The bridge to possible

AI 시대를 위한 초저지연·고성능 네트워크:

Cisco Nexus로 구현하는 AI Networking

임규현 이사, 시스코코리아

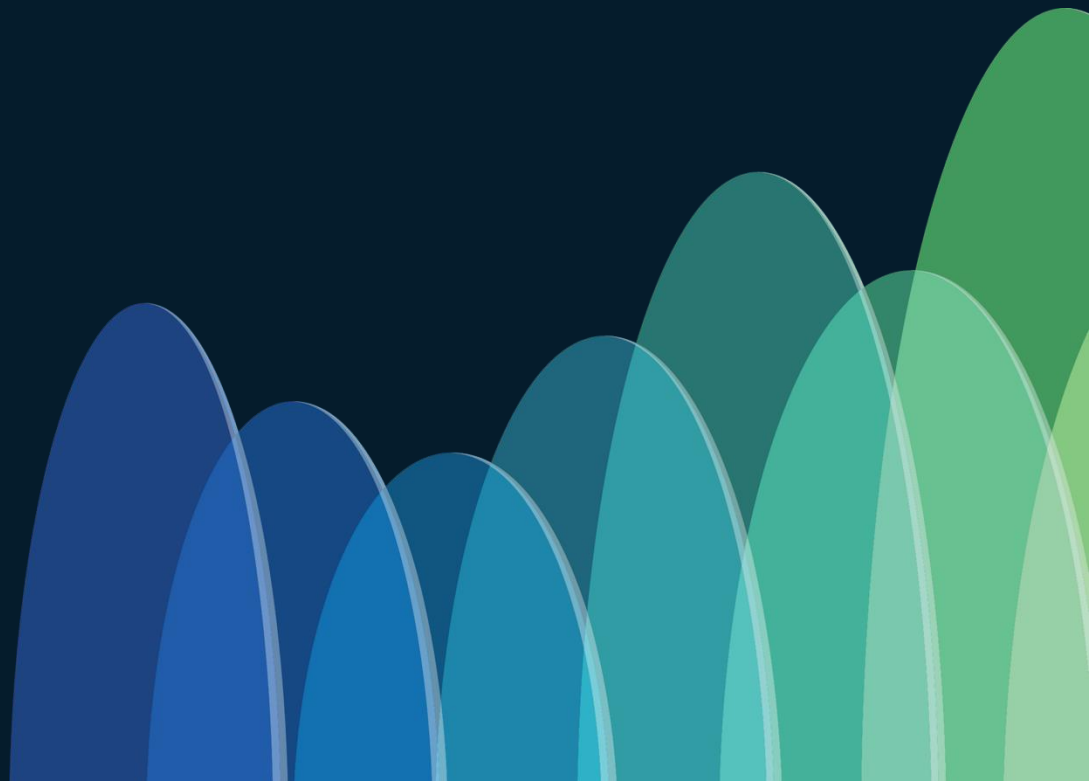
Cloud & AI Infrastructure team

cisco Connect

Agenda

- AI Networking 특징
- AI Networking 요구사항
- Cisco Nexus가 해결하는 AI Networking의 숙제

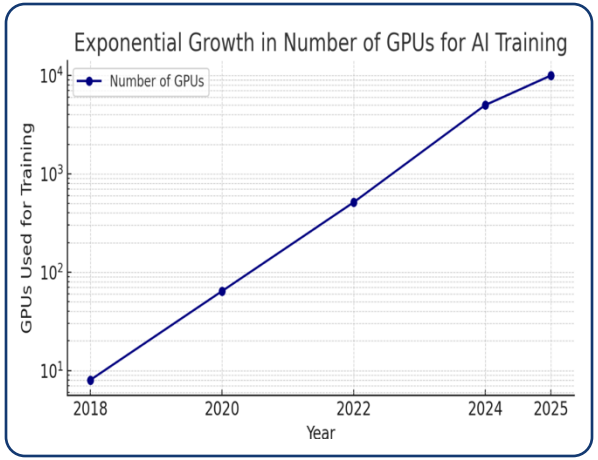
AI Network 특징



네트워크가 AI 성능에 미치는 영향도

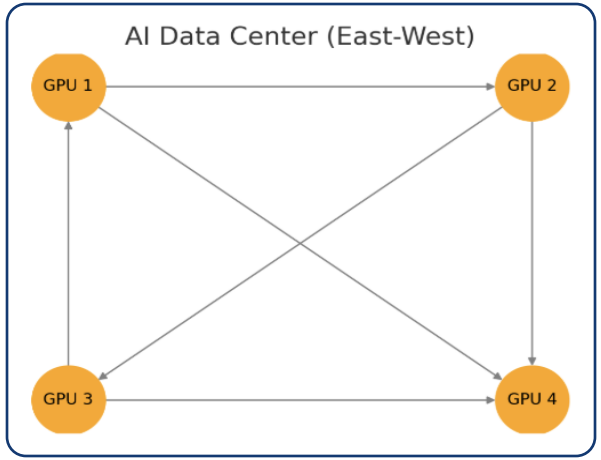
대규모 AI 클러스터를 수용하면서도, 병목을 제거하는 것이 핵심 목표

AI 훈련에 요구하는 GPU 클러스터 규모증가



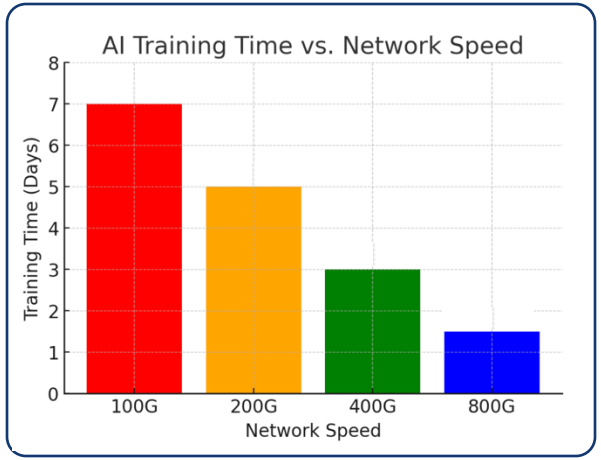
기하 급수적으로 증가하는 GPU Cluster 사이즈
(OpenAI GPT-4 / xAI Grok 3 : 100,000)

