CCR/14/07/p

Commissione Calcolo e Reti dell'INFN

Gruppo di lavoro su "Evoluzione delle esigenze di rete dell'INFN"

Evoluzione delle esigenze di rete geografica dell'INFN negli anni 2008-2011 e prospettive offerte dal progetto GARR-X

V. 26 giu 2007

Giugno 2007

Il gruppo di lavoro

1. Gianpaolo Carlino (INFN - Sez. di Napoli)
2. Roberto Gomezel (INFN - Sez. di Trieste)
3. Gaetano Maron (INFN- LNL)
4. Alberto Masoni (INFN - Sez. di Cagliari)
5. Mauro Morandin (INFN - Sez. di Padova)
6. Davide Salomoni (INFN - CNAF)
7. Stefano Zani (INFN - CNAF)

Sommario

Riassunto

Lo sviluppo di adeguati collegamenti di rete geografica fra le sedi dell'INFN e fra queste e i principali laboratori con cui l'Ente collabora è stata essenziale negli anni recenti per permettere ai propri ricercatori di svolgere efficacemente le proprie attività.

Nel prossimo futuro tale legame risulterà essere ancora più essenziale. La necessità di distribuire su base geografica le sempre crescenti risorse di calcolo necessarie per la simulazione e l'elaborazione dei dati scientifici, sia in Italia che su scala internazionale, richiederanno collegamenti in grado di sostenere elevatissimi volumi di traffico con grande affidabilità. La rete costituirà la base per il funzionamento di tale infrastruttura di calcolo distribuita e rivestirà quindi una importanza vitale per le comunità scientifica dell'INFN. Curiosamente, anche la tendenza opposta, ovvero quella di migrare servizi e applicazioni tradizionalmente residenti nelle strutture locali verso soluzioni centralizzate, accrescerà la criticità dei collegamenti verso i siti dove risiedono i servizi centrali che dovranno fornire possibilità di accesso ininterrotta da tutte le sedi sparse sul territorio nazionale.

Per far fronte alla crescenti esigenze, è in corso la definizione di un progetto di evoluzione della rete GARR, che sulla scia di quanto già avvenuto per altre reti della ricerca in Europa ed altrove, sfrutterà l'attuale disponibilità sul territorio di fibre ottiche spente per realizzare una nuova rete ottica proprietaria (GARR-X).

Alla luce del nuovo scenario che si sta prefigurando, la Commissione Calcolo e Reti dell'INFN ha ritenuto opportuno operare una ricognizione sulle esigenze di sviluppo della rete che le attività scientifiche dell'Ente potrebbero richiedere nei prossimi anni.

L'analisi riportata in questa nota descrive in modo dettagliato le necessità future dei progetti già approvati, così come esse vengono oggi stimate dai gruppi coinvolti. Comprende anche i requisiti relativi ad alcuni progetti in corso di definizione che si ritiene possano risultare di impatto significativo per la rete. Occorre osservare che, in alcuni casi, i modelli di calcolo su cui si basano tali estrapolazioni non sono stati ancora verificati sul campo e questo è in particolare vero nel caso della sperimentazione al LHC che rappresenta l'utenza INFN considerata di gran lunga predominante nei prossimi anni. Le proiezioni che ne conseguono sono quindi soggette ad ampie incertezze, tanto più consistenti quanto più proiettate nel futuro, e saranno soggette quindi ad aggiornamenti periodici su scala annuale.

Il quadro complessivo che ne deriva è comunque ben definito e vede il CNAF di Bologna, dove è ospitato il centro Tier1 per LHC, come principale centro di aggregazione del traffico di dati scientifici a livello nazionale ed internazionale. Tuttavia di notevole impatto risultano essere anche le prospettive di crescita di altri centri di secondo livello che, a partire dal 2008, necessiteranno di

accesso alla rete geografica con velocità di trasferimento tali da superare le capacità degli attuali circuiti. Inoltre, la distribuzione geografica di tali centri, che risulta essere piuttosto uniforme sul territorio nazionale, comporta la necessità di una dorsale di rete ottica che vada estesa lungo tutti gli assi principali del Paese.

La possibilità, anche per i centri di secondo livello, di accedere alla rete con bande trasmissive dell'ordine dei 10 Gb/s, così come realizzabile con la rete GARR-X, permette poi di aumentare sensibilmente il livello di disponibilità dei dati agli utenti e di poter far fronte a necessità impreviste (come quelle derivanti da perdite improvvise dei dati) con maggiore efficacia.

L'avvento di collegamenti ottici proprietari, non implica però solo maggiori larghezza di banda, ma fornisce anche la possibilità di implementare nuovi servizi con una flessibilità nelle modalità di fornitura che non poteva in precedenza essere ottenuta dai provider tradizionali. I due fattori costituiscono le premesse per la sviluppo di opportunità nuove nell'organizzazione dei modelli di calcolo scientifico degli esperimenti. L'ultima parte di questo rapporto riporta alcune idee in forma preliminare per indicare quali possono essere i possibili impatti di tali sviluppi sulle caratteristiche della rete GARR-X.

1 Il punto di partenza

1.1 La rete GARR oggi e il suo utilizzo da parte dell'INFN

La rete GARR attuale è rappresentata nella seguente figura che riporta un quadro riassuntivo dei collegamenti che la compongono e di quelli che forniscono l'accesso verso l'esterno. La rete è strutturata su GigaPOP posizionati nei baricentri di traffico della rete accademica e di ricerca del territorio nazionale.

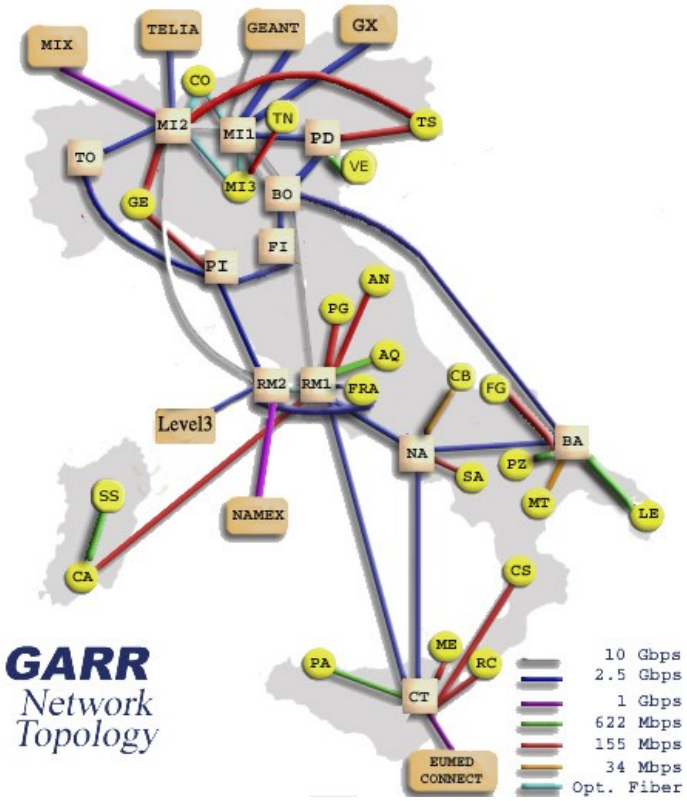

Fig. 1.1: Topologia rete GARR attuale

Viene classificato come Core Backbone la dorsale con topologia ad anello che interconnette i Giga-POP di concentrazione e trasporto relativi a Milano1, Milano2, Bologna e Roma.

A questa dorsale si connette il backbone di accesso a 2.5 Gb/s con topologia magliata che prevede almeno due link a 2.5 Gb/s da ogni POP (Torino, Padova, Pisa, Firenze, Napoli, Bari, Catania, Milano, Frascati e Roma), di cui uno connesso direttamente con il Core Backbone.

Troviamo un terzo livello di collegamenti tra i Mega-POP di concentrazione e il Backbone realizzati con collegamenti a singolo link SDH multipli di 34Mbps o 155Mbps che raccolgono tutti gli altri siti

italiani.

I collegamenti internazionali si attestano e sono generalmente collocati nelle sedi dei Giga-POP GARR.

