The bridge to possible

AI Network 디자인 고려 사항 및 기술 Deep dive

엄현학 이사, 시스코 코리아

Cloud AI Infrastructure Team

cisco Connect

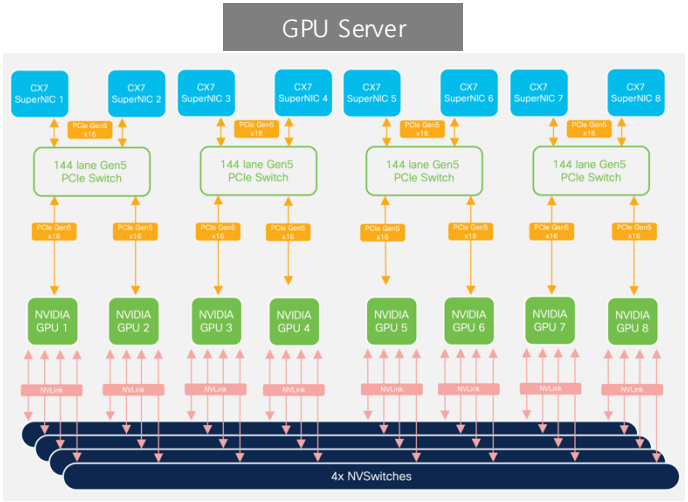
Agenda

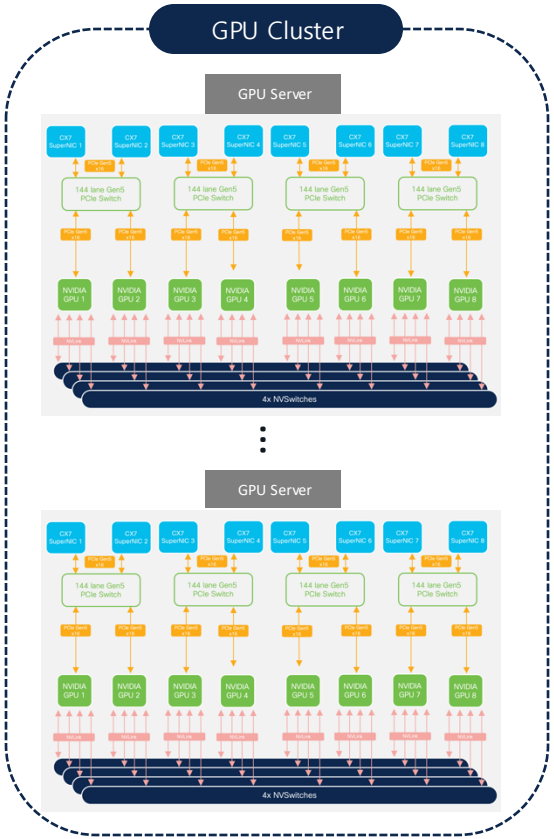
- AI Network 디자인 고려 사항
- Lossless Ethernet을 구현을 위한 기술
- AI Network에서의 부하 분산 기술

AI Network 디자인 고려 사항

GPU Cluster를 통한 연산 성능 향상

GPU Cluster를 통해 높은 성능 향상을 가져 올 수 있습니다.



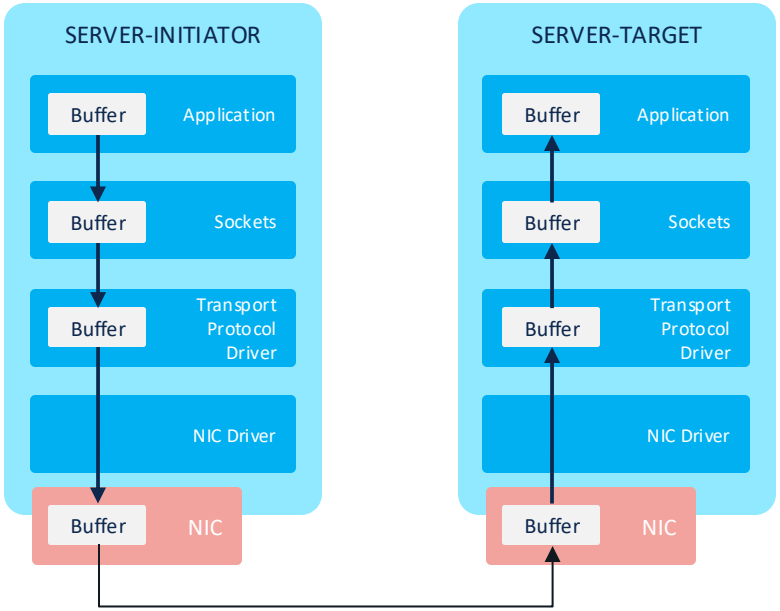


cisco Connect

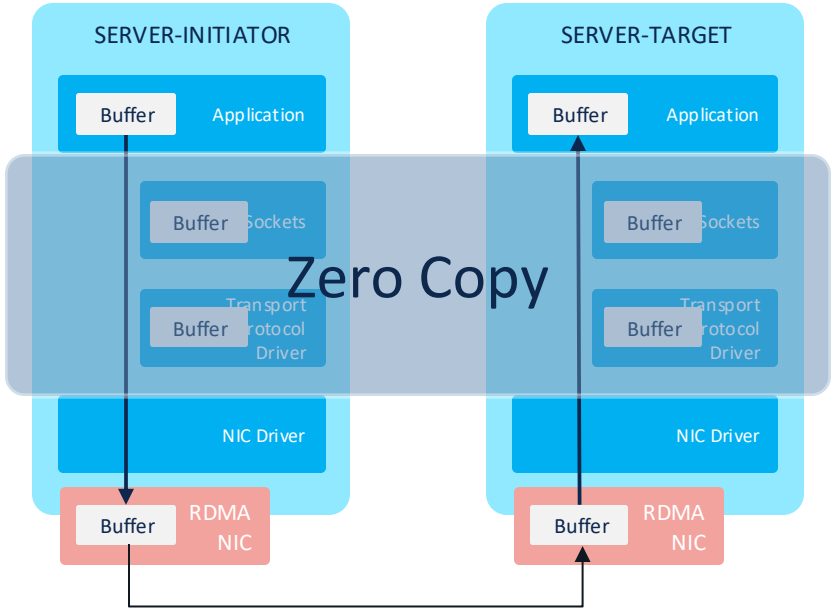
RDMA를 이용한 서버간 통신 방법

GPU Cluster의 성능 향상을 위해서는 RDMA 통신 방법이 필요합니다.

전통방식의 Server to Server 통신

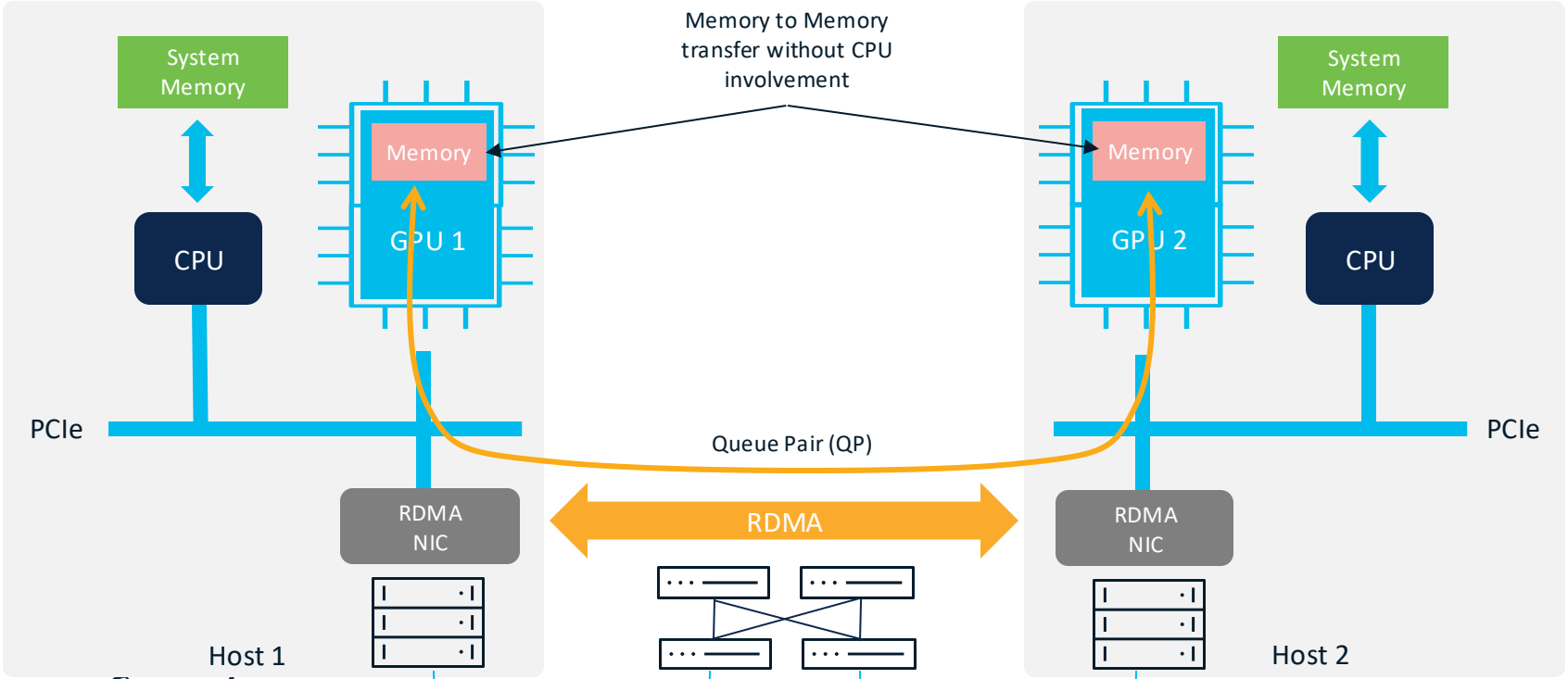


RDMA 기반 Server to Server 통신



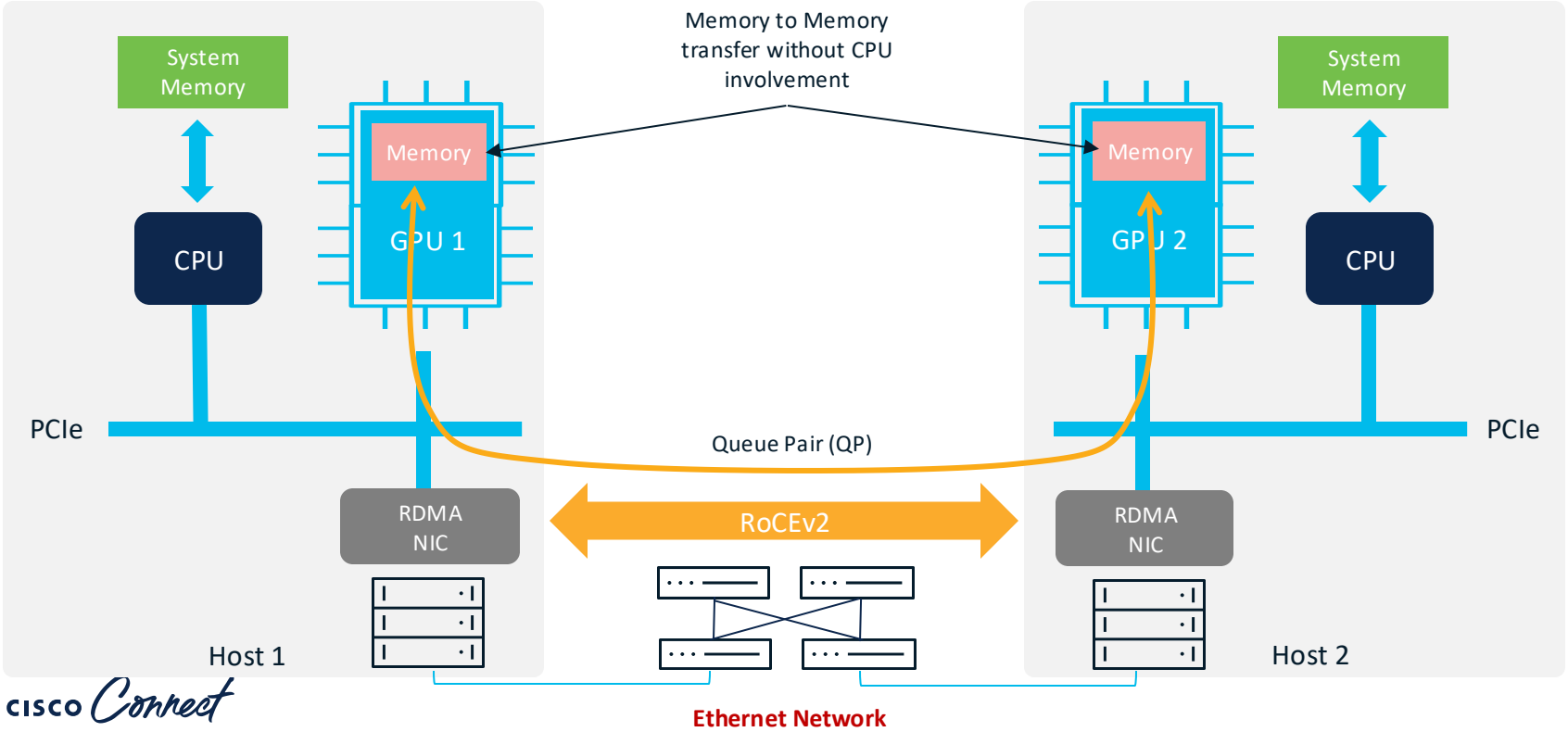
RDMA(Remote Direct Memory Access) 동작

RDMA를 이용하여 GPU Memory에 Direct로 job을 적재 할 수 있습니다.