GPU 클러스터 간 East-West 트래픽 증가



단일 서버가 아닌,
수천~수만개의 GPU를 활용한 병렬 처리

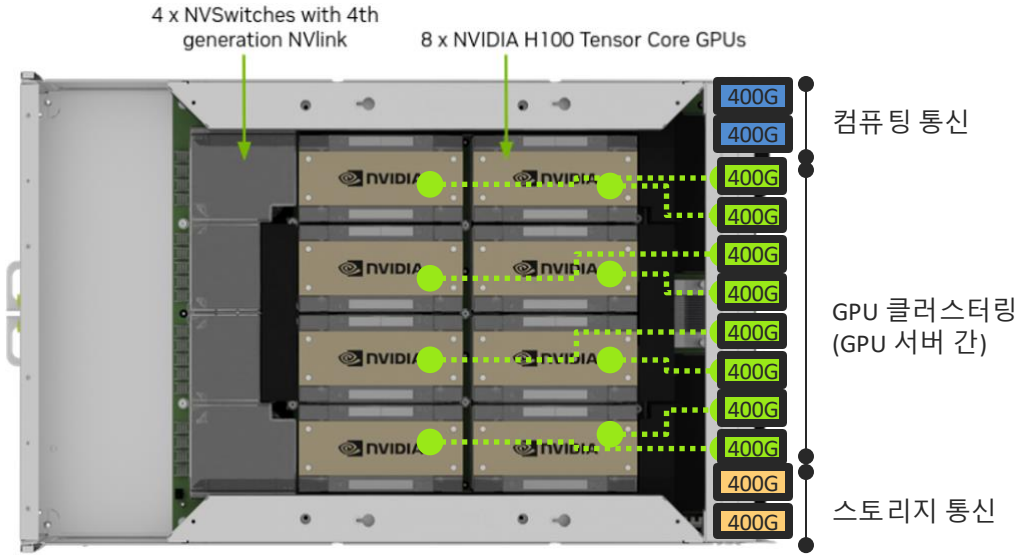
AI 학습시간 네트워크속도 증가 따라 단축



네트워크 속도증가 → 학습시간의 획기적 단축

AI의 시작 : 네트워크장비와 연결되는 GPU 서버

AI Network 이해를 위한 GPU 서버 구조 확인



전력과 상면

- Power Consumption : 10.2 kW Max (46A)

400G 스위치 : 3A(Max 기준)

- 5 RU(Rack Unit)

네트워크 연결 인터페이스

GPU 클러스터링을 위한 네트워크

- 400G x 8포트(OSFP)

컴퓨팅/스토리지 통신을 위한 네트워크

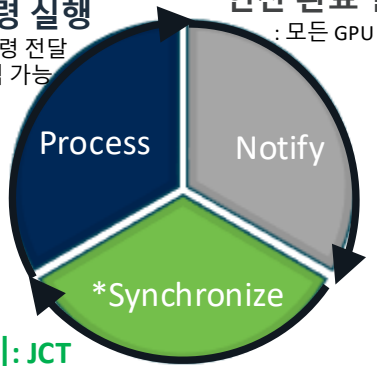
- 400G x 4 포트(QSFP-DD)

GPU Cluster 가 발생시키는 Burst 트래픽 패턴

AI Training Process 요약

GPU 명령 실행
: GPU 클러스 내 다른 GPU에게 명령 전달
: Network Link 잠식 가능

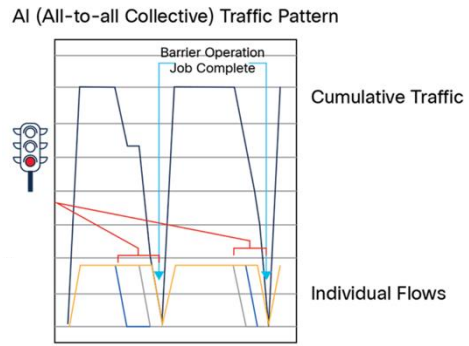
연산 완료 결과 전달
: 모든 GPU 병렬 처리 수행



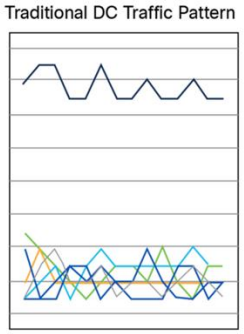
연산 완료 결과 수신 대기: JCT
: GPU 클러스터 내 모든 GPU로 부터 대기
: 기존 트래픽 패턴에 없었던 AI환경의 추가된 패턴

정확한 모델의 학습을 위해 위 프로세스를 수 차례 반복

AI 트래픽 패턴	VS	일반적인 데이터센터 트래픽 패턴
트래픽의 ***급격한 증감*** 반복		비교적 ***일정한 흐름***



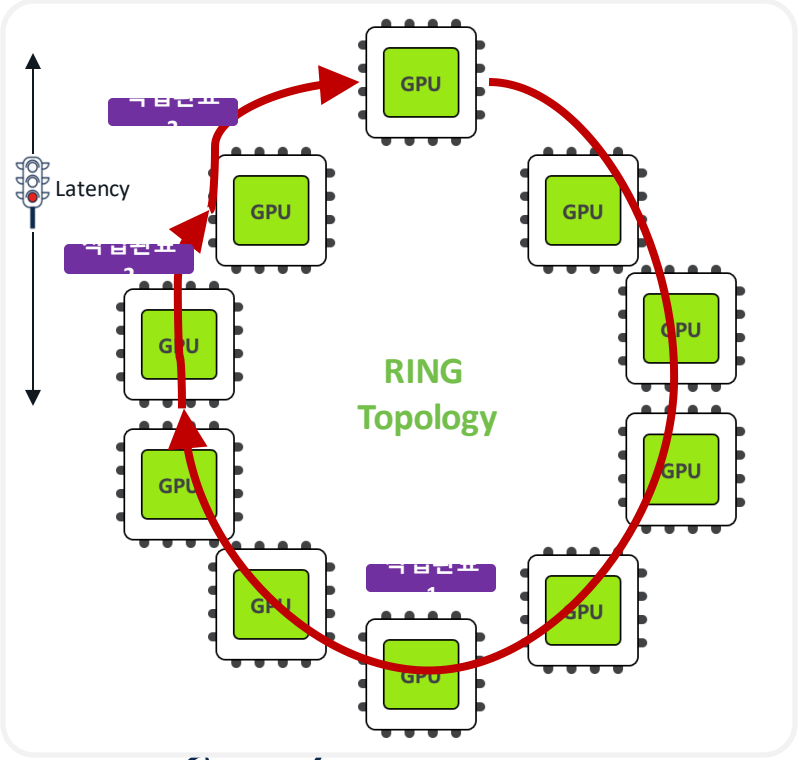
Few synchronous high BW flows
Synchronization magnifies long tail latency and bad load balancing decisions