Questa infrastruttura permette di interconnettere già ora i centri di calcolo principali dell'INFN e, in particolare, i siti Tier1 e Tier2 con una banda trasmissiva di almeno 1 Gb/s (10 Gb/s nel caso del Tier1), permettendo in questo modo di realizzare un canale di comunicazione adeguato alle esigenze di trasferimento di dati per le applicazioni attualmente in uso.

In questo modello di rete gli accessi dei siti INFN italiani risultano inseriti nella infrastruttura di rete nazionale e sono strettamente legati allo sviluppo della dorsale. Non è possibile allo stato attuale configurare interconnessioni dirette tra siti che eventualmente potrebbero necessitare di uno scambio di dati end-to-end, ma queste devono necessariamente seguire i percorsi fissati dalla topologia fisica esistente.

Questo, sino ad oggi, non ha rappresentato un limite all'utilizzo della rete, né è stata di impedimento alle applicazioni e ai flussi di traffico esistenti tra le sezioni INFN e tra queste e i siti internazionali di ricerca di interesse specifico per l'INFN.

L'aumento della larghezza di banda disponibile e la possibilità di disporre di servizi differenziati di trasporto permette già in questo momento un approccio di erogazione di servizi di connessione che potrebbe gestire la priorità differenziata del trasporto con l'adozione di reti private virtuali sulla infrastruttura fisica esistente.

Diversa appare delinearsi la situazione oggi con la configurazione dei nuovi centri Tier2 e di altri poli che si preannunciano come sorgenti e destinazioni di un elevato flusso di traffico di dati; si impone sempre più chiara la necessità di poter disporre di una maggiore flessibilità nella definizione dei circuiti e il bisogno di poter gestire una priorità differenziata del trasporto con l'adozione di reti private virtuali sulla infrastruttura fisica esistente.

Diventa sempre più urgente sottrarsi dalla rigida topologia classica attuale che, pur essendo capace di rispondere alla richiesta di aumento di banda trasmissiva, difficilmente consente di adattare dinamicamente le interconnessioni ai mutamenti di flussi di traffico che cambiano rapidamente nel tempo.

Lo sviluppo progressivo di farm locali che si possono trovare a dover soddisfare richieste di trasferimento consistenti e variabili in maniera dinamica nel corso del tempo e la sempre crescente diffusione delle applicazioni della GRID rendono sempre più urgente il passaggio a un modello di rete che consenta una maggior flessibilità della topologia di interconnessione.

1.2 L'evoluzione verso la rete GARR-X

L'evoluzione della rete, concepita per soddisfare le esigenze future dell'Università e della Ricerca

in Italia, prevede il passaggio ad una nuova infrastruttura, denominata GARR-X, basata sull' impiego massiccio di dark fiber che potranno essere illuminate in base alle esigenze dei diversi utilizzatori.

Con l'utilizzo di queste tecnologie trasmissive quali la ampia diffusione di fibre ottiche e l'adozione di apparati DWDM, il GARR sarà in grado di aumentare sensibilmente la capacità trasmissiva tra tutti i siti verso i GigaPOP, ma soprattutto sarà in grado di riservare lambda per applicazioni specifiche sull'intera dorsale, permettendo di attivare all'occorrenza dei veri e propri light path end-to-end.

Si prevede l'utilizzo della multiplazione di lambda sul backbone con meccanismi di instradamento che consentano di minimizzare il numero di hop tra i diversi X-POP e con la possibile attivazione di connessioni di backup tra i siti.

L'infrastruttura prevede una gerarchia a tre livelli (come riassunto nella Fig. 1.2) per raccogliere l'utenza in base alle loro esigenze di interconnessione e in base alla tecnologia di accesso al backbone.

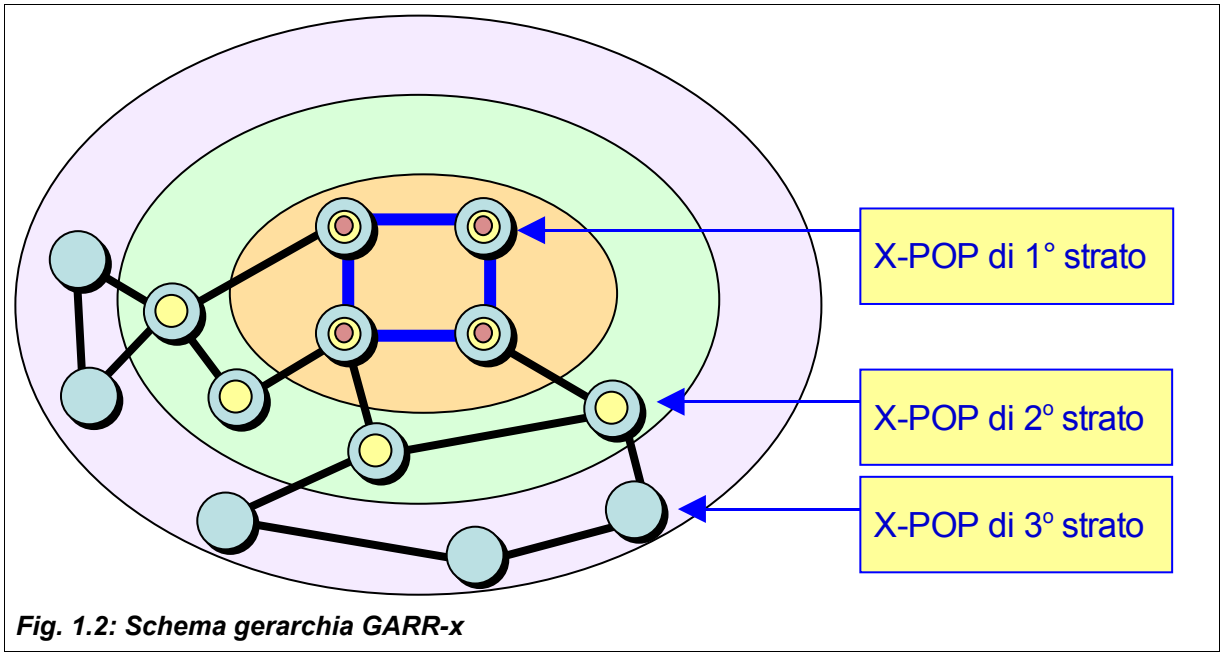

Fig. 1.2: Schema gerarchia GARR-x

Questo modello di rete permetterà di superare i limiti attualmente evidenziati nel paragrafo precedente che sono legati alla interconnessione esistente.

Il modello di calcolo dei prossimi anni richiederà una dinamicità di configurazione dei nodi di calcolo e di storage; questo potrebbe generare dei flussi di traffico che non sono sempre prevedibili e che potrebbero necessitare di riconfigurazioni di accesso più flessibili di quelle attuali.

Il modello di comunicazione dei Tier1 con i Tier2 potrebbe richiedere una rete di comunicazione

riservata per le applicazioni LHC distinta da quella utilizzata per il traffico general purpose.

L'utilizzo di lambda distinte consentirebbe una configurazione di interconnessione quasi privata tra i diversi Tier che garantirebbe i flussi di dati che le applicazioni LHC richiedono.

La possibilità di disporre di un numero di lambda configurabili dinamicamente aprirebbe la possibilità di estendere una rete a livello 2 tra due o più sedi INFN, consentendo di condividere sistemi di storage o altri dispositivi che potrebbero essere dislocati nelle diverse sedi interconnesse.

L'aspetto rilevante della nuova infrastruttura appare essere la plasticità che potrebbe portare a una topologia che cambia dinamicamente in base alle esigenze dell'utenza INFN e per periodi di tempo variabili.

2 Le esigenze dei prossimi anni

2.1 Introduzione

Per ottenere una quantificazione dei requisiti che la futura rete GARR-X potrebbe dover soddisfare per l'INFN nei prossimi anni sono state utilizzati come base di partenza le esigenze derivanti dai seguenti tre tipi principali di trasferimenti, in cui si è ritenuto utile, date le specifiche caratterizzazioni, suddividere le previsioni di traffico:

- trasferimenti di dati scientifici, che costituiscono la modalità dominante di utilizzo della rete per l'INFN;
- trasferimenti per l'accesso a servizi di base (E-mail, servizi Web, ecc.);
- trasferimenti dovuti all'uso di sistemi di audio/video-conferenza su IP.