RDMA over Converged Ethernet (RoCEv2)

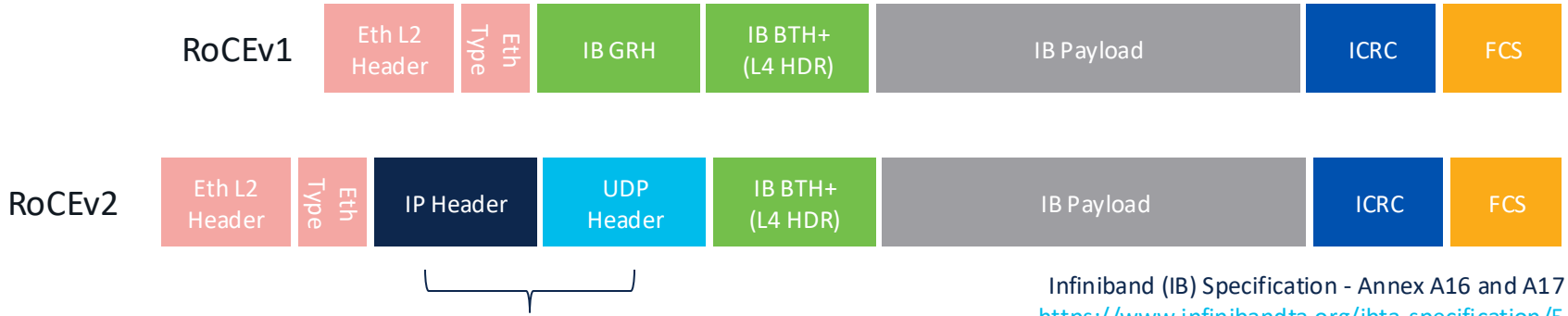
RDMA를 범용적으로 사용하기 위해 이더넷 기반의 RoCE가 출현하였습니다.



RoCEv2 상세

UDP 기반의 RoCEv2로 인해 기존 RoCEv1보다 더 많은 장점을 가지게 되었습니다.

- 최초의 RoCE는 RDMA 패킷에서 이더넷 헤더만 포함된 RoCEv1으로 Release 됨
- RoCEv2는 IP 헤더와 UDP 헤더가 추가 되었으며 UDP Destination Port 는 4791입니다.
- IP 헤더가 추가된 RoCEv2의 경우 ECN bit를 이용하여 End-to-End Congestion Control이 가능하며 라우팅 기반의 GPU 서버 네트워크 디자인이 가능합니다.



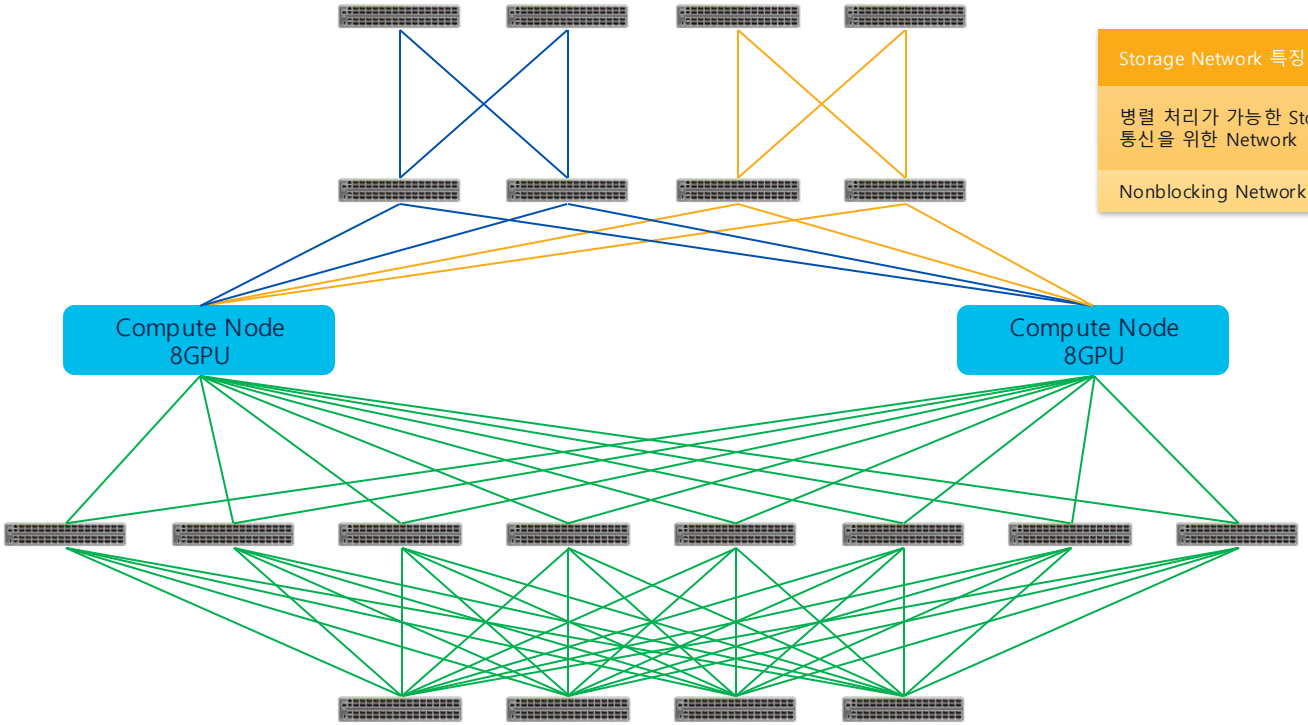
Infiniband (IB) Specification - Annex A16 and A17
https://www.infinibandta.org/ibta-specification/5

AI Network 디자인시 고려 사항

AI Network 구성 시 Front-End, Back-End, Storage Network가 필요합니다.

Front-End Network 특징
AI 시스템에 입력될 데이터를 수신하고 처리하며 외부 사용자와의 상호 작용을 수행하는 Network
Loosely Applications 특성
TCP (Low Bandwidth Flows)
높은 Jitter를 허용
Oversubscription Network 구조 허용

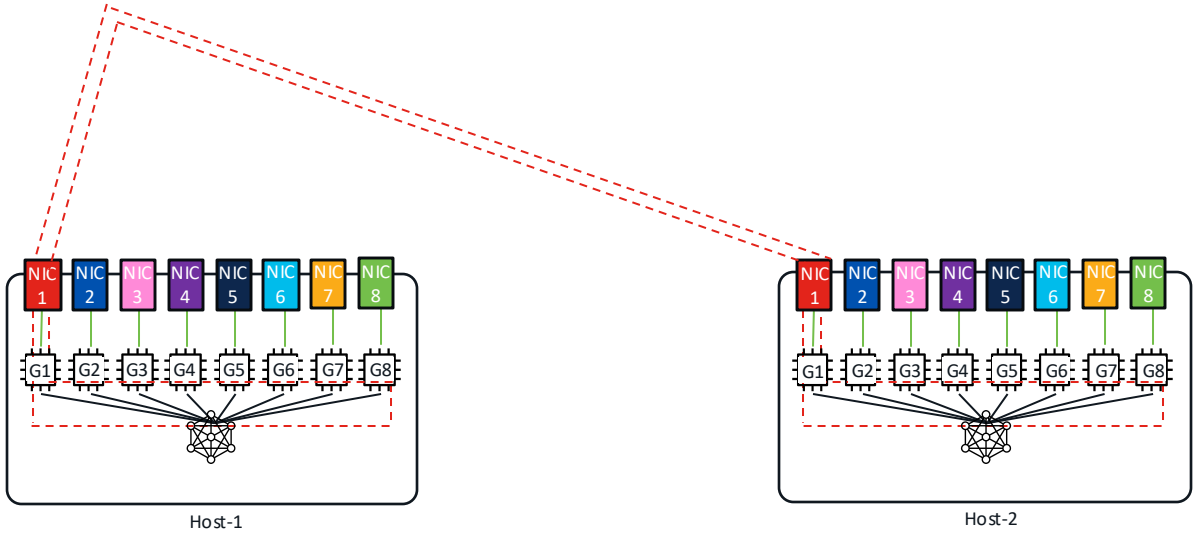
Back-End Network 특징
모델링을 작성하고 추론을 수행하며 분석을 수행하는 Network
Tightly Coupled Processes
RoCE
낮은 Jitter 필요
Nonblocking Network 토폴로지



Storage Network 특징
병렬 처리가 가능한 Storage와의 통신을 위한 Network
Nonblocking Network 토폴로지 권장

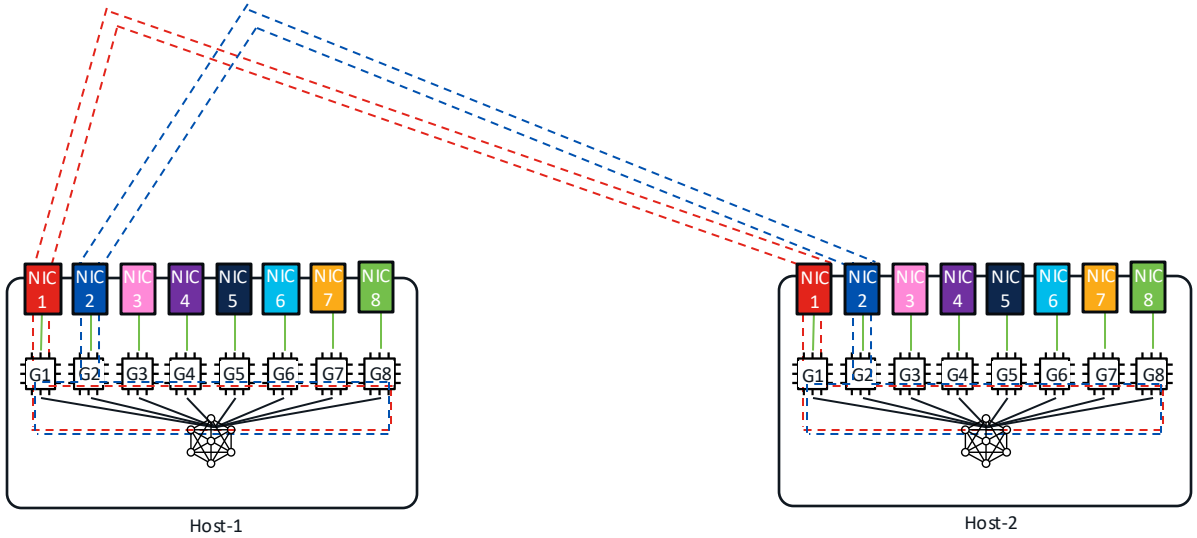
Rail을 통한 GPU 서버간 통신 방법

일반적으로 동일한 인터페이스를 이용하여 GPU 클러스터를 구성합니다.



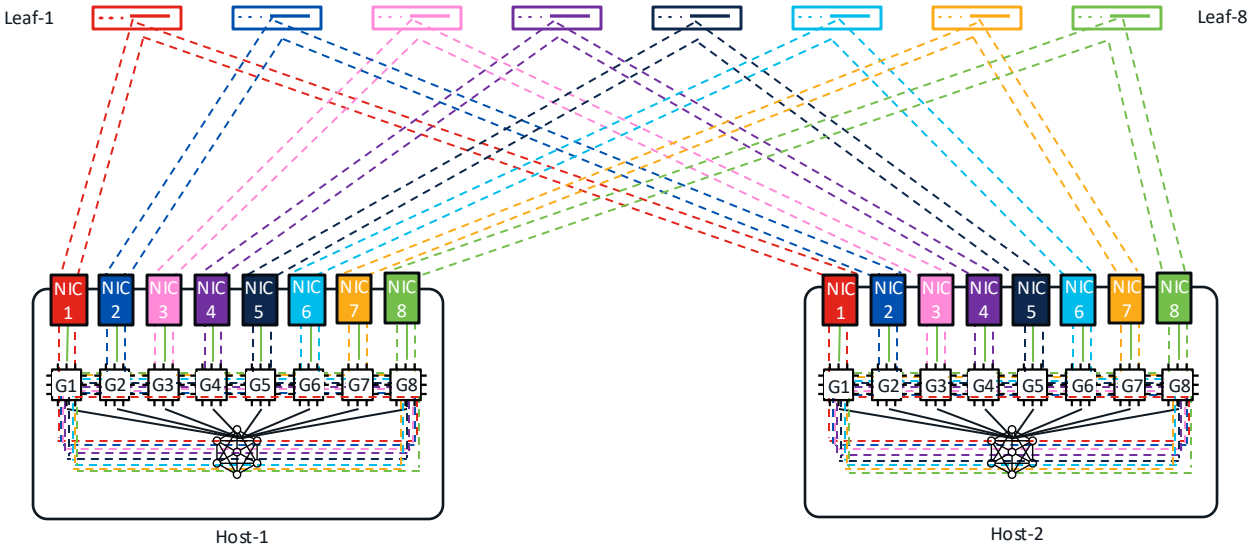
Rail을 통한 GPU 서버간 통신 방법

일반적으로 동일한 인터페이스를 이용하여 GPU 클러스터를 구성합니다.