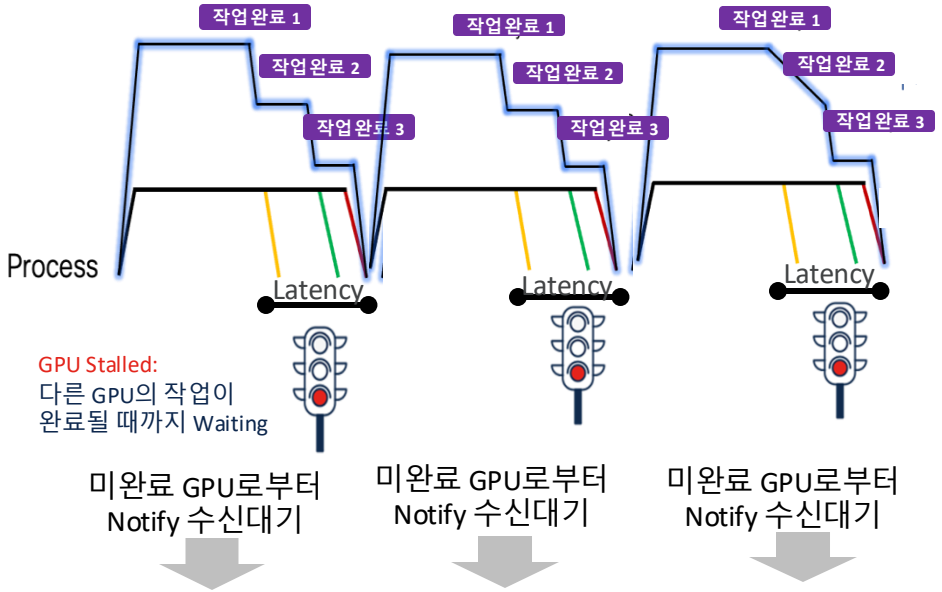
Many asynchronous small BW flows
Chaotic pattern averages out to consistent load

