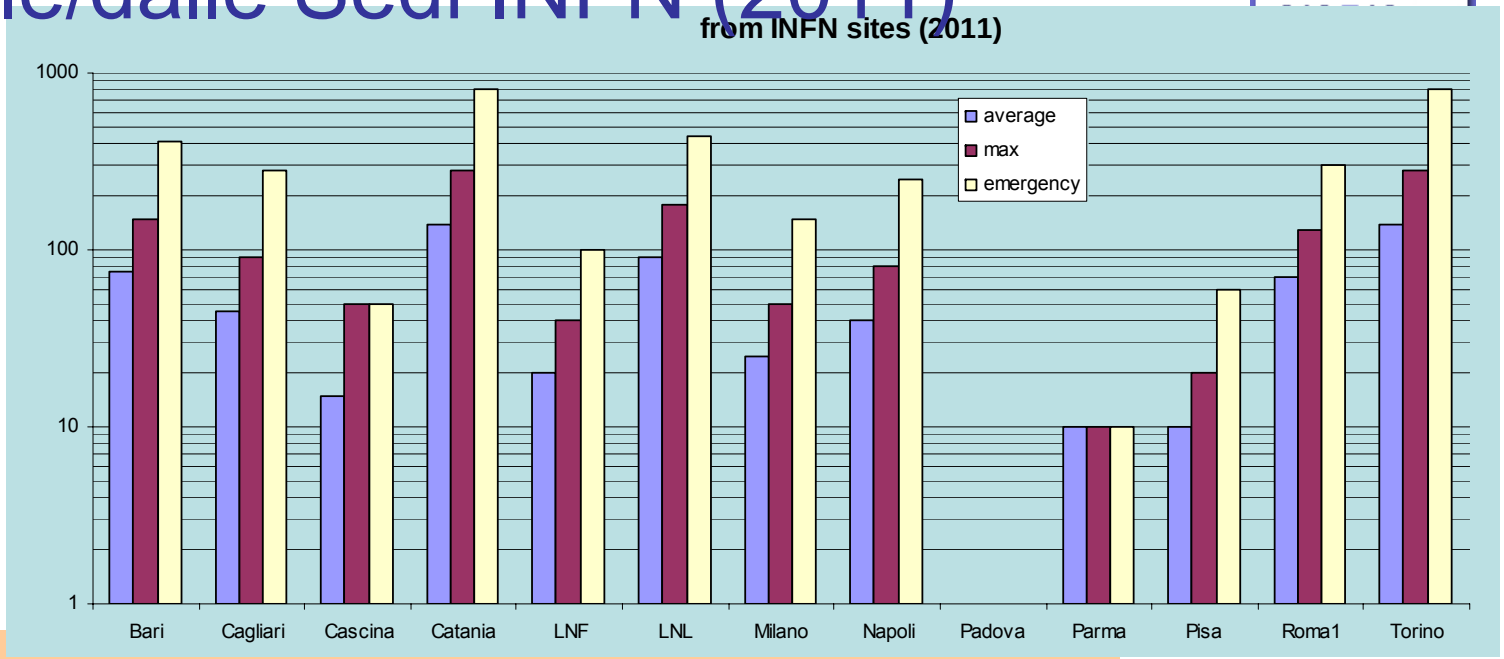


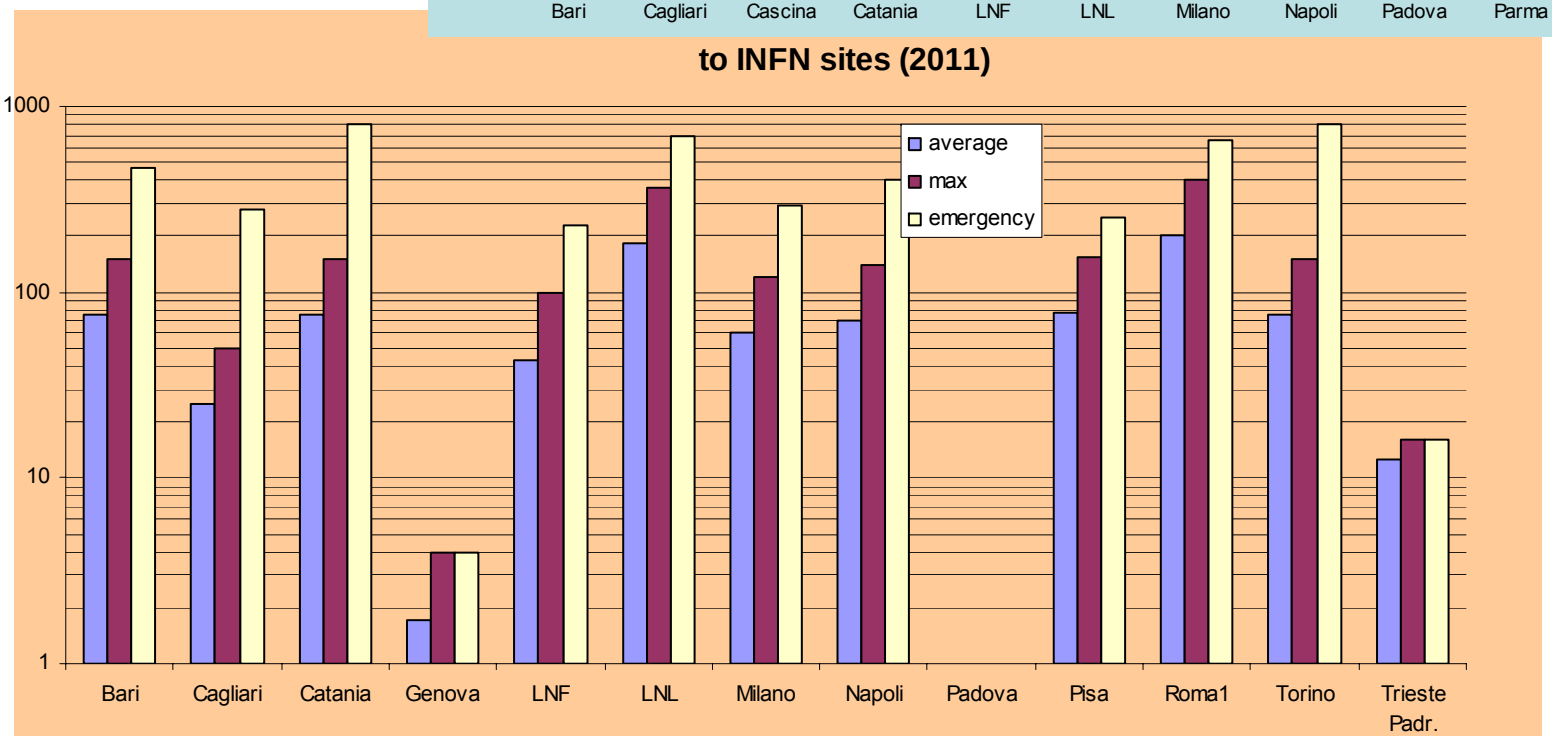
Accesso alle/dalle Sedi INFN (2008)





Accesso alle/dalle Sedi INFN (2011)





Requisiti per Comunicazioni Audio e Video

I sistemi per comunicazione interattiva audio ed audio/video su IP per funzionare al meglio richiedono di prestare particolare attenzione a parametri di rete quali *Ampiezza di banda, delay, jitter e Loss.* Dove per *delay* si intende il ritardo end to end fra i due sistemi, per *jitter* si intende la variazione del delay nel tempo (Δ Delay/δt) e per *Loss* si intende la percentuale di pacchetti persi.

