ISTITUTO NAZIONALE DI FISICA NUCLEARE

INFN/CCR-10/05
7 Settembre 2010

CCR-39/2010/P

VALUTAZIONE DI APPARATI DI AGGREGAZIONE E DI "EDGE ROUTING" PER I TIER2 INFN

A. Brunengo[1], A. De Salvo[2], D. Di Bari[3], G. Donvito[3], R. Gomezel[4], P. Lo Re[5], G. Maron[6], E. Mazzoni[7], M. Morandin[8], A. Spanu[2], S. Zani[9].

[1)] *INFN-Sezione di Genova, Via Dodecaneso 33 - 16146 Genova*
[2)] *INFN-Sezione di Roma, P.le Aldo Moro 2 - 00185 Roma*
[3)] *INFN-Sezione di Bari, Via E. Orabona 4 - 70126 Bari*
[4)] *INFN-Sezione di Trieste, Via A. Valerio 2 - 34127 Trieste*
[5)] *INFN-Sezione di Napoli, Complesso Univ. di Monte Sant'Angelo, Via Cintia - 80126 Napoli*
[6)] *INFN-Laboratori Nazionali di Legnaro, Viale dell'Università 2 - 35020 Legnaro (Padova)*
[7)] *INFN-Sezione di Pisa, Edificio C, Polo Fibonacci, Largo Bruno Pontecorvo 3 - 56127 Pisa*
[8)] *INFN-Sezione Padova, Via F. Marzolo 8 - 35131 Padova*
[9)] *INFN-CNAF, Viale Berti Pichat 6/2 - 40127 Bologna*

Abstract

Scopo di questo documento è di riportare ai siti ospitanti i centri di analisi LHC di secondo livello (Tier2) e ai relativi servizi calcolo il lavoro fatto dal gruppo NetArch della CCR nell'ambito dell'evoluzione a 10 Gbps della rete delle farm dei suddetti centri. Viene anche affrontata la problematica dell'accesso a 10 Gbps di questi centri T2 all'infrastruttura ottica della rete della ricerca denominata Garr-X.

Questo documento focalizza quindi l'attenzione sul centro T2 e sulle possibili architetture di rete in grado di connettere nei modi più efficaci il set di worker nodes con i relativi server di disco garantendo al tempo stesso un flusso duplex a 10 Gbps verso la WAN.

Il documento non si occupa invece della topologia di connessione tra i vari T2 con il T1 nazionale e con gli altri siti Tier internazionali. Queste informazioni sono riportate in "Proposta INFN per la rete dei Tier2 di LHC in GARR-X" (documento CCR-37/2010/P), prodotto sempre dal gruppo NetArch.

Published by ***SIS–Pubblicazioni***
Laboratori Nazionali di Frascati

1 ARCHITETTURA DI AGGREGAZIONE E REQUISITI DELLO SWITCH CORE DELLA FARM DEL T2

La fig. 1 illustra in modo schematico una tipica topologia di rete di un centro T2. Questa topologia ha diverse varianti implementative spesso dovute ad esigenze locali del centro o alla sua "storia", ma la tendenza a passare ad una aggregazione centrale di "core" con linee a 10 Gbps è confermata in tutte le sedi T2.

