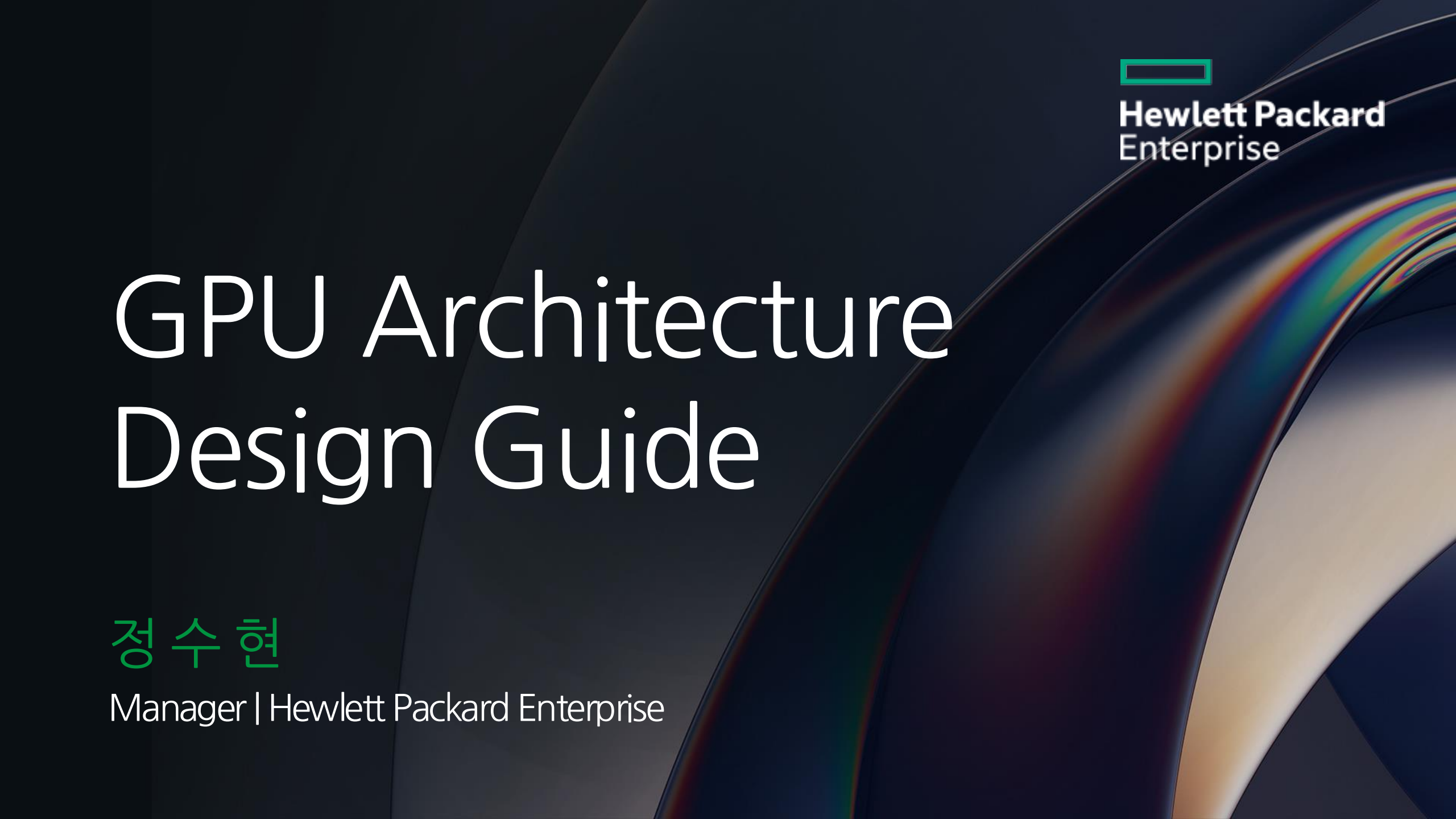


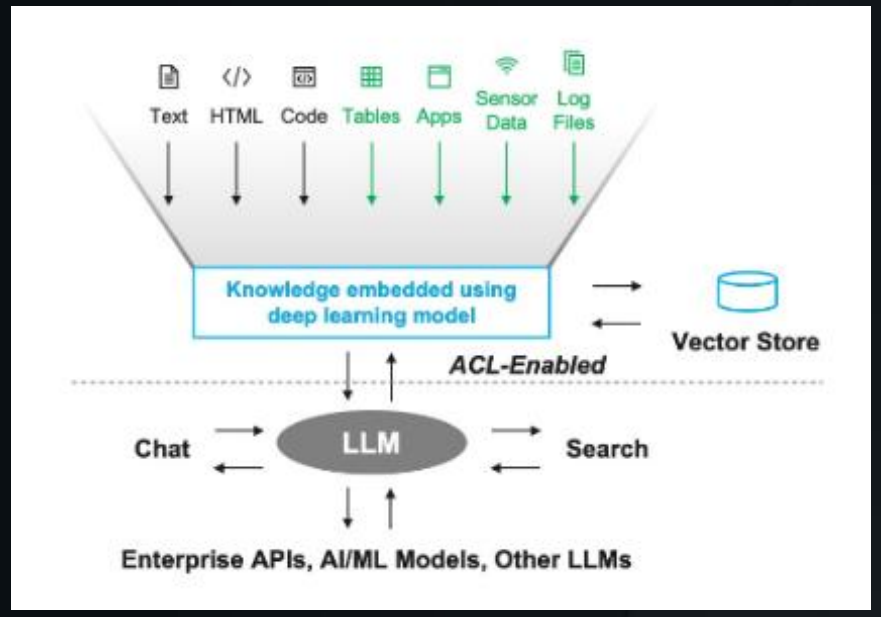
GPU Architecture Design Guide

정 수 현

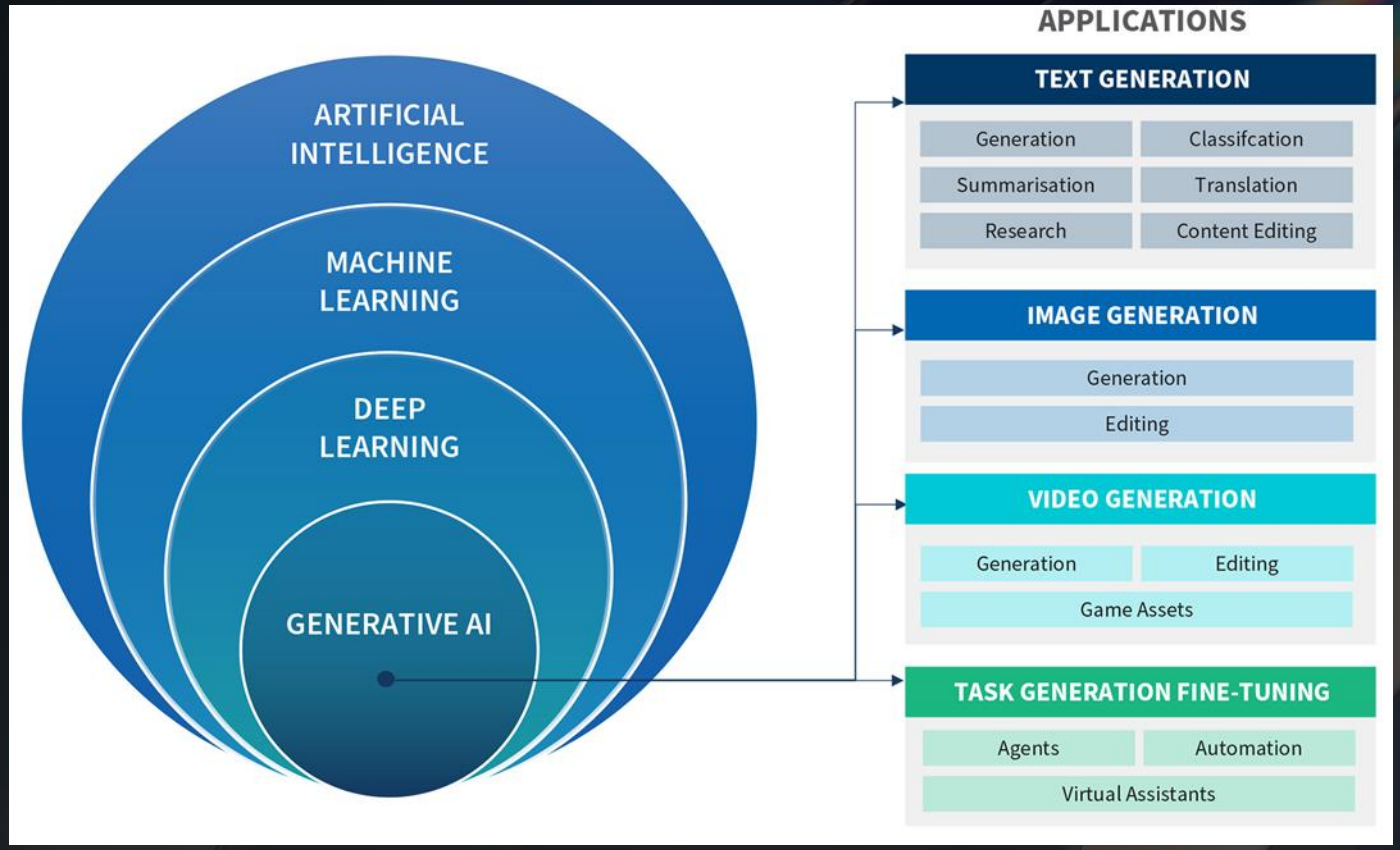
Manager | Hewlett Packard Enterprise

Gen AI

생성형 AI는 텍스트, 이미지, 음악, 코드 등 새롭고 혁신적인 콘텐츠를 제작하기 위한 광범위한 AI 시스템을 아우르는 포괄적인 분야



LLM은 텍스트 기반 데이터에 초점을 맞춘 특정 유형의 생성 AI 모델을 구성



Should you build or buy

데이터와 데이터 주권의 관련성에 따라 Gen AI 솔루션을 개발 및 배포할 때 온사이트 또는 COLO 솔루션이 사용됩니다. Gen AI 솔루션을 구매해야 할까요, 아니면 구축해야 할까요?

답은 여러 요인에 따라 달라집니다. :

- 전체 목표
- 규모
- 복잡성
- 솔루션의 고유성
- 사용 사례
- 비즈니스 전략
- 원하는 투자 수준
- 위험 허용 범위
- 데이터 준비 상태
- Training 또는 Inference

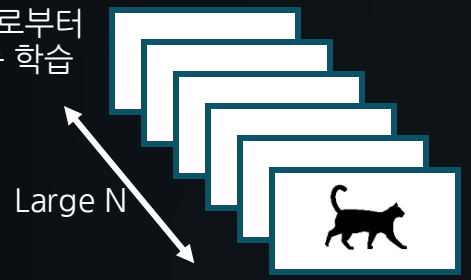
위의 요소에 따라 구축 또는 구매를 결정하든, HPE Aruba와 HPE GreenLake는 고객의 요구 사항을 충족하는 GEN AI 솔루션을 제공할 수 있습니다.

What model do you need?