	Banda	Delay	Jitter	Loss
Audio	< 100 Kbps	< 100 ms	40 ms	< 0.1 %
Video	0.1 – 30 Mbps	< 100 ms	40 ms	< 0.1 %

Applicazioni Real Time (I)

- Lo sviluppo e l'adozione di sistemi di controllo remoto di apparati strumentali e/o industriali accessibili in rete geografica (WAN) e' sempre stato frenato, oltre che da problemi di sicurezza, dalla difficolta' di avere caratteristiche di comunicazione in grado di garantire l'esecuzione di un data processo o di rispondere ad una richiesta di attenzione dell'apparato remoto (per esempio un allarme) entro un ben preciso intervallo di tempo.

- I parametri di rete che influenzano questo tipo di applicazioni real time sono sostanzialmente gli stessi delle trasmissioni video (vedi slide precedente)

- La possibilita' di avere percorsi ottici end-to-end di livello 2 in cui parametri come il delay e il jitter sono garantiti rende possibile il controllo remoto anche di apparati complessi come sistemi di accelerazione, rivelatori molto remoti (per esempio sottomarini), ecc.

	Banda	Delay	Jitter	Loss
Real Time	Dipende dall'applicazione	O (ms)	10 % delay	Deve garantire il delay richiesto

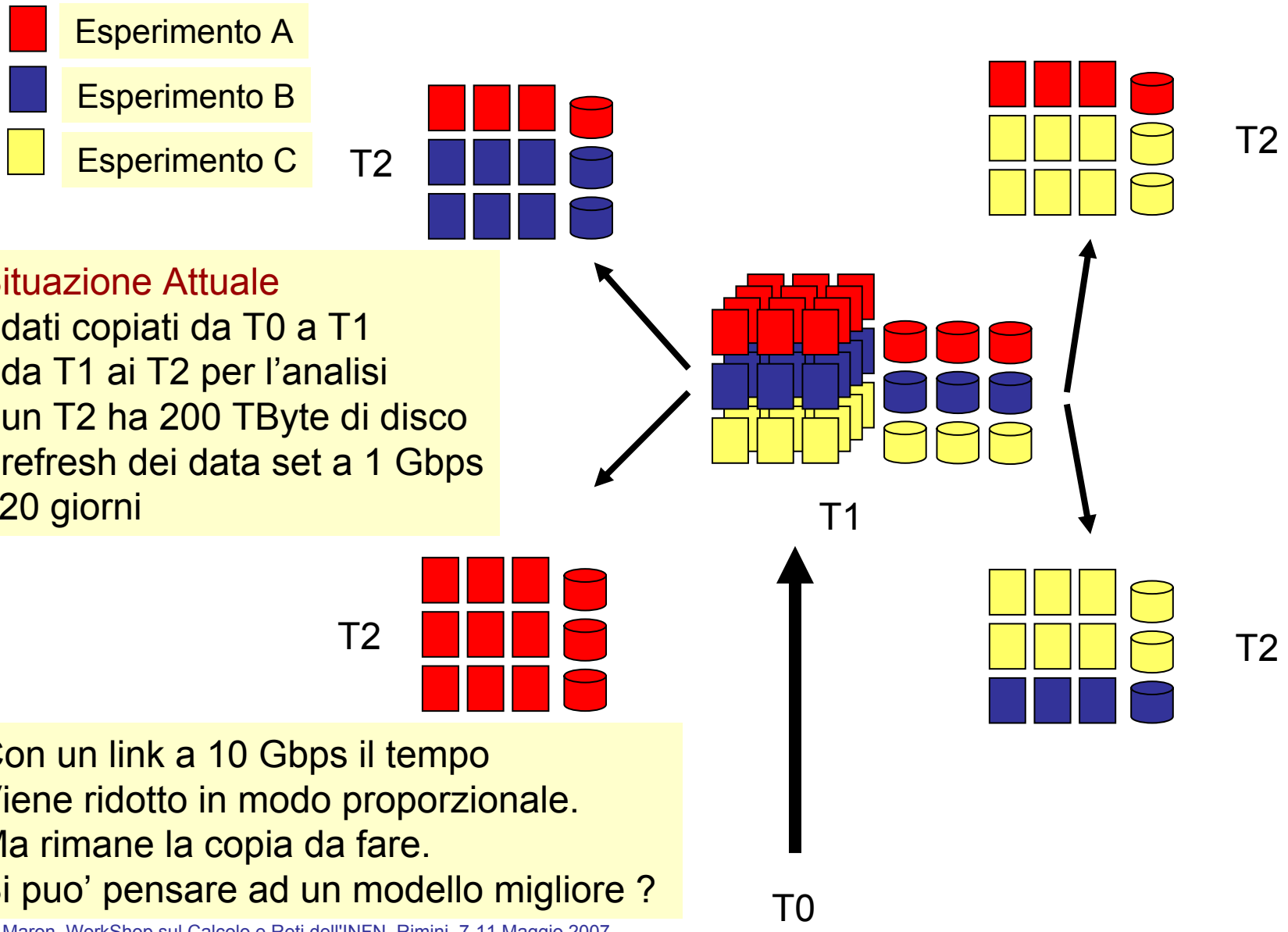
Applicazioni Real Time (II)

