

Unified Fabric

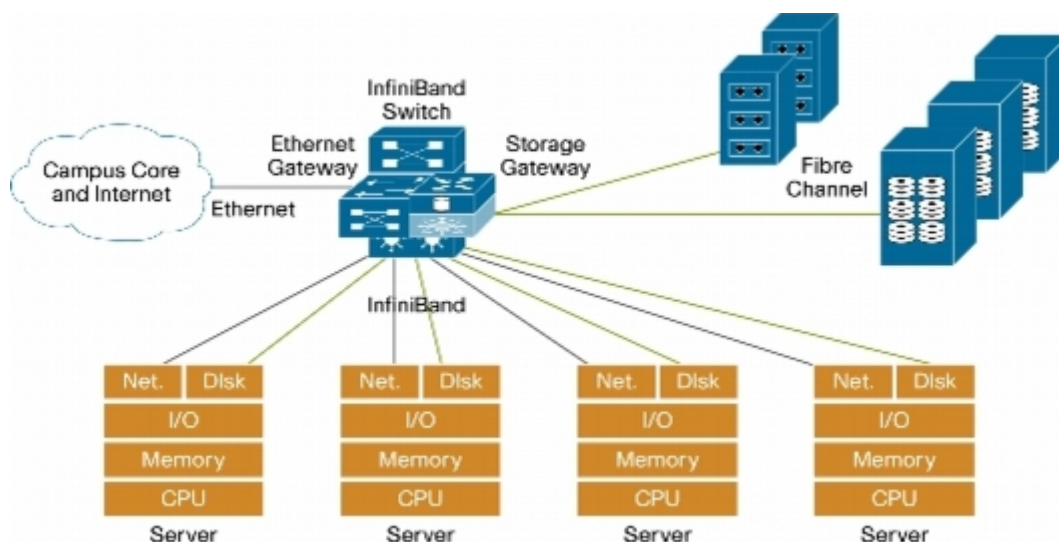
Innovazioni Cisco per le reti del Data Center

L'architettura Unified Fabric comprende nuovi concetti, come il Data Center Bridging dell'IEEE, che migliorano la robustezza di Ethernet riducendo il costo operativo e migliorandone la reattività abilitando il consolidamento delle reti. Cisco mantiene l'impegno nel supporto dello sviluppo di reti per il DC basate su Ethernet per fornire supporto ai requisiti delle nuove applicazioni.

Introduzione

Ethernet è la scelta predominante per l'interconnessione delle risorse all'interno del Data Center. E' universalmente diffusa e ben compresa dagli specialisti di rete, e sviluppatori. Ethernet ha resistito a prove nel tempo contro tecnologie sfidanti che tentavano di destituirlo dalla sua popolare posizione di rete per ambienti DC.

Esigenze emergenti da parte di applicazioni richiedono nuove capacità alle infrastrutture di rete, risultando nell'impiego di reti multiple, separate e dipendenti dalle applicazioni stesse. Comunemente all'interno dei DC vengono impiegate reti Ethernet per l'IP, una o due SAN per il traffico Fibre Channel in modalità di trasferimento a blocco, e spesso una rete InfiniBand per il cluster computing ad alte prestazioni. I costi operativi e d'investimento combinati per l'impiego e la gestione delle singole infrastrutture distinte, sono alti e creano opportunità per il consolidamento delle reti in una singola Unified Fabric.



Valutate tecnicamente le tre tipologie di rete, Ethernet risulta essere quella che riesce ad incontrare meglio le caratteristiche richieste dalle tre tipologie distinte, ma occorre che vengano aggiunte alcune funzionalità. L'IEEE con Data Center Bridging fornisce degli avanzamenti basati su standard ad Ethernet tali da renderla capace del consolidamento delle varie infrastrutture di rete.

Data Center Bridging dell'IEEE è una collezione di estensioni architetturali all'Ethernet progettate per migliorare ed estendere il ruolo della gestione e del networking Ethernet nel

DC. Ci sono due aspetti fondamentali presi in considerazione dal Data Center Bridging: estensioni all'Ethernet per il supporto del consolidamento delle operazioni di I/O su Unified Fabric, con separazione e mantenimento di distinte classi di traffico attraverso la Fabric; ed il supporto del servizio di trasporto "no-drop" così da poter garantire al traffico che lo richiede di poter essere trasportato in una Fabric di tipo "lossless".