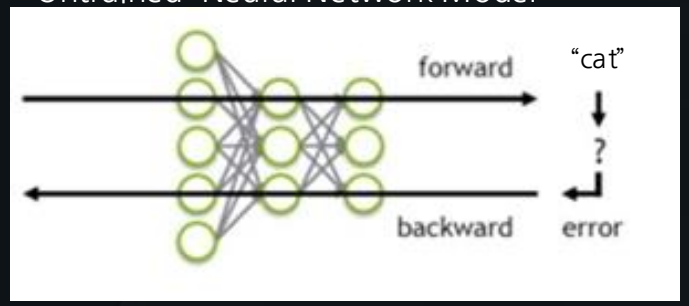
AI 학습과 추론

Training

기존 데이터로부터 새로운 기능 학습



Untrained Neural Network Model



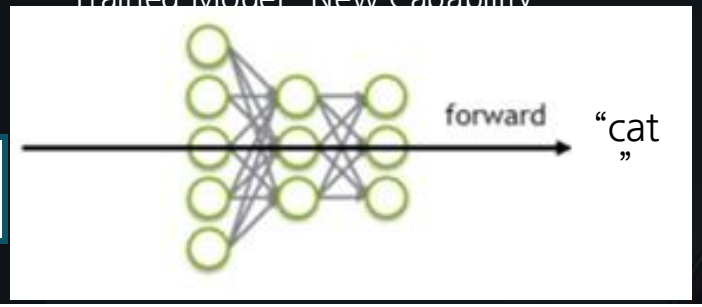
Labels

데이터로부터 모델을 학습하고 최적화하는 프로세스

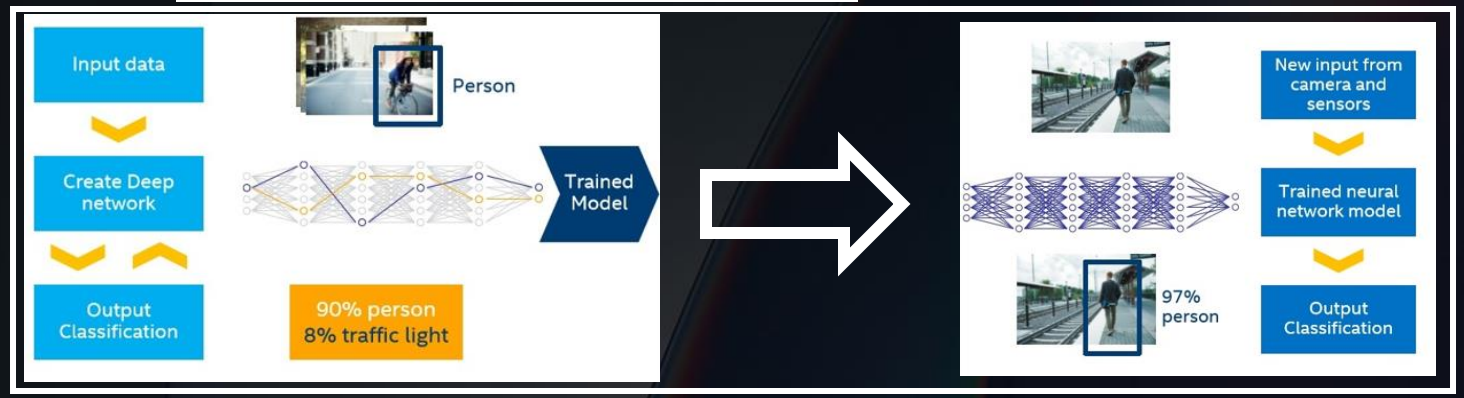
Inference

새로운 데이터에 학습된 기능 적용

Trained Model New Capability



효율적인 전개에서 새로운 관찰 결과를 예측/추정하기 위해 훈련된 모델 사용



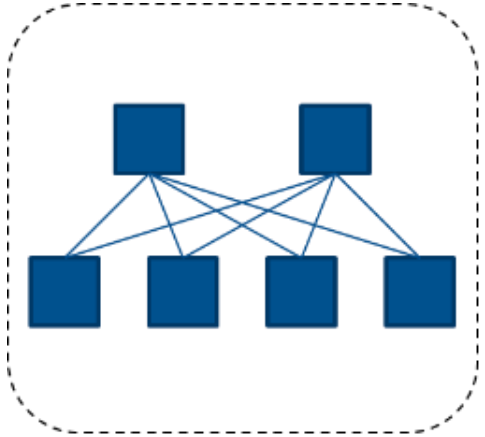
Workload solution parameters

		AI – Inference		AI – Training						
	Non-AI	Batch	Real-time	Expert Systems	Reinforcement Learning	Deep Learning - Speech	Deep Learning - Computer Vision	Deep Learning - NLP	Generative AI - Small	Generative AI - Large
	Operational data processing	*Batch execution per model logic*	*Real-time execution per model logic*	*Execute under rules / constraints*	*Optimize across wtd. decision-tree*	*Identify patterns across text, audio, image or video to synthesize, summarize, classify or trigger actions*			*Generating code, text, audio, image, video or synthetic data when prompted*	
AI model characteristics										
Model techniques	*NA*	*NA*		Knowledge Graph, Rules/Constraints	Weights / Biases, RLHF	Discriminative ML (SVM, Random Forest) Deep Learning / Neural Nets (cNN, ffNN, rNN)			Gen ML(GAN), Encoder-Decoder, Diffusion	Transformers, Transfer Learning, RAG
Examples	*NA*	*NA*		DeepBlue, XCON	AlphaGO	Kaldi	ResNet, Waymo	BERT	Dolly, DALL E-2	GPT-3, Llama-2
# parameters	*NA*	*NA*		<10K	<500 M				1 – 20 B	>40 B
# GPUs[1]	*NA*	0-1		0-1	1-2		1-2		1-10	>100
Trg. data size[2]	*NA*	*NA*		<5 GB	<20 GB		0.5 – 5 TB		0.1 – 1 TB	1 – 15 TB
Storage characteristics										
Capacity[3]	*NA*	*[Depends on scale]*		<1 TB	1 - 500 TB				0.5 - 5 PB	
Max. Latency[4]	<100 ms	<50 ms	<20 ms	<50 ms	<20 ms				<10 ms	
# parallel clients	*NA*	*NA*		<100	500 – 1,000		1,500 – 2,000		1,500 – 2,000	
IOPS[5]	<10 K	<10 K		<10 K	10 – 20 K				>20 K	>40 K
Storage energy	2 – 3 watt/TB	0.5 – 1.5 watt/TB		2 – 3 watt /TB	0.5 – 1.5 watt/TB				0.2 – 0.5 watt/TB	

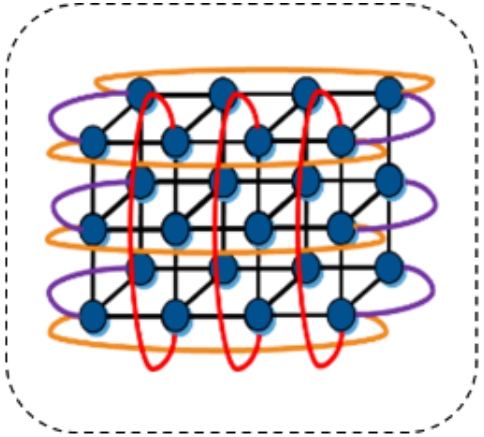
HPE ProLiant | Apollo | HPE Aruba Networking | Nvidia