Burst 발생이유와 성능저하 요인 이해

GPU Cluster의 Ring Topology 트래픽 플로우



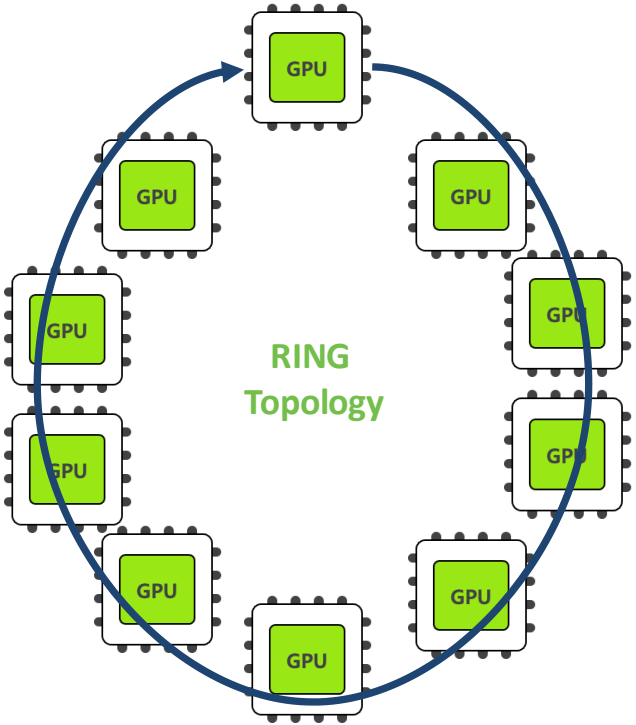
GPU Cluster에서 발생하는 Latency 이해



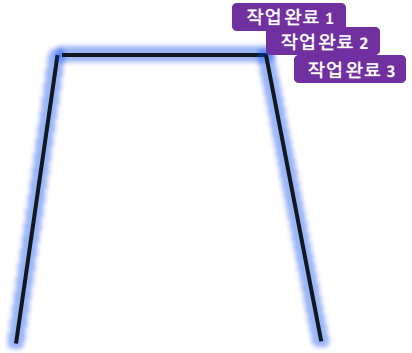
데이터 불일치 / 성능 저하 / 병렬처리 장점 감소

AI Network의 목표

AI Network의 목표

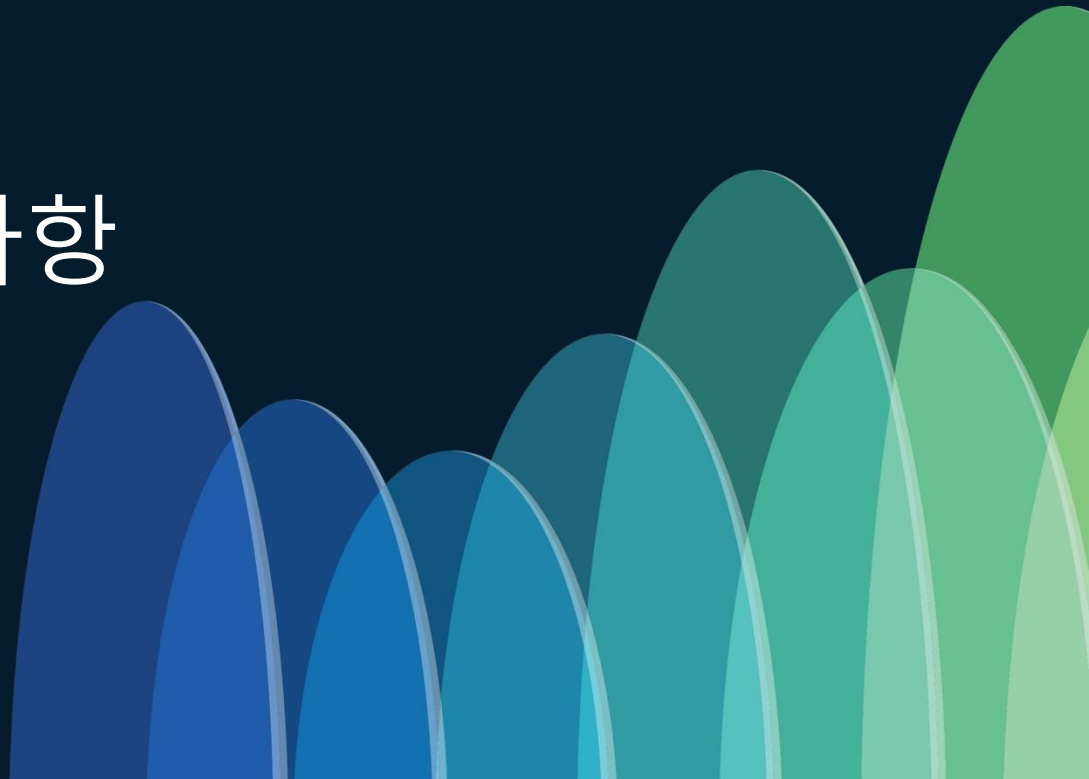


최적의 AI Network 통신 완료 그래프



미완료 GPU가 없기 때문에,
Notify 수신대기 없음

AI Network 요구사항



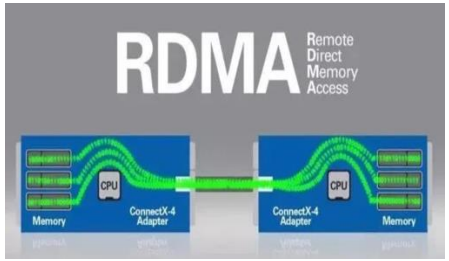
cisco Connect

AI Network 핵심 요구사항

Bandwidth

800/400G 기반
Non-Blocking Network

Latency



초저지연, Lossless Network

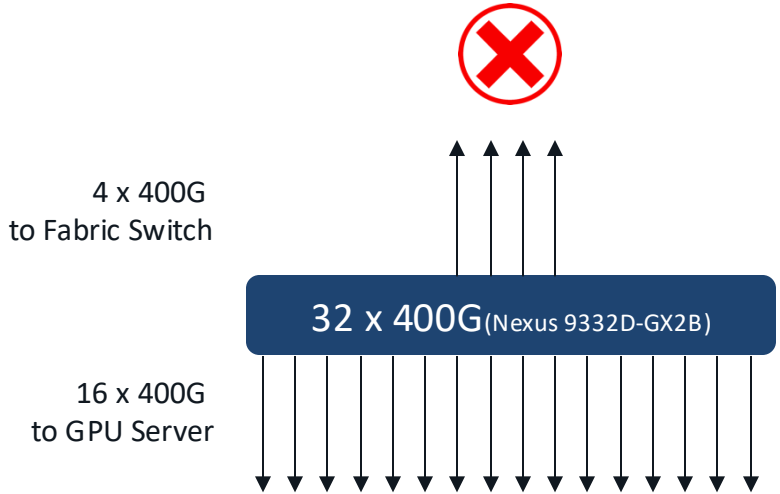
Advanced Load Balancing

기존 방식을 뛰어넘는
Load Balancing

Non Blocking Network 인터페이스 설계

Uplink : Downlink Oversubscription 없는 1:1 비율

일반적인 1:4 Oversubscription Ratio



Non Blocking을 위한 1:1 Oversubscription Ratio

초저지연/고성능을 위한 RoCEv2

핵심 컨셉 : CPU를 미경유 → CPU 보호 및 성능의 극대화

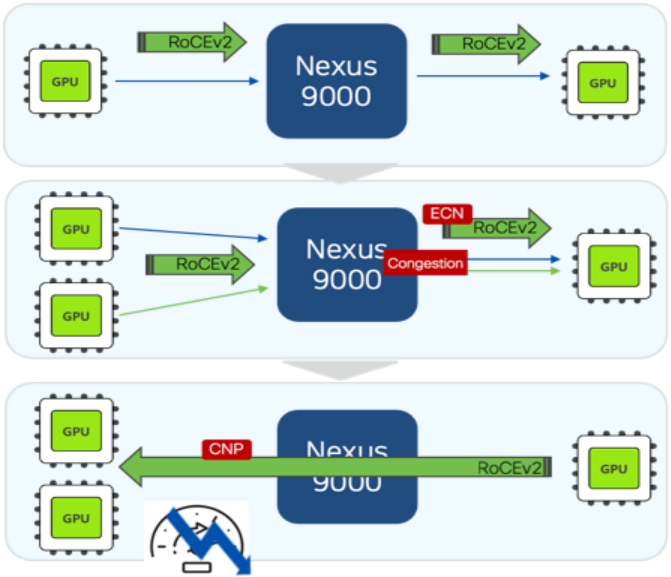
Data Center Quantized Congestion Notification (DCQCN)

- DMA(Remote Direct Memory Access) : 장치들이 CPU를 거치지 않고, 메모리에 직접 접근할 수 있는 기술
- RDMA(Remote Direct Memory Access) : DMA를 네트워크를 통해 원격지(Remote)로 확장하는 기술
- *RoCE(RDMA Over Converged Ethernet) : RDMA 통신을 전달하는 이더넷 기술