Uno dei benefici economici del consolidamento delle reti è il risparmio sui costi. Una infrastruttura Ethernet omogenea è operativamente più semplice, attingendo all'abilità esistente dei tecnici di rete e risultando in impiego di minori strumenti di gestione e minor tempo di inizializzazione di nuove reti. In aggiunta, una rete consolidata del DC fornisce tutte le funzionalità esistenti delle reti di livello 2 che va a rimpiazzare. Per Ethernet ciò include il supporto del traffico multicast e broadcast, le VLAN, l'aggregazione dei collegamenti; per Fibre Channel include i servizi quali zoning e name server, ed il supporto delle virtual SAN (VSAN), il routing tra le VSAN.

Evoluzione della rete del Data Center

Ethernet continua ad avanzare e la rete del Data Center continua ad evolvere, modellata dal modo in cui le applicazioni utilizzano la rete come risorsa. Le richieste alla rete sono cambiate, essa non viene solo più usata semplicemente per il trasferimento di transazioni tra client e server. Ad esempio: l'impiego di server in clusters sta aumentando, incrementando il traffico tra i server. Anche il grid computing incrementa il traffico tra i server. L'incremento del traffico periodico di backup dei dati porta ad una crescita del traffico tra le server farm ed i dispositivi di storage sulle SAN. In aggiunta, le operazioni di backup serverless tra i dispositivi di storage sono molto comuni incrementando il traffico da disco a disco e da disco a nastro. Il traffico dei dati nel Data Center quindi si sta spostando dal client-server a server-server, server-storage e storage-storage.

L'incremento del traffico complessivo e il cambio negli schemi di scambio del traffico dei dati ha portato ad affidarsi maggiormente alla rete per ottenere la capacità di trattamento dei dati necessaria al supporto delle applicazioni in cluster di server. Le prestazioni delle applicazioni sono adesso misurate assieme alle prestazioni della rete, il che significa che latenza e banda sono entrambe rilevanti. Sono da considerare anche le differenze tra le tipologie di traffico trasmesso. Client-server e server-server, impiegano trasmissioni brevi e intermittenti, mentre la maggior parte del traffico server-storage o di applicazioni puramente storage, richiedono flussi di messaggi lunghi e continui. Ciò comporta che l'architettura di rete sia sufficientemente flessibile da poter supportare, scoprire e rispondere ai cambiamenti nelle dinamiche della rete.

Vanno anche considerate le differenze nelle capacità da parte delle applicazioni di trattare correttamente la perdita di pacchetti. Gli effetti della perdita dei pacchetti nel flusso di traffico tra le applicazioni ha effetti differenti sui vari protocolli, con applicazioni che quindi reagiscono in modo differente: alcune applicazioni possono tollerare la perdita dei pacchetti ed includono dei meccanismi di recupero e ritrasmissione. Ethernet supporta questi casi specifici, ma altre applicazioni non possono tollerare la perdita dei pacchetti richiedendo quindi ai mezzi trasmissivi la consegna garantita del traffico impiegando servizi di trasmissione "no-drop". Il traffico Fibre Channel trasportato su Ethernet ne è un esempio. Occorre per le reti Ethernet stabilire un metodo per implementare il servizio di tipo lossless

in grado di supportare adeguatamente tali tipologie di applicazioni. Ciò avviene con l'adozione del Data Center Bridging dell'IEEE.

Nel Data Center occorre inoltre che vengano accomodate topologie piatte e larghe, basate su domini di livello 2 estesi. Ciò è dovuto alla continua espansione delle reti a supporto del Data Center, con la continua aggiunta di apparati di switching e di server.

Altre opzioni per il consolidamento delle reti

Ethernet non è l'unica opzione per il consolidamento delle reti nel Data Center. Ma ha la maggior possibilità di aver successo se comparata alle altre tecnologie. Il traffico FC richiede un trasporto affidabile con servizio di tipo no-drop, senza perdita di pacchetti durante i periodi di congestione della rete. Una sfida per Ethernet era quella di garantire tale servizio per il trasporto del traffico FC. PFC, priority-based flow control abilita i meccanismi di trasmissione lossless in Ethernet, ed FCoE, Fibre Channel over Ethernet, abilita al trasporto del traffico FC in Ethernet.

iSCSI

Small Computer System Interface su IP (iSCSI), basato su Ethernet è stato considerato un rimpiazzo del Fibre Channel, per permettere il consolidamento del trasferimento del traffico storage a blocchi su Ethernet. Sebbene iSCSI continui ad essere popolare in molte applicazioni storage, in particolare in reti di piccole e medie dimensioni, non sta sorpassando o rimpiazzando la vasta diffusione dell'adozione di Fibre Channel per l'impiego in trasferimenti critici di dati in applicazioni storage in ambito enterprise. Una causa è la scarsa volontà di affidare il trasferimento dei dati critici ad un'infrastruttura Ethernet che non possa garantire un servizio di tipo no-packet-drop. Una seconda ragione per cui iSCSI non ha sostituito FC è che iSCSI non supporta i servizi nativi che gli amministratori delle SAN sono abituati ad utilizzare con l'utilizzo di FC.