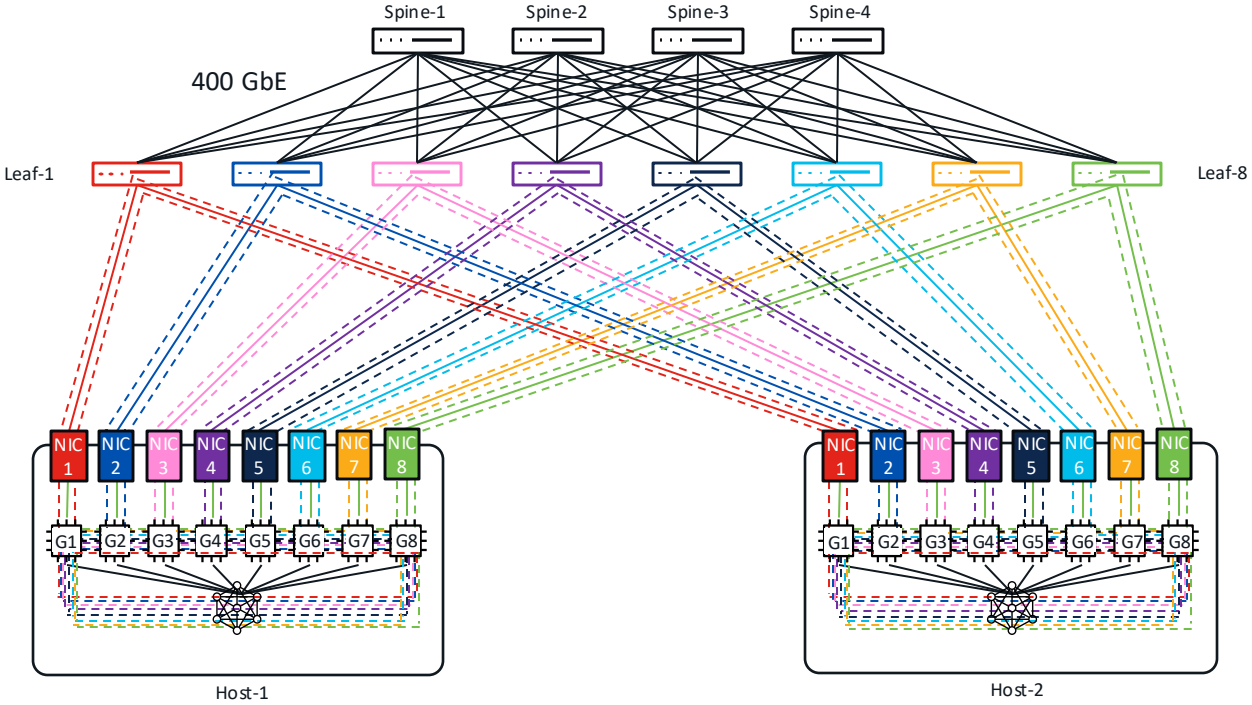
Rail을 통한 GPU 서버간 통신 방법

일반적으로 동일한 인터페이스를 이용하여 GPU 클러스터를 구성합니다.



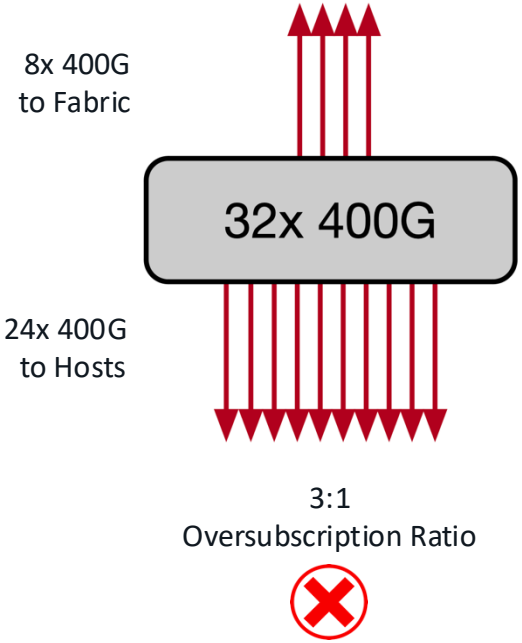
Rail을 통한 GPU 서버간 통신 방법

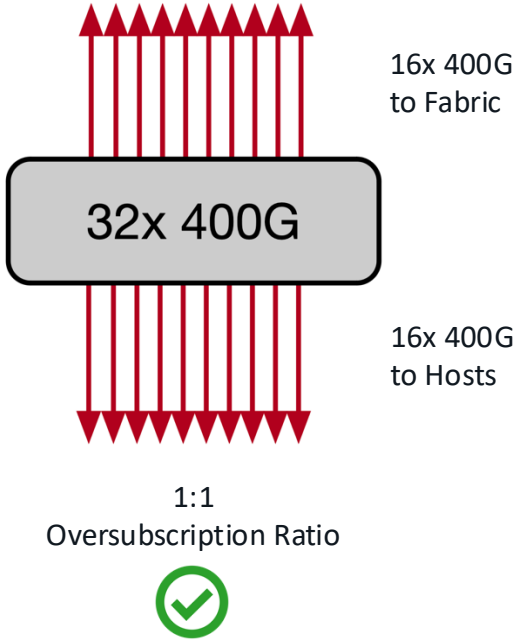
모든 경우의 수에 대한 GPU 클러스터를 구성 할 수 있습니다.



Spine/Leaf Back-end 네트워크 고려 사항

서버가 연결되는 인터페이스와 Spine이 연결되는 인터페이스의 비율을 1:1로 디자인 해야 합니다.





Spine/Leaf Back-end Network 디자인 예시

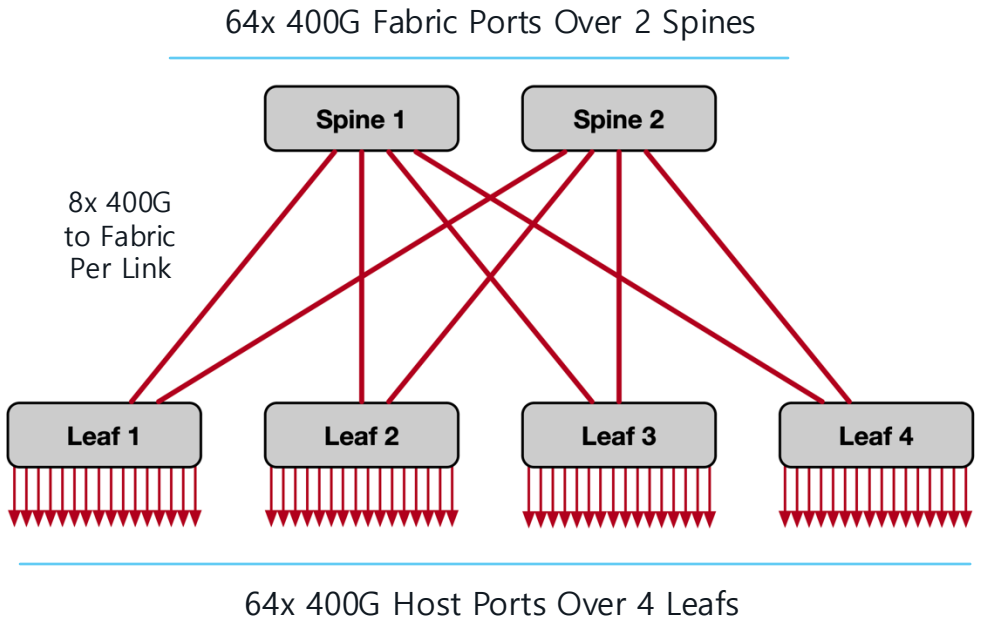
서버가 연결되는 인터페이스와 Spine이 연결되는 인터페이스의 비율을 1:1로 디자인 해야 합니다.

Compute Cluster 상세:

- 16x Nodes
  - 4x GPUs per Node
  - 4x 400G Interfaces

Back-end Fabric 상세:

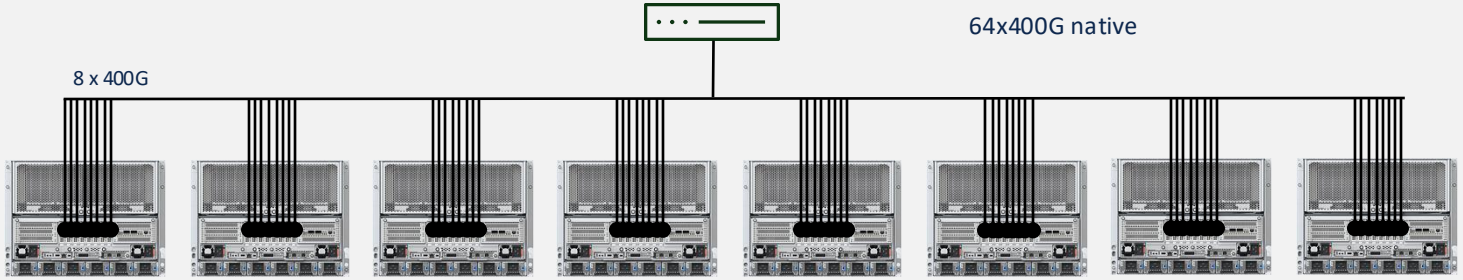
- 2x N9K-C9332D-GX2B Spines
- 4x N9K-C9332D-GX2B Leafs



소규모 Back-End Network Design

최대 64EA의 GPU를 Cluster로 구성 할 수 있습니다.

1x Leafs:
Nexus 9364D-GX2A

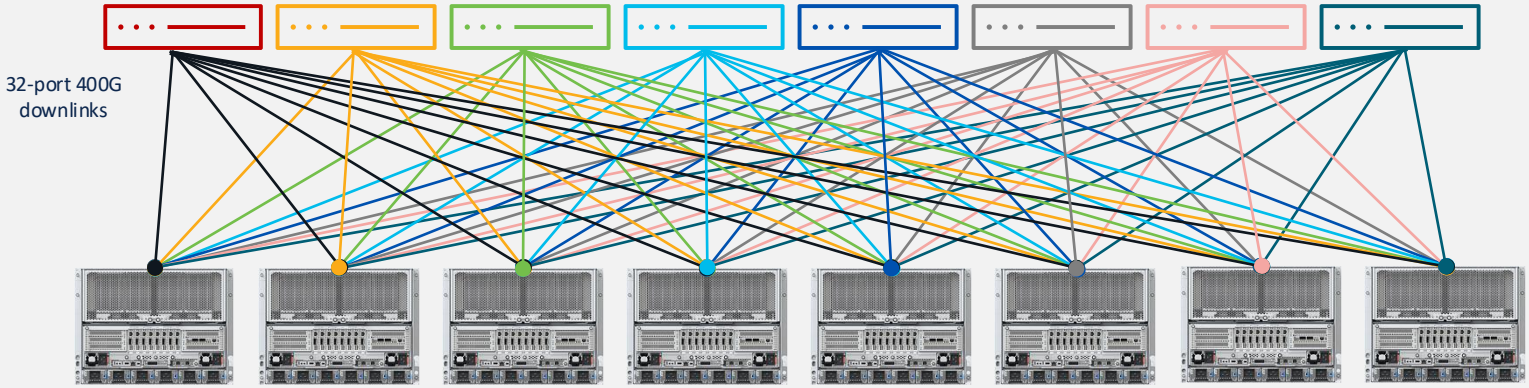


8 GPU Server, 64 GPUs

중소규모 Back-End Network Design

동일한 Rail만을 사용하는 구조이며 효율적 구성이 가능하지만 개발팀과의 협업 및 여러 고려 사항이 필요합니다.