Sono state quindi ricavati da tali dati preliminari, i requisiti per la rete futura in termini di:

- le componenti di trasmissione per i trasferimenti di dati scientifici, con la specificazione della relativa ampiezza di banda per:
  - ciascun tragitto end-to-end fra coppie di siti INFN,
  - il tragitto su fibra ottica dedicata tra il CERN ed il CNAF e
  - i tragitti da siti INFN e la rete della ricerca internazionale;
- indicazioni di qualità del servizio per le varie tipologie di trasferimento;
- componenti di accesso per ciascun sito INFN con specificazione della ampiezza massima di banda aggregata necessaria a soddisfare le relative componenti di trasmissione.

2.2 Analisi dei requisiti

2.2.1 Requisiti degli esperimenti per trasferimento dati scientifici

Per la valutazione delle esigenze di trasferimento dei dati scientifici, sono state richieste, attraverso gli osservatori delle Commissioni Scientifiche Nazionali in Commissione Calcolo e Reti, le previsioni di utilizzo della rete per gli anni 2007-2011. Per praticità sono stati considerati solo trasferimenti oltre una soglia costituita da una media di almeno 1 TB al giorno per almeno un mese di attività all'anno. Le stime comprendono anche possibili attività future di interesse dell'INFN non ancora approvate, o partecipazioni di siti che potrebbero non realizzarsi. Sono quindi da intendersi come delle estrapolazioni di massima che, in parte, potrebbero non concretizzarsi nei tempi e nei modi descritti.

I risultati di tale sondaggio sono riportati in tabelle dettagliate disponibili, in forma di foglio

elettronico Excel, all'indirizzo xxx.xxx..x.x. Esse riportano indicativamente, per ciascuna esperimento o attività di calcolo teorico, e per ogni trasferimento da sito INFN ad altro sito:

1. il periodo dell'anno in cui l'attività di trasferimento si prevede sia in corso;
2. la banda trasmissiva nominale media prevista;
3. la banda trasmissiva massima, derivata tenendo conto del margine necessario per il recupero di fisiologiche interruzioni e ritardi;
4. la banda trasmissiva auspicabile per poter far fronte a eventi straordinari, come la perdita o la invalidazione di una grossa parte dei dati residenti nei centri Tier2 di LHC, che richiedano un incremento temporaneo della

Nel caso delle federazioni Tier2 INFN a servizio degli esperimenti al LHC, sono stati riportati per i siti già approvati dall'INFN, i relativi volumi di trasferimento previsti per il loro funzionamento. Sono stati altresì inclusi anche i volumi di trasferimento per altri siti INFN, dove vi sono attualmente significative risorse di calcolo a disposizione degli esperimenti stessi, e che si potrebbe dover sviluppare in futuro. In questi casi la banda associata con ciascun sito è stata determinata assumendo, per ciascun esperimento, una ripartizione a traffico globale costante, e diminuisce quindi all'aumentare dei siti presi in considerazione.

In generale le stime fornite per gli esperimenti al LHC sono consistenti con le estrapolazioni più recenti riassunte nella cosiddetta "megatable" (disponibile all'indirizzo ...), tranne che nel caso della collaborazione Alice in cui il valori sono in generale superiori. Ciò è stato motivato dal gruppo italiano sulla base della presunta obsolescenza dei dati che compaiono nella citata megatable.

Nelle due tabelle che seguono è riportata una sintesi dei dati raccolti (elementi 3 e 4 della lista precedente), rispettivamente per gli esperimenti al LHC e per tutti gli altri.

Tab. 2.1: Velocità di trasferimento richieste dagli esperimenti LHC

Esperimento/attività	da	a	rate di trasferimento max previsto (MB/s)					rate di trasferimento stimato per esigenze straordinarie e temporanee (MB/s)				
			2007	2008	2009	2010	2011	2007	2008	2009	2010	2011
CSN1-LHC												
ALICE												
da T0 a T1	CERN	CNAF	**20**	**200**	**400**	**400**		**20**	**200**	**400**	**400**	
da T1 a CNAF	Σ T1 ALICE non-INFN	CNAF	**30**	**60**	**90**	**90**		**30**	**60**	**90**	**90**	
da CNAF a T1	CNAF	Σ T1 ALICE non-INFN	**10**	**20**	**30**	**30**		**10**	**20**	**30**	**30**	
da CNAF a T2	CNAF	Catania	**28**	**80**	**100**	**100**		**105**	**210**	**300**	**300**	
	CNAF	Torino	**28**	**80**	**100**	**100**		**105**	**210**	**300**	**300**	
	CNAF	Bari	14	40	50	50		53	105	150	400	
	CNAF	LNL	14	40	50	50		53	105	150	400	
	CNAF	Cagliari	10	40	40	40		140	280	120	280	
da T2 a CNAF	Catania	CNAF	**50**	**144**	**180**	**180**		**50**	**144**	**180**	**180**	
	Torino	CNAF	**50**	**144**	**180**	**180**		**50**	**144**	**180**	**180**	
	Bari	CNAF	25	72	90	90		25	72	**90**	**90**	
	LNL	CNAF	25	72	90	90		25	72	**90**	**90**	
	Cagliari	CNAF	18	72	72	72		18	72	**72**	**72**	
Note: Nel periodo di presa dati Pb-Pb ci si attende un raddoppio del flusso dei dati T0->T1 e un incremento del 30% del flusso T1-T1 - per il traffico T1-T2 e' stata considerata un' attivita' media												
ATLAS												
da T1 a T2	CNAF	Roma1	**60**	**100**	**140**	**140**		**150**	**250**	**400**	**400**	
	CNAF	Napoli	**60**	**100**	**140**	**140**		**150**	**250**	**400**	**400**	
	CNAF	Milano	40	70	100	100		100	175	275	275	
	CNAF	LNF	27	50	70	70		67	125	200	200	
da T2 a T1	Roma1	CNAF	**40**	**60**	**80**	**80**		**40**	**60**	**80**	**80**	
	Napoli	CNAF	**40**	**60**	**80**	**80**		**40**	**60**	**80**	**80**	
	Milano	CNAF	30	40	50	50		30	40	50	50	
	LNF	CNAF	20	30	40	40		20	30	40	40	
da T0 a T1	CERN	CNAF	**180**	**180**	**200**	**200**		**180**	**180**	**200**	**200**	
Σ T1 da/a CNAF	Σ T1 ATLAS non-INFN	CNAF	**180**	**320**	**540**	**540**		**180**	**320**	**540**	**540**	
	CNAF	Σ T1 ATLAS non-INFN	**220**	**380**	**640**	**640**		**220**	**380**	**640**	**640**	
Note: rate di trasferimento in emergenza determinato come quello necessario per ricopiare completamente in un sito la collezione completa degli AOD in un tempo di circa 10 giorni.												
CMS												
da T1 a T2	CNAF	LNL	**32**	**56**	**64**	**104**		**128**	**224**	**256**	**416**	
	CNAF	Roma1	**32**	**56**	**64**	**104**		**128**	**224**	**256**	**416**	
	CNAF	Pisa	16	32	40	60		64	128	160	240	
	CNAF	Bari	8	16	20	28		32	64	80	112	
da T2 a T1	LNL	CNAF	**10**	**21**	**31**	**41**		**10**	**21**	**31**	**41**	
	Roma1	CNAF	**10**	**21**	**31**	**41**		**10**	**21**	**31**	**41**	
	Pisa	CNAF	6	10	15	21		6	10	15	21	
	Bari	CNAF	4	4	6	10		4	4	6	10	
da T1 a T2	Σ T1 CMS non-INFN	LNL	**50**	**88**	**100**	**163**		**100**	**175**	**200**	**325**	
	Σ T1 CMS non-INFN	Roma1	**50**	**88**	**100**	**163**		**100**	**175**	**200**	**325**	
	Σ T1 CMS non-INFN	Pisa	25	50	63	94		50	100	125	188	
	Σ T1 CMS non-INFN	Bari	13	25	31	44		25	50	63	88	
Σ T2 a/da T1 CNAF	CNAF	Σ T2 CMS	200	240	360	360		400	1000	1000	1000	
	Σ T2 CMS	CNAF	21	41	62	82		21	41	62	82	
Σ non-INFN T2 a/da T1 CN	CNAF	Σ T2 CMS non-INFN	**136**	**128**	**232**	**152**		**544**	**512**	**928**	**608**	
	Σ T2 CMS non-INFN	CNAF	**0**	**0**	**0**	**0**		**0**	**0**	**0**	**0**	
Σ T1 da/a CNAF	Σ T1 CMS non-INFN	CNAF	**180**	**200**	**300**	**300**		**360**	**400**	**600**	**600**	
	CNAF	Σ T1 CMS non-INFN	**160**	**180**	**260**	**260**		**320**	**360**	**500**	**500**	
T0 a T1 CNAF	CERN	CNAF	**125**	**200**	**300**	**300**		**200**	**300**	**500**	**500**	
Note: rate di trasferimento in emergenza determinato come quello necessario per ricopiare completamente in un sito la collezione completa degli AOD in un tempo di qualche giorno.												
LHCb												
da T0 a T1	CERN	CNAF	**60**	**100**	**100**	**100**		**60**	**100**	**100**	**100**	
da T1 a T1	CNAF	Σ T1 LHCb non-INFN	**100**	**150**	**150**	**150**		**100**	**150**	**150**	**150**	
	Σ T1 CMS non-INFN	CNAF	**100**	**150**	**150**	**150**		**100**	**150**	**150**	**150**	