Si può pensare che l'introduzione su Ethernet della capacità trasmissiva a 10 Gigabit possa consentire a iSCSI di sostituire FCoE su base puramente prestazionale. Secondo questa prospettiva, i 10 Gbps aggiungono comunque maggior velocità anche quando viene trasportato il traffico FCoE. L'affidabilità e l'integrità dei dati sono in per il Fibre Channel molto più importanti della velocità. Per questa ragione, l'implementazione di una rete Ethernet di tipo lossless è interessante. Infine, un'altra sfida per l'iSCSI è stato l'incremento nel carico imposto dall'utilizzo del TCP/IP sulle CPU dei server. iSCSI si affida al TCP/IP per la realizzazione del trasporto ordinato e affidabile del traffico storage, Anche se sono stati implementati dei motori di offload del TCP sulle schede di rete, ma è un approccio che aumenta il costo delle schede di rete richiedendo specifici circuiti integrati ASIC.

InfiniBand

Anche InfiniBand è stata posizionata come tecnologia potenzialmente candidata al consolidamento delle reti nel DC. Sebbene InfiniBand fornisca dei gateways verso reti FC e Ethernet, richiede comunque la costruzione di una nuova rete parallela, e con l'alta penetrazione delle reti Ethernet nei DC è improbabile che i dipartimenti di IT spostino le loro infrastrutture basate su Ethernet su InfiniBand; non sarebbe efficiente dal punto di vista del costo in quanto comporterebbe una costruzione incrementale con conseguente ulteriore sforzo di gestione, ed inoltre non potrebbe essere operativamente supportata senza l'ausilio di formazione estensiva degli staff tecnici che normalmente lavorano con IP ed Ethernet. Un

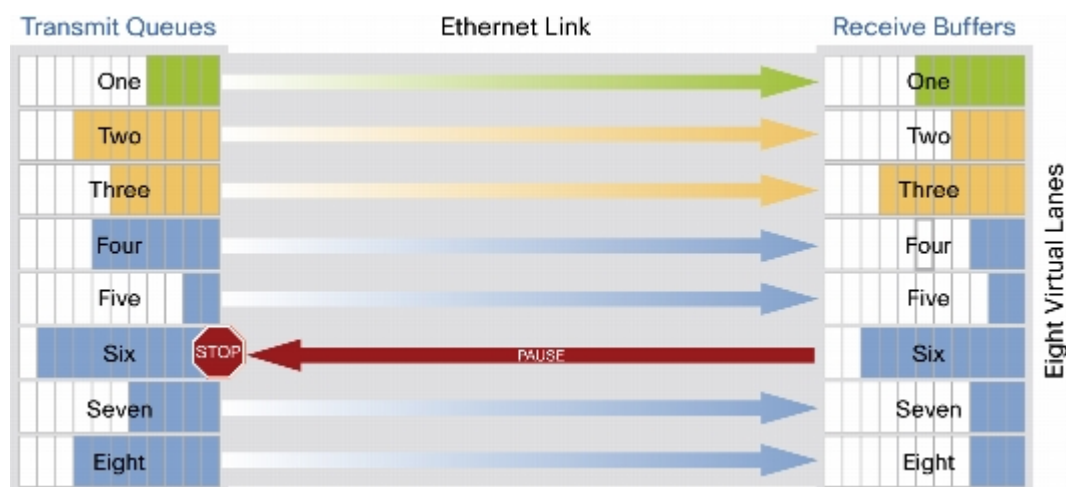
altro ostacolo che InfiniBand deve affrontare è la scarsità di opzioni per la connettività tra le varie sottoreti InfiniBand, l'interconnessioni delle sottoreti richiede l'impiego di router per InfiniBand oggi praticamente inesistenti. Considerando che almeno un 80% di tutti i cluster di server oggi è formato attraverso le infrastrutture Ethernet, c'è una buona probabilità che Ethernet possa essere adattata ai requisiti necessari per servire una ancor maggiore percentuale di cluster di server e di applicazioni di grid computing. La creazione di drivers per il supporto di RDMA (Remote Direct Memory Access) su 10 Gigabit Ethernet è uno sviluppo che sembra inevitabile. E' necessario disporre di connessioni Ethernet a bassa latenza ed alta capacità trasmissiva direttamente realizzate tra le risorse di memoria dei computer. Le classi di trasporto in modalità lossless del Data Center Bridging IEEE, diventano vantaggiose anche per le applicazioni su cluster di server.