- L'infrastruttura ottica di GARR-X e' anche adeguata per sistemi di acquisizione dati in rete.
- Le tecniche di "Remote Memory Access" RMA o di RDMA (Remote Direct Memory Access) su DWDM consentono l'accesso alla memoria degli apparati come in locale.
-
- In questo contesto si puo' sfruttare anche la possibilita' di effettuare MULTICAST ottici su degli endpoints definiti.
  - E' un meccanismo estremamente efficiente quando si vuole distribuire i dati acquisiti sia per monitoraggio, ma anche per immagazzinare i dati in piu posti per ragioni di back up o di distribuzione dei data set
- In questi casi il parametro fondamentale e' la banda disponibile e, per il caso degli accessi in memoria, anche il delay

REALTA' VIRTUALE

- Nuove tecniche si stanno affermando ("teleimmersion") nel campo dell'interazione uomo-macchina (strumento).
- Si basano sulla ricostruzione virtuale dell'ambiente che si sta controllando (per esempio la control room dell'esperimento o la console di comando dell'acceleratore) permettendo all'utente del sistema un completa "telepresence" sul sito remoto.
- Servono risoluzioni molto alte O(100 Mpixel)
- Queste tecniche impattano principalmente sui parametri di "delay" e "jitter", e sull'ampiezza di banda.
- Con 10 – 20 frame/s servono bande O (10 Gbps)

Verso dei Tier Virtuali (TV) ?