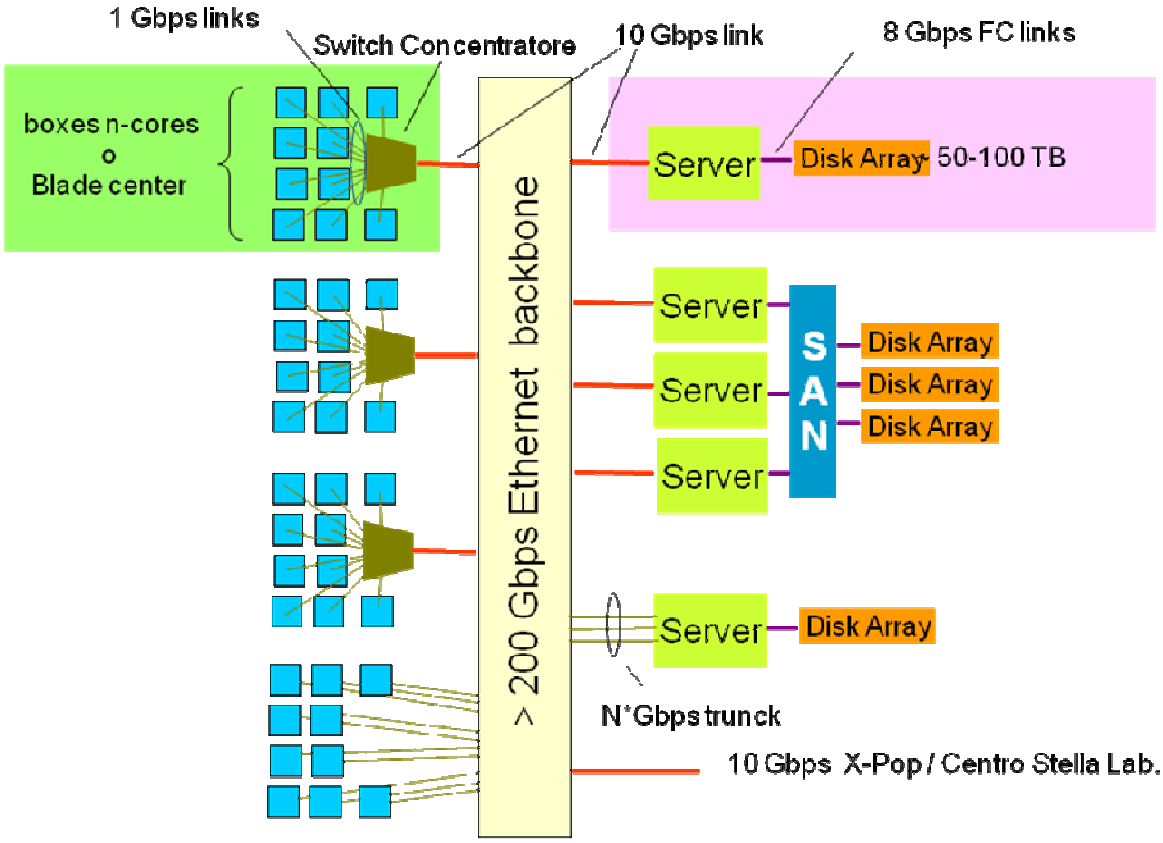

Fig.1: Esempio di rete per la farm di un centro T2

I requisiti dello switch di core variano in funzione della dimensione del T2 e quindi dal numero di esperimenti supportati dal centro. Dall'esperienza fatta e dalle previsioni fornite dagli esperimenti possiamo affermare che:

- la banda complessiva richiesta può variare da 150 a 300 Gbps a seconda che il sito abbia uno o due esperimenti;

- il numero di porte a 10 Gbps varia grosso modo da 24 a 48 sempre in funzione del numero di esperimenti supportati dal T2 (massimo 2 per INFN);

- Per la funzione di aggregazione è sufficiente uno switch di livello 2. Funzionalità di livello 3 sono richieste solo se lo switch è anche edge router verso Garr-X (vedi paragrafo seguente);

- Esistono core switch compatti (1 o 2 U) a 10 Gbps con al massimo una cinquantina di porte non blocking e con un numero significativamente maggiore se si accetta un over-subscription. Il range di prezzo è decisamente interessante se confrontato con gli apparati modulari. L'utilizzo di switch stand-alone con porte GE verso i WN ed uplink a 10 GE verso il core switch compatto permette una configurazione della rete con aggregazioni a livello di rack, semplificando il cablaggio al costo di una minore flessibilità sul posizionamento delle connessioni dei singoli WN;

- Una configurazione non blocking è chiaramente da preferire ad una che prevede un qualche grado di oversubscription. Come abbiamo visto, configurazioni fino a 50 porte non blocking sono possibili con gli switch compatti. Mettendo più switch in stack si supera questo limite, ma si introduce oversubscription che, in ultima analisi, limita la banda netta disponibile al nodo di calcolo o al server di disco. Va quindi analizzato attentamente il grado di oversubscription accettabile dal sistema. Mettendo due switch in stack avremo rischio di oversubscription solo quando le connessioni che arrivano al primo switch devono comunicare con quelle che si attestano nel secondo. Va quindi determinata la banda complessiva che in media deve passare dal primo al secondo switch e confrontata con quella permessa dal bus di comunicazione tra i due switch;

- La tecnologia Fiber Channel over Ethernet (FCoE) sembra molto promettente e di sicuro interesse per l'evoluzione delle reti dei T2 che hanno in questo modo la possibilità di convergere su un unico standard (ethernet) sia per le comunicazioni tra nodi che tra nodi e sistemi di storage (san). Gli switch utilizzati per FCoE devono però essere compliant ai protocolli ethernet di nuova generazione (Priority Flow Control – 802.1Qbb, Enhanced TRansmission Selection – 802.1Qaz e Congestion Notification & Congestion Management – 802.1Qau) che permettono appunto al FC di essere mappato su una rete ethernet. E' chiaro che uno switch di core che includa anche queste caratteristiche rappresenta un notevole valore aggiunto e un potenziale investimento per il futuro. Va tuttavia sottolineato che l'utilizzo di tale protocollo richiede che l'infrastruttura di rete lo supporti per tutta la catena di connessioni tra il disco e il disk server, che si rivalutino il numero di porte 10 GE necessarie, che si

valuti l'opportunità di separare logicamente o fisicamente l'infrastruttura Ethernet per la SAN da quella per IP; un tale approccio richiede quindi una riprogettazione complessiva della rete.