IEEE Data Center Bridging

Il Data Center Bridging dell'IEEE è stato ponderato per trarre vantaggio dai punti di forza dell'Ethernet classica, aggiunge molte estensioni cruciali per fornire l'infrastruttura di prossima generazione per la rete del DC. Il resto di questo documento delinea le componenti del DCB e ne descrive le funzionalità di ciascuna e come queste contribuiscano alla creazione di una architettura Ethernet robusta al fine di incontrare le moderne e crescenti necessità delle applicazioni e rispondere a future esigenze del Data Center.

Priority-based Flow Control: IEEE 802.1Qbb

Per il consolidamento dell'I/O la corretta condivisione del collegamento è cruciale. Affinché la condivisione di collegamento sia efficace le grosse trasmissioni intermittenti di una tipologia di traffico non devono incidere su altri tipi di traffico, grosse code di traffico di un tipo non devono strozzare il traffico di risorse di altro tipo, e l'ottimizzazione di un tipo di traffico non deve creare eccessiva latenza per i messaggi brevi di traffico di altra natura. Il meccanismo di Pause dell'Ethernet può essere utilizzato per controllare gli effetti di un traffico sugli altri. Il Priority-based Flow Control (PFC) è una miglioria del meccanismo di Pause.

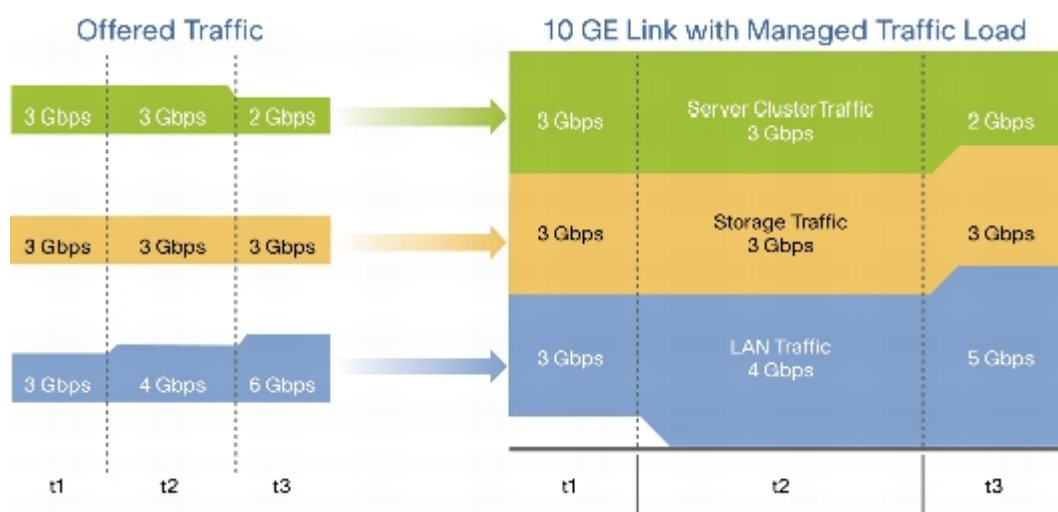


L'attuale meccanismo di Pause ferma tutto il traffico sul collegamento; impone una pausa trasmissiva riguardante tutto il collegamento. Il PFC crea otto collegamenti virtuali separati sul singolo collegamento fisico e permette a ciascuno di essi di essere fermato e ripartire indipendentemente. Questo approccio permette alla rete di creare delle classi di servizio di tipo no-drop per collegamenti virtuali che possono coesistere con altre tipologie di traffico sulla stessa interfaccia.

Il PFC permette politiche di gestione della QoS (quality-of-service) differenziate per gli otto collegamenti virtuali. Gioca anche un ruolo determinante quando funziona come arbitro dentro la switch fabric per collegare le porte in ingresso con quelle d'uscita.