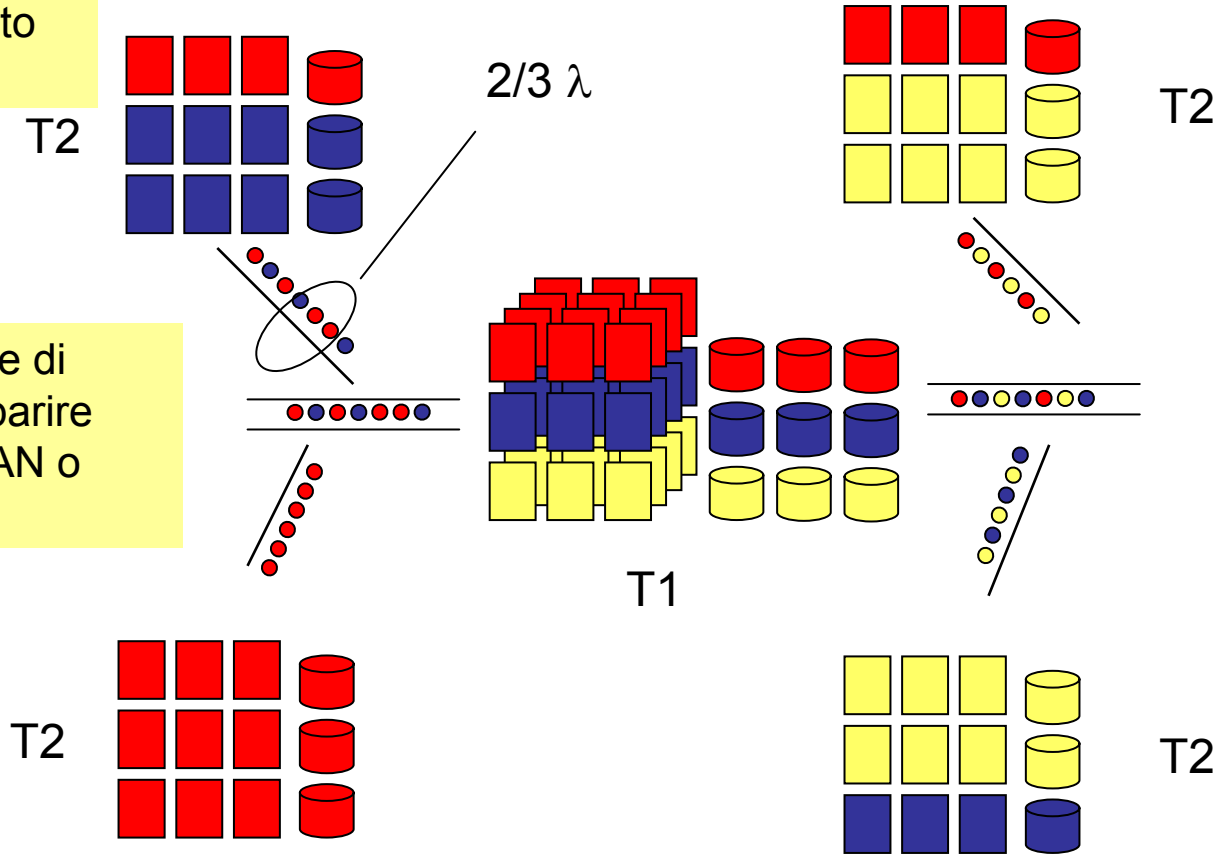
Situazione Attuale
- dati copiati da T0 a T1
- da T1 ai T2 per l'analisi
- un T2 ha 200 TByte di disco
- refresh dei data set a 1 Gbps
  20 giorni

Con un link a 10 Gbps il tempo
Viene ridotto in modo proporzionale.
Ma rimane la copia da fare.
Si puo' pensare ad un modello migliore ?