HPE Cray EX / XD - Supercompute/HPC Ethernet | InfiniBand | Cray Slingshot

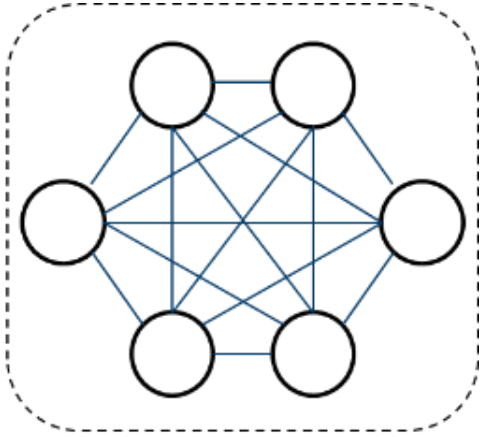
Cluster Architectures



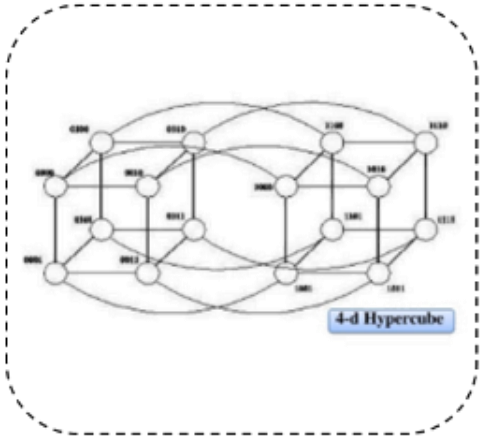
Fat Tree



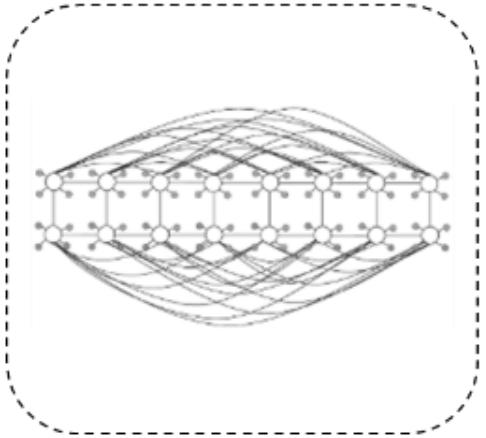
Torus



Dragonfly

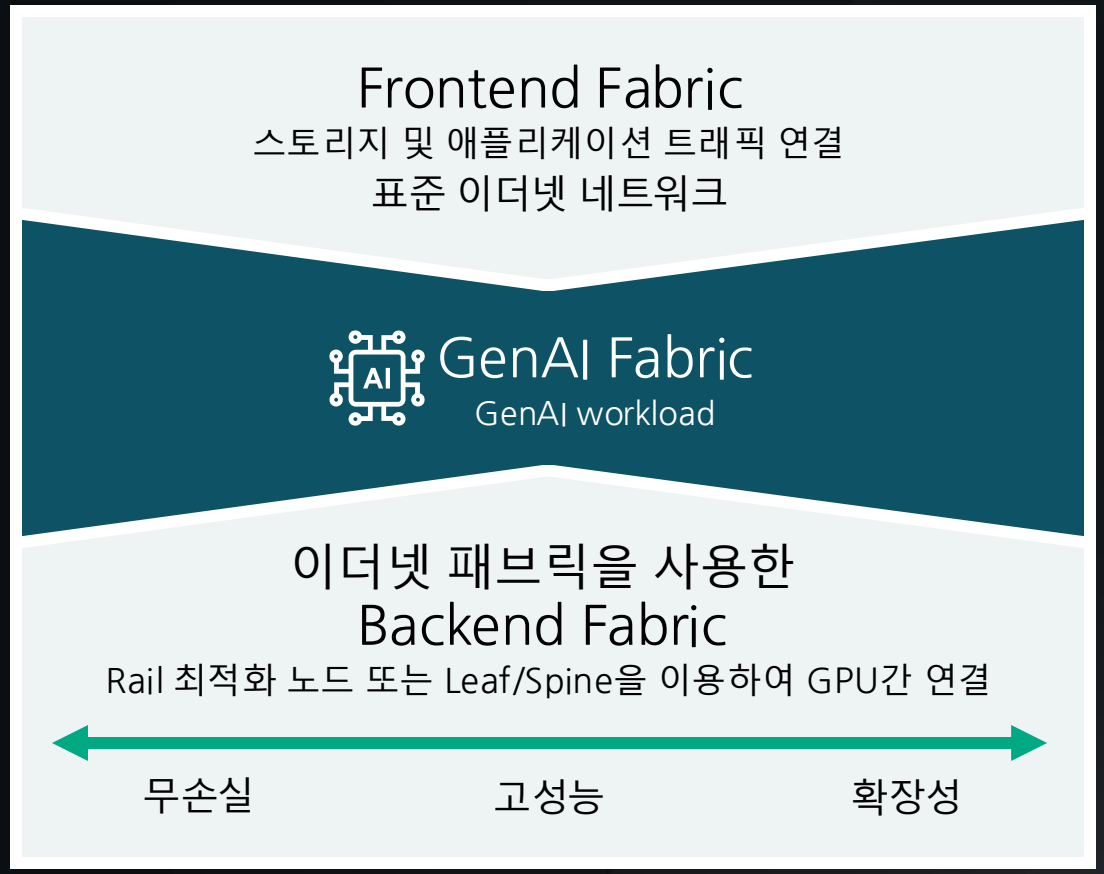


Hypercube



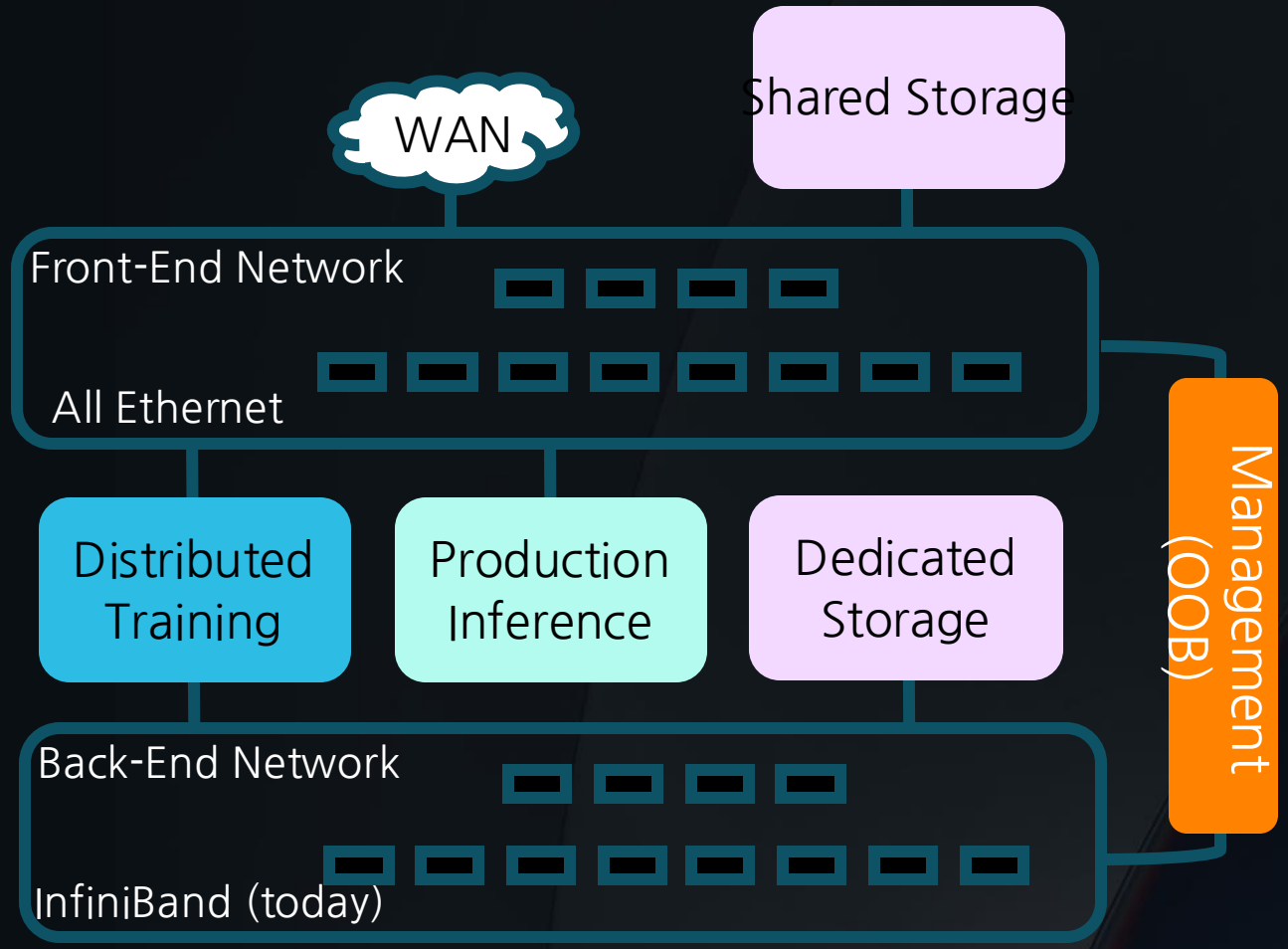
HyperX

AI 클러스터 네트워크



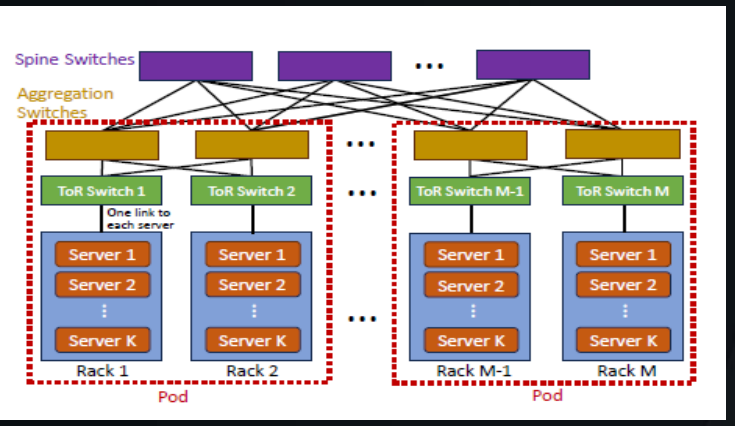
- **Frontend** 네트워크
  - 기존 이더넷 스위치를 사용하여 구축
  - 범용 서버/워크로드에 사용
  - 공유 스토리지를 지원하기 위해 무손실용으로 설계
- **Gen AI** 패브릭
  - GPU가 내부 PCIe 버스와 백엔드 패브릭을 모두 사용하여 워크로드 트래픽을 이동하는 워크로드로 구성
- **Backend** 네트워크
  - AI 서버/워크로드를 서로 연결하는 데 사용
  - 전용 스토리지도 포함될 수 있음
  - 워크로드와 패브릭 간에 Link Oversubscription 없는 상태로 구축
  - Low Latency, DCB, RoCE 등을 사용하여 GPU 워크로드에 최적화

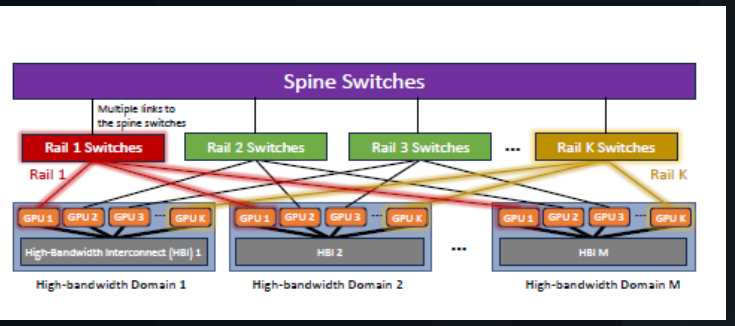
AI 클러스터 및 네트워크 요구사항

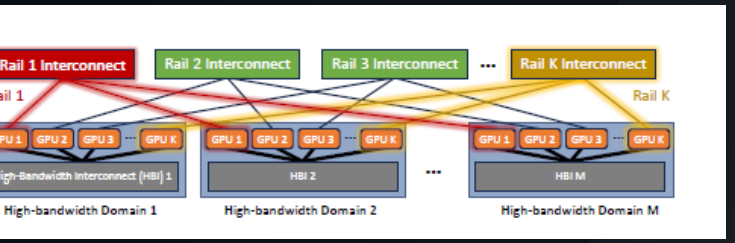


- **Frontend** 네트워크 요구사항
  - 100G/200G 구성
  - 외부 연결
  - 멀티테넌트
- **Backend** GPU 네트워크 요구사항
  - GPU 당 200G/400G 구성
  - 혼잡 회피(1:1 Non-blocking fabric, Dynamic Load-balancing)
  - RDMA
- **Backend** Storage 네트워크 요구사항
  - 100G/200G 구성
  - 혼잡관리

백-엔드 패브릭 네트워크 설계 유형







Multi-layer Clos:

- ToR 스위치는 각 서버를 연결하고 집선 스위치를 통해 다른 랙에 연결을 제공
- Spine 스위치는 다른 POD에 대한 연결을 제공
- CPU 사용량이 많은 워크로드에 가장 적합

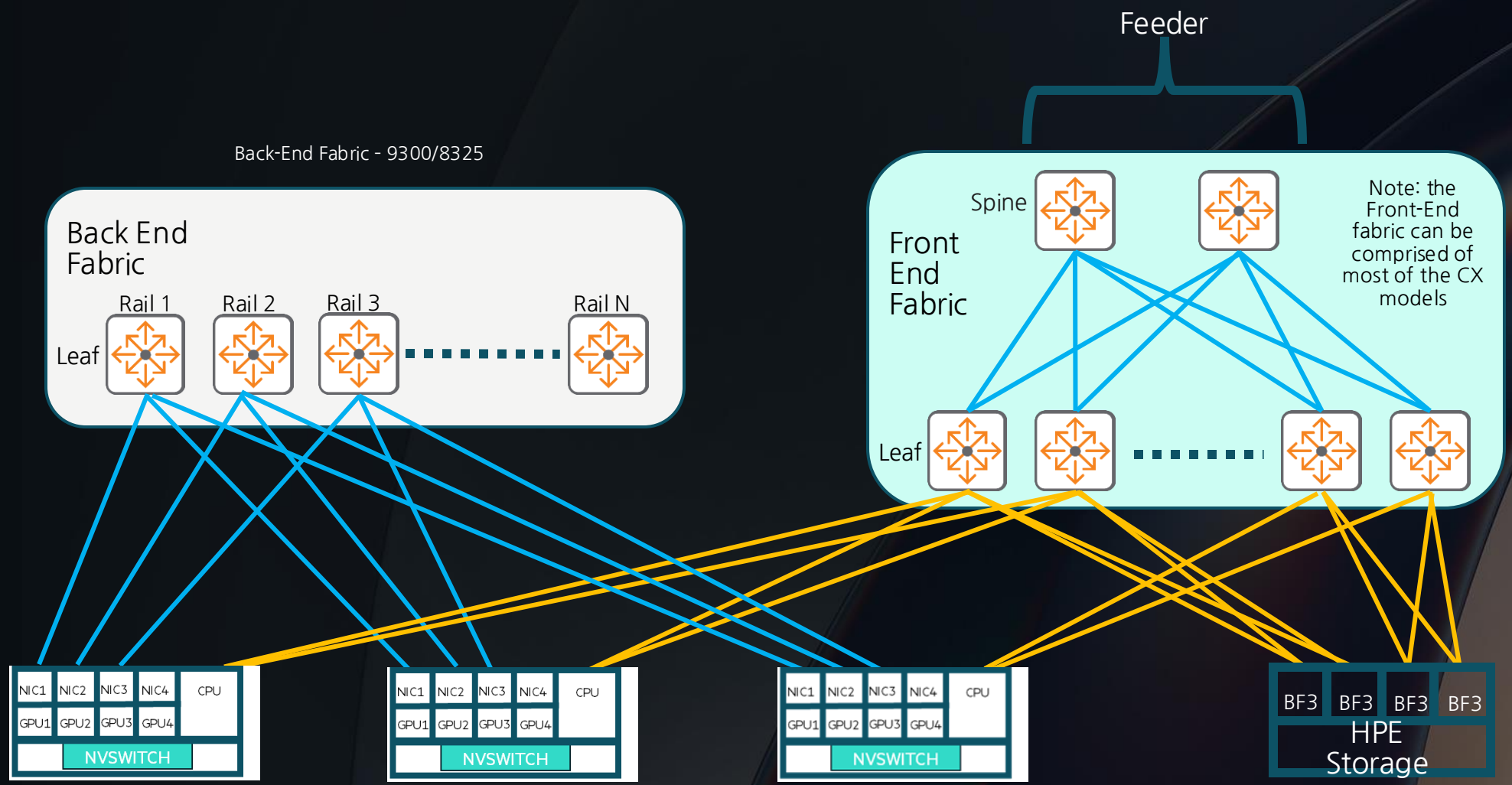
Rail Optimized

- 각 GPU에 두 개의 서로 다른 통신 경로가 있는 GPU 중심 클러스터를 기반으로 구성
- 한 경로는 NVIDIA® NVSwitch(고대역폭을 지원하지만 단거리 상호 연결을 지원)를 통해, 다른 경로는 레일 스위치를 통해 연결
- 개별 서버의 NVIDIA NVSwitch는 고속 상호 연결을 생성하여 고 대역폭(HB) 도메인을 형성
- 이러한 레일 스위치는 Spine 스위치에 연결되어 전체 구간, Any-to-any Clos 네트워크 토폴로지를 형성

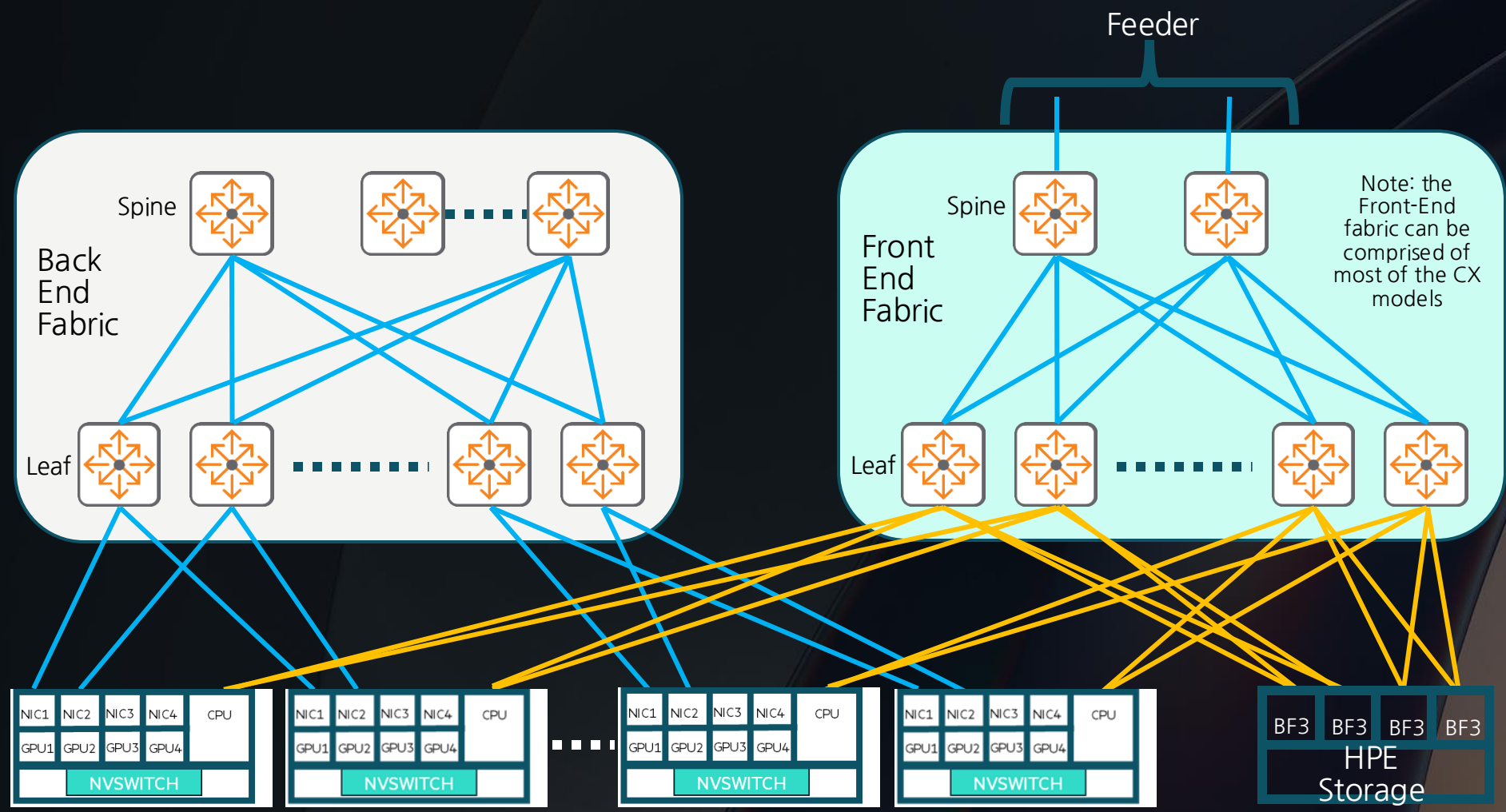
Rail Only

- Rail Optimized와 유사
- 대신, Rail 안의 다른 등급의 GPU간 네트워크 연결 제거
- HB 도메인을 통해 데이터를 전달하여 통신은 여전히 가능

Rail-Only Architecture



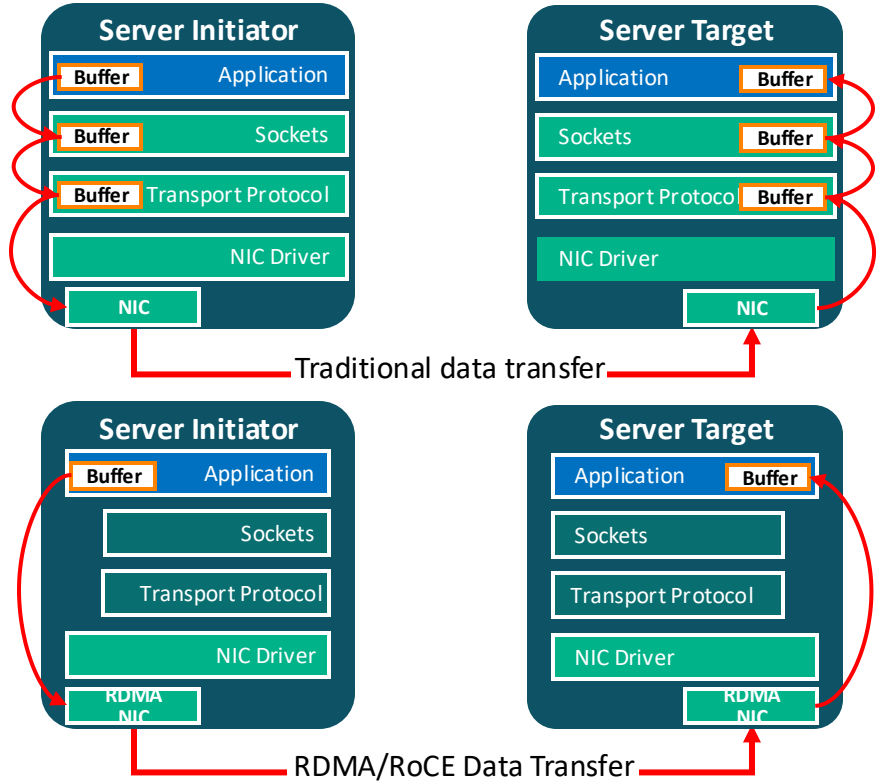
Rail-Optimized Architecture

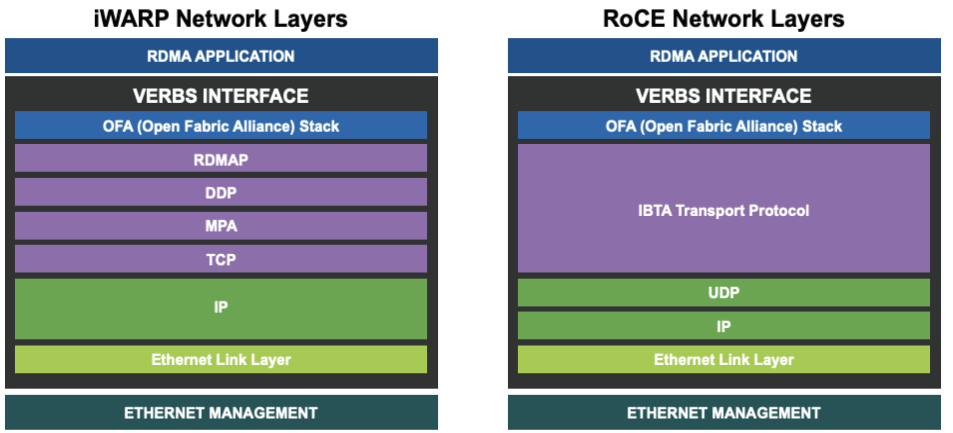


Remote Direct Memory Access (RDMA)

IB, iWARP & RoCE

- **InfiniBand (IB)**
  - 처음부터 기본적으로 RDMA를 지원하는 프로토콜
  - 해당 기술을 지원하는 전용 NIC 및 스위치가 필요
  - 순수 InfiniBand 링크는 RoCE 또는 iWARP보다 뛰어난 성능 지원
- **RDMA Over Converged Ethernet (RoCE)**
  - 무손실 패브릭을 제공하는 DCB 스위치 필요
  - NIC는 RoCE에 대한 오프로드 필요
  - 패킷 유실을 보호하기위한 이더넷 메커니즘 (PFC: 트래픽 유실 방지 / ETS: 클래스 보호)
  - RoCEv1에는 DCB 무손실 패브릭(L2) 권고
  - RoCEv2에는 DCB 무손실 패브릭(L2/3) 권고
    - 동일 포트에 IP ECN 및 PFC 구현
- **Internet Wide Area RDMA Protocol (iWARP)**
  - RDMA over TCP – RDMA 운영 전송에서 TCP 프로토콜 사용
  - DCB 이더넷은 iSCSI와 마찬가지로 도움이 되지만 필수는 아님.
  - NIC의 RDMA 하드웨어 오프로드 & HW 가속의 TCP 재전송을 사용 – 혼잡 처리, 확장 및 오류 처리에 비효율적





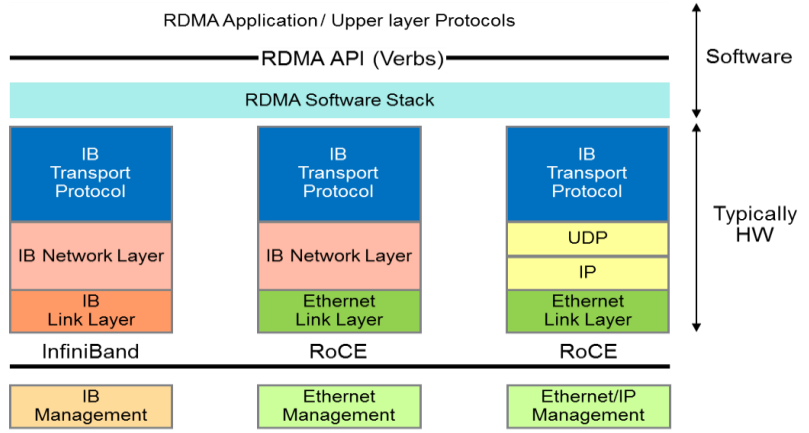
RoCE

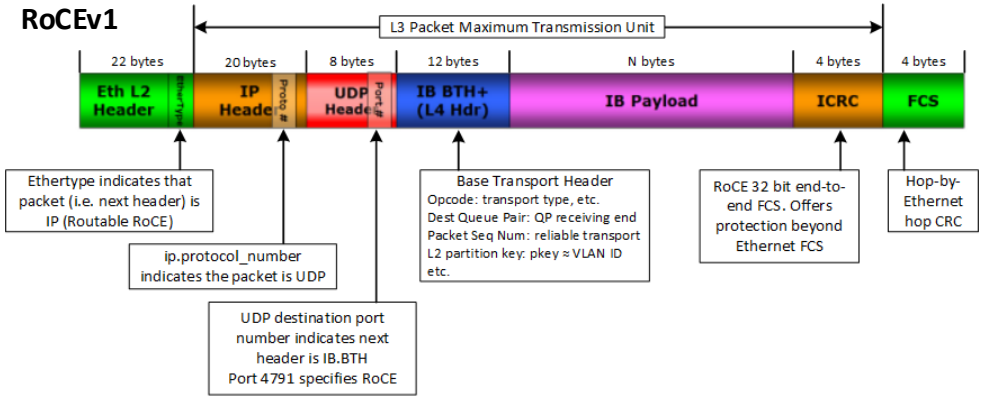
RDMA over Converged Ethernet

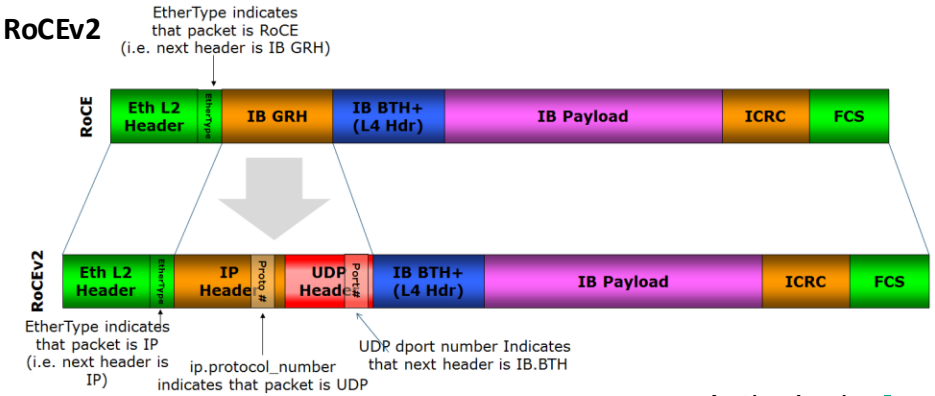
RoCE는 IBTA에서 정의

InfiniBand Trade Association은 InfiniBand 사양을 정의하고 유지 관리하는 표준 조직

- 엔드-투-엔드 솔루션
- 서버 어뎁터 부분
  - RoCE 지원
  - 오프로드 지원 (many buffer memory copies & significant CPU overhead)
- 네트워크 부분
  - 무 손실 이더넷 패브릭((Lossless Fabric) 제공하는 DCB 스위치 필요
  - 패킷 유실을 보호하기위한 이더넷 메커니즘 (PFC: 트래픽 유실 방지 / ETS: 클래스 보호)
  - 트래픽 유실 복구에서 재전송 메커니즘 사용
  - 엔드 투 엔드 솔루션 – 어뎁터 스택 지원 및 네트워크 지원 필요
  - RoCEv1는 EtherType에 RoCE/IP 로 색인(IB GRH - (0x8915))하고 IB GRH 헤더와 IB 전송 부분을 이더넷 프레임으로 생성
  - RoCE v2 : IB GRH 를 IP / UDP 헤더로 대체하고 IB L4 헤더와 IB 전송 부분을 캡슐화(라우팅 지원) & 무손실 작동을 위한 DCB 스위치의 트래픽 명시(식별)에서 PFC & ECN 사용
  - 모든 이더넷 속도에서 사용 가능 : 10/25/40/50/100/200/400 GbE







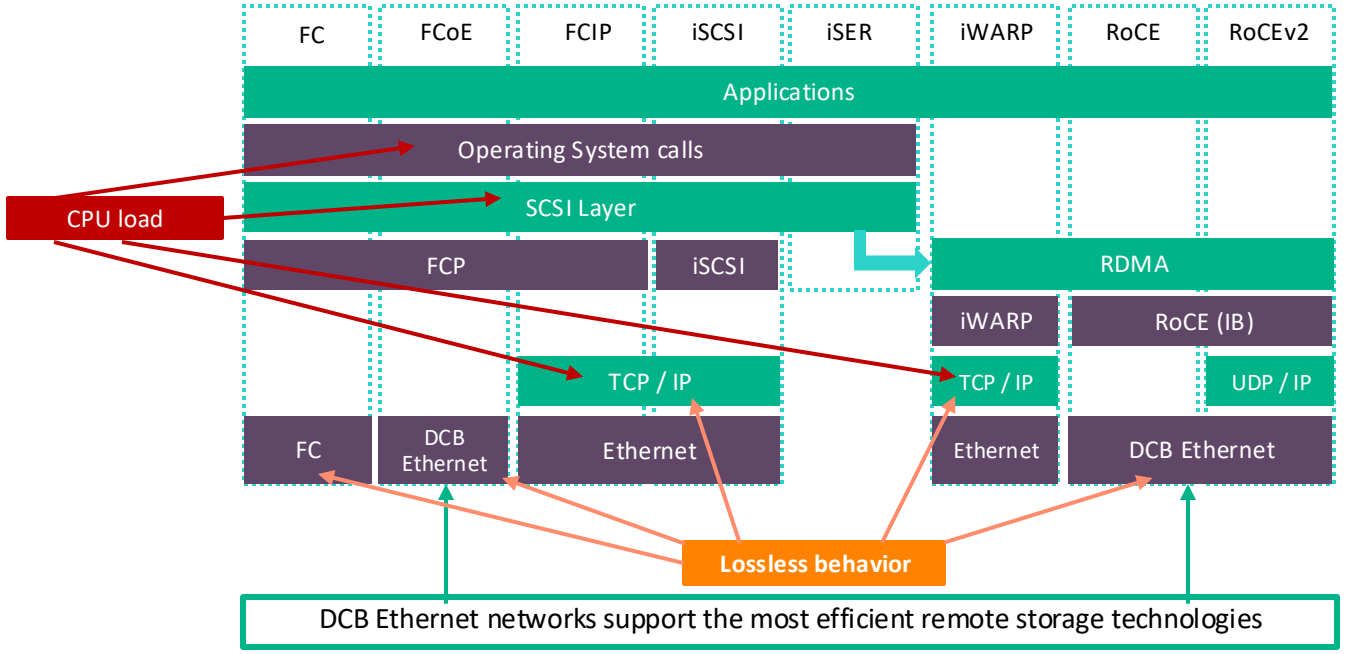
RoCE

a lossy or a lossless network

- lossless 가 RoCE 의 본질은 아니나, 특히 부하가 심한 경우, DCB 환경에서 잘 작동

Fabric	Lossy(TCP)	Lossy w/ QoS	Lossless(RoCE)
802.1Q Tagging	**No**	**No**	**Yes**
ECN(혼잡 알림)	Yes	Yes	**Yes**
PFC(흐름 제어)	**No**	**No**	**Yes**

- RoCEv1 및 RoCEv2 – PFC 사용
  - 모든 서버 및 이더넷 포트에 기능 구현
  - 802.1p CoS 사용
- RoCEv2 혼잡 관리 (RCM)는 IP ECN (Explicit Congestion Notification)을 지원. - 가능한 경우 일시 중지를 방지 / 감소시키는 데 도움.
  - RoCE와 마찬가지로 RoCEv2의 기본 네트워크는 무손실로 구성되어야 하나 무손실은 패킷이 전혀 손실되지 않는다는 의미는 아님
  - RoCEv2 전송 프로토콜에는 패킷 재전송 논리가 내장된 안정적인 엔드-투-엔드 전달 메커니즘 지원(일반적으로 HW에서 구현되며 소프트웨어 스택의 개입 없이 손실된 패킷을 복구하기 위해 작동)
  - 기본 무손실 네트워크에 대한 요구 사항은 패브릭 병목으로 인해 RoCEv2 패킷 손실을 방지하는 것을 목표



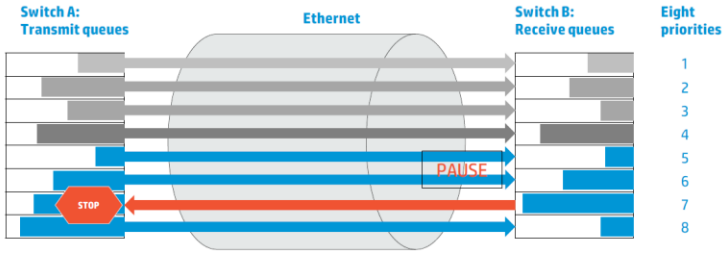
Feature	RoCEv1	RoCEv2	Comware7	AOS-CX
Priority-based Flow Control (PFC)	**Yes**	**Yes**	Yes	9300/10000/8325/8360/8100
Enhanced Transmission Selection (ETS)	Yes	Yes	Yes	9300/10000/8325/8360/8100
Data Center Bridging Exchange (DCBX)	Yes	Yes	Yes	9300/10000/8325/8360/8100
Quantized Congestion Notification (QCN)	Yes	No	Yes	No
IP Explicit Congestion Notification (ECN)	No	Yes	Yes	9300/10000/8325/8360/8100

DCB

Data Center Bridging

IEEE 802.1Qbb - Priority-based Flow Control (PFC)

- PFC는 각 우선 순위를 독립적으로 제어 할 수 있는 링크 레벨의 흐름 제어 메커니즘(DCB 스위치의 HW 큐 단위의 pause 활성화)
- DCB 네트워크의 정체로 인한 손실을 방지
- 802.3x flow control 와 비교 향상된 이더넷 일시 정지 프레임을 제공
- 단일 물리적 링크에서 최대 8 개의 가상 링크를 별도로 일시 중지하거나 다시 시작할 수 있어 제로 패킷 손실.



IEEE 802.1Qaz: Enhanced Transmission Selection (ETS)