Enhanced Transmission Selection: IEEE 802.1Qaz

Il PFC crea otto tipi distinti di collegamento virtuale su un collegamento fisico, e può essere quindi vantaggioso avere delle classi di traffico definite distintamente entro ogni collegamento virtuale. Il traffico all'interno della stessa classe IEEE 802.1p può essere raggruppato insieme e ancora trattato differentemente all'interno di ogni gruppo. ETS, Enhanced Transmission Selection, fornisce l'elaborazione prioritizzata e basata sull'allocazione di banda, con bassa latenza o best-effort, risultando nell'allocazione delle classi di traffico per gruppo. Estendendo il concetto di virtual-link, la scheda NIC (network interface controller), fornisce code virtuali di interfaccia: una per ciascuna classe di traffico. Ogni coda virtuale di interfaccia è responsabile della gestione della propria banda assegnata per il proprio gruppo di traffico, ma ha la flessibilità all'interno del gruppo di gestire dinamicamente il traffico. Ad esempio, il collegamento virtuale 3 per la classe di traffico IP può avere un designazione ad alta priorità e di tipo best-effort all'interno della stessa classe di servizio, ottenendo che la classe del collegamento virtuale 3 condivida una percentuale di collegamento totale con le altre classi di traffico. L'ETS permette la differenziazione tra il traffico della stessa classe di priorità, creando così i gruppi di priorità.



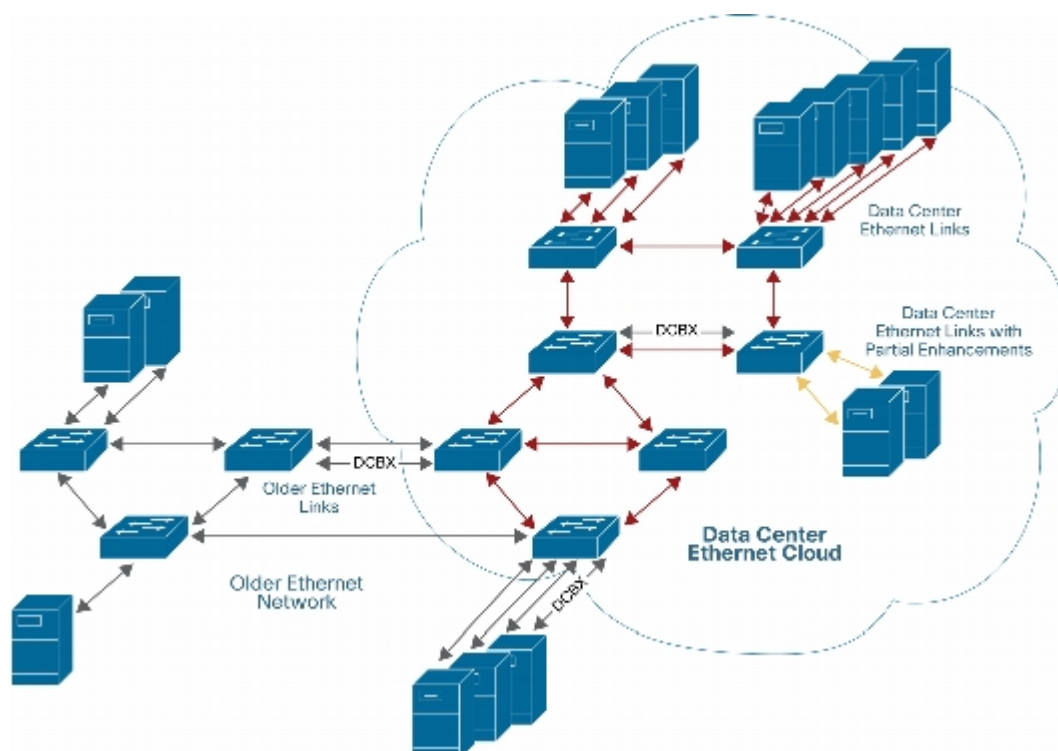
Nei meccanismi classici implementati secondo l'IEEE 802.1p viene definita una schedulazione rigida delle code in base alla priorità. Con l'ETS, una schedulazione flessibile, esente da perdita di pacchetti, per la coda può prioritizzare il traffico in base alle classi definite dal 802.1p nonché alla gerarchia di trattamento del traffico definita all'interno di ciascun gruppo di priorità. La capacità di applicare trattamenti differenti a traffico

differenziato all'interno della stessa classe di priorità è abilitata dall'implementazione del ETS.