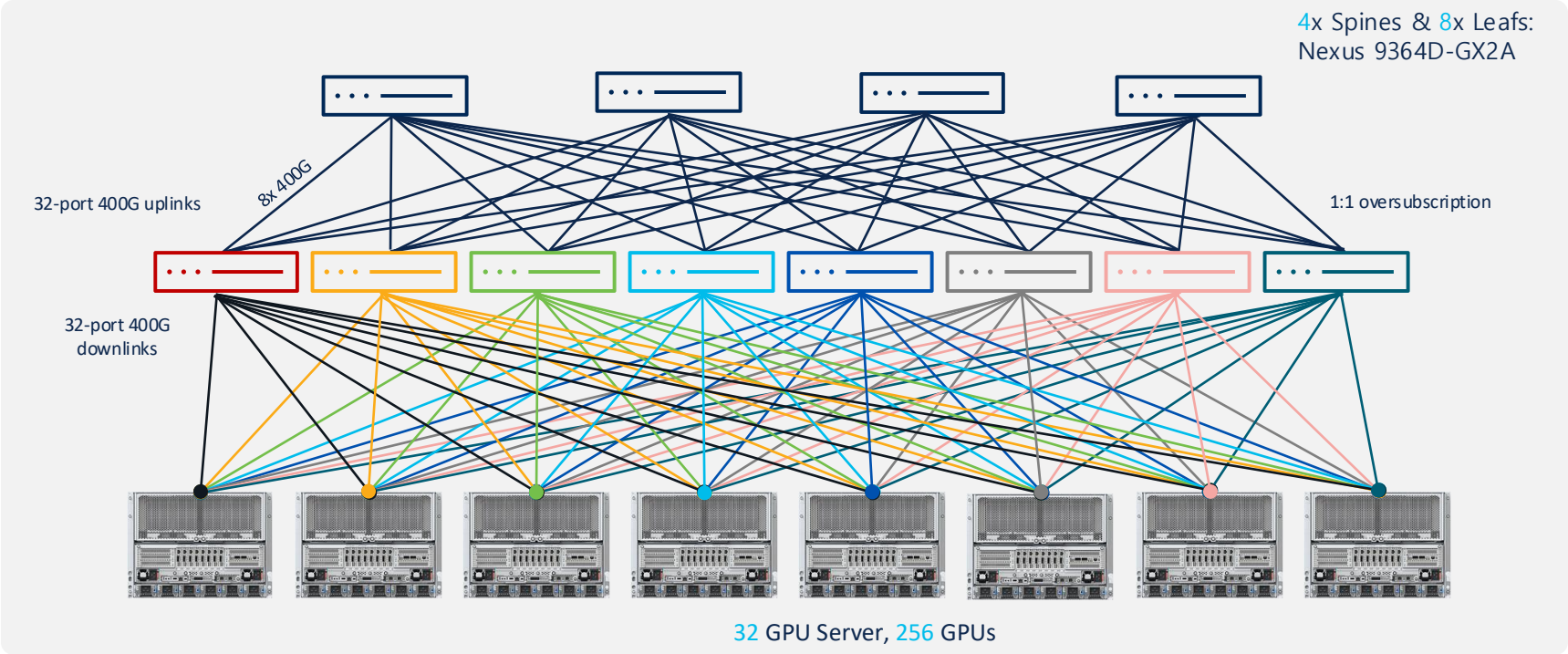
8x Leafs:
Nexus 9332D-GX2B



32 GPU Server, 256 GPUs

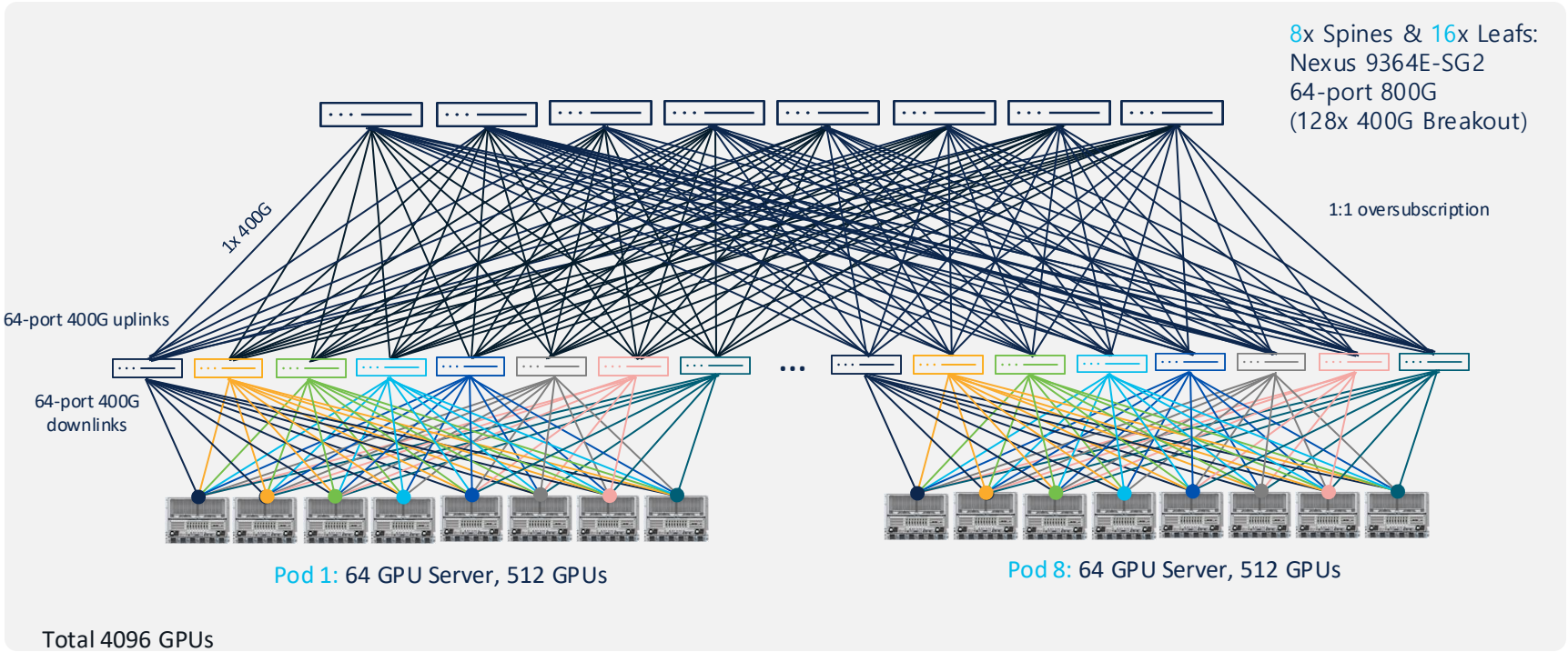
중소규모 Back-End Network Design

모든 Rail을 사용하는 구조이며 모든 조건을 수용 할 수 있는 디자인입니다.



대규모 Back-End Network Design

모든 Rail을 사용하는 구조이며 모든 조건을 수용 할 수 있는 디자인입니다.



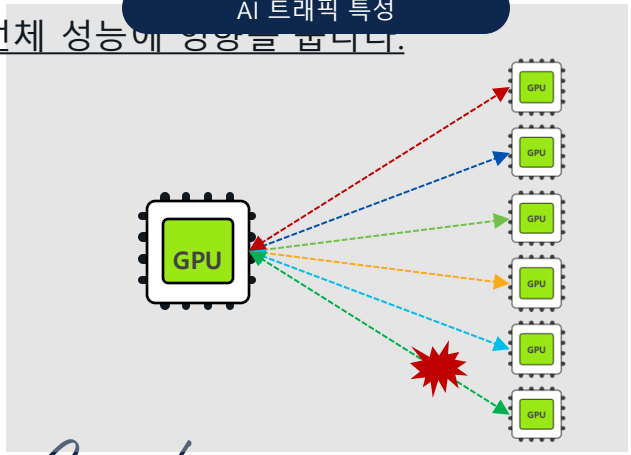
Lossless Ethernet 구현을 위한 기술

cisco Connect

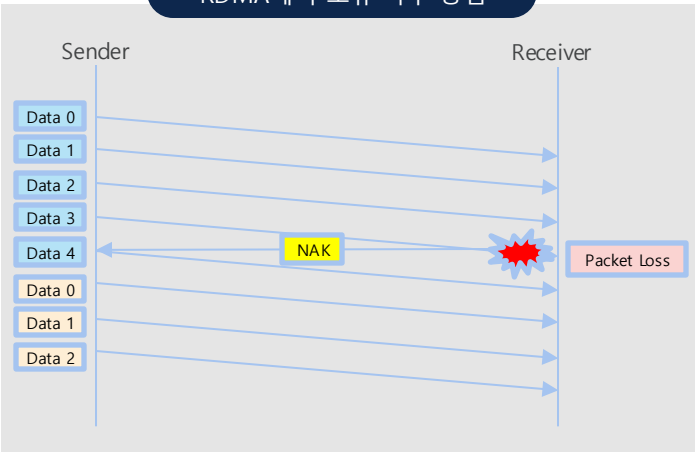
패킷 손실에 따른 GPU 연산의 성능 저하

- RoCE는 UDP 기반의 통신으로 TCP와는 다르게 패킷을 재전송하는 메커니즘을 가지고 있지 않습니다.
- RDMA에서 패킷의 Lost가 발생하면 재 전송으로 인한 딜레이가 발생하고 이로 인해 성능 저하가 발생합니다.
- 모든 GPU에서 분산하여 Job을 처리 한 결과값을 합치는 과정에서 특정 GPU의 결과값 수신에 문제가 발생하면
  전체 성능에 영향을 줍니다.

AI 트래픽 특성

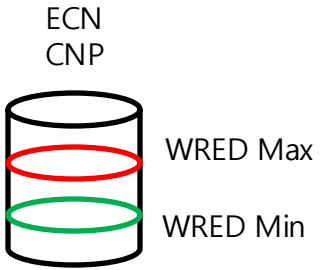


RDMA에서 오류 복구 방법



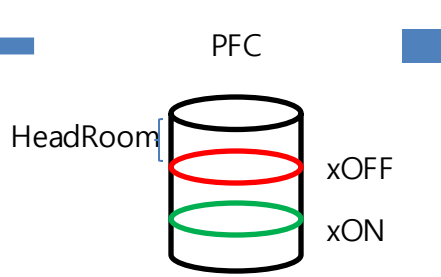
RoCE에서 Lossless Network 구현 방법

Buffer의 사용량에 따라서 PFC/ECN이 Trigger되며 이를 이용하여 Lossless Network를 구현합니다.

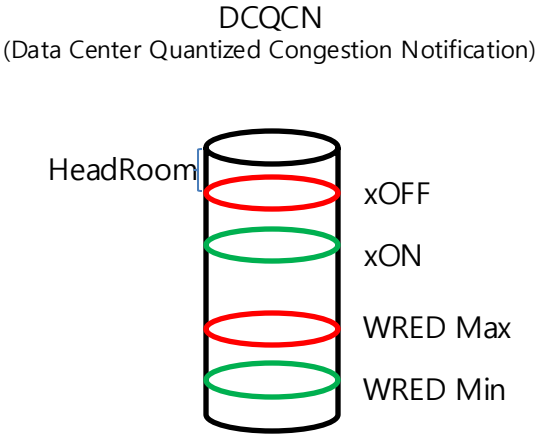


WRED MIN ~ MAX : 설정에 의한 ECN Packet 생성

WRED MAX 이상 : 100% ECN Packet



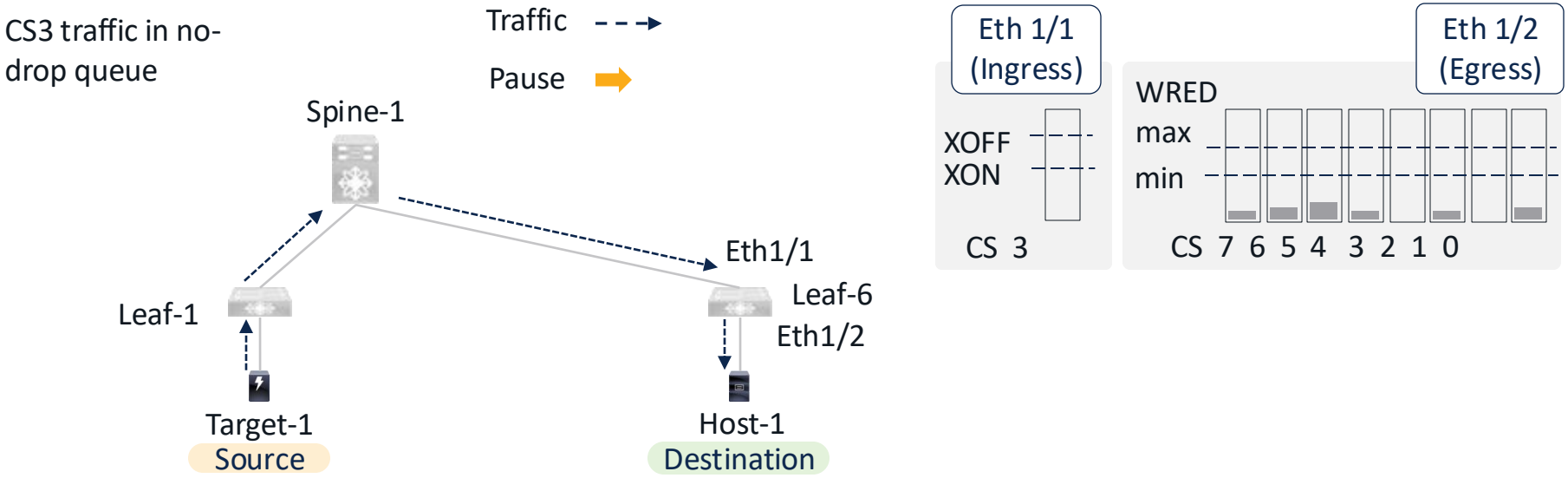
xOFF 이상 임계치 Pause Frame 생성



- Switch Buffer 임계치 값이 초과되면 PFC(Priority Flow Control)를 사용하여 트래픽 전송을 중단 시킴
- Weight Random Early Detection 기능을 통해 ECN 패킷 전송 송신 호스트에 혼잡을 알려 트래픽 전송 속도를 낮춰 줌
- 2가지를 모두 사용하는 DCQCN 환경을 통한 Lossless Network 구현

