NVMe Over Fabric: Cosa c’e’ di nuovo?

Cosa è e che differenza c’e’ tra NVMe over Fabric e SCSI device

Nel mondo dell’IT non c’è cosa peggiore dell’utilizzo degli acronimi e dell’utilizzo degli stessi in modo scorretto. Questo spesso genera delle vere e proprie leggende metropolitane se non veri e propri fraintendimenti. A cominciare dal vecchio e caro **FC (Fiber Channel)** usato insieme all’acronimo **SCSI (Small Computer System Interface)** per indicare tutto, la rete, la SAN, il protocollo, il tipo di connettività. Insomma, diciamo FC ed indichiamo l’universo della rete di accesso allo storage.

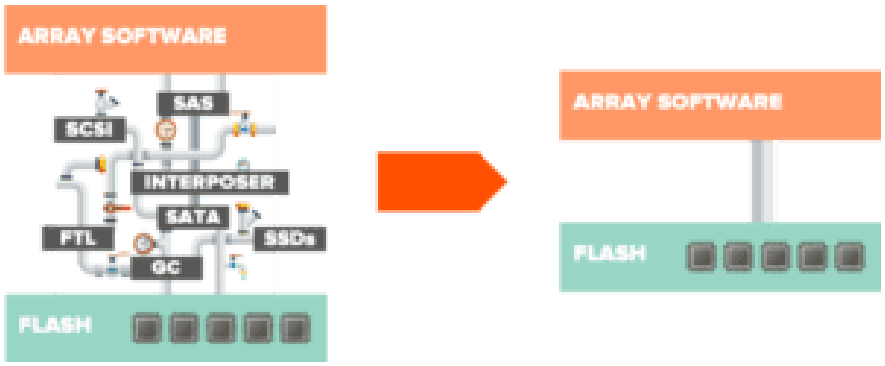
Poi abbiamo lo SCSI, il protocollo seriale per eccellenza, l’ultimo baluardo del modo tradizionale di pensare allo storage. Tanto per chiarire quanto sia *legacy* lo SCSI ricordo che supporta comandi come REWIND, READ REVERSE, SEEK ossia assolutamente associati ad un mondo di dischi rotativi e di dischi magnetici in cui, il posizionamento delle testine era la parte più pregiata da attuare.

Ad oggi stiamo utilizzando storage super performanti con un protocollo pensato per dischi la cui migliore proprietà non era certo la prestazione ma l’affidabilità meccanica. Ora dobbiamo usare protocolli di accesso

che concentrino tutto il lavoro sulla ricerca di prestazione.

Lo SCSI serviva, e serve ancora, per far sì che un *device* (**l'host o server**) inviasse comandi ad un altro *device* (**storage array**) per recuperare informazioni secondo un modello di allocazione dei blocchi tipico dei dischi (Volume, Traccia, Settore).

Lo schema che segue semplifica l'idea.



Abbiamo bisogno di eseguire questo salto senza influire sul risultato applicativo.

Il mercato quindi ha chiesto, e l'industria ha risposto con uno standard nuovo per far sì che i *device host* possano indirizzare lo spazio di memoria in modo univoco. Si passa, quindi da un modello di richiesta / risposta ad un modello di accesso allo spazio di indirizzamento. Dall'utilizzo di una coda di comandi unica ad un modello che permette di indirizzare fino a 64.000 code parallele, ognuna in grado di trasportare 64.000 comandi, e soprattutto da un protocollo nato per i dischi (HDD) ad uno che lavora ed è pensato solo su memorie (flash).

NVMe è uno standard, over Fabric è la modalità con cui i comandi NVMe vengono trasportati su una rete.

Il disegno del protocollo rende indipendente il protocollo stesso dalle reti sottostanti: Ethernet, FC, Infiniband, fino ad arrivare al TCP (Trasmission Control Protocol).

Si tratta di una vera rivoluzione sostenibile.

Cosa cambia nel datacenter con l'adozione di NVMe oF?

Perché la definiamo una rivoluzione sostenibile? Perché il modo in cui i vendor (Pure in primis) prevedono l'adozione dell'NVMe, è graduale, non invasivo e in modo da essere contiguo con le attuali tecnologie, permettendo un'adozione per passi successivi.

Quindi alla domanda: cosa cambia in un datacenter? La risposta potrebbe essere "niente" in prima istanza e "tutto" in seconda battuta. Niente in termini di impatto. È possibile prevedere l'adozione dell'NVMe beneficiando della tecnologia per migliorare l'accesso al flash. Quindi confinando l'adozione agli array di storage.

Tanto per svelare un segreto, tutti i clienti FlashArray (FA) di Pure beneficiano di questa

tecnologia praticamente da sempre.

La prima edizione di un FA con NVMe risale alla serie //M (2015), e luglio 2018 per la fornitura di un sistema totalmente NVMe ossia la serie //X. La serie //X è quella che i nostri clienti hanno ricevuto da quella data in poi anche non chiedendo esplicitamente NVMe.

