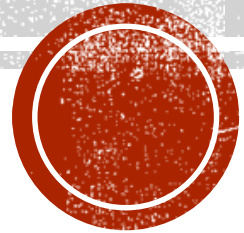
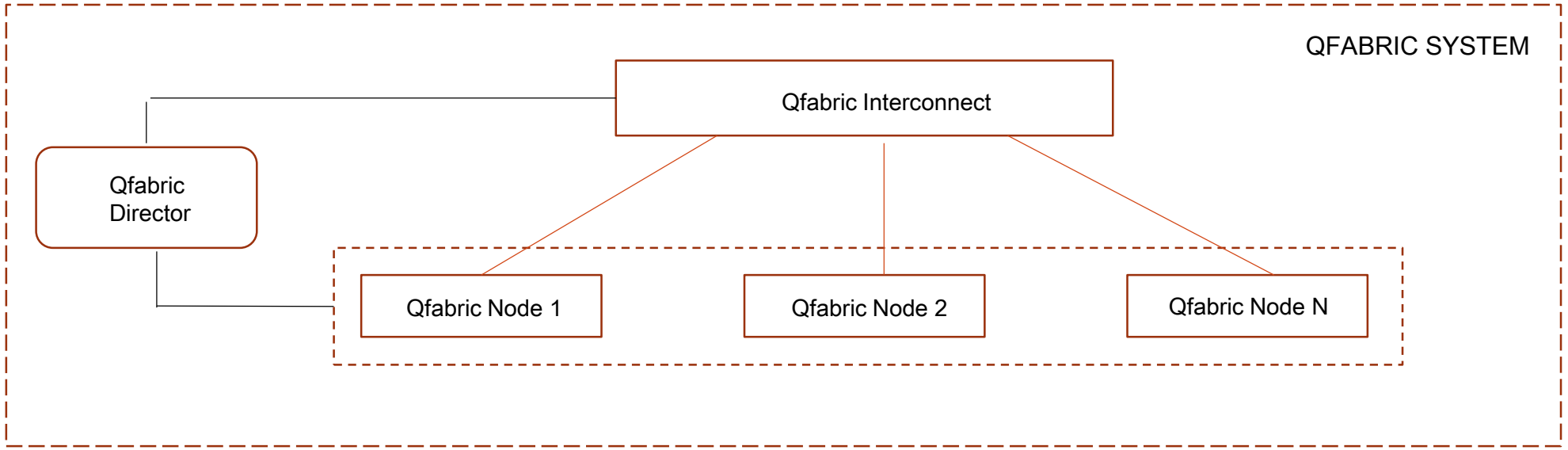
QFABRIC

Massimiliano Sbaraglia



QFABRIC COMPONENTS

- Node:
  - line card I/O modules
  - agisce come punto di ingresso e di uscita dalla Qfabric
- Interconnect:
  - interconnette tutti i Nodes (backplane modular switch)
- Director:
  - provvede alla controllo ed alla gestione dei servizi
  - offre una finestra di configurazione per la gestione di tutti i componenti come un singolo devices via Junos OS CLI, system log ed SNMP



QFABRIC PLANE

- Data Plane:
  - i Nodes ed l'Interconnect sono il piano di forwarding del sistema Qfabric;
  - tutto il traffico dati tra i servers e storage è trasportato attraverso il data plane;
  - assenza di STP spanning tree protocol (active link end-points);
  - L2, L3 e FCoE traffic è bilanciato attraverso singoli o multipli path tra i node e l'Interconnect system.

- Control Plane:
  - separato dal data plane;
  - tutto il traffico di controllo e segnalazione e trasportato attraverso il control plane;
  - è gestito via Director utilizzando una rete out-of-band di collegamento con il sistema Interconnect ed i Nodes;
  - la rete oobm è usata per auto-discovery di tutti gli elementi, provisioning, image upgrades, configurazione automated.

- Management Plane:
  - il Director garantisce tutto il sistema di gestione comunicando direttamente con i Node ed l'Interconnect devices.