Tab. 2.2: Velocità di trasferimento dati degli esperimenti non-LHC

Esperimento/attività	da	a	rate di trasferimento max previsto (MB/s)					rate di trasferimento stimato per esigenze straordinarie e temporanee (MB/s)				
			2007	2008	2009	2010	2011	2007	2008	2009	2010	2011
CSN1-nonLHC												
BABAR												
Skim per analisi	SLAC	CNAF	**30**	**40**	**40**			**30**	**40**	**40**		
Dati raw	SLAC	Padova	**40**	**70**	**70**			**40**	**70**	**70**		
Dari ricostruiti	Padova	SLAC	**30**	**60**	**60**			**30**	**60**	**60**		
CDF												
	FNAL	CNAF	**25**	**25**	**25**	**25**		**25**	**25**	**25**	**25**	
	CNAF	FNAL	**12**	**12**	**12**	**12**		**12**	**12**	**12**	**12**	
COMPASS												
Replica Dati per l'analisi	CERN	Trieste Padr.	**9**	**9,2**	**9,4**	**9,8**	**10,4**	**9**	**9,2**	**9,4**	**9,8**	**10,4**
Montecarlo in grid	MONDO	Trieste Padr.	**2,4**	**2,8**	**3,5**	**4,5**	**5,6**	**2,4**	**2,8**	**3,5**	**4,5**	**5,6**
CNS2												
ARGO	Pechino	CNAF	**9**	**15**	**25**	**25**	**25**	**9**	**15**	**25**	**25**	**25**
VIRGO	Cascina	CNAF	**15**	**15**	**30**	**30**	**50**	**15**	**15**	**30**	**30**	**50**
VIRGO	CNAF	Lyon	**15**	**15**	**30**	**30**	**50**	**15**	**15**	**30**	**30**	**50**
NEMO/KM3	Portopalo	Catania		**100**	**100**	**500**	**500**	**0**	**100**	**100**	**500**	**500**
CNS3												
AGATA	LNL	Milano		**20**	**20**	**20**			**20**	**20**	**20**	
AGATA	LNL	Lyon		**100**	**100**	**100**			**100**	**100**	**100**	
AGATA	GSI	Milano					**20**					**20**
AGATA	GSI	LNL					**20**					**20**
AIACE	JLAB	Genova	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**
AIACE	JLAB	LNF	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**
HERMES	DESY	LNF	**13**	**13**	**13**	**13**	**13**	**13**	**13**	**13**	**13**	**13**
PANDA	GSI	LNF		**13**	**13**	**13**	**13**		**13**	**13**	**13**	**13**
Note: GSI->LNL sta per GSI->LNL/Padova												
CSN4												
APE	Roma1	Cnaf	**10,0**	**10,0**	**10,0**	**10,0**	**10,0**	**10**	**10**	**10**	**10**	**10**
3D	PARMA	BERLINO	**10,0**	**10,0**	**10,0**	**10,0**	**10,0**	**10**	**10**	**10**	**10**	**10**
Flusso1: Trasferimento risultati numerici simulazioni APE tra APEnext Center e CNAF T1.												
Flusso2: Trasferimento risultati numerici 3D simulazioni di stelle di neutroni tra Parma e AEI di Potsdam (1dataset=1Tbyte)												

2.2.2 Requisiti per l'accesso ai servizi di base su intranet ed internet

Una stima diretta dell'utilizzo di base della rete (mail, Web, ecc.) si può estrarre facilmente per i siti GARR in cui esso costituisce la maggior componente di traffico, e dove si evidenzia come variazione ciclica giorno-notte in diretta dipendenza della presenza del personale nelle sedi. Esempio tipico sono le Università italiane, come si può vedere ad esempio nel caso dell'Università di Padova, il cui andamento del traffico verso la rete GARR è rappresentato nella seguente Fig. 2.1

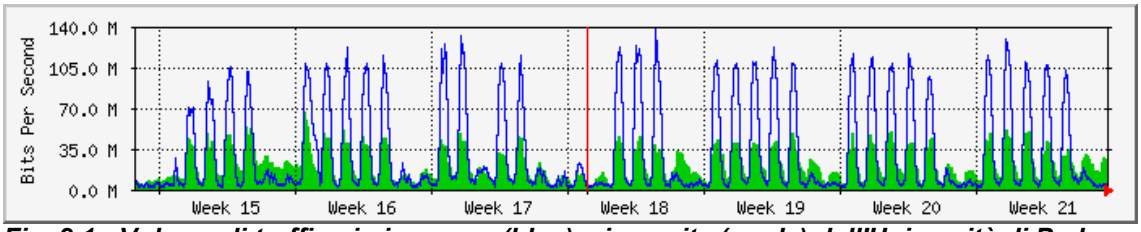

Fig. 2.1.: Volume di traffico in ingresso (blue) e in uscita (verde) dall'Università di Padova registrato dal router del GARR dal 13/4/7 al 27/6/07

Per quasi tutte le sedi INFN, tali variazioni cicliche non sono visibili nei grafici di traffico in quanto dominati da ben più consistenti e irregolari trasferimenti di dati scientifici. Tuttavia dall'ispezione di alcuni grafici dove quest'ultima componente è meno presente, si può ricavare una stima pur molto approssimativa di tale traffico e correlarlo con il numero di utenti che, per semplicità viene assunto pati al numero complessivo di dipendenti ed associati INFN. Si ottiene che per una sede con 150 utenti, il traffico di punta è dell'ordine dei 10 MB/s nei primi mesi del 2007.

L'estrapolazione del volume di traffico agli anni oggetto di questo studio, è stata ricavata anch'essa, come indicazione di massima, dai grafici di traffico aggregato del GARR; è stato quindi assunto un valore corrispondente ad un raddoppio del traffico ogni due anni. Con questi dati si sono ricavati le esigenze previste dell'INFN, come riassunti nella seguente Tab. 2.3, in funzione della dimensione del sito:

Tab. 2.3.: Stima traffico di base nei siti INFN (valori di picco)

da/a	utenti	Volume di traffico				
		2007 MB/s	2008 MB/s	2009 MB/s	2010 MB/s	2011 MB/s
LNF	600	4,0	5,7	8,0	11,3	16,0
grande sez.	350	2,3	3,3	4,7	6,6	9,3
media sez.	220	1,5	2,1	2,9	4,1	5,9
piccola sez.	120	0,8	1,1	1,6	2,3	3,2
gr. coll.	50	0,3	0,5	0,7	0,9	1,3

2.2.3 Requisiti per comunicazioni audio e video

I sistemi per comunicazione interattiva audio ed audio/video sono destinati a sfruttare sempre più le connessioni di rete ad elevata velocità. La qualità del servizio dipende in questo caso non solo dall'ampiezza di banda disponibile ma anche da altri parametri quali *delay (D), jitter (J) e loss (L).* Per *delay* si intende il ritardo end-to-end fra due terminali, per *jitter* si intende la variazione del delay nel tempo e per *loss* si intende la percentuale di pacchetti persi.