Lambda Networks per i Tier

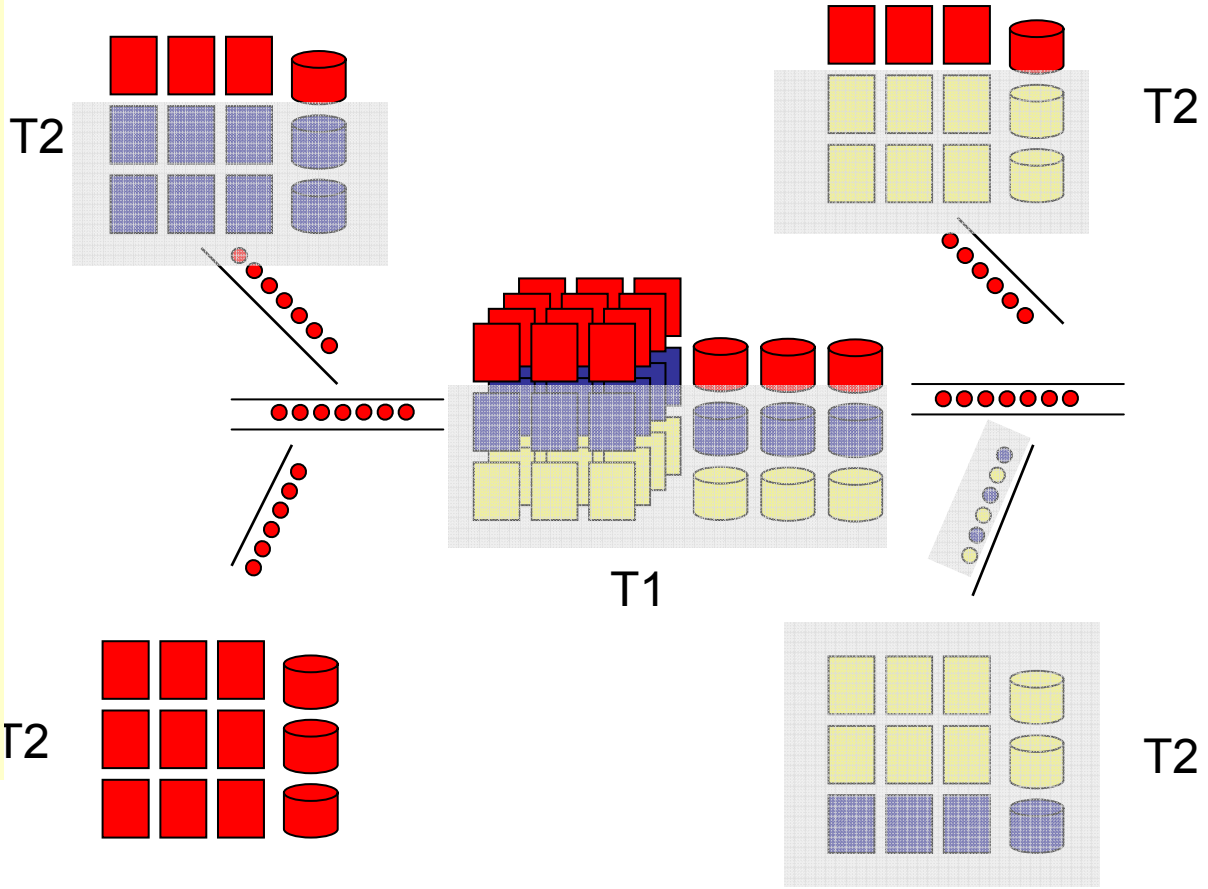
GARR-X fornisce l'infrastruttura di rete Per un accoppiamento "stretto" tra i Tier

Le risorse di calcolo e di storage possono apparire come nella stessa LAN o nella stessa SAN



Un Tier 2 Virtuale per ogni VO

- Trasparenze di accesso a tutte le risorse della VO distribuite nei Tier fisiciù
- acceso “diretto” ai data set, non piu necessita di copia
- la rete puo essere condivisa dinamicamente, le risorse possono essere ridistribuite ad altre VO se necessario
- possibilita’ di ottimizzazione d’uso delle risorse, fault tollerance
  back up funzionali.



Astrazione delle risorse

- L'infrastruttura ottica di GARR-X e' necessaria per la virtualizzazione dei Tier, ma non e' sufficiente
- Serve una nuova astrazione delle risorse di calcolo e storage che definisca il Tier virtuale e che tenga presente le caratteristiche dinamiche della rete.
- L'astrazione a la GRID delle risorse e' fondamentale per garantire una facile gestione delle risorse che, pur logicamente attigue, appartengono a centri fisici diversi con team di gestione e con responsabilita diverse
- Ci sono gia' diversi progetti sperimentali in questa direzione, alcuni attivi da piu anni.
- Se questo approccio risulta interessante bisogna investire nello studio di questi progetti e nella definizione di un progetto di fattibilita' che tenga conto di tutte le ns condizioni al contorno (non ultimo che il nostro deve essere un progetto di produzione)

Estensione delle SAN per un TV (I)