Data Center Bridging Exchange Protocol

Il DCBX, Data Center Bridging Exchange Protocol è un protocollo di scambio di scoperta dei dispositivi e scambio dei parametri sviluppato da Cisco, Nuova e Intel e viene utilizzato nell'ambito dell'architettura IEEE Data Center Bridging per scoprire i dispositivi vicini e scambiare informazioni di configurazione tra essi. I valori dei parametri elencati di seguito possono essere oggetto di scambio mediante il protocollo DCBX:

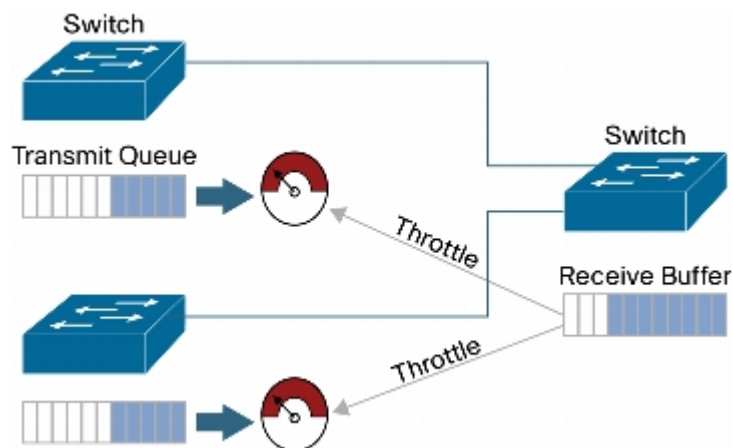
- Priority groups in ETS
- PFC
- Congestion Notification
- Applications
- Logical link-down
- Network interface virtualization



Congestion Notification: IEEE 802.1Qau

La tecnica di congestion notification è parte della gestione del traffico e spinge la congestione ai bordi della rete indicando a dei limitatori del traffico di effettuare operazioni di shaping del traffico che provoca la congestione. Il gruppo di lavoro del IEEE 802.1Qau ha accettato la proposta di Cisco per la gestione delle notifiche di congestione, che definisce un'architettura per la gestione attiva dei flussi di traffico per evitarne il blocco trasmissivo.

La congestione viene misurata nel punto di congestione e se rilevata, operazione di limitazione della velocità oppure di back-pressure vengono imposte nel punto di reazione per effettuare shaping del traffico e quindi ridurre l'effetto della congestione sul resto della rete. In questa architettura uno switch di livello d'aggregazione può inviare trame di controllo a due switch di livello d'accesso, chiedendo loro di regolare l'invio del loro traffico. Questo approccio mantiene l'integrità del core della rete e incide solo sulle parti della rete che causano la congestione, verso il punto sorgente del traffico.



Standard relativi a Unified Fabric

In aggiunta al Data Center Bridging di IEEE, gli switch Cisco Nexus per il DC includono altri miglioramenti come il layer 2 multi-pathing basato su standard e FCoE, Fibre Channel over Ethernet, così come l'implementazione della fabric di tipo lossless al fine di consentire la realizzazione di una Unified Fabric per il consolidamento delle reti e dell'I/O del Data Center.

Layer 2 Multi-pathing

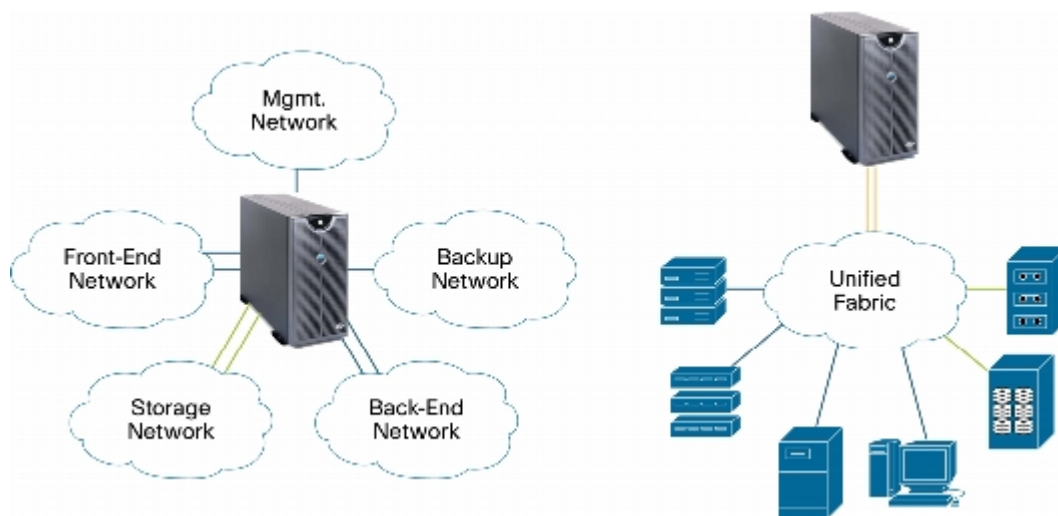
L'instradamento su percorso multiplo con ugual costo viene realizzato tramite i protocolli di livello 3. Le organizzazioni per le standardizzazioni stanno proponendo diverse alternative per la realizzazione della funzionalità ma a livello 2. TRILL, Transparent Interconnection of Lots of Links, è una soluzione proposta dall'IETF, mentre l'IEEE propone il SPB, Shortest-path Bridging, IEEE 802.1aq.