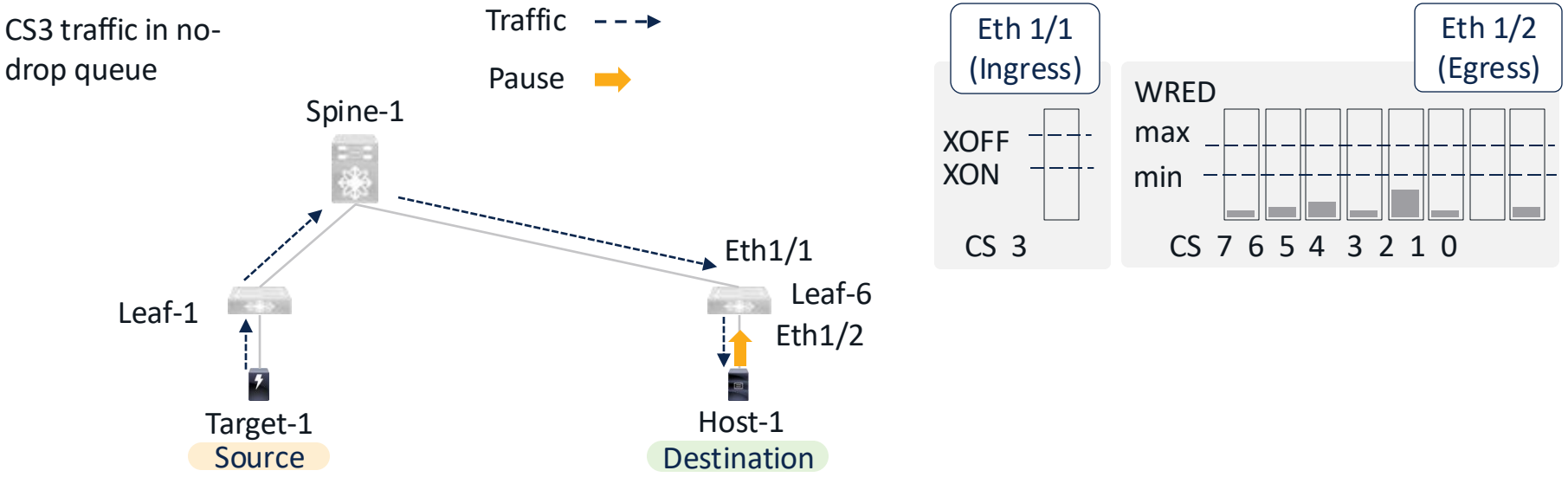
RoCEv2 Congestion Management 동작

혼잡이 발생하기 않는 경우 안정적으로 RoCE 트래픽이 전송됩니다.



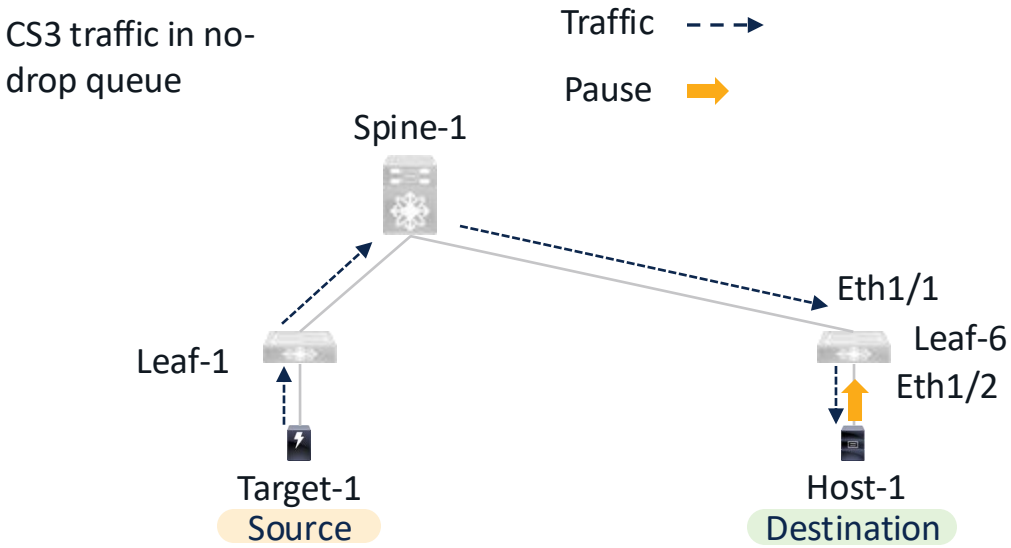
RoCEv2 Congestion Management 동작

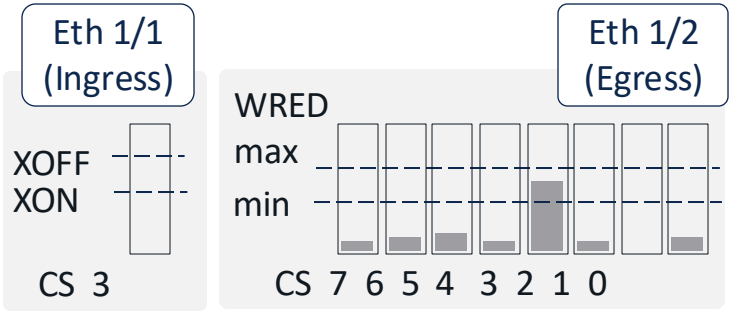
Host-1의 RDMA 트래픽 처리 지연으로 인해 PFC를 전송하여 특정 시간 동안 트래픽 송신을 멈추라고 Leaf-6에 요청합니다.



RoCEv2 Congestion Management 동작

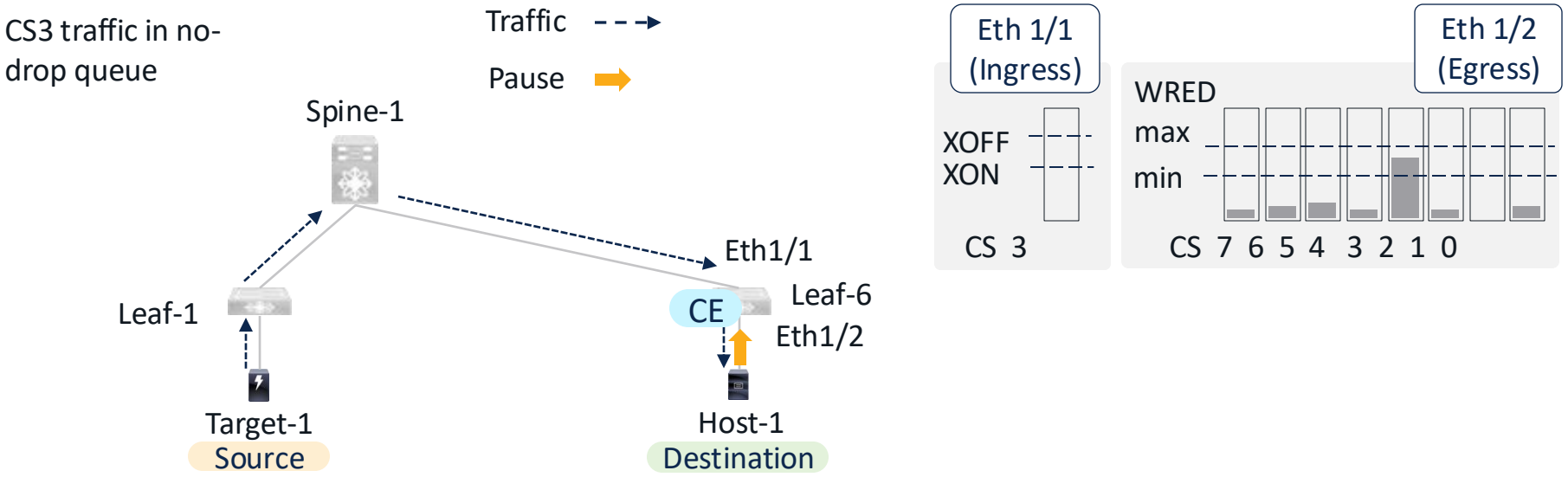
PFC를 수신한 후 Leaf-6은 특정 시간 동안 트래픽을 전송 할 수 없기 때문에 CS3의 Queue에 트래픽이 버퍼링되고
WRED min 임계치를 초과하게 됩니다.





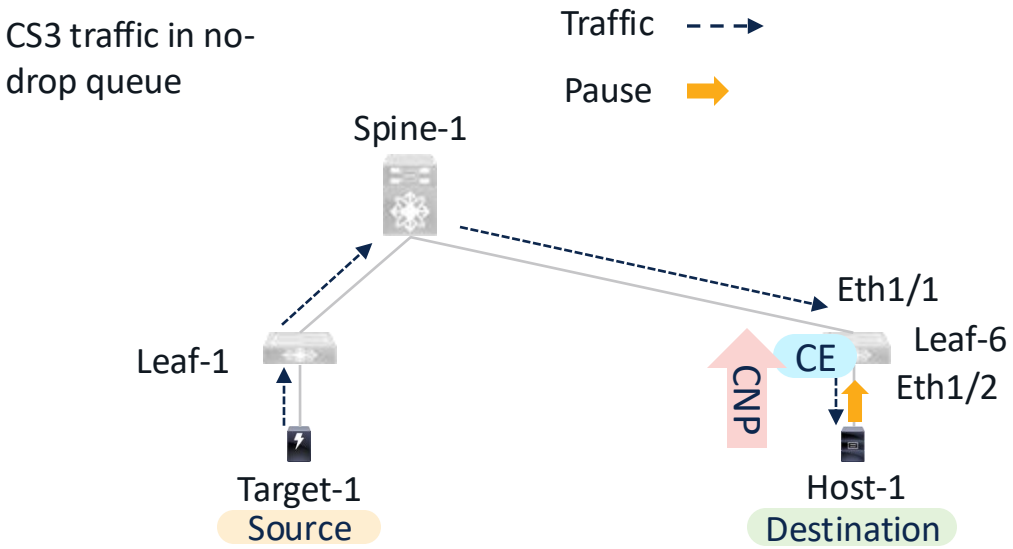
RoCEv2 Congestion Management 동작

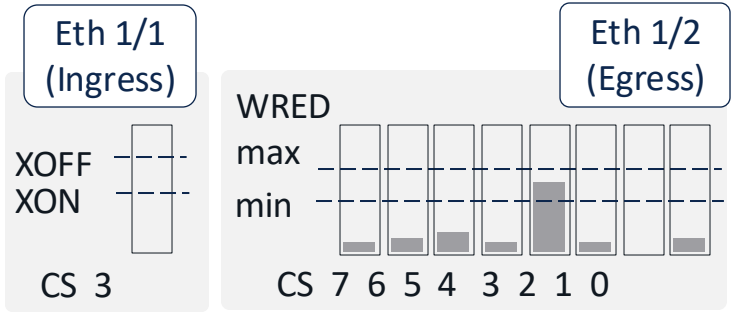
Leaf-6은 랜덤하게 ECN bit를 11로 마킹하여 Host-1에 혼잡을 알립니다.



RoCEv2 Congestion Management 동작

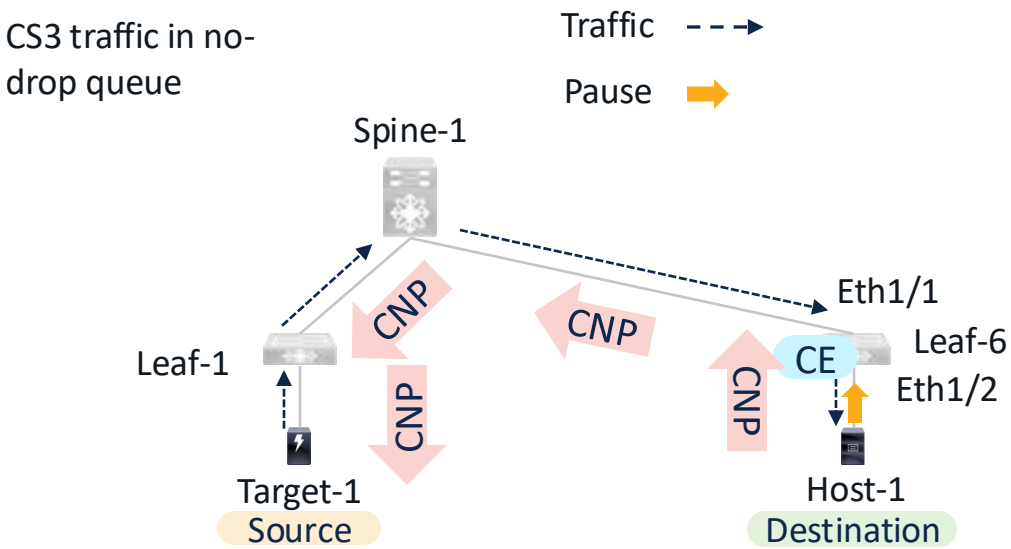
ECN bit 11을 인지한 Host-1은 혼잡 상태임을 인지하고 Target-1에 CNP를 전송합니다.