2.2.3.1 Ampiezza di banda per tipo di comunicazione

Per quanto riguarda le applicazioni di videoconferenza le bande in gioco variano da un minimo di 64kb/s per videoconferenze di bassissima qualità (di solito per collegamenti estremi da siti remoti) ad un massimo di 30 Mb/s nel caso di applicazioni limite come DVTS (Full PAL video + 2 x 44KHz audio).

In generale le videoconferenze più comunemente utilizzate nell'INFN variano fra i 384 Kb/s ai 2Mb/s. Dovendo identificare un valore medio si può stabilire **1Mb/s per ogni cliente collegato in videoconferenza.**

Per quanto riguarda le audio (o fono) conferenze, le richieste di banda variano da un massimo di 64Kb/s a 6,5 Kb/s in base all'algoritmo di compressione utilizzato. Si può assumere, con un approccio conservativo, che sia richiesta una banda di **64Kb/s per ogni cliente collegato in audio conferenza.**

2.2.3.2 Requisiti su Delay, Jitter, Loss:

Sia per applicazioni Audio che Audio/Video è auspicabile per preservare una sufficiente qualità del

segnale che i parametri delay, jitter e loss non eccedano rispettivamente i seguenti limiti

- Delay: D ≤ 100 ms
- Jitter: J ≤ 40 ms
- Loss: L ≤ 0,1 %

Garantire questi parametri di rete a livello di trasporto non dovrebbe essere difficile fino ad un utilizzo inferiore al 75-80% della banda disponibile come banda di accesso di un sito INFN.

Dovrebbero essere identificati nella nuova Rete della Ricerca gli strumenti necessari a garantire i parametri sopra citati anche in situazioni critiche a livello di utilizzo della banda (QoS, Ip premium, ecc).

Naturalmente i parametri sopra descritti devono essere garantiti (end-to-end) e quindi anche all'interno delle reti locali che andranno dimensionate adeguatamente. Sistemi di gestione della banda simili a quelli che si richiedono per la rete geografica potrebbero essere poi necessari anche sugli apparati di rete locale.

2.2.4 Ampiezza di banda richiesta per sede INFN

Estrapolando i dati di uso attuali e considerando l'aumento dell'utilizzo di audio/videoconferenze presumibile all'approssimarsi della presa dati di LHC, sia quello dovuto alla disponibilità di nuovi e più efficaci strumenti collaborativi, si può considerare che nel 2008 un numero massimo pari al 10% degli utenti INFN di una sede facciano contemporaneamente uso di uno strumento di teleconferenza.

Di questo 10% si assume che 1/3 faccia uso di videoconferenza e 2/3 utilizzi sistemi per fono conferenze su IP.

Per una ipotetica sezione con 100 utenti, l'utilizzo di banda massimo risulta quindi essere pari a 3,5 Mb/s (nel seguito, approssimato per uniformità a 0,35 MB/s)

Essendo quasi tutti i sistemi di audio/video conferenza adottati dall' INFN basati su traffico unicast, occorre considerare che al CNAF (sede che ospita i sistemi di multi conferenza), sarà necessario riservare molto più ampia, dell'ordine dei 10 MB/s o più per questo tipo di traffico.

Si può pensare che vi sia un aumento nell'utilizzo di applicazioni Audio/Video fra il 2009 ed il 2011 che possa portare la percentuale di utenti contemporaneamente connessi dal 10 al 15-20 %. Essendo però questo tipo di applicazioni in rapidissima evoluzione, sarà necessario verificarne anno per anno l'effettivo utilizzo di banda.

L'esplosione di applicativi multimediali basati su tecnologie peer-to-peer potrà inoltre modificare notevolmente lo scenario dei pattern di traffico Video/Voce che ci troveremo a dover veicolare sulle nostre reti.

Una stima dell'impatto delle esigenze sulla banda di accesso dei siti INFN viene riportata quindi nella seguente Tab. 2.4:

Tab. 2.4: Stima traffico per audio/video-conferenze nei siti INFN

sito	utenti	Volume di traffico				
		2007 (MB/s)	2008 (MB/s)	2009 (MB/s)	2010 (MB/s)	2011 (MB/s)
LNF	600	1,4	2,0	2,8	4,0	5,6
grande sez.	350	0,8	1,2	1,6	2,3	3,3
media sez.	220	0,5	0,7	1,0	1,5	2,1
piccola sez.	120	0,3	0,4	0,6	0,8	1,1
gr. coll.	50	0,1	0,2	0,2	0,3	0,5
CNAF MCU		10	14,1	20,0	28,3	40,0

2.3 Sommario delle prestazioni richieste

Dalle stime dei volumi di traffico, si sono quindi ottenute:

- le bande aggregate relative alla componente dominante rappresentata dal trasferimento dei dati scientifici, per le varie combinazioni terminali dei collegamento (componenti di trasmissione),
- l'ampiezza totale di banda di accesso per ciascun sito (componenti di accesso).

2.3.1 Componenti di trasmissione

La seguente tabella riporta per ciascun collegamento la banda richiesta per i trasferimenti di dati scientifici. Nel caso dell'utilizzo temporaneo di emergenza, l'aggregazione è stata ottenuta considerando, sempre per ciascun collegamento, il maggiore trasferimento temporaneo fra quelli possibili e sommandovi il valore di banda corrispondente agli altri trasferimenti calcolati al rate massimo.