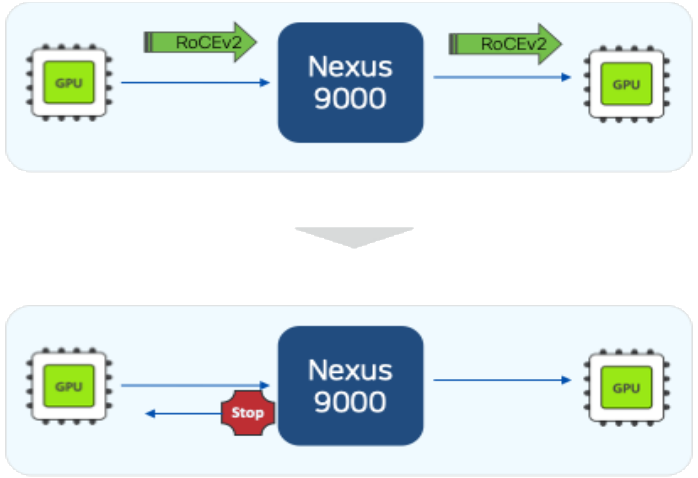
RoCEv2 Flow Control : ECN, PFC

Lossless Network 구현을 위한 RoCE의 Flow Control

1차 방어 - Explicit Congestion Notification(ECN)



2차 방어 - Priority Flow Control(PFC)



AI Network를 위한 Lossless Ethernet, Non-Blocking 구성

1) Non-Blocking 1:1 Oversubscription

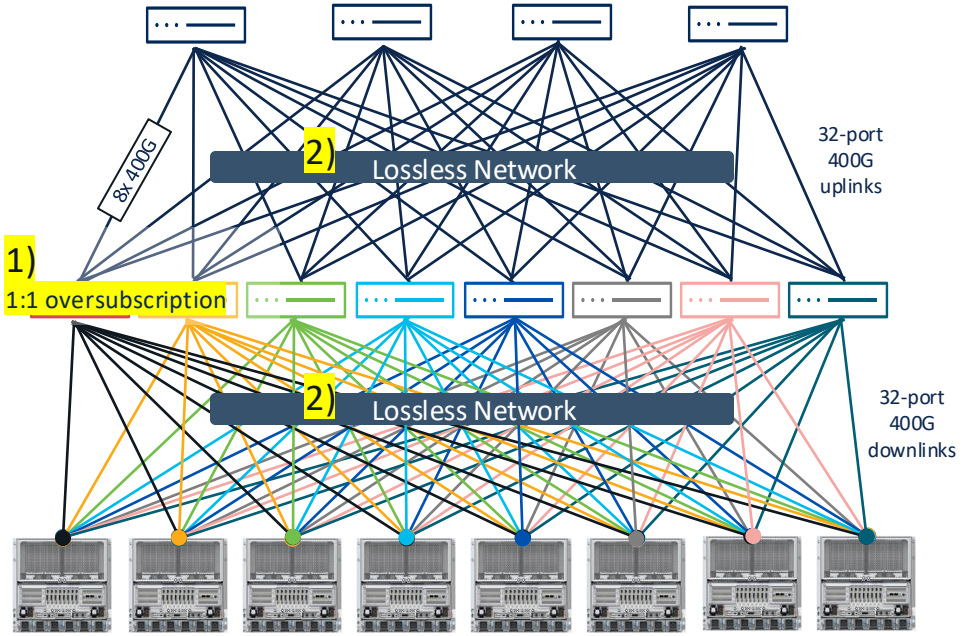
Network Switch는 Up/Down Link 1:1 Subscription 구성
► 32포트 업링크(To Spine) / 32포트 다운링크(To Server)

Non Blocking 네트워크 구성

2) Lossless Network

(RoCEv2) RDMA Over Ethernet v2를 위한 Lossless 구성
►AI Fabric을 위한 ECN, PFC Flow Control