2 ARCHITETTURA DI ACCESSO A GARR-X E REQUISITI DI ROUTING

Ogni T2 INFN è inserito nella rete locale della sede che lo ospita. Con l'avvento di Garr-X ogni T2 avrà un flusso dedicato a 10 Gbps, separato dal traffico ordinario di 1 Gbps della sezione o laboratorio che lo ospita. Si pone quindi il problema dell'integrazione di questi due accessi.

Abbiamo individuato questi casi:

- Il flusso T2 e il flusso ordinario procedono all'interno della LAN di sezione in modo separato, hanno due router distinti che comunicano tra loro tramite una back door. Vedi fig. 2. In questo caso è anche possibile concentrare le funzionalità di router del T2 nell'apparato di core switch (livello 3). Vedi fig. 3.

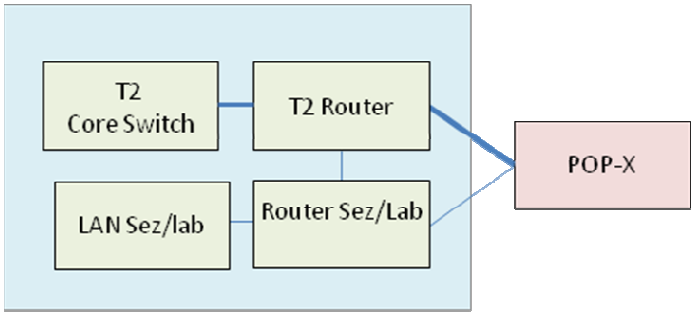

Fig. 2: Flusso T2 e ordinario con router separati per accedere al POP-X

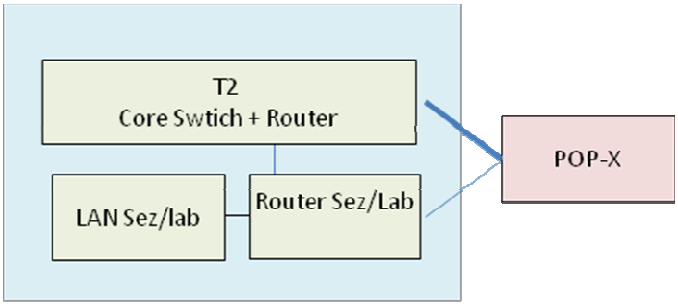

Fig.3: Come in Fig. 2 , ma con il core switch del T2 con funzionalità di router

- Il flusso T2 e il flusso ordinario confluiscono nel router di sezione o laboratorio che si occupa di incanalare i flussi verso il pop di riferimento su interfacce distinte, a 10 Gbps e a 1 Gbps. Vedi fig. 4.

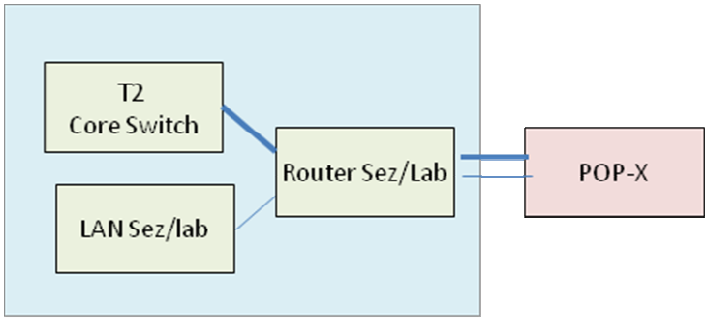

Fig. 4: Flusso T2 e ordinario confluiscono nello stesso router di sezione o laboratorio

- Per esigenze di ridondanza può anche essere possibile uscire con due flussi a 10 Gbps dove far passare sia il traffico T2 che quello ordinario. E' il caso per esempio delle sedi dislocate ad alcuni km dal POP-X (LNL e Pisa) e connesse ad esso in dark fiber dedicata.

Per quanto riguarda i requisiti degli apparati di routing va tenuto presente che l'accesso al POP-X e quindi a Garr-X sarà in tecnologia IP over ethernet (vedi il sopra citato documento "Proposta INFN per la rete dei Tier2 di LHC in GARR-X") con la definizione di VLAN separate per i percorsi verso il T1 nazionale e verso gli IP internazionali e quindi verso i centri Tier di altri Paesi.