Tab. 2.5: Sommario requisiti per trasferimento dati scientifici

tipo collegamento		*Banda Garantita di Trasferimento (MB/s)*					*Banda massima per uso Temporaneo (MB/s)*				
da	*a*	*2007*	*2008*	*2009*	*2010*	*2011*	*2007*	*2008*	*2009*	*2010*	*2011*
Collegamenti dedicati											
CERN	CNAF	**0**	**385**	**680**	**1000**	**1000**	**0**	**460**	**780**	**1200**	**1200**
Collegamenti da siti INFN a siti non italiani											
CNAF	ext	**27**	**653**	**900**	**1354**	**1282**	**27**	**1061**	**1284**	**2050**	**1738**
ext	CNAF	**64**	**570**	**820**	**1130**	**1105**	**64**	**750**	**1020**	**1430**	**1405**
LNL	ext	**0**	**100**	**100**	**100**	**0**	**0**	**100**	**100**	**100**	**0**
Roma1	ext	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Pisa	ext	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Bari	ext	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Parma	ext	**10**	**10**	**10**	**10**	**10**	**10**	**10**	**10**	**10**	**10**
Padova	ext	**30**	**60**	**60**	**0**	**0**	**30**	**60**	**60**	**0**	**0**
ext	LNL	**0**	**50**	**88**	**100**	**183**	**0**	**100**	**175**	**200**	**345**
ext	Roma1	**0**	**50**	**88**	**100**	**163**	**0**	**100**	**175**	**200**	**325**
ext	Pisa	**0**	**25**	**50**	**63**	**94**	**0**	**50**	**100**	**125**	**188**
ext	Bari	**0**	**13**	**25**	**31**	**44**	**0**	**25**	**50**	**63**	**88**
ext	Padova	**40**	**70**	**70**	**0**	**0**	**40**	**70**	**70**	**0**	**0**
ext	Trieste P.	**11**	**12**	**13**	**14**	**16**	**11**	**12**	**13**	**14**	**16**
ext	Milano	**0**	**0**	**0**	**0**	**20**	**0**	**0**	**0**	**0**	**20**
ext	Genova	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**	**4**
ext	LNF	**17**	**30**	**30**	**30**	**30**	**17**	**30**	**30**	**30**	**30**
Collegamenti fra siti INFN											
CNAF	Bari	**0**	**22**	**56**	**70**	**78**	**0**	**60,5**	**121**	**170**	**428**
CNAF	Cagliari	**0**	**10**	**40**	**40**	**40**	**0**	**140**	**280**	**120**	**280**
CNAF	Catania	**0**	**28**	**80**	**100**	**100**	**0**	**105**	**210**	**300**	**300**
CNAF	LNF	**0**	**27**	**50**	**70**	**70**	**0**	**67**	**125**	**200**	**200**
CNAF	LNL	**0**	**46**	**96**	**114**	**154**	**0**	**142**	**264**	**306**	**504**
CNAF	Milano	**0**	**40**	**70**	**100**	**100**	**0**	**100**	**175**	**275**	**275**
CNAF	Napoli	**0**	**60**	**100**	**140**	**140**	**0**	**150**	**250**	**400**	**400**
CNAF	Pisa	**0**	**16**	**32**	**40**	**60**	**0**	**64**	**128**	**160**	**240**
CNAF	Roma1	**0**	**92**	**156**	**204**	**244**	**0**	**188**	**324**	**464**	**556**
CNAF	Torino	**0**	**28**	**80**	**100**	**100**	**0**	**105**	**210**	**300**	**300**
Bari	CNAF	**0**	**29,3**	**76,1**	**96,2**	**100,3**	**0**	**29,3**	**76,1**	**96,2**	**100,3**
Cagliari	CNAF	**0**	**18**	**72**	**72**	**72**	**0**	**18**	**72**	**72**	**72**
Catania	CNAF	**0**	**50,4**	**144**	**180**	**180**	**0**	**50,4**	**144**	**180**	**180**
LNF	CNAF	**0**	**20**	**30**	**40**	**40**	**0**	**20**	**30**	**40**	**40**
LNL	CNAF	**0**	**35,5**	**92,5**	**120,8**	**131**	**0**	**35,5**	**92,5**	**120,8**	**131**
Milano	CNAF	**0**	**30**	**40**	**50**	**50**	**0**	**30**	**40**	**50**	**50**
Napoli	CNAF	**0**	**40**	**60**	**80**	**80**	**0**	**40**	**60**	**80**	**80**
Pisa	CNAF	**0**	**6,2**	**10,3**	**15,4**	**20,5**	**0**	**6,2**	**10,3**	**15,4**	**20,5**
Roma1	CNAF	**10**	**60,3**	**90,5**	**120,8**	**131**	**10**	**60,3**	**90,5**	**120,8**	**131**
Torino	CNAF	**0**	**50,4**	**144**	**180**	**180**	**0**	**50,4**	**144**	**180**	**180**
Cascina	CNAF	**15**	**15**	**30**	**30**	**50**	**15**	**15**	**30**	**30**	**50**
LNL	Milano	**0**	**20**	**20**	**20**	**0**	**0**	**20**	**20**	**20**	**0**
Portopalo	Catania	**0**	**100**	**100**	**500**	**500**	**0**	**100**	**100**	**500**	**500**

2.3.1.1 Parametri di qualità del servizio

Per caratterizzare la qualità del servizio si sono utilizzati i parametri e le relative classificazioni riportate nella seguente tabella:

Tab. 2.6: Parametri di qualità del servizio

Parametri di qualità del servizio		
Disponibilità	***base***	98% mediato sull'anno
	standard	99,5% mediato sull'anno
	premium	99,8% mediato sull'anno
	mission-critical	99,99% mediato sull'anno
Tempo di ripristino	***base***	guasto bloccante risolto nel 95% entro NBD, 100%, entro NtNBD
	standard	guasto bloccante risolto entro 8h lavorative nel 95%, 12h lavorative nel 100%
	premium	guasto bloccante risolto entro 4h lavorative e 8h solari nel 95%, 8h lavorative e 12 h solari nel 100%
	mission-critical	guasto bloccante risolto entro 4h solari nel 95%, 6h solari nel 100%
Classi di servizio	***best effort***	D < 500 ms; L < 5%
	base	D < 200 ms; L < 1 %
	standard	D < 100 ms; L < 0,1%
	audio/video conf.	D < 100 ms; L < 0,1%; J < 40 ms
	real time	D < 40 ms; L < 0,1 %; J < 10 ms

Per i trasferimenti, raggruppati in alcune categorie principali, sono stati quindi individuati i requisiti di qualità di servizio ritenuti opportuni, come indicato nella seguente tabella. In particolare tenendo presente quanto previsto nel modello di accordo di servizio elaborato per i centri ci calcolo LHC[1]:

Tab. 2.7: Qualità dei servizi di trasferimento

Da/a		disponibilità	tempo di ripristino	classe di servizio
Trasferimento dati	CERN <-> CNAF	**premium**	**premium**	**standard**
	CNAF <-> siti Tier1	**premium**	**premium**	**standard**
	CNAF <-> siti Tier2	**standard**	**standard**	**base**
Servizi base	accesso a siti che gestiscono servizi centrali (nel 2007: CNAF e LNF)	**standard**	**premium**	**base**
	fra sito e sito INFN e fra siti INFN e siti serviti dalle reti della ricerca	**standard**	**standard**	**base**
	fra siti INFN e internet	**base**	**standard**	**base**
Servizi per audio/video-conferenza	far sito e sito INFN e fra siti INFN e siti serviti dalle reti della ricerca	**standard**	**standard**	**audio/video-conf.**

1 v. http://lcg.web.cern.ch/lcg/C-RRB/MoU/WLCGMoU.pdf

2.3.2 Componenti di accesso

La seguente tabella riporta infine il sommario dei requisiti di accesso per ogni sede INFN, ottenuta aggregando le componenti di trasferimento relative ai tre utilizzi considerati in questo studio e riportando il valore massimo fra quelli di ingresso ed uscita[2].

Si può notare come, già nella stime delle bande di accesso massime per trasferimenti continuativi, ci sia una previsione di superare nel 2008 la soglia del Gb/s (~ 100 MB/s) per alcune sedi di centri di calcolo degli esperimenti LHC. Per il CNAF i valori aggregati crescono invece rapidamente oltre i 20 Gb/s, dove, come si può notare dalla Tab. 2.5, vi è una prevalenza dei trasferimenti su link internazionali rispetto a quelli nazionali.

Se poi si prendono in considerazione le bande di accesso richieste per far fronte a delle emergenze temporanee, il numero di siti INFN che richiedono velocità di trasmissione superiori al Gb/s crescono ulteriormente. Queste estrapolazioni sono perciò consistenti con uno scenario in cui nel prossimo futuro vengano messi a disposizione dei principali centri INFN disposti su tutto il territorio nazionale dei collegamenti di rete a livello di 10 Gb/s.