Lossless 네트워크 구성



32 GPU Server, 256 GPUs 예

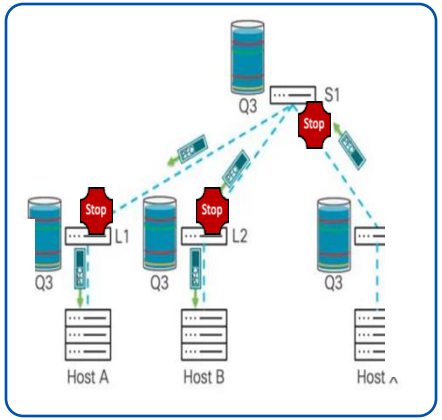
Cisco Nexus로 해결하는 AI Networking의 숙제

cisco Connect

여전히 해결해야 할 네트워크의 숙제

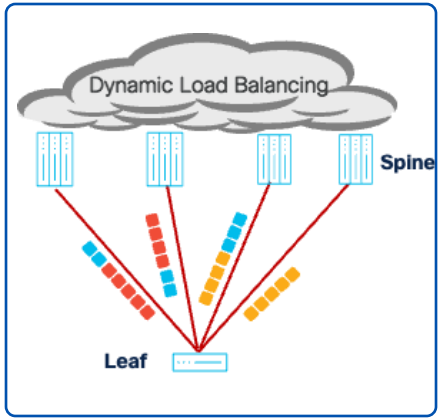
AI Network 주요 Challenge

전 Network 영역 Stuck 방지



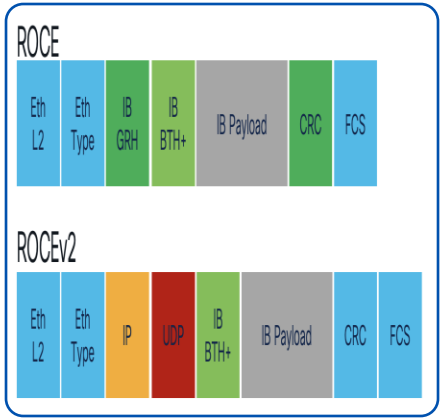
Lossless Network
Flow Control의 부작용

혼잡인지 기반 Load Balancing



Link 사용량에 관계없이
Forwarding 시
Back pressure 부작용

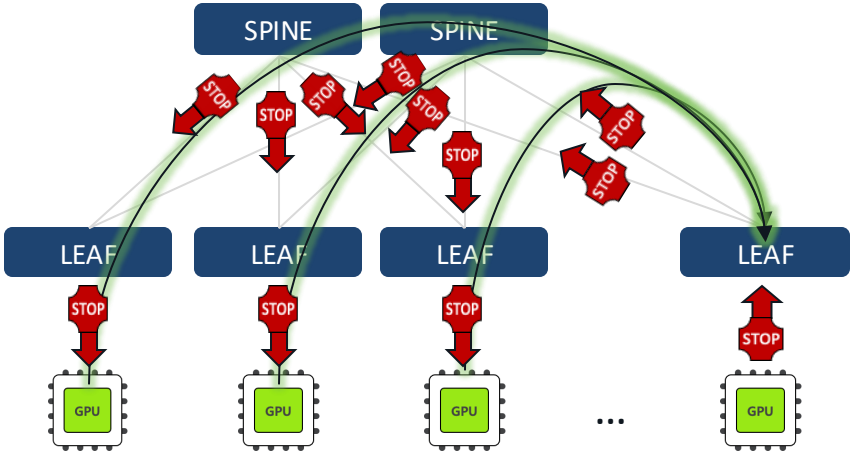
기존과 다른 RoCEv2의 상이한 패킷 헤더



IP/Port Hashing 기반
Load balancing의 부작용

AI Network 전체 Stuck 발생 이슈

PFC Storm에 의한 네트워크 전 영역 Stuck 방지 필요



PFC Storm 이슈

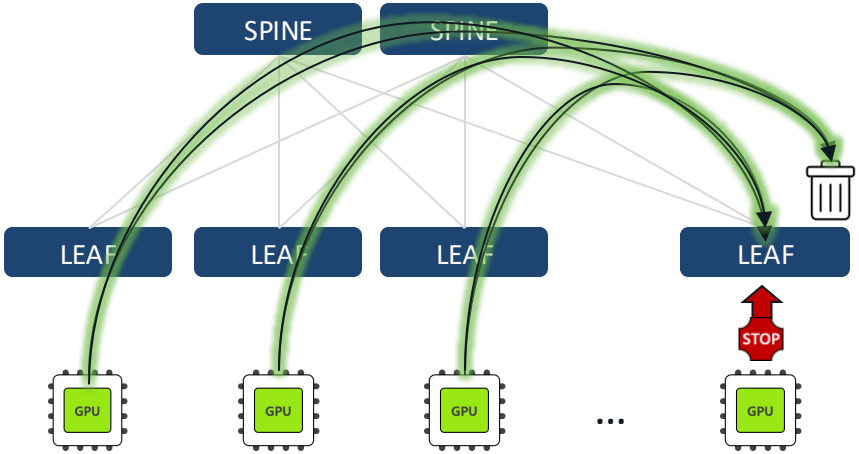
- Congestion 발생 시 PFC 동작
- PFC Storm 발생으로 네트워크 트래픽 전체 Stuck

Tx Frames	Rx Frames	Tx L1 Rate (bps)	Rx L1 Rate (bps)
213,029,928	0	9,999,999,744.000	0
213,029,928	0	9,999,999,744.000	0
213,029,927	0	9,999,999,744.000	0
213,029,928	0	9,999,999,744.000	0

수신 Rate : Zero (Network Stuck)

Nexus PFC Watchdog

PFC Storm에 의한 네트워크 전 영역 Stuck 방지



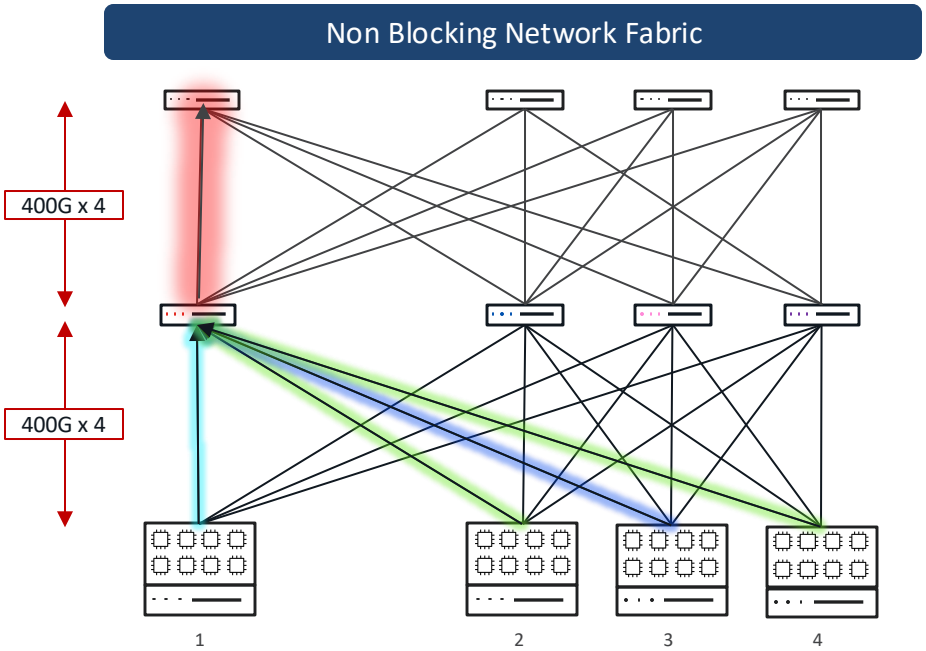
네트워크 Stuck을 방지하는 PFC Watchdog

- 과도한 PFC Storm 발생으로
  네트워크 트래픽 전체 Stuck를 사전에 차단

- Option 1. No drop Q에 Buffering 중인 트래픽 **강제 Drop**
- Option 2. Traffic이 쏠리는 포트 **강제 Shutdown**