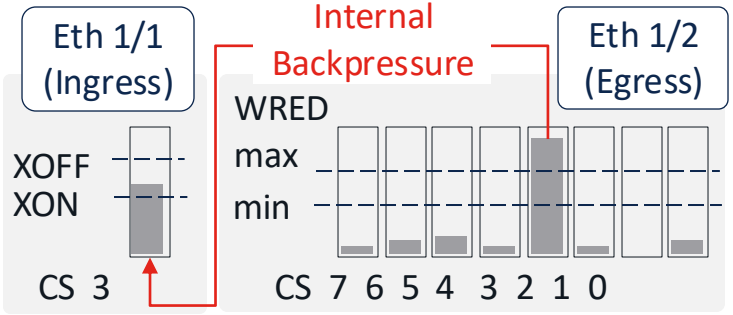




RoCEv2 Congestion Management 동작

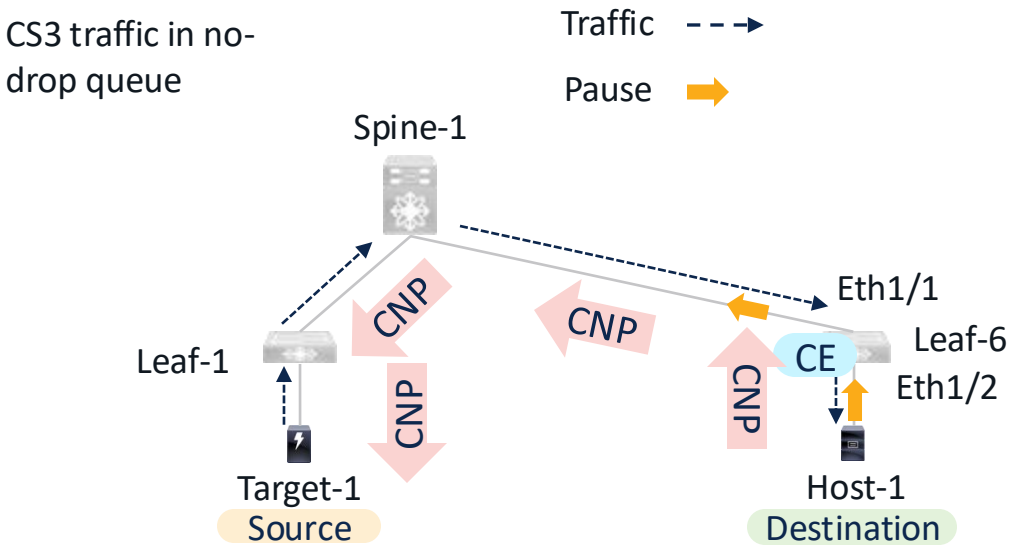
Target-1은 CNP를 수신하고 Host-1로 송신되는 트래픽 전송량을 줄입니다.

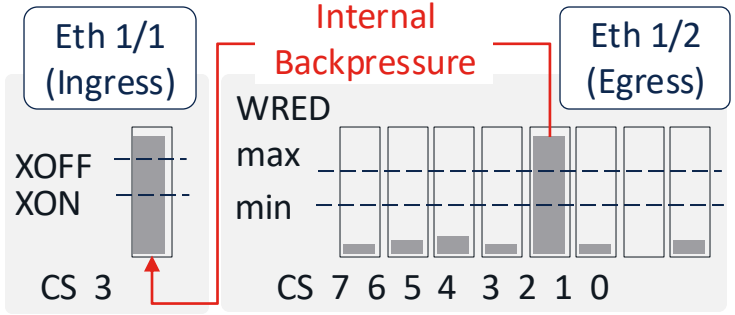




RoCEv2 Congestion Management 동작

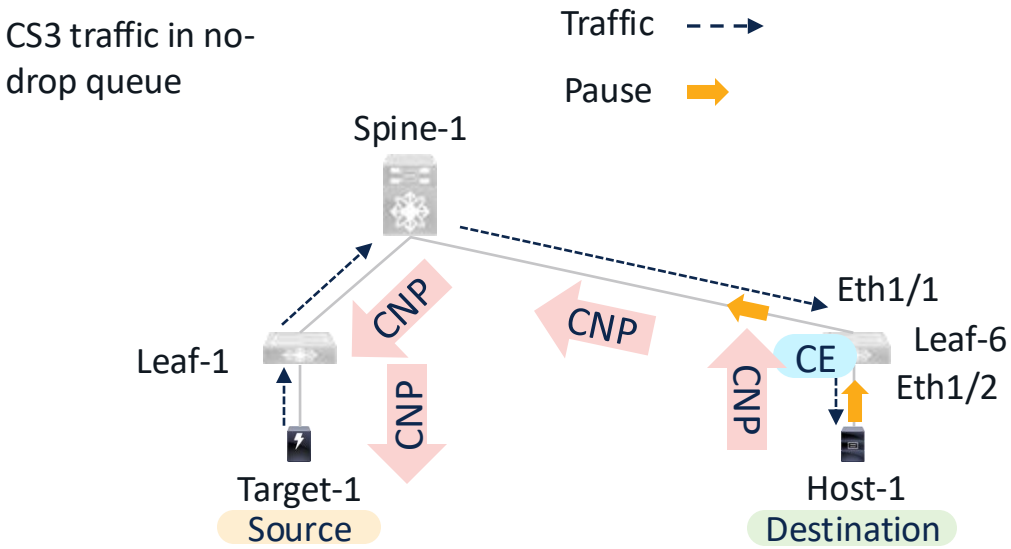
Target-1은 CNP 수신 이 후 Host-1로 송신되는 트래픽 전송을 줄임에도 불구하고 Leaf-6의 Ingress Port (Eth1/1)의 XOFF 값을 초과하게 되면 PFC 패킷을 Spine-1로 전송합니다.

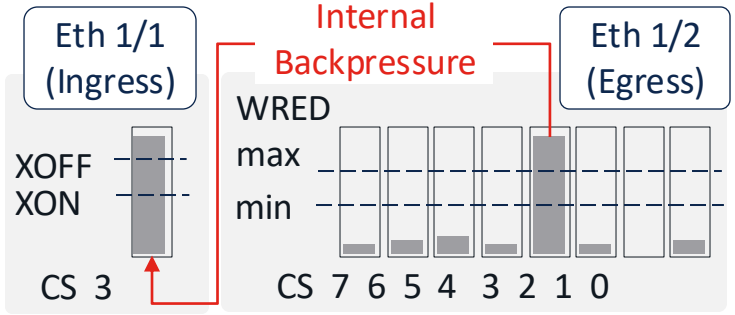




RoCEv2 Congestion Management 동작

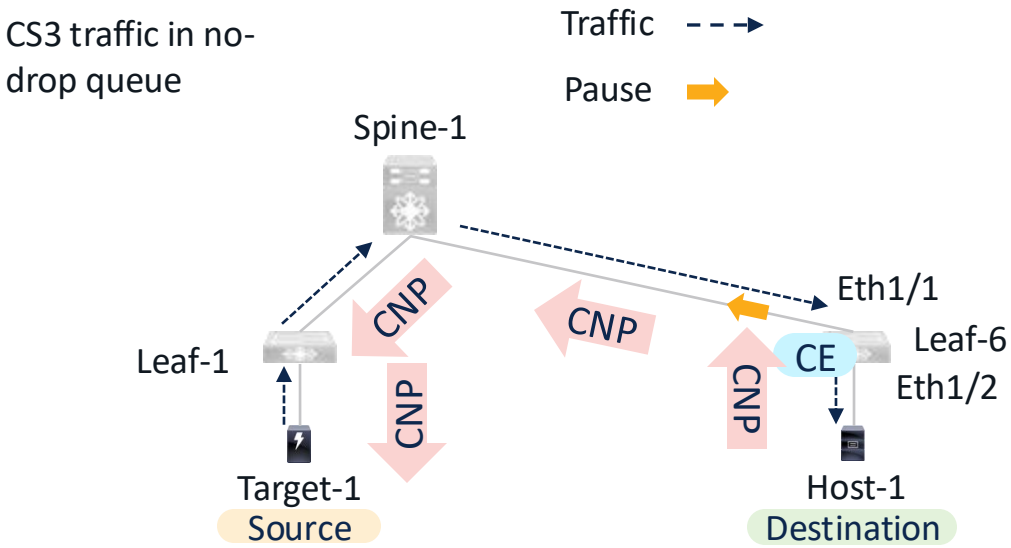
PFC를 수신한 Spine-1은 Leaf-6의 CS3 queue로 전송하는 트래픽을 중단합니다.

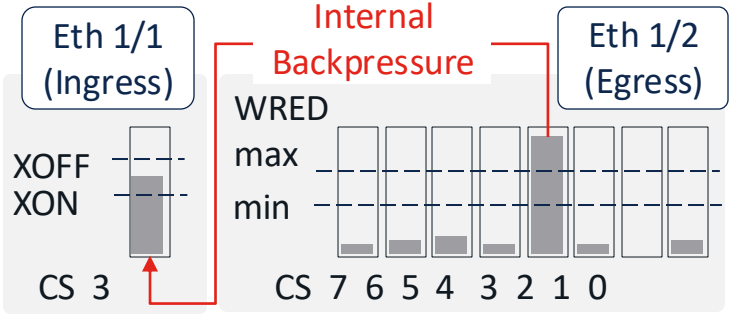




RoCEv2 Congestion Management 동작

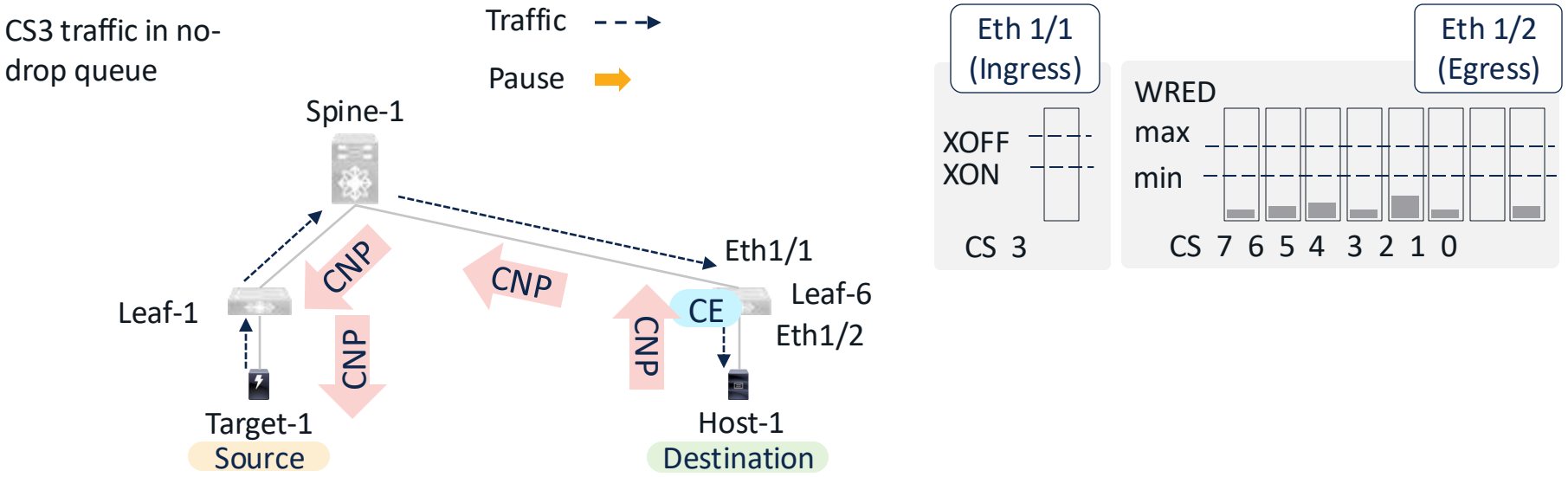
Leaf-6은 CS3 queue 버퍼링된 트래픽을 Host-1에 전송하여 Ingress Buffer의 사용률이 XOFF값 이하로 내려갑니다.





RoCEv2 Congestion Management 동작

Leaf-6은 Congestion 상황이 해소 됨으로 인해 PFC 전송을 중단하고 Egress Buffer의 상태도 정상으로 복구됩니다.



AI 환경에서의 부하 분산 기술

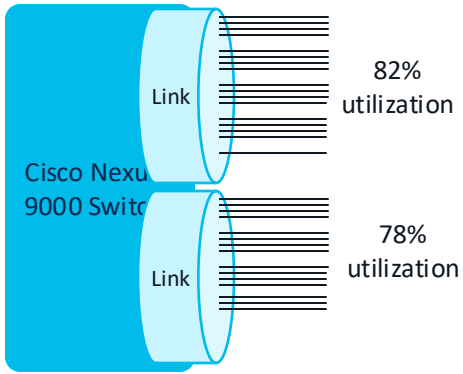
cisco Connect

Cisco에서 제공하는 RoCEv2 부하 방법

Cisco는 크게 3가지의 부하 분산 (ECMP with UDF, Static Pinning, DLB) 방법을 제공합니다.