L2MP, layer-2 multipath, aumenta la banda bidirezionale abilitando molteplici percorsi paralleli tra i nodi, ottenendo più capacità trasmissiva nella rete di interconnessione con una minor latenza. In base ai modelli di traffico indotti dall'impiego di grosse server farm, il L2MP aumenta le prestazioni delle reti. Il bilanciamento del carico attraverso percorsi alternativi con ugual costo migliora le prestazioni applicative e anche la resilienza della rete. Impiegando il layer 2 multi-path e l'architettura Data Center Bridging si vanno ad utilizzare tutti i collegamenti disponibili tra i nodi di rete, si elimina l'utilizzo dello Spanning Tree ed i vincoli che esso impone, e si consente agli operatori del Data Center di effettuare modifiche di topologia anche dinamicamente senza preoccupazioni degli effetti della convergenza causata dalla rimozione o aggiunta di percorsi di collegamento.

I/O Consolidation

La continua espansione di Ethernet a 10 Gigabit supporta il messaggio di tipi diversi di traffico tra i server e le reti composte da switch. Le estensioni di Ethernet introdotte dall'IEEE con il Data Center Bridging (PFC, ETS, DCBX e Congestion Notification), abilitano una connessione Ethernet a 10 Gigabit al supporto simultaneo di più tipologie di traffico, preservando i rispettivi trattamenti del traffico. Con queste estensioni lo stesso collegamento Ethernet a 10 Gigabit viene anche utilizzato per trasportare il traffico storage del Fibre Channel, offrendo funzionalità di trasporto no-drop al traffico FCoE. Il consolidamento dell'I/O sul server che supporta FCoE permette agli host di avere accesso alle risorse di storage su una stessa fabric unificata.

I cambiamenti nelle architetture dei server stanno influenzando l'adozione di una fabric unificata. In particolare l'adozione di PCI Express (PCIe), ha permesso ai server di superare il collo di bottiglia di I/O del bus PCI. Questo cambiamento fondamentale permette ai server di utilizzare appieno le capacità trasmissive dell'interfaccia Ethernet a 10 Gbit. Allo stesso tempo, i server stanno impiegando chips a più alta densità, quad cores e piattaforme multiprocessore, con il risultato di incrementare la richiesta di capacità trasmissiva in ingresso ed in uscita dai server. Con processori multipli, cores, e macchine virtuali esistenti su singoli server, Ethernet a 10 Gbit verrà più ampiamente adottata, ed un metodo per gestire molteplici tipologie di traffico simultaneamente diventerà essenziale per permettere la condivisione del traffico su connessioni I/O consolidate.



Un collegamento di I/O consolidato può presentare alla fabric unificata traffico multiprotocollo su un singolo cavo. Una fabric unificata realizza un trasporto singolo, multiscopo, in Ethernet, che permette il transito di IP e Fibre Channel contemporaneamente attraverso la stessa interfaccia e lo stesso sistema di switching, mantenendo le caratteristiche di differenziazione introdotte dalle classi di servizio. Casi d'impiego includono il trasporto multiprotocollo, FCoE, e il Remote Direct Memory Access (RDMA) su ethernet a bassa latenza.

Lossless Fabric

Sebbene non sia parte della definizione dell'IEEE con il Data Center Bridging, uno switch per il Data Center deve essere in grado di implementare una architettura lossless per

assicurare che la trasmissione della classe di servizio lossless non perda una trama. Per supportare FCoE una fabric di tipo lossless è quindi obbligatoria per assicurare al traffico storage trasmissioni secondo le modalità di un servizio no-drop. Per la creazione di una fabric Ethernet lossless con supporto multiprotocollo, sono richiesti due elementi: un meccanismo di pausa basato su priorità (PFC), ed un meccanismo intelligente di arbitraggio all'interno della fabric di switching che leghi le porte del traffico in ingresso con quelle d'uscita per onorare i requisiti imposti dai meccanismi di pausa e di controllo del flusso.