ECMP의 함정: 균등해 보이지만 균등하지 않다

Default ECMP 방식의 한계점

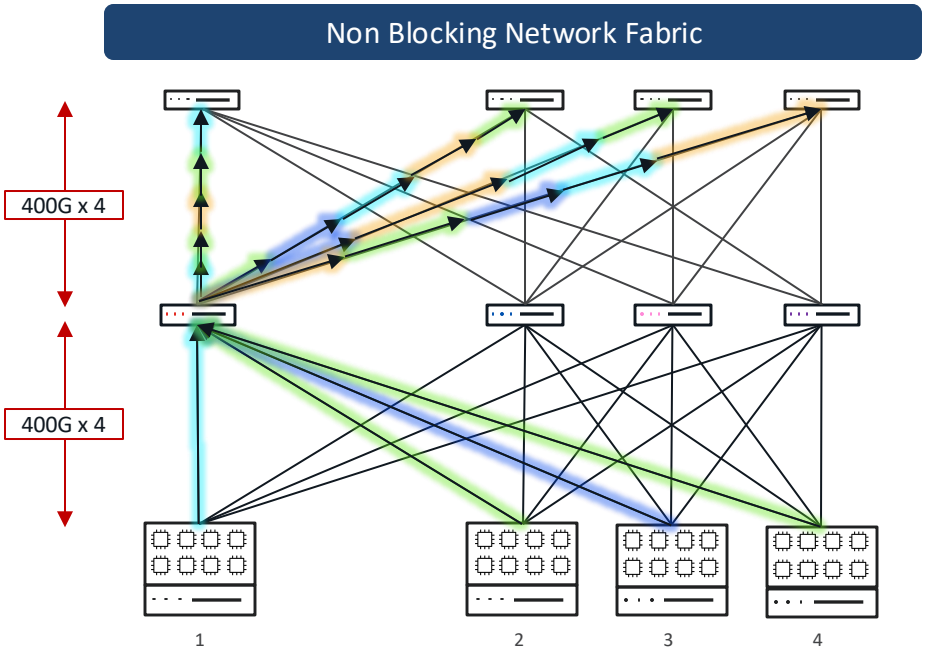


ECMP(Equal Cost Multi Path)의 한계

- 고대역폭 장비 투자로 Non Blocking Fabric 구축
- ECMP는 Equal Flow가 아니므로, 트래픽 쏠림

ECMP의 함정: 균등해 보이지만 균등하지 않다

Per Packet Load Balancing과 Packet Reordering



만약, Flow를 per packet 별로 Load Balancing 한다면?

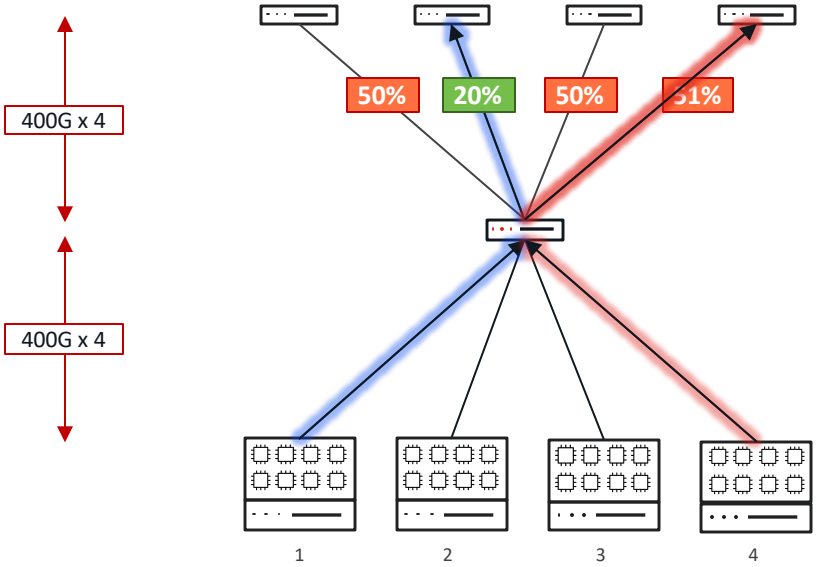
- Traffic 쏠림 현상 없음
- 하지만, **Packet Reordering이 발생**

도착지 스위치 or 서버 NIC 패킷 재조합 수행

- **Latency 증가**
- **Packet Drop(Out of Order)**

Dynamic Load Balancing(DLB)

DLB 강점 : Congestion에 의한 패킷 드랍 방지, 패킷 순서 보장을 통한 Reordering 현상 최소화



Flowlet 방식의 DLB

Tx Load Aware

- Link Utilization을 실시간으로 인지
- Tx Load가 가장 낮은 인터페이스로 Forwarding

Flowlet Switching

- Forwarding Flow에 대한 인터페이스 기록
- Aging Time 내 생성된 동일한 Flow는 동일한 인터페이스로 Forwarding 함으로써 Reordering 방지

AI Network 환경의 IP 기반 Hash의 한계

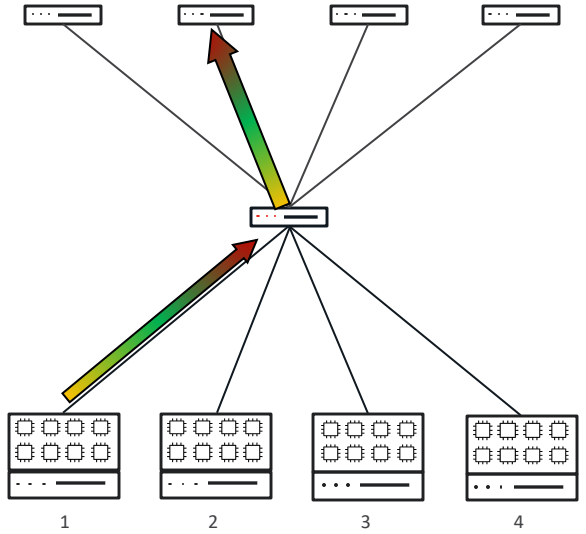
RoCE 트래픽은 동일한 5 Tuple Flow 이지만, Unique한 Q-Pair가 존재

RoCE 패킷플로우

Dst Port(UDP)		Src Port(UDP)	Dst IP	Src IP
Q-Pair : 1번	4791	65321	2.2.2.2	1.1.1.1
Q-Pair : 2번	4791	65321	2.2.2.2	1.1.1.1
Q-Pair : 3번	4791	65321	2.2.2.2	1.1.1.1
Q-Pair : 4번	4791	65321	2.2.2.2	1.1.1.1