Gli apparati di routing dei T2 non avranno quindi caratteristiche particolari se non quelle di essere dotati di un numero adeguato di interfacce a 10 Gbps ed essere in grado di gestire il protocollo BGP oltre a quelli ordinari di livello3. Nel caso di link in ridondanza verso il POP-X è richiesto il protocollo 802.1ag (ethernet OAM). La dimensione della tabella di routing è valutata essere di $O(10^3)$. L'insieme di queste caratteristiche definiscono una classe di apparati relativamente semplici che, pur richiedendo prestazioni "wire speed", non necessitano dei servizi e delle caratteristiche dei costosi router classe enterprise.

3 APPARATI VALUTATI

In base alle considerazioni qui sopra riportate e ai requisiti esposti, il gruppo NetArch ha preso in considerazione un insieme di apparati di aggregazione a 10 Gbps e di routing che crediamo di interesse per i centri T2. Riportiamo nella tabella che segue le caratteristiche

peculiari di questi apparati, rimandando all'appendice per i puntatori a descrizione più dettagliate.

Marca	Modello	Porte 10 Gbps	Aggregazione	Router per Garr-X	FCoE	Note
CISCO	4900M	24	Si	Si	No	
	45XX	4-44	Si	Si	No	
	Nexus 5000	20-40	Si	No	Si	
JUNIPER	MX80	4-6	No	Si	No	
	EX2500	24	Si	No	No	
Extreme	X650	24 >24 stack	Si	Si	No	Stack a max 512 Gbps
HP	84XX	48	Si	No	No	L3 senza BGP
	6600	24	Si	No	No	L3 senza BGP
Allied Telesyn	X908	16	Si	Si	No	

4 CONCLUSIONI

Gli apparati di aggregazione individuati rispondono ai requisiti descritti nei paragrafi precedenti e permettono ad un centro T2 di dotarsi di un core switch adeguato all'evoluzione delle risorse di calcolo e storage previste per i prossimi 3-4 anni. I più interessanti da un punto di vista rapporto prezzo/prestazioni sono gli apparati compatti a 1 o 2 U con un numero massimo di 50 porte a 10 Gbps, tutte non blocking. Alcuni di questi apparati inoltre offrono la possibilità di configurazione a stack con un bus da 128 a 512 Gbps. Questo dà la possibilità, in caso di necessità, di aumentare il numero di porte a scapito di un certo grado di oversubscription. Spesso questi apparati includono anche un L3 completo in modo da renderli adatti anche alla funzione di edge router per il T2. L'adozione di questi apparati permette un'aggregazione a livello di rack in modo da concentrare tutte le linee a 1 Gbps dei worker nodes in uno switch intermedio che poi con uno o più uplink a 10 Gbps si connette all'aggregatore centrale. Va notato anche che la tendenza attuale di aumentare il numero di core per macchina (parliamo già di una ventina di core per macchina nella generazione di

motherboard in uscita tra qualche settimana) porterà i worker node ad avere connessioni a 10 Gbps embedded nella motherboard, quindi in prospettiva ci aspettiamo una rete “flat” , cioè senza aggregatori intermedi, tutta a 10 Gbps.

Sono state anche individuate le caratteristiche minime che un apparato di routing deve avere per accedere a Garr-X a 10 Gbps e sono stati identificati gli apparati che sembrano più adatti a questo scopo. Ne risulta un apparato con esigenze prestazionali elevate, ma abbastanza convenzionale nei protocolli di routing standard, con una tabella di routing piccola e senza l’esigenza di servizi di rete sofisticati. Queste caratteristiche si possono facilmente trovare implementate su qualsiasi switch con una buona implementazione di L3 e non serve ricorrere agli apparati di classe enterprise che hanno, a parità di prestazioni, un range di prezzo di un ordine di grandezza superiore. Queste considerazioni sugli apparati di routing per un T2 aprono la possibilità di utilizzare, dove la cosa è possibile, lo stesso switch di aggregazione della farm T2 come edge router.

A seconda delle situazioni locali, il centro T2 può adottare varie configurazioni o combinando apparati di aggregazione e di routing o individuando un apparato che possa accorpare entrambe le funzioni o infine scegliendo un aggregatore che possa interfacciarsi al router di sede. Qualora il router di sede fosse inadeguato per l’accesso a 10 Gbps crediamo che gli apparati presentati possano essere dei buoni candidati anche per questo ruolo.

Su scala geografica, nel medio periodo, è ragionevole ritenere che un modello di collegamento basato su tecnologia 10GE sia sufficiente per i Tier2. Questo però può non essere del tutto vero per l'architettura della LAN che risulta essere molto più soggetta alle indeterminazioni del modello computazionale degli esperimenti. Alla luce di queste considerazioni la scelta di tenere separate le funzionalità di routing da quelle di concentrazione, ove economicamente possibile, potrebbe rivelarsi un’utile forma di protezione dell'investimento fatto.