Il PFC, così come il meccanismo di Pause dell'Ethernet, è in grado di rendere lossless il collegamento, ma non è sufficiente a rendere lossless la rete. Occorre un metodo per legare il meccanismo di pausa delle porte in ingresso alle risorse delle porte in uscita attraverso la fabric interna dello switch utilizzando il PFC. Ogni switch deve associare le risorse del collegamento in ingresso a quelle del collegamento in uscita, realizzando tra loro un collegamento logico che lega la disponibilità delle risorse trasmissive in uscita alle porte di ingresso del traffico creando l'arbitraggio necessario a permettere che il traffico possa essere trasmesso senza perdita di trame. In questo modo, con questo meccanismo intra-switch basato sul comportamento lossless dell'Ethernet si riesce ad emulare il sistema di gestione basato sul sistema dei crediti di buffer impiegato negli switch FC.

Fibre Channel over Ethernet (FCoE)

Per trasportare il traffico storage Fibre Channel o altre applicazioni che richiedono servizi lossless su rete Ethernet e quindi ottenere una fabric unificata è richiesta una classe di servizio lossless. Il traffico Fibre Channel richiede funzionalità trasmissive no-drop. Una classe di servizio no-drop può essere creata utilizzando l'architettura IEEE Data Center Bridging impiegando switch con fabric di tipo lossless Ethernet.

Il protocollo FCoE mappa le trame Fibre Channel native su Ethernet, indipendentemente dallo schema di inoltro nativo dell'Ethernet. Permette un approccio evolutivo al consolidamento dell'I/O mantenendo la composizione originale del Fibre Channel.

Lo standard FCoE è uno sviluppo del comitato T11 del INCITS (organismo internazionale dell'ANSI), in questo ambito è definito che il trasporto del FC su Ethernet richieda una implementazione lossless dell'Ethernet. Quindi occorre garantire sia l'impiego dei meccanismi di pausa e di controllo del flusso, come il PFC, oltre che l'impiego di switch in grado di correlare i buffers delle porte collegate ai link su cui il traffico entra con i buffer delle porte in uscita e gestirle legandole tra loro con i meccanismi di controllo del flusso e di pausa del traffico.

E' una funzionalità che viene implementata da Cisco sulle macchine della serie 5000 Nexus adesso, ed in futuro sulla serie Nexus 7000.

Conclusione

La fabric unificata fornisce l'architettura che risponde alla promessa fatta dalla visione di Cisco del Data Center 3.0. Ethernet è la scelta ovvia per una singola rete convergente che può supportare differenti tipi di traffico. Ethernet può contare su una vasta base di esperienza operativa e di ingegnerizzazione a livello globale derivante dalla sua presenza costante nei data center di tutto il mondo. FCoE è il primo caso d'impiego di una fabric unificata. Le estensioni del Data Center Bridging del IEEE forniscono funzionalità Ethernet di tipo lossless che rispondono ai requisiti di trasmissione no-drop del Fibre Channel.

La creazione di servizi lossless su fabric basata su Ethernet avvantaggia FCoE, RDMA, iSCSI e applicazioni come il video real-time al fine di ottenere consegna garantita del traffico da parte di collegamenti virtuali di tipo no-drop, miscelati con altre applicazioni contendenti. L'innovazione basata sul L2MP avvantaggia ogni rete di Data Center a prescindere dal fatto che si sia fatto convergere su di essa il traffico della SAN. La capacità trasmissiva incrementata con diminuzione della latenza grazie all'impiego di percorsi multipli paralleli, frutta un guadagno prestazionale per tutte le applicazioni. Con le nuove estensioni architetturali delle reti dei Data Center, i dipartimenti di IT ottengono svariati benefici: un nuovo e flessibile metodo di consolidamento dell'I/O su Ethernet sulla stessa fabric di rete, in contrapposizione al supporto di differenti reti separate; un metodo per il trasferimento del traffico in modalità lossless; più efficace utilizzo della capacità trasmissiva della rete grazie all'impiego del multi-pathing di livello 2.

Per maggiori informazioni

- **IEEE Data Center Bridging:** <http://www.cisco.com/en/US/netsol/ns783/index.html>
- **Priority Flow Control IEEE 802.1Qbb:**
<http://www.ieee802.org/1/pages/802.1Qbb.html>
- **Enhanced Transmission Selection IEEE 802.1Qaz:**
<http://www.ieee802.org/1/pages/802.1Qaz.html>
- **Congestion Notification IEEE 802.1Qau:**
<http://www.ieee802.org/1/pages/802.1Qau.html>
- **DCBX explanation:** <http://www.ieee802.org/1/files/public/docs2008/az-wadekar-dcbcxp-overview-rev0.2.pdf>
- **IETF Transparent Interconnection of Lots of Links (TRILL):**
<http://www.ietf.org/html.charters/trill-charter.html>