31-24 bit		23-16 bit			15-8 bit	7-0 bit
OpCode		S	M	Pa Version	Partition Key	
ECN	Reserved	Destination QP				
A	Reserved	Packet Sequence Number				

RoCE 패킷으로 **IP Hash 기반** Load balancing 수행 시



동일한 Flow이기 때문에, 동일한 Link로 Forwarding Link를 지속적으로 잠식

UDF(User Defined Field) 기반 ECMP

Cisco Nexus는 Header 내 Field 값을 기준으로 ECMP 수행 가능

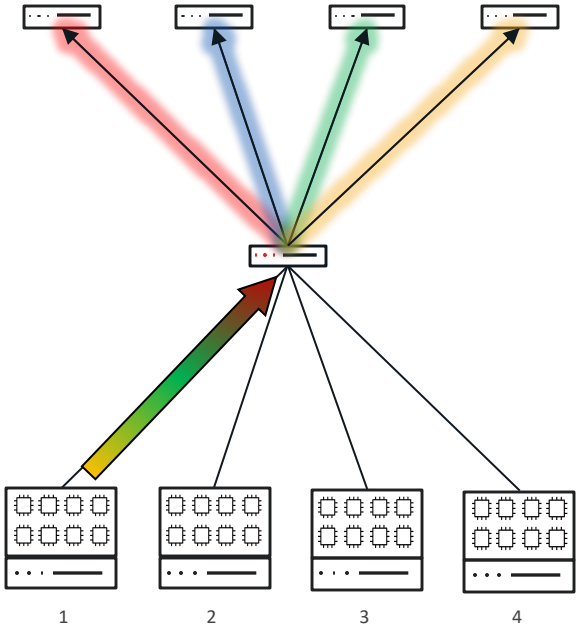
Nexus의 ECMP Forwarding을 위한 고급 Lookup 기능

- 단순 헤더가 아닌, 헤더 내 Field의 Q-Pair값을 기준으로 Load Balancing 수행

RoCE 패킷플로우

Dst Port(UDP)		Src Port(UDP)	Dst IP	Src IP
Q-Pair : 1번	4791	65321	2.2.2.2	1.1.1.1
Q-Pair : 2번	4791	65321	2.2.2.2	1.1.1.1
Q-Pair : 3번	4791	65321	2.2.2.2	1.1.1.1
Q-Pair : 4번	4791	65321	2.2.2.2	1.1.1.1

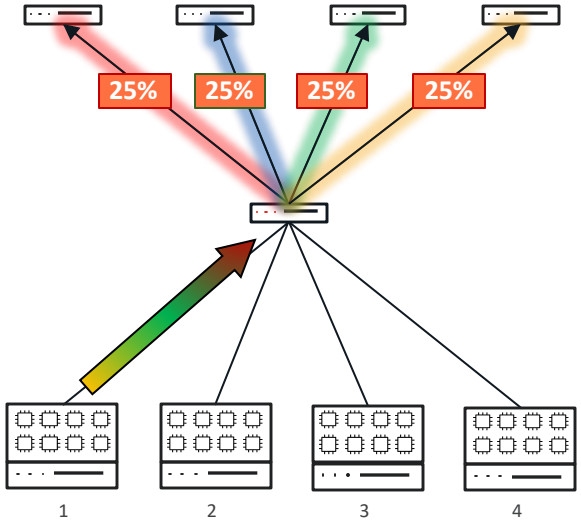
RoCE 패킷으로 **UDF 기반** Load Balancing 수행 시



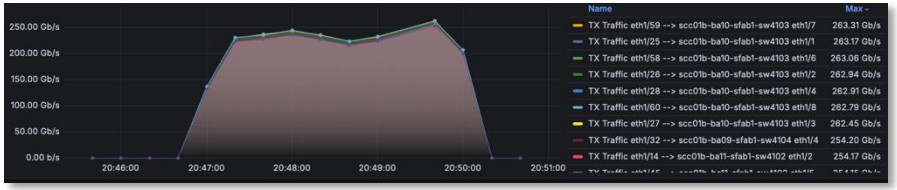
AI Network에 최적화된 기술조합(UDF + DLB)

UDF ECMP와 DLB기능 적용

- UDF ECMP : 동일 Flow 내에서 ***고유 Flow를 분리***
- DLB : UDF를 통해 ***고유하게 분리된 Flow***를
  Tx Load가 가장 낮은 Link로 Forwarding
  (Re-ordering 방지기술 포함)

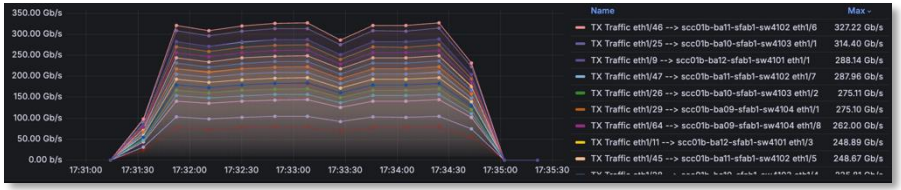


UDF+DLB 적용 후 : 트래픽 쏠림 현상 제거로 균등한 성능



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

UDF+DLB 적용 전 : 트래픽 쏠림 현상 발생으로 큰 성능 저하



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

Summary

- **고성능 AI Network은**

  'RoCE 기술 적용, Non Blocking을 위한 대규모 장비투자' 외에도 고려해야 할 사항이 많습니다.

  ✓ 네트워크 전체 Stuck

  ✓ 기본 ECMP기반에서는 Congestion 발생

- **시스코 넥서스는 AI Network에서 요구하는 Advanced Load balancing 을 제공**

  ✓ **Dynamic Load balancing : Tx Link 사용량 기반 Load balancing 수행** (Flowlet방식으로 packet reordering 원천 방지)

  ✓ **UDP based Hash : 동일한 5-Tuple Flow 내에서, 고유 패킷플로우를 인지하여 Load balancing 수행**


cisco
The bridge to possible
Thank you
cisco Connect