2 Nel foglio di calcolo sono riportati anche i dati disaggregati.

Tab. 2.8: Componenti di accesso per i siti INFN

Sede	Atttuali parametri bande di accesso		Banda Garantita di Accesso							
			Continuativo				Temporaneo			
	BGA (MB/s)	BEA (MB/s)	2008 (MB/s)	2009 (MB/s)	2010 (MB/s)	2011 (MB/s)	2008 (MB/s)	2009 (MB/s)	2010 (MB/s)	2011 (MB/s)
AC - Frascati	(LNF)	(LNF)	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Alessandria	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Bari	45,0	100,0	34,5	81,0	101,3	121,8	73,0	146,0	201,3	471,8
Bologna	5,0	15,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Brescia	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Cagliari	1,6	3,0	18,0	72,0	72,0	72,0	140,0	280,0	120,0	280,0
Catania-Cittadella	3,2	10,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Catania	6,0	30,0	128,0	180,0	600,0	600,0	205,0	310,0	800,0	800,0
CNAF	100,0	200,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
CNAF-LCG Bologna	150,0	700,0	1097,9	1945,3	2706,0	2729,0	1305,4	2165,8	3028,5	3034,0
Cosenza	0,8	1,6	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Ferrara	2,5	3,4	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Firenze	2,0	3,2	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Genova	1,2	3,2	4,0	4,0	4,0	4,0	4,0	4,0	4,0	4,0
GGI Arcetri (FI)	0,0	0,2	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
L'Aquila	0,8	3,4	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Lecce	0,8	1,6	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
LNF	7,0	20,0	56,7	80,0	100,0	100,0	96,7	155,0	230,0	230,0
LNGS	5,0	15,5	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
INFN – LNGS (sale sp. per Backup)	0,1	0,2	0,0	0,0	0,0	0,0	0,2	0,2	0,2	0,2
LNL	40,0	100,0	155,5	212,5	240,8	336,5	192,0	351,5	406,0	686,5
LNS	1,5	3,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Messina	0,2	0,4	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Milano	5,0	35,0	60,0	90,0	120,0	120,0	120,0	195,0	295,0	295,0
Milano Bicocca	2,0	3,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Napoli	10,0	20,0	60,0	100,0	140,0	140,0	150,0	250,0	400,0	400,0
Padova	26,5	100,0	70,0	70,0	0,0	0,0	70,0	70,0	0,0	0,0
Parma	0,8	1,6	10,0	10,0	10,0	10,0	10,0	10,0	10,0	10,0
Pavia	1,5	3,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Perugia	1,2	2,4	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Pisa	35,0	100,0	41,0	82,0	102,5	153,8	89,0	178,0	222,5	333,8
Presidenza	1,0	10,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Presidenza-backup - Roma	0,2	0,2	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Roma1	20,0	40,0	142,0	243,5	304,0	406,5	238,0	411,5	564,0	718,5
Roma2	1,0	7,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Roma3	5,0	10,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Salerno	0,4	1,6	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
ISS	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Siena	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Torino	10,0	100,0	50,4	144,0	180,0	180,0	105,0	210,0	300,0	300,0
Trento	0,2	1,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Trieste Miramare	0,2	0,4	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Trieste Padr.	5,0	10,0	12,0	12,9	14,3	16,0	12,0	12,9	14,3	16,0
Trieste – Dipartimento	0,8	0,8	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Udine	1,0	3,2	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
			>15	>100			>15	>100		

3 Nuove modalità di utilizzo su base geografica offerte da un'infrastruttura ottica proprietaria

Una infrastruttura ottica proprietaria come quella proposta da GARR-X non rappresenta solo una moltiplicazione della banda disponibile, ma un notevole salto tecnologico che può trasformare le reti geografiche in una sorta di estensione di quelle locali. Questo, se da una parte apre nuovi scenari e possibilità, dall'altra suggerisce di riconsiderare modelli di calcolo consolidati negli anni e che si poggiavano sull'assunzione che il trasporto fosse fatto solo su IP condiviso tra tutti i nodi interconnessi alla rete. La semplificazione nelle procedure di accesso alle risorse di calcolo e di storage che si può intravedere e la maggiore efficienza di trasferimento rendono questa nuova visione di grande stimolo verso un'accelerazione della definizione di nuovi modelli. Tale evoluzione ai modelli più avanzati potrà avvenire in modo semplice e graduale, visto che comunque GARR-X garantirà il tradizionale routing IP su cui si basano i modelli attuali, e darà anche accesso fin da subito a bande di trasmissione pia grandi (10 Gb/s).

Inoltre la possibilità, a livello di rete geografica, di avere cammini ottici dedicati end-to-end o di stabilire reti private virtuali a parametri di rete garantiti, rende possibile un affidabile controllo remoto di strumentazione scientifica e di apparati complessi come gli acceleratori di particelle. Anche in questo caso vengono infrante modalità di lavoro consolidate (p.es.: la control room locale) in favore di una virtualizzazione dell'accesso delle risorse strumentali (e.g. Virtual control room) e si aprono interessantissimi scenari di cooperazione remota, flessibilità nell'utilizzo delle risorse umane, ecc.

In entrambi i casi ci sarà un impatto sul modello organizzativo che abbiamo (una LAN estesa, per esempio, costringe una cooperazione molto stretta tra i siti coinvolti) e che andrà valutato in modo opportuno, anche sul piano organizzativo e sociologico.

Qui di seguito proviamo ad analizzare con qualche maggior dettaglio i due casi menzionati, tentando di dare un'ordine di grandezza dell'impatto sulla rete (sia in termini di banda che di parametri di rete) di queste nuove modalità di utilizzo della stessa.

3.1.1 Verso dei Tier Virtuali ?

La rete dei Tier per l'analisi degli esperimenti LHC è costituita da centri autonomi di calcolo connessi tra la loro con link veloci su rete Ip a commutazione di pacchetto e layer 3. La topologia logica di connessione prevede una gerarchia a tre livelli in modo da definire un modello di calcolo dove i data set vengono dapprima copiati dal CERN (livello 0, Tier 0) verso i Tier 1 (centri di livello 1) e da qui, in accordo alle esigenze di analisi, ai centri di secondo livello (Tier 2) dove risiede la potenza di calcolo per effettuare l'analisi richiesta. Mentre la copia dal livello 0 ai livelli 1 risponde anche ad esigenze di back up dei dati originali, la copia dai livelli 1 ai livelli 2 viene effettuata solo per garantire una sufficiente velocità di accesso ai dati da parte delle CPU presenti nei T2. I T2

sono infatti solitamente connessi al T1 di riferimento con un link a 1 Gb/s e sono dotati di una quantità ragguardevole di disco in modo da fornire una cache locale a questi dati duranti il periodo di elaborazione. Il refresh completo dei dati, stimando che un T2 abbia dell'ordine di qualche centinaio di Tbyte, richiede dell'ordine di una decina di giorni ed è quindi un'operazione che non può essere fatta di frequente. I centri Tier2 inoltre producano anche analisi MC e quindi c'è un flusso dati di una certa rilevanza nella direzione da T2 a T1.

I centri, tra loro autonomi sia dal punto di vista fisico che di amministrazione dei sistemi, sono "federati" tra loro tramite un middleware di Grid che definisce per ogni esperimento una "Virtual Organization" (VO). In questo modo le risorse di calcolo e storage di una data VO distribuite tra i vari centri Tier sono virtualizzate e diventano accessibili in modo sicuro e trasparente all'utente membro di quella VO, indipendentemente dalla locazione fisica di quelle risorse.

La situazione italiana rispecchia esattamente questo quadro, avendo al CNAF il Tier 1 di riferimento per i quattro esperimenti mentre i relativi Tier 2 sono dislocati in una decina di sedi e laboratori dell'INFN.

Questa situazione, se pur soddisfacente dal punto di vista logico, è limitata dalla banda disponibile e dalla tecnologia trasmissiva che non permette una totale virtualizzazione delle risorse. Per esempio:

- i data set da analizzare e risiedenti nel Tier1 sono in genere di dimensioni notevoli. Sarebbe molto più efficiente potervi accedere direttamente dalle CPU dei Tier2 invece che copiarli nella cache del disco locale dei T2 stessi. Grazie ai servizi di GRID e al middleware sviluppato dagli esperimenti l'operazione di copia è trasparente all'utente, ma richiede tempi notevoli e quindi limita la capacita di analisi dei dati e la flessibilità di tutto l'insieme. Un T2 ha tipicamente qualche centinaio di Tbyte di disco locale e per ricopiare interamente tutti i data set del T2 servono, ad 1 Gb/s, una decina di giorni. Il passaggio a 10 Gb/s riduce il tempo di copia a qualche giorno, ma sarebbe molto più efficiente poter lavorare remotamente sui dischi.
- le risorse di calcolo all'interno di un centro Tier possono essere utilizzate in modo flessibile anche da VO diverse definendo delle politiche di accesso che sfruttino le proprietà dei job scheduler del centro. In principio questo è anche possibile tra centri diversi sfruttando la capacità dei broker di GRID di indirizzare le richieste di analisi laddove la potenza di calcolo è disponibile. Questo però è nella realtà limitato dal fatto che i data set da analizzare sono molto grandi e non è pensabile spostarli con la stessa facilità con cui si può spostare un job. È difficile quindi, allo stato attuale, definire delle politiche di ottimizzazione delle risorse, fault tolerance e back up funzionali tra i centri Tier italiani.