Tornando agli impatti, possiamo solo commentare e far notare che l'adozione end to end della tecnologia NMVe anche per la parte over Fabric (ossia di trasporto) permette di traguardare **latenze nell'ordine dei microsecondi** (100 ca). Permette quindi di ragionare end to end in un ordine di grandezza simile a quello dei sistemi DAS (Direct Attached Storage). Permette quindi di disegnare e architettare il proprio data center in una ottica di consumo di risorse distribuito invece che pensarlo in ottica di colli di bottiglia che devono essere ottimizzati.

Si passa da un disegno che deve prevedere forme di tuning, ad un disegno che prevede il massimo delle prestazioni, sempre ed in ogni condizione.

CPU e RAM / Rete / Storage

Cosa succede nell'ambito delle infrastrutture di Data Center? Come cambia e se cambia la modalità di acquisto di CPU, oppure della RAM, della rete e anche dello stesso storage?

Il cambiamento consiste nel modo in cui si potrà distribuire il carico di lavoro.

Oggi, solitamente, nel disegnare la parte di computing di una infrastruttura si considerano fattori determinanti per il dimensionamento:

1. Carico di I/O;
2. Quantità gestita di RAM per unità di CPU;
3. E solo in ultimo si considera il carico elaborativo.

Per quanto riguarda la parte di rete, ad oggi sono disponibili modelli architetturali che permettono di collassare nella stessa fabric sia le efficienze necessarie per il traffico nord-sud che quello est-ovest in un datacenter. Nonostante questo, i requisiti per il trattamento della parte di Storage e di Backup sono sempre gestiti in modo separato.

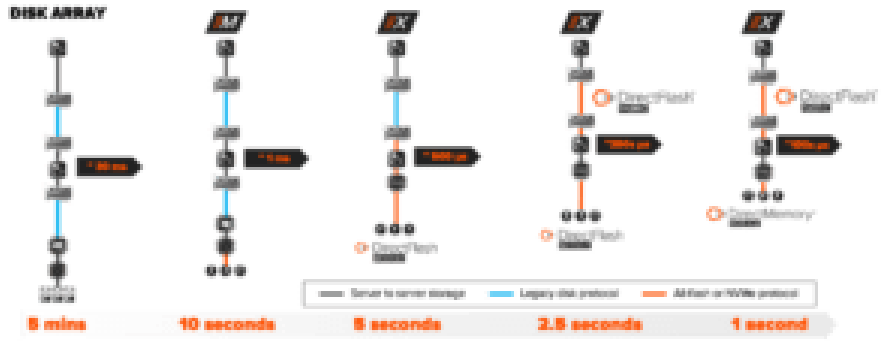
I parametri di dimensionamento di una rete di storage erano:

1. Fan In / Fan Out (ratio)
2. Banda
3. Latenza

Adottando NVMe over Fabric, in fase di disegno, il data center si semplifica e migliora nella distribuzione dei carichi di lavoro per quanto riguarda la componente I/O.

Adottando NVMe over Fabric possiamo, finalmente, lasciare ai server la parte computazionale e allo storage la parte di memorizzazione estraendo il meglio dai due mondi, attraverso una rete efficiente ed efficace sia essa FC, Ethernet o qualsivoglia.

Questo bilanciamento delle fasi di I/O produce diretti benefici misurabili nel data center.



L’adozione dell’NVMe prima a livello di Array e poi a livello di fabric (over Fabric) porta sostanziali benefici prestazionali avvicinando le prestazioni all’ordine di grandezza di accessi DAS.

In primo luogo, nella piena libertà di ogni IT Manager, la rete Storage diventa un *overlay* o meglio diventa indipendente dalla Fabric sottostante. Questo ovviamente semplifica e crea un miglior modello di gestione. In secondo luogo, i driver I/O diventano meno fondamentali nella ricerca delle prestazioni. Il carico di lavoro viene fortemente bilanciato tra le varie componenti del data center.

La difficoltà dell’adozione sta nel pieno lavoro della filiera: dal sistema operativo in giù. L’intero ecosistema deve collaborare all’adozione di un’architettura NVMe over Fabric.

Finisce qui? Tutto qui? Cosa ci aspetta oltre il 2020 per l’adozione di queste tecnologie?

Ci aspetta un vero salto quantico. L’adozione dell’NVMe over Fabric over TCP. L’utilizzo del TCP come protocollo per il trasporto dell’NVMe è stato adottato nel 2018 all’interno dell’iniziativa NVMe. Ancora meno vincoli a livello di fabric (qualche implicazione sugli MTU per bilanciare gli IOPS) ma comunque un ulteriore livello di libertà per il disegno dei moderni data center.

Siamo finalmente arrivati al momento della totale evoluzione del data center: **come per tutti i cloud provider, i server si possono concentrare sulla fase computazionale, lo storage a memorizzare i dati e la rete a massimizzare i throughput e le latenze.**

Finalmente sarà possibile disegnare i data center massimizzando i ritorni degli investimenti senza imbarcarsi in strane alchimie tra computing, storage e rete.

Il futuro è agile e soprattutto semplice.

ALFREDO NULLI// Principal System Engineer office of CTO, EMEA | Pure Storage

T: @alfcloud