QFABRIC PLATFORM AND INTERFACE

- Interconnect:
  - QFX3800-I
  - QFX3600-I

- Node:
  - QFX3500
  - QFX3600

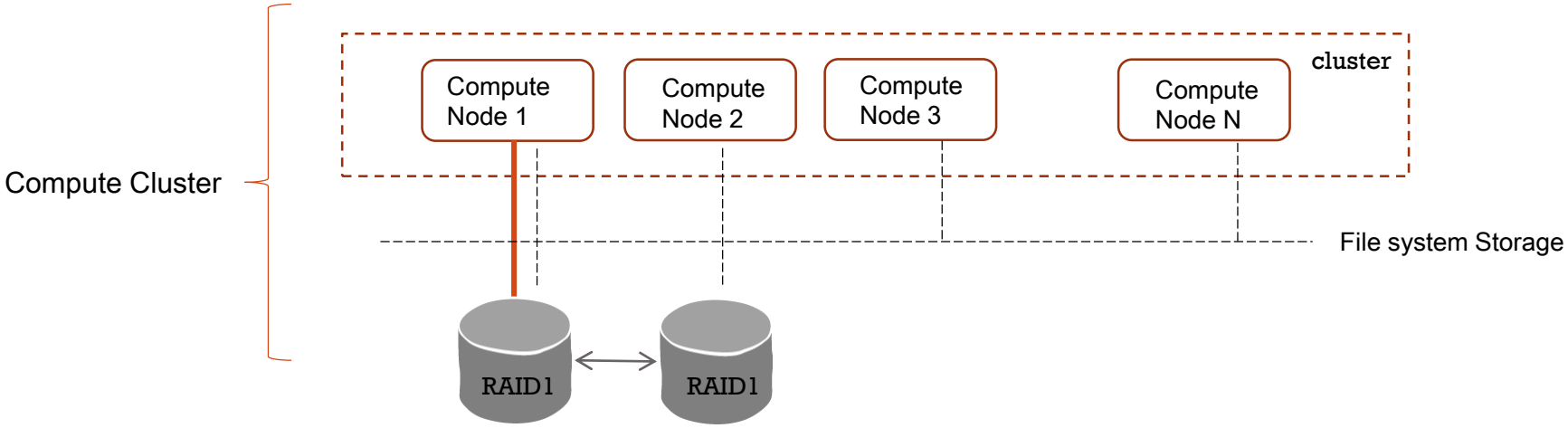
- Director:
  - QFX3100

- Interfaces:
  - Ethernet
  - IP
  - FCoE
  - FC

QFABRIC COMPUTE CLUSTER (MANAGEMENT PLANE)

Qfabric Director Compute Cluster è composto:

- Compute node
  - Due dei compute node servers sono composti da Disk subsystem direttamente connessi (gli altri node sono diskless);
  - Il cluster compute node è di tipo self-assembled at system boot-up dalla images memorizzata nei disk.
- Disk subsystem
  - Ogni disk subsystem è composto da due TB dischi RAID1 (Redundant Array of Inexpensive Disks) mirror configuration ed i contenuti sono blocchi sincroni replicati per ridondanza;
  - La partizione del disco contiene la boot-image Junos system software.
- File System
  - Un singolo file system è stratificato on top a questo storage subsystem ed esportato agli altri node servers via NFS (Network File System), assicurando cosi che i dati possono essere controllati e/o modificati da una sola persona/sistema per volta.