L' accesso ad un'infrastruttura ottica multi-lambda di tipo layer2 come proposta da GARR-x, garantirebbe ai centri Tier italiani un'adiacenza molto più stretta delle loro risorse di calcolo che

potrebbero quindi apparire contigue, cioè come se appartenessero alla stessa LAN o SAN. La disponibilità poi di banda a basso prezzo renderebbe la velocità di accesso a tutte queste risorse, indipendentemente dalla loro posizione geografica, comparabile a quelle all'interno di una LAN, modulo naturalmente i tempi di trasmissione. Inoltre la possibilità di definire i percorsi ottici di collegamento e la banda da dedicare ad essi permette configurazione dinamiche delle risorse.

Questo accoppiamento stretto tra le risorse dei Tier fisici può portare alla definizione di un Tier virtuale composto appunto dalle risorse di calcolo e storage di quelli fisici. Per esempio tutti i Tier di un esperimento potrebbero comporre il Tier virtuale di quell'esperimento. Grazie alla possibilità di configurazione dinamica della rete, le risorse di un Tier virtuale possono essere trasferite rapidamente in favore di un altro Tier che in quel momento ne ha piè bisogno. Per esempio se il Tier virtuale di Atlas è scarico potrebbe "affittare" le proprie CPU a quello di CMS e questo indipendentemente da dove siano localizzate fisicamente le CPU e da dove si trovino i data set da analizzare. In questo modello infatti la banda a disposizione e la tecnologia trasmissiva permette un accesso diretto allo storage remoto da parte delle CPU. La virtualizzazione del Tier riguarda solo le risorse italiane e quindi per la propria VO esso si presenta all'esterno (e quindi alla GRID di LHC) come un normale Tier (o come un insieme di Tier).

Questo approccio richiede però una nuova astrazione di tutte le risorse di calcolo e storage che definisca l'ambiente del Tier virtuale e che tenga presente le caratteristiche dinamiche della nuova infrastruttura di rete. L'astrazione a la GRID delle risorse di calcolo distribuite è fondamentale per garantire una facile gestione di risorse che, pur logicamente attigue, appartengono a centri diversi tra loro con team di gestione e responsabilità diverse. Nuovi servizi di GRID (o Web Service in un'architettura SOA), da sviluppare o da evolvere da quelli oggi disponibili, devono essere quindi progettati per poter definire questa nuova astrazione.

Sistemi di cluster virtuali su infrastruttura ottica sono già stati sperimentati e la loro esperienza può essere presa a riferimento per definire un progetto di fattibilità del Tier virtuale. Una delle esperienze più fortunate è quella di "OptiPuter" (http://www.optiputer.net/) . In questo caso l'astrazione delle risorse viene fatta tramite un "Distributed Virtual Computer (DVC)" middleware. I componenti principali di questa astrazione sono servizi e librerie per la:

- gestione delle risorse
  - ha lo scopo di allocare/rilasciare le risorse disponibili
- gestione del nome delle risorse
  - definisce un nome unico, all'interno del computer virtuale, delle risorse
- gestione delle comunicazioni
  - interagisce con i controller della rete ottica per negoziare, allocare e riservare i percorsi ottici necessari. Da inoltre accesso alle nuove proprietà della rete ottica quali la

definizione di virtual private network (vpn) ottiche, multicast ottico e LambdaRAM con "remore memory access (RMA)"

- sicurezza
- esecuzione dei job

Da un punto di vista più strettamente fisico vanno anche notate le esperienze, ormai consolidate, di estensione di SAN su area geografica. Anche queste esperienze, opportunamente mediate da un servizio di astrazione, possono contribuire alla condivisione delle risorse necessarie per la definizione di un Tier virtuale.

Sono stati individuati, al momento, tre possibili casi di interesse:

- trasporto di Fiber Channel o SCSI su IP
  - e' il caso più semplice e, in teoria, già realizzabile 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 può 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 più 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 modalità. Crediamo questo sia un punto da studiare attentamente con gli esperti del GARR.
- File System Distribuiti
  - Questa e' un'opzione molto sofisticata e ancora poco chiara (almeno per quanto riguarda la scalabilità), 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'eventuale fattibilità tecnica di un suo utilizzo su GARR-X

3.1.2 Controllo remoto di apparati strumentali e acquisizione remota di dati

Lo sviluppo e l'adozione di sistemi di controllo remoto di apparati strumentali e/o industriali accessibili in rete geografica (WAN) è sempre stato frenato, oltre che da problemi di sicurezza, dalla difficoltà di avere caratteristiche di comunicazione in grado di garantire l'esecuzione di un dato 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 che caratterizzano le trasmissioni video con l'eccezione, nel caso del solo controllo, dell'ampiezza di banda che in questi casi solitamente non è un importante requisito.

La possibilità offerta da GARR-X (e della sua estensione a livello europeo) di stabilire percorsi ottici end-to-end di livello 2 garantendo su questi percorsi parametri di qualità minima (delay, jitter, loss) apre alla rete della ricerca lo scenario dei controlli remoti e del controllo di processo a distanza.

L'INFN è chiaramente interessata anche a questo tipo di utilizzo della rete avendo strumentazione remota (apparati sperimentali) sparsa sostanzialmente su tutto il pianeta. In particolare però una rete a parametri garantiti permette per la prima volta di controllare apparati e strumenti remoti come fossero in una rete locale. Questo dovrebbe estendere le possibilità di controllo a distanza di apparati che solitamente vengono controllati solamente in locale come i sistemi di acceleratori, sistemi di controllo e monitor di fascio, magneti, ecc.

I requisiti per i sopra menzionati parametri dipendono molto dal tipo di applicazione e dal tipo di controllo, però l'interesse è chiaramente rivolto a ordini di grandezza dell'ordine dei tempi di trasmissione e con un jitter molto piccolo. La perdita di pacchetti può non essere un parametro cruciale alla condizione che il protocollo di recupero dei pacchetti persi non infici il delay prefissato.

Un altro requisito importante è che il controllo di questi parametri di rete sia gestito dall'applicazione stessa (p.es.: Il middleware di controllo) con meccanismi semplici di accesso a

servizio (p.es.: SOA) messi a disposizione dal gestore della rete.

Un'altra caratteristica della rete che può essere sfruttata in questo contesto è la possibilità di effettuare multicast ottici su endpoints definiti. Questo può essere un meccanismo estremamente efficiente quando si vuole distribuire dati acquisiti da uno strumento o da un esperimento a più utenti contemporaneamente.

Il controllo di strumentazione prevede una continua interazione uomo-strumento attraverso quadri sinottici, interfacce grafiche, plots e charts, ecc. Il sistema uomo-strumento deve garantire tempi di interazione con lo strumento stesso almeno dell'ordine di quelli umani (frazioni di secondo). Questo si riflette ancora una volta nel parametro “delay” della rete. In questo caso però il “jitter” può avere valori intorno al 40-50 % continuando a garantire una buona interattività con lo strumento.

Nuove tecniche si stanno affermando (“teleimmersion”) nel campo dell'interazione uomo-macchina (strumento). Esse 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. Queste tecniche impattano principalmente sui parametri di “delay” e “jitter”, ma anche sull'ampiezza di banda. I requisiti per questi parametri sono tipicamente quelli per le applicazioni video.

3.1.3 Requisiti per le nuove modalità di uso della rete

Per fornire un'indicazione di come gli scenari descritti in precedenza possono incidere nella definizione dei parametri di qualità del servizi che la rete GARR-X potrebbe essere chiamata a offrire, riassumiamo nella tabella seguente i valori di riferimento relativi alle nuove modalità di utilizzo della rete.

Tab. 3.1: Requisiti derivanti da possibili nuovi utilizzi della rete GARR-X

	Banda	*Delay*	*Loss*
Applicazioni Real Time	**Dipende dall'applicazione**	**O(ms) + tempo di trasmissione**	**Non deve influire sul delay richiesto**
Virtual Reality (teleimmersion)	**O(10 Gbps)**	**O(ms) + tempo di trasmissione**	**Non deve influire sul delay richiesto**
Virtual Tier	**2/3 10 Gb/s per ogni tier**	**Non e' critico**	**Non deve influire sulla banda richiesta**