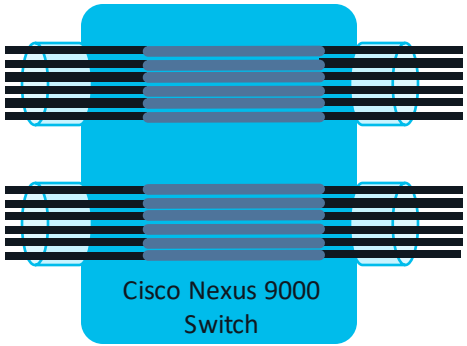
ECMP with UDF

IB의 Queue pair를 이용하여 부하 분산을 수행하는 방법



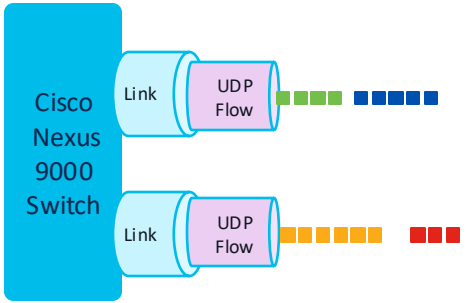
Static Pinning

Ingress Port와 Egress Port를 1:1로 매핑 시키는 방법



Dynamic Load Balancing(DLB)

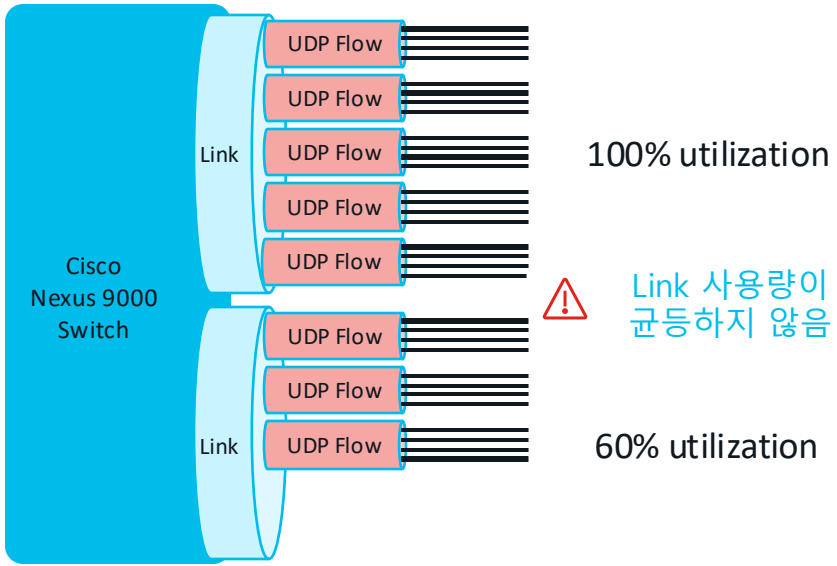
Flowlet 기반의 LB 방법으로 Link의 사용량을 기준으로 가장 낮은 사용량을 가진 인터페이스로 부하 분산을 수행하는 방법



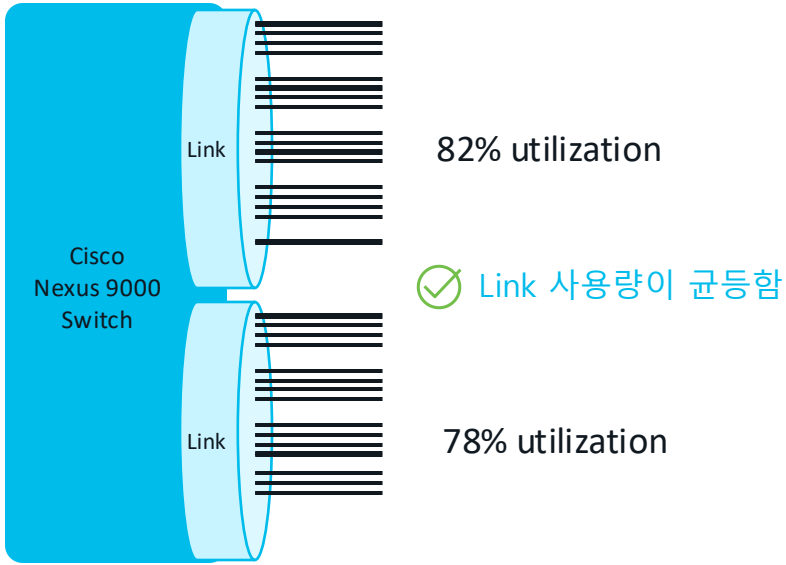
RoCE 환경에서의 UDF with ECMP의 필요성

RDMA 트래픽은 동일한 UDP Flow (Source IP, Destination IP, protocol (UDP), source port, destination port)내에서 다수의 Destination Queue-pair를 가질 수 있습니다.

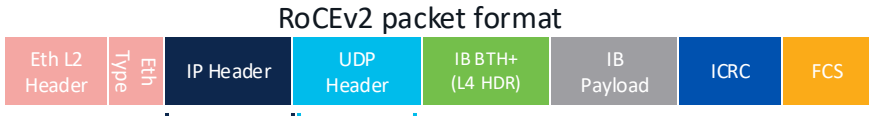
RoCE 환경에서 Default ECMP Load-Balancing 사용



IB의 Destination Queue-pair 기반의 ECMP 사용



RoCE의 IB Queue Pair



IPv4 header: 20 bytes

UDP header: 8 bytes

Location of QP in IB header: 5 bytes

QP field size: 3 bytes

```
No.   Time          Source        Destination   Protocol  Length Info
  348 102.9157652…  10.1.15.226   10.1.15.230   RRoCE        106 RC RDMA Write Only QP=0x00c96d
  349 102.9157749…  10.1.15.230   10.1.15.226   RRoCE         62 RC Acknowledge QP=0x000f98
  350 102.9185232…  10.1.15.226   10.1.15.230   RRoCE       4078 RC RDMA Write Only Immediate QP=0x00c96f
  351 102.9185301…  10.1.15.226   10.1.15.230   RRoCE        106 RC RDMA Write Only QP=0x00c96d

> Frame 350: 4078 bytes on wire (32624 bits), 4078 bytes captured (32624 bits) on interface ps-inb, id 0
> Ethernet II, Src: c4:70:bd:0a:ed:ca (c4:70:bd:0a:ed:ca), Dst: cc:36:cf:86:44:c3 (cc:36:cf:86:44:c3)
v Internet Protocol Version 4, Src: 10.1.15.226, Dst: 10.1.15.230
    0100 .... = Version: 4
    .... 0101 = Header Length: 20 bytes (5)
  > Differentiated Services Field: 0x62 (DSCP: CS3, ECN: ECT(0))
    Total Length: 4064
    Identification: 0xf559 (62809)
  > 010. .... = Flags: 0x2, Don't fragment
    ...0 0000 0000 0000 = Fragment Offset: 0
    Time to Live: 64
    Protocol: UDP (17)
    Header Checksum: 0x0188 [validation disabled]
    [Header checksum status: Unverified]
    Source Address: 10.1.15.226
    Destination Address: 10.1.15.230
v User Datagram Protocol, Src Port: 53676, Dst Port: 4791
    Source Port: 53676
    Destination Port: 4791
    Length: 4044
  > Checksum: 0x0000 [zero-value ignored]
    [Stream index: 5]
  > [Timestamps]
    UDP payload (4036 bytes)
v InfiniBand
  v Base Transport Header
      Opcode: Reliable Connection (RC) - RDMA WRITE Only with Immediate (11)
      0... .... = Solicited Event: False
      .1.. .... = MigReq: True
      ..00 .... = Pad Count: 0
      .... 0000 = Header Version: 0
      Partition Key: 65535
      Reserved: 00
      Destination Queue Pair: 0x00c96f
      1... .... = Acknowledge Request: True
      .000 0000 = Reserved (7 bits): 0
      Packet Sequence Number: 57
  > RETH - RDMA Extended Transport Header
  > IMMDT - Immediate Data Extended Transport Header
    Invariant CRC: 0xc9d4528f
> Data (4000 bytes)
```

IB Queue Pair 기반의 LB 구성

User-Defined Field(UDF)를 이용하여 IB Queue Pair 기반의 LB를 구성 할 수 있습니다.

기본 LB 구성

```
N9K# show ip load-sharing
IPv4/IPv6 ECMP load sharing:
Universal-id (Random Seed): 564667391
Load-share mode : address source-destination port source-destination
Exclude L3 proto from ECMP hashing : Disabled
Rotate: 32
```

IB queue pair 기반의 LB 구성

```
(config)# ip load-sharing address source-destination udf offset 33 length 24
```

Offset in bytes from L3 header

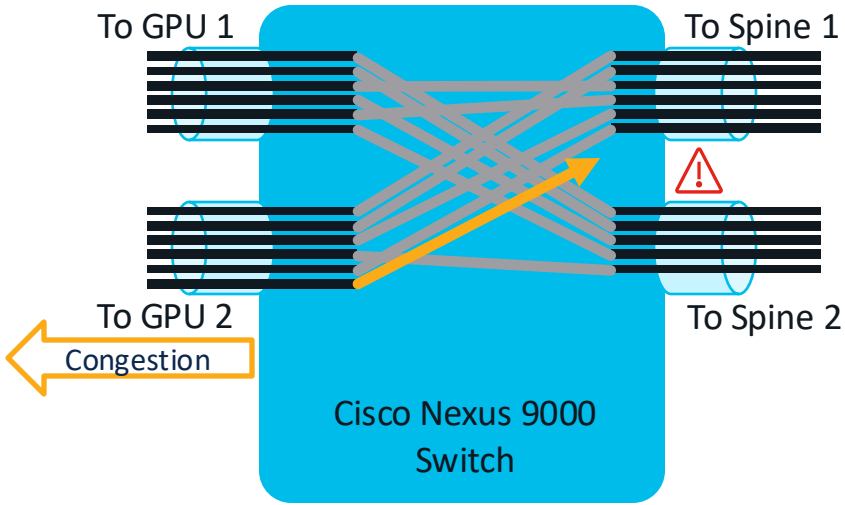
UDF length in bits

```
N9K# show ip load-sharing
IPv4/IPv6 ECMP load sharing:
Universal-id (Random Seed): 4003426154
Load-share mode : address source-destination udf offset 33 length 24
Exclude L3 proto from ECMP hashing : Disabled
Rotate: 32
```

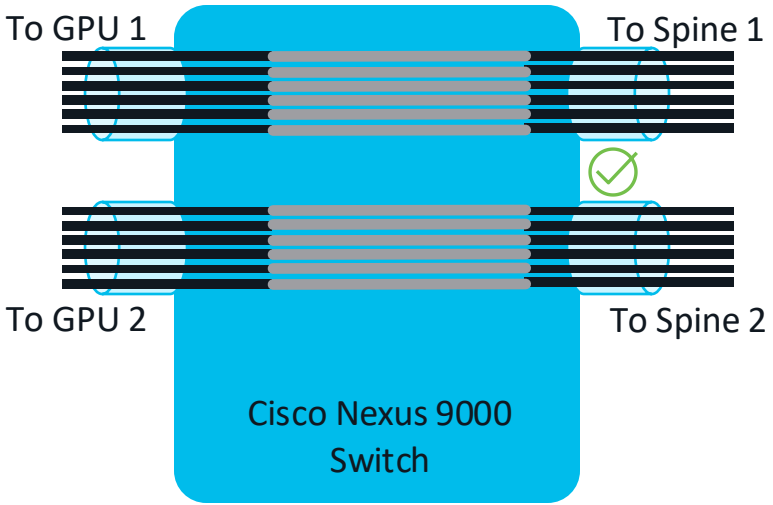
Static Pinning

Static Pinning을 통해서 GPU와 Spine간의 연결을 1:1로 매핑 할 수 있습니다.