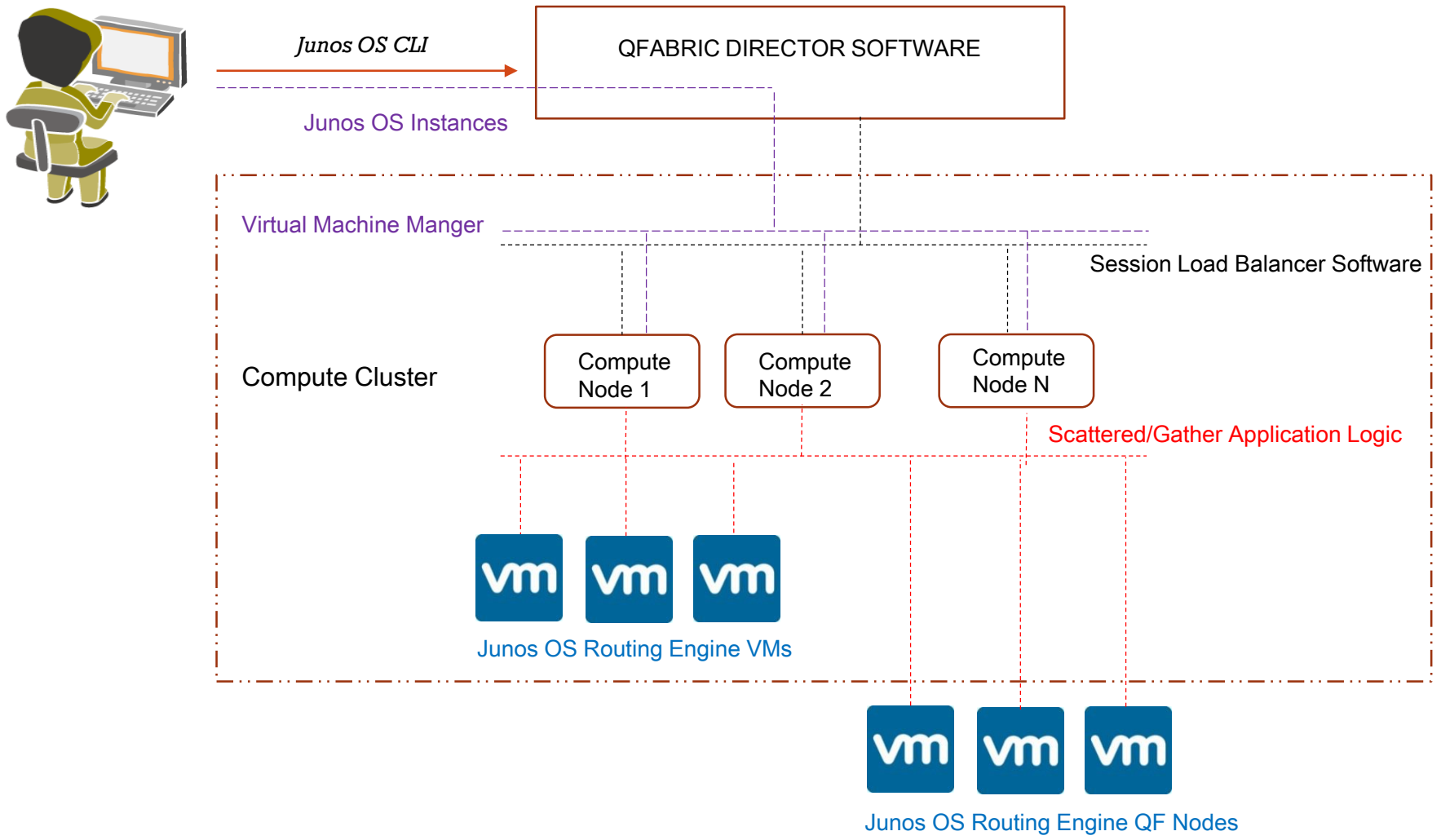


QFABRIC COMPUTE CLUSTER (MANAGEMENT PLANE)

Multiple software modules risiedono attraverso il Director Compute Cluster e sono gestiti via Junos OS CLI command (questo permette di gestire un modo unico tutto il Qfabric System)

- Virtualized Compute Node:
  - sono virtuali;
  - permettono multiple instanze Junos OS Routing Engine, consolidati in piccoli set di servers, offrono grande efficienza attraverso il Director;
- Junos OS instances:
  - opera come una VM (Virtual Machine) ed è chiamata clustered application attraverso i compute node;
  - queste VM sono organizzate in active/passive in coppia attraverso disuniti compute node offrendo resilienza.
- Director Software
  - contiene le applicazioni logiche che performano una distribuzione delle funzioni via CLI Junos (una singola Junos OS CLI command è sparsa attraverso i node
- Session Load Balancer
  - consiste in una applicazione che permette il bilanciamento delle sessioni sui compute node disponibili del cluster;
  - come benefici abbiamo una riduzione del carico di ogni CPU dei node compute, ed una tolleranza in caso di fault di alcuni servers;
  - solo le sessioni attraverso i server failed occorrono di restart del sistema (tutti gli altri node compute continuano a lavorare).

QFABRIC COMPUTE CLUSTER (MANAGEMENT PLANE)



QFABRIC CONTROL PLANE

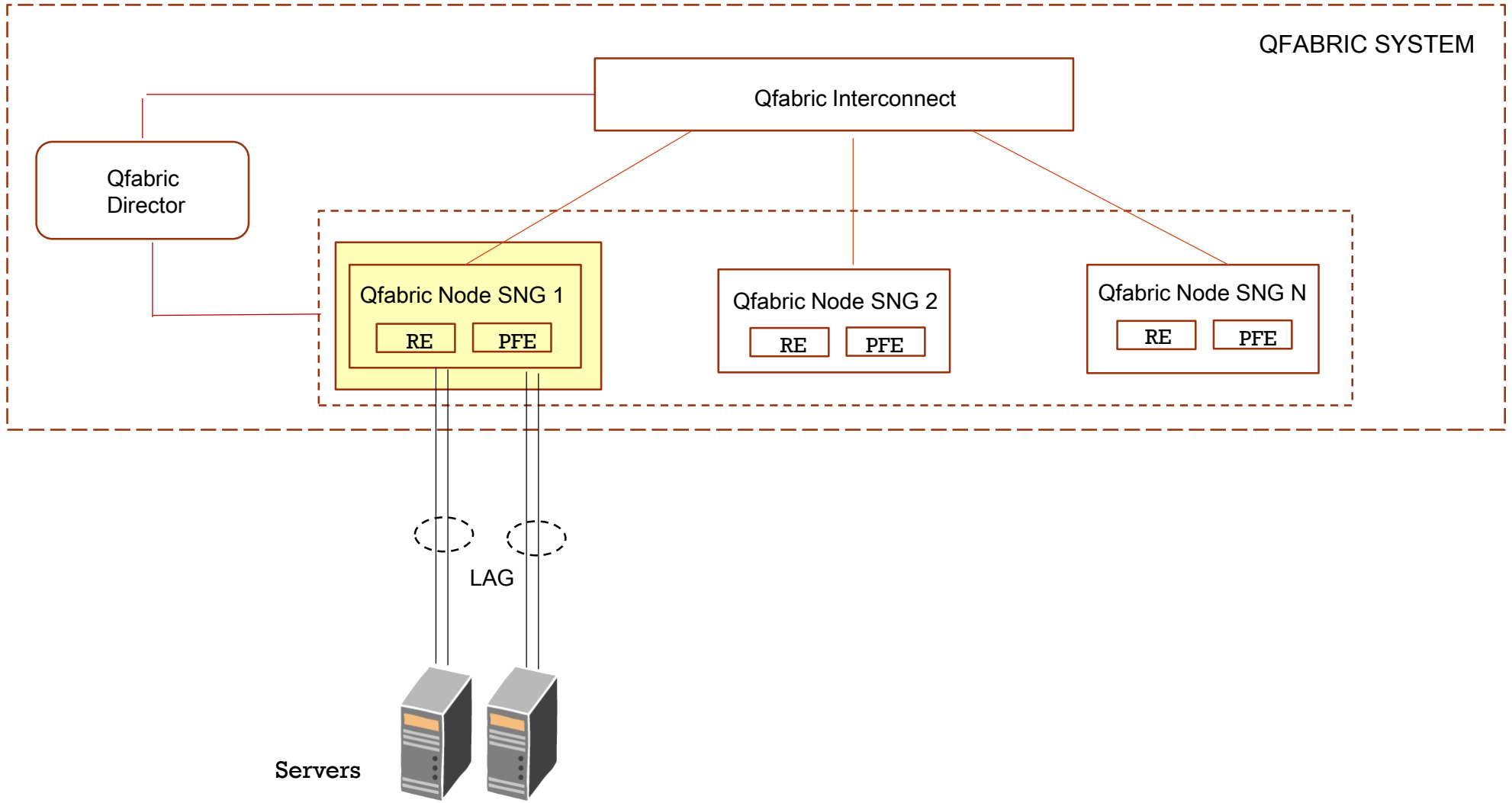
L’architettura Qfabric è di tipo switch/router L2/L3 distribuita, con tre differenti modi di configurazione:

- Server Node Group (SNG):
  - Qualsiasi Edge Node visto come un’unica entità logica è un SNG;
  - Gli SNG connettono servers e storage endpoint alla Qfabbric system;
  - Member di un LAG (link aggregation group) connessi tra un SNG ed un server offrono ridondanza tra i servers ed la QF
  - Tutti i Nodes QF sono attivati (boot up) come un SNG di default.

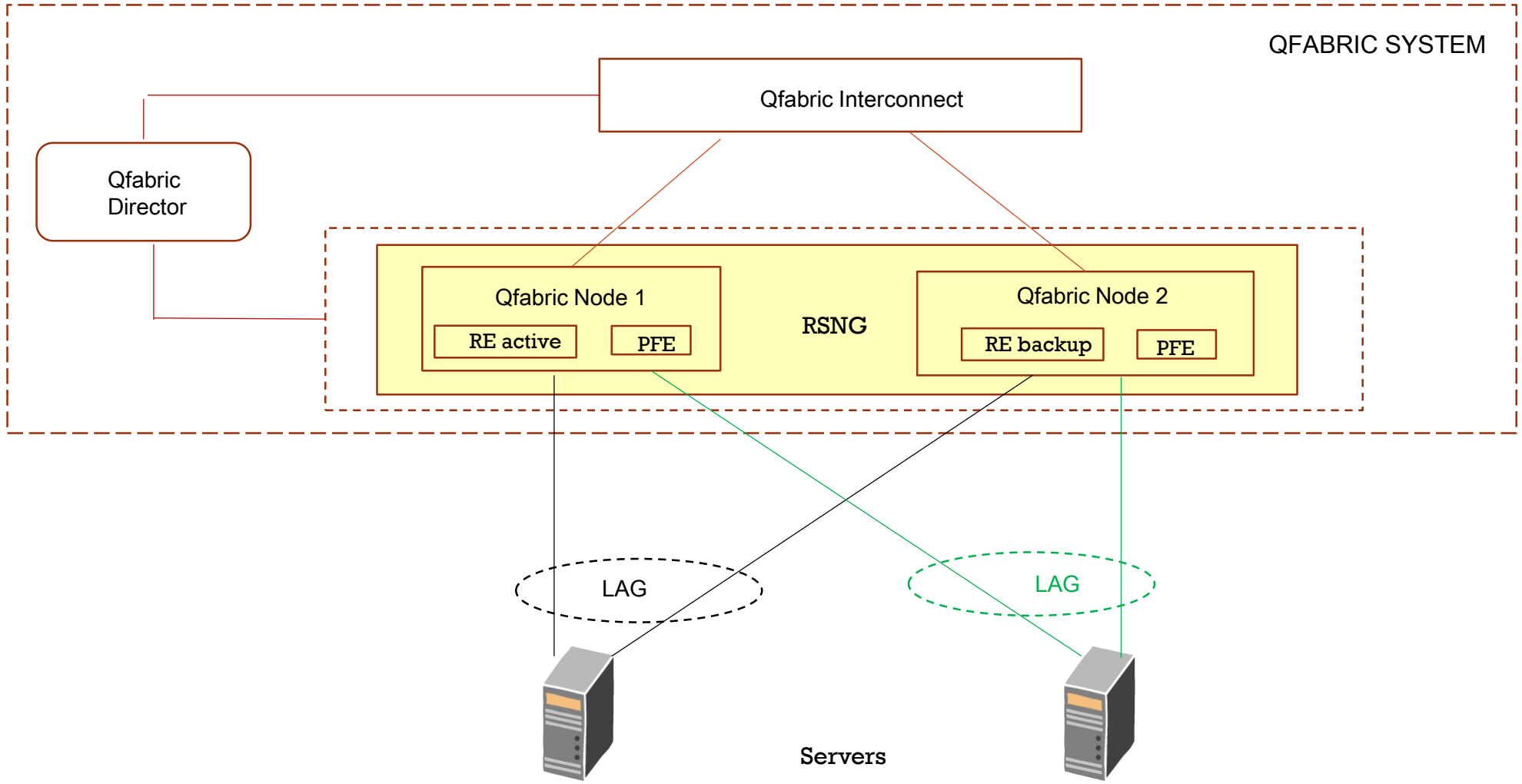
- Redundant Server Node Group (RSNG):
  - Una coppia di Nodes QF vista come un’unica entità logica è chiamata RSNG;
  - Member di un LAG sono distribuiti attraverso il RSNG ed i server in modo da offire ridondanza tra i servers ed la QF;
  - RSNG’s Routing Engine è attivo solo su uno switch della coppia;
  - CPU role performa sia la Routing Engine (RE) che la Packet Forwarding Engine (PFE) all’interno della coppia di switch RSNG

- Network Node Group (NNG):
  - Un set di QF Nodes opera a livello di protocollo quale STP, OSPF, PIM e BGP verso external devices come routers, switches, firewall e loab-balancer;
  - Solo un NNG all’interno QF può operare per volta;
  - CPU role performa solo le funzioni PFE (Packet Forwarding Engine) mentre la RE è delegata esternamente al NNG sulla QF Director cluster.

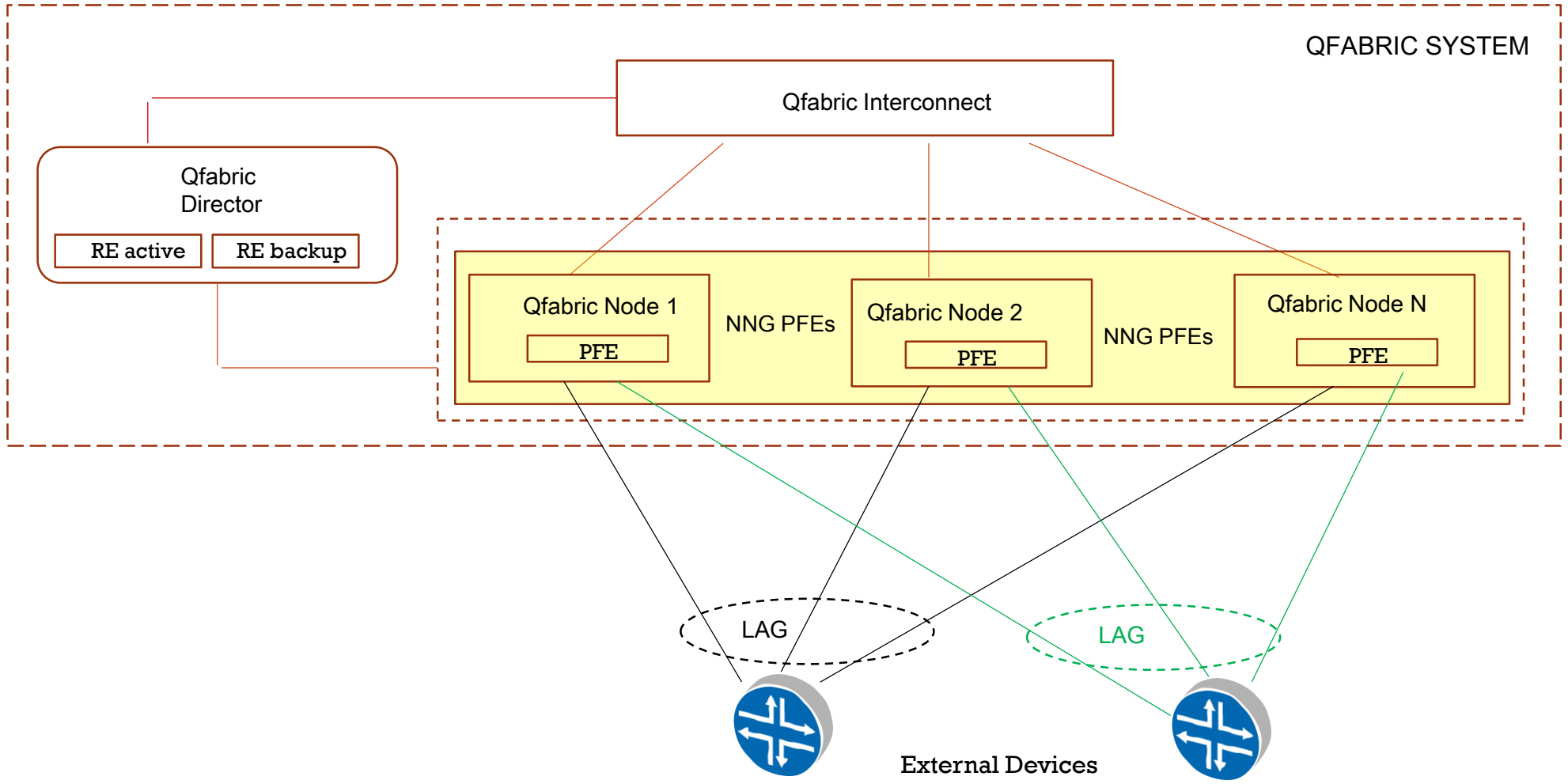
SERVER NODE GROUP DESIGN CONFIGURATION (SNG)



REDUNDANT SERVER NODE GROUP DESIGN CONFIGURATION (RSNG)



NETWORK NODE GROUP DESIGN CONFIGURATION (NNG)



ROUTING ENGINE PROPERTIES

All'interno del contesto Qfabric Node configurations ci sono differenti personalità della RE che comprendono il piano di controllo:

- Server Node Group Routing Engine:
  - SNGRE performa le funzionalità Routing Engine di un R(SNG) QF Nodes;
  - Opera come only-master all'interno di un SNG ed opera come master-backup all'interno di un RSNG;

- Network Node Group Routing Engine:
  - NNGRE performa le funzionalità Routing Engine in modalità active-backup attraverso una coppia di VMs via Compute Node della QF Director cluster.

- Fabric Manager Routing Engine (FMRE):
  - Mantiene le informazioni di inventory e topology all'interno di un database centralizzato:
  - FMRE ottiene le informazioni di connettività direttamente dai componenti QF Nodes e QF Interconnect via periodici Hello messages trasmessi attraverso gli Interconnect links;
  - FMRE calcola pesi per controllare la distribuzione di traffico degli uplinks di ogni QF Nodes basandosi dalla informazioni del database
  - FMRE lavora in modalità active-backup attraverso una coppia di VMs via Compute Node della QF Director cluster.

- Fabric Control Routing Engine:
  - FCRE controlla lo scambio di routes tra differenti Routing Engine del sistema QF
  - FCRE lavora in modalità active-backup attraverso una coppia di VMs via Compute Node della QF Director cluster.

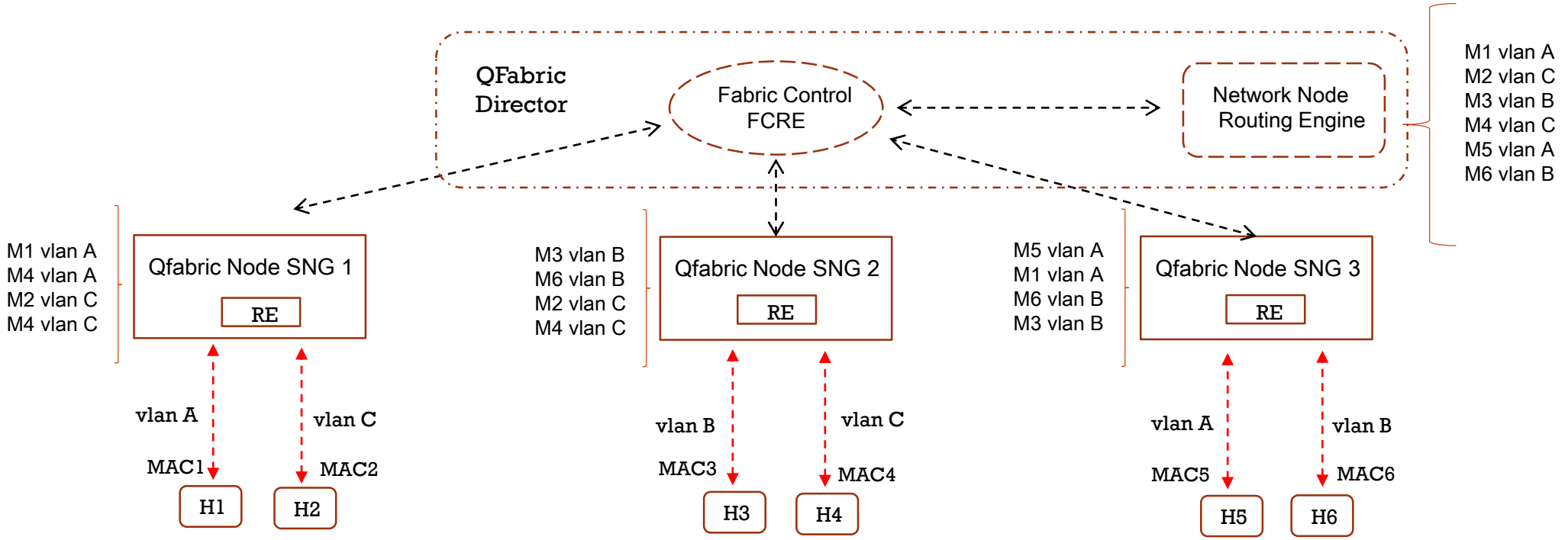
QFABRIC ARCHITECTURE INTERNAL PROTOCOL

- Qfabric System Discovery Protocol (logical connectivity)
  - Il sistema di discovery è basato su IS-IS routing protocol;
  - IS-IS messages sono scambiati attraverso il control plane via QF Nodes, QF Interconnect e QF Director avente come risultante una interna LAN dove risiedono tutti i componenti;
  - FMRE assegna private IP address ad ogni componente QF su questa interna LAN, permettendo cosi a tutti i sistemi membri di comunicare l'uno con gli altri via TCP-IP
  - Questo stabilisce una topologia logica del sistema QF (non una topologia fisica di come sono connessi i componenti tra loro)

- Qfabric Topology Discovery Protocol (physical connectivity)
  - L' architettura Qfabric manca di una connessione di tipo internal-to-chassis e pertanto necessità di un protocollo che sia in grado di fare un discovery fisico tra i QF Nodes ed i QF Interconnect ;
  - Questo protocollo è basato su una porzione IS-IS per il discovery neighbor
  - IS-IS permette lo scambio di Hello messages tra QF Nodes e QF Interconnect attraverso links a 40Gbps;
  - Ogni discovery forma una informazione destinata verso l'FMRE (Fabric Manager database) all'interno del Qfabric Director dove il data plane è assemblato
  - Una volta creato il database l'FMRE programma specifici pesi per i QF Nodes bilanciando in modo equo il traffico attraverso tutti i disponibili QF Interconnect
  - In caso di event-failure il sistema multipath di QF è abilitato dalla Fabric Manager centralizzata.

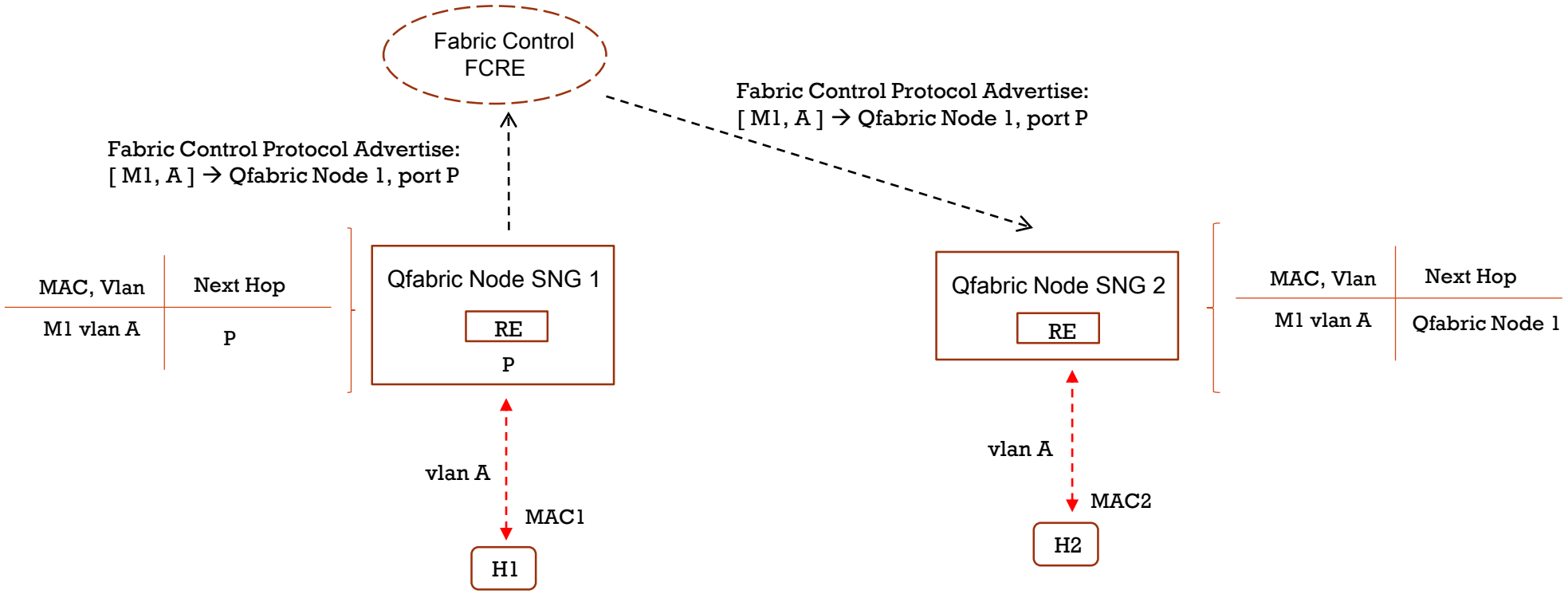
QFABRIC ARCHITECTURE INTERNAL PROTOCOL

- Qfabric Control Protocol
  - Dato il consistente numero di elementi indipendenti quali SNG e NNG Routing Engine, Qfabric system ha bisogno di un meccanismo di scambio network state tra loro;
  - In considerazione di eventuali e possibili differenze di versione Junos OS in differenti partizioni, questo meccanisco/protocollo deve avere multiple versioni Junos OS supportate;
  - Questo meccanisco è basato sul protocollo BGP con funzionalità di Router Reflector per la sua scalabilità; Juniper ha aggiunto una nuova versione di address-family to multi-protocol BGP, permettendo di trasportare **MAC router** oltre che IP e VPN router.
  - Il meccanismo di RD (Route Distinguiscer) per pormettere overlapping di indirizzi ed RT (Route Target) permettendo filters di import ed export routes all'interno della comune infrastruttura Qfabric, offre una seletiva condivisione network dell'infrastruttura stessa.
  - A differenza di una comune L3VPN rete dove l'utente deve configurare parametri RD ed RT, in Qfabric è necessario solo definire vlan e routing-instances per lo scopo, ed il meccanisco RD ed RT rimane completamente trasparente.



QFABRIC PACKET FORWARDING

- L2 unicast forwarding:



QFABRIC PACKET FORWARDING

- Data Path for Layer 2 Unicast Forwardind
  - Ogni Qfabric Node ha una tabella per i Remote Qfabric Node attraverso la quale contiene un entry per ciascuno di essi;
  - Considera N percorsi disponibili per il raggiungimento del Remote Node attraverso la Fabbrica;
  - Fabric Topology Manager programma la tabella Remote Qfabric Node in ciascun Qfabric Node.

  - Nell'esempio della precedente slide:
    - Qfabric Node 2 consulta la sua tabella di raggiungibilità dei remote Nodes e verifica quanti path sono disponibili per raggiungere il Node 1
    - Seleziona uno di questi path su base flow-hash, antepone un'intestazione nel pacchetto e lo spedisce al corrispettivo Qfabric Interconnect
    - L'intestazione identifica il pacchetto destinato al Node 1
    - Quando il pacchetto arriva al Qfabric Interconnect, questi guarda la sua tabella di raggiungibilità Qfabric Nodes e selezione uno dei disponibili path scegliendone uno su base flow-hash
    - Notare solo che il Qfabric Interconnect guarda solo l'intestazione inserita dal Node 2, e non il MAC Address, semplificando l'interconnessione intra-fabbrica
    - Quando il pacchetto arriva al Node 1, subisce un MAC address lookup che risulta via la egress port e trasmesso.

- Data Path for Layer 3 Unicast Forwardind
  - Consiste in una associazione tra l'indirizzo IPv4 ed il MAC address, Vlan e Port (IPv4 → < MAC, Vlan. Port >);
  - Quando un host è collegato ad un Qfabric Node, il suo ARP default gateway viene considerato dalla RE del QF Node formando un'associazione L3 entry IP address → < MAC, Vlan, Port >
  - La RE trasmette questa L3 entry a tutte le altre Routing Engine, le quali settano questa entry dentro la loro packet processor;
  - Pertanto il remote Qfabric Node associa il destination Qfabric Node con il solo IP address, ed il destination QF Node mantiene l'associazione IP address → < MAC, Vlan, Port >

QFABRIC BROADCAST, UNKNOW UNICAST AND MULTICAST FORWARDING

Ci sono tre punti dove una multi-destination (broadcast, unknow unicast e multicast) data stream possono essere replicati all'interno di una struttura Qfabric:

- Ingress Qfabric Node:
  - replica attraverso le sue local port e attraverso uno o pià dei suoi uplink diretti verso Qfabric Interconnect
- Qfabric Interconnect:
  - replica attraverso se stessi verso tutti i Qfabric Nodes (ad eccezione della porta di ingresso);
  - Solo una copia del pacchetto è trasmesso ad ogni QF Nodes;
  - Il carico della replica di un particolare QF Nodes (ad es il numero di Nodes al quale un particolare QF Interconnetc replica) è basato su come il piano di controllo (control plane) costruisce il multi-destination tree per quel determinato gruppo (o vlan)
- Egress Qfabric Node:
  - Riceve una singola copia del pacchetto da un particolare Qfabric Interconnect eppoi replica il pacchetto attraverso le sue local port.