- trasporto di Fiber Channel or SCSI su IP
  - e' il caso piu semplice e, in teoria, gia possibile adesso. Per essere efficiente e poter lavorare con bande sul Gbyte necessita di schede di rete sui server equipaggiate con “TCP offloading engine”. Non ci sono particolari esigenze di rete per questo caso se non quello della banda.
- Estensione della SAN
  - In questo caso il protocollo Fiber Channel (FC) viene trasportato su una lambda del DWDM stabilendo una connessione nativa FC con il sito remoto. E' una soluzione molto efficiente. Un sito T2 puo' richiedere tipicamente da 2 a 3 lambda per esperimento per accedere allo storage del T1.
  - Il protocollo FC e' un protocollo con backpressure (a differenza di ethernet) dove sono definiti dei timeout. Questo introduce dei limiti al possibile ritardo di trasmissione e quindi alla distanza massima raggiungibile. Questi limiti sono in parte superabili introducendo degli apparati di buffer. Distanze tipiche sono intorno ai 100 km, ma ci sono apparati che garantiscono fino a piu di 300 km. Purtroppo questo e' un numero relativamente piccolo se consideriamo la topologia della rete dei T2-T1. Solo Legnaro, Milano e forse Pisa rientrano nel range di possibile utilizzo di FC in questa modalita'. Crediamo questo sia un punto da studiare attentamente con GARR-X

Estensione delle SAN per un TV (II)

- File System Distribuiti
  - Questa e' un'opzione molto sofisticata e ancora poco chiara (almeno per quanto riguarda la scalabilita'), anche se molto attraente. In questo caso i T2 potrebbero federarsi a livello di file system con il T1 di riferimento. I T2 dovrebbero essere connessi in FC con il T1 (vedi punto precedente) e sopra questa infrastruttura hardware andrebbe montato un file system distribuito. Le lamba occupate dovrebbero essere, come nel caso precedente, da 2 a 3 per sito T2. Candidati naturali a questo tipo di soluzione sono GPFS dell'IBM http://www-03.ibm.com/systems/clusters/software/gpfs.html e LUSTRE http://www.lustre.org/

- Vicino a FC si e' affiancato lo standard Infiniband (IB) http://www.infinibandta.org/home come protocollo per SAN. IB e' stato definito e accettato come standard qualche anno fa, ma la sua presenza nel mercato e' ancora limitata. Sono state fatte alcune dimostrazioni di estensione di IB su WAN sia a SC05 che a SC06. Crediamo che potrebbe essere interessante esplorare l'eventaule fattibilita' tecnica di un suo utilizzo su GARR-X

Configurazione dinamica della rete GARR-X

- L'alta configurabilita' delle reti lamda deve essere rese disponibile all'utenza (INFN) tramite opportuni servizi (possibilmente in architettura SOA o GRID)
- Deve essere possibile:
  - Allocare/de-allocare percorsi end-to-end e virtual private network
  - Settare i parametri di QoS dei percorsi come banda minima, delay massimo, loss, ecc.
  - Definire dei percorsi multicast
  - Accedere a servizi di monitor della rete (sempre in architettura SOA) e dei suoi parametri. Questo per ogni lambda allocata

Conclusioni

- Requisiti Fisici
  - Con il modello di calcolo attuale si ottiene:
    - Il T1 ha bisogno di n x 10 Gbps dove n > 3
    - I T2 abbisognano di link a 10 Gbps a partire dal 2008 per gli interventi di "emergenza" e dal 2009 per soddisfare gli accessi massimi
  - Applicazioni multimediali e real time
    - Delay da O(ms) a O(100 ms)
    - Jitter da 10% a 50%
    - Loss < 0.1 % o tale da non inficiare il delay
  - Realta' virtuale richiede una banda O (10 Gbps) per applicazione
  - Virtual Tier 2-3 lambda per Tier 2

- Requisiti di infrastruttura
  - Configurazione della rete "on demand"
  - Monitoraggio della rete "on demand"