Default ECMP Load-Balancing

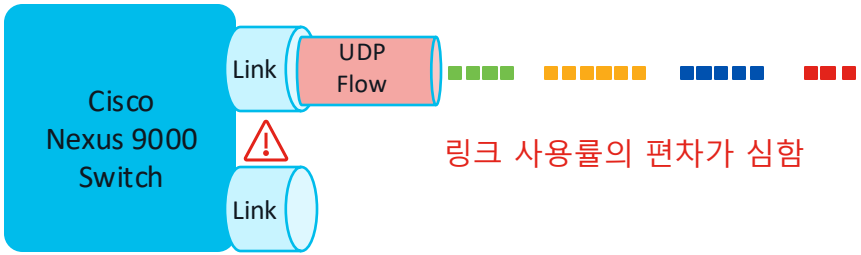


트래픽의 Input 포트와 Output 포트를 정적으로 매핑

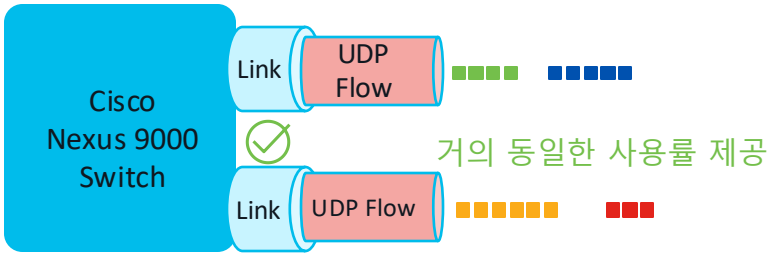


Dynamic Load Balancing (DLB)

Default ECMP Load-Balancing

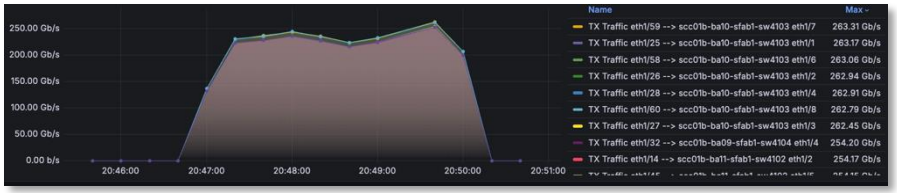


패킷 간의 갭으로 식별하는 Flowlet을 기반으로 한
실시간 사용량 기반의 Load-Balancing 방법



Utilization of all equal-cost paths on a leaf switch

AllReduce 256 GPUs
(Max: 327.22 Gbps, Min: 76.68 Gbps)



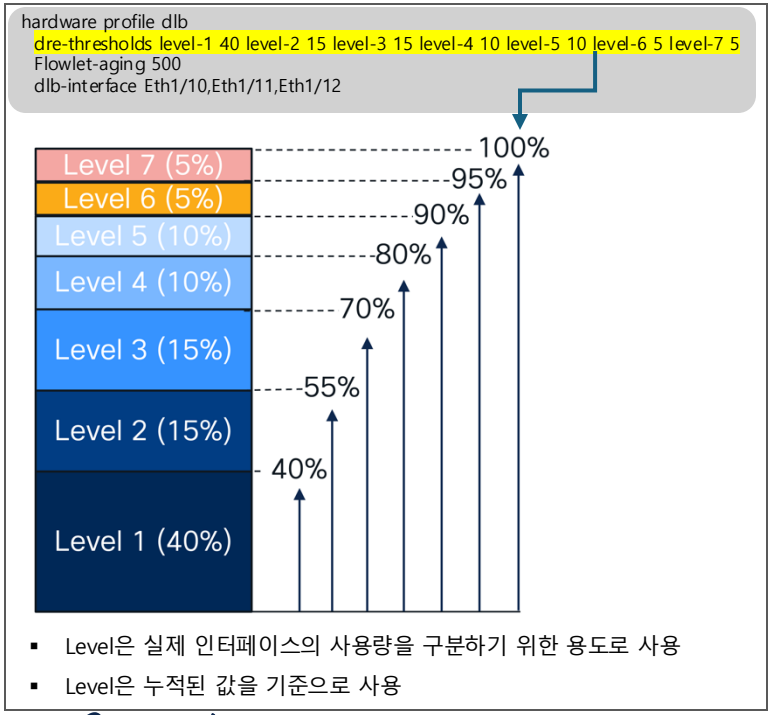
AllReduce 256 GPUs
(Max: 263.31 Gbps, Min:253.66 Gbps)

거의 동일한 사용률

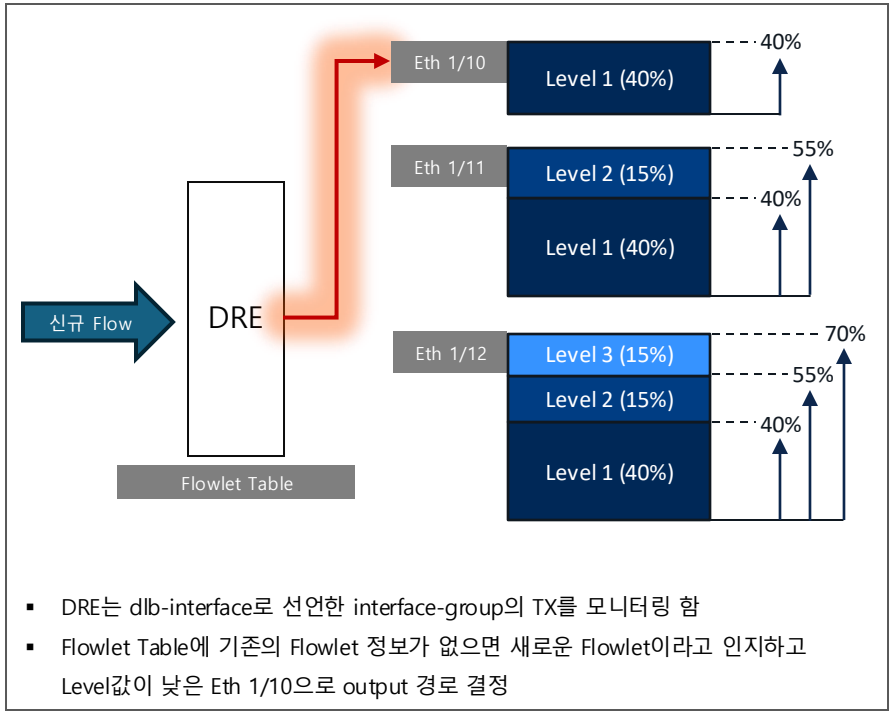
cisco Connect

Dynamic Load Balancing의 동작 방법

Flowlet Table에 정보가 없는 신규 Flowlet이 인입되면 Level값이 가장 낮은 인터페이스로 스위칭합니다.



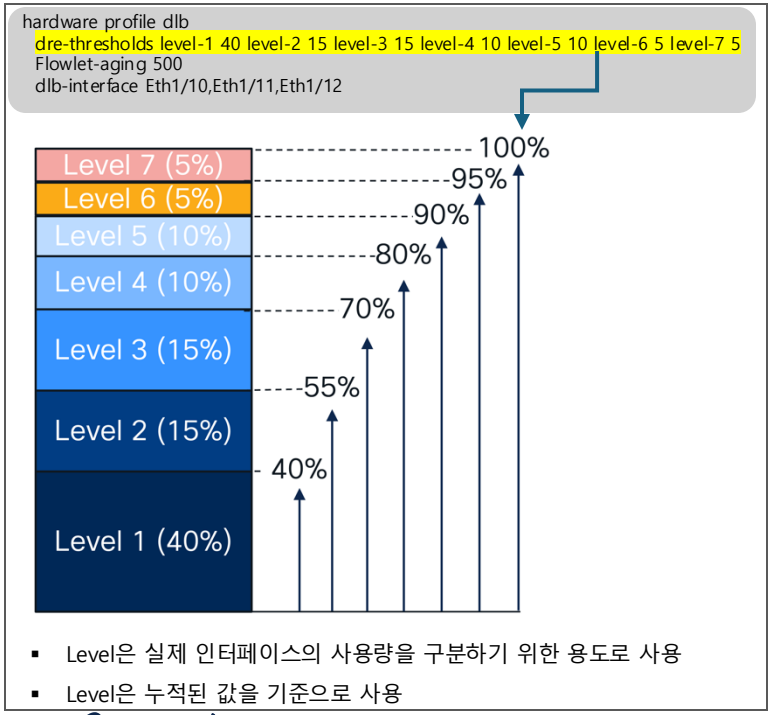
- Level은 실제 인터페이스의 사용량을 구분하기 위한 용도로 사용
- Level은 누적된 값을 기준으로 사용



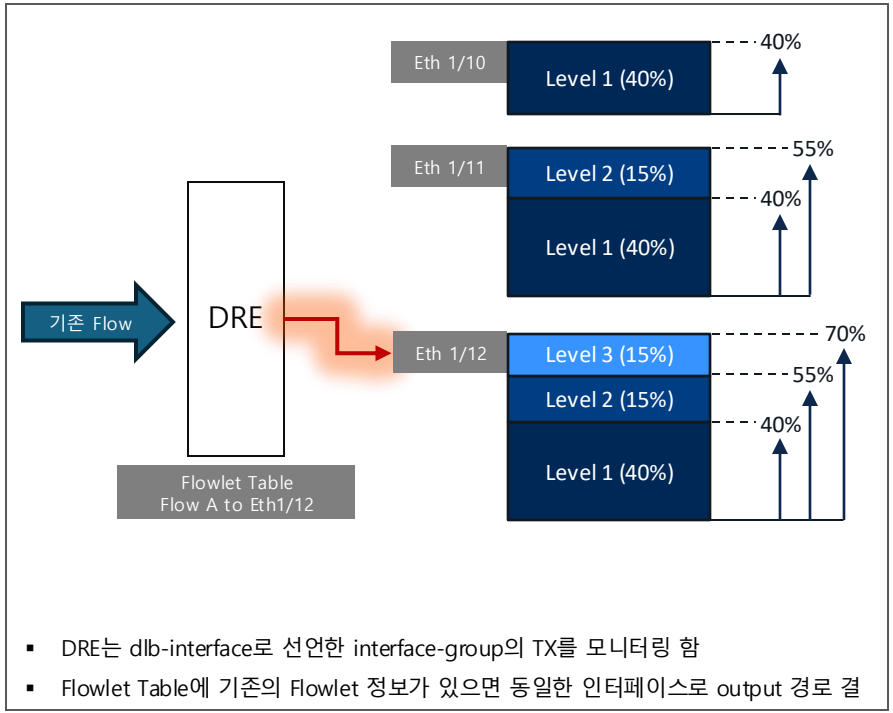
- DRE는 dlb-interface로 선언한 interface-group의 TX를 모니터링 함
- Flowlet Table에 기존의 Flowlet 정보가 없으면 새로운 Flowlet이라고 인지하고 Level값이 낮은 Eth 1/10으로 output 경로 결정

Dynamic Load Balancing의 동작 방법

Flowlet Table에 정보가 있는 Flowlet이 인입되면 Level값을 무시하고 Flowlet Table을 참조하여 기존 포트로 스위칭합니다.



- Level은 실제 인터페이스의 사용량을 구분하기 위한 용도로 사용
- Level은 누적된 값을 기준으로 사용



- DRE는 dlb-interface로 선언한 interface-group의 TX를 모니터링 함
- Flowlet Table에 기존의 Flowlet 정보가 있으면 동일한 인터페이스로 output 경로 결정

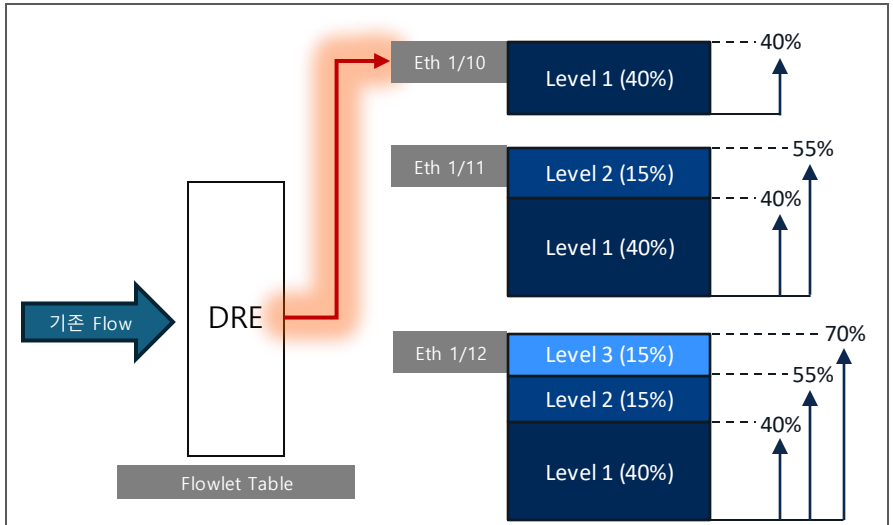
Dynamic Load Balancing의 동작 방법

Flowlet Table의 정보가 Flowlet-aging이 지나면 기존의 Flowlet table을 삭제 하고
Level값이 가장 낮은 인터페이스로 스위칭합니다.

```
hardware profile dlb
  dre-thresholds level-1 40 level-2 15 level-3 15 level-4 10 level-5 10 level-6 5 level-7 5
  Flowlet-aging 500
  dlb-interface Eth1/10,Eth1/11,Eth1/12
```

Flowlet Table
Flowlet A to Eth1/12

- Flowlet-aging time이 지나면 기존에 작성한 Flowlet table을 삭제 하고 재 작성
- Flowlet-aging time은 다른 경로를 이용하더라도 목적지에서 re-ordering이 일어나지 않음을 보장하는 시간을 의미



- DRE는 dlb-interface로 선언한 interface-group의 TX를 모니터링 함
- Flowlet Table에 기존의 Flowlet 정보가 없으면 새로운 Flowlet이라고 인지하고 Level값이 낮은 Eth 1/10으로 output 경로 결정

Summary

- GPU Server Cluster를 위해 RDMA를 사용하며 이를 Ethernet에 사용할 수 있게 하는 것이 RoCE 입니다.
- AI 네트워크를 구성하기 위해서는 Front-end, Back-end, Storage Network이 필요합니다.
- RoCE에서는 Lossless Ethernet을 구현하기 위해 PFC+ECN을 이용합니다.
- Cisco는 RoCE 네트워크의 효율적인 부하 분산을 위해 ECMP with UDF, Static Pinning, Dynamic Load Balancing(DLB) 기능을 제공합니다.


cisco
The bridge to possible
Thank you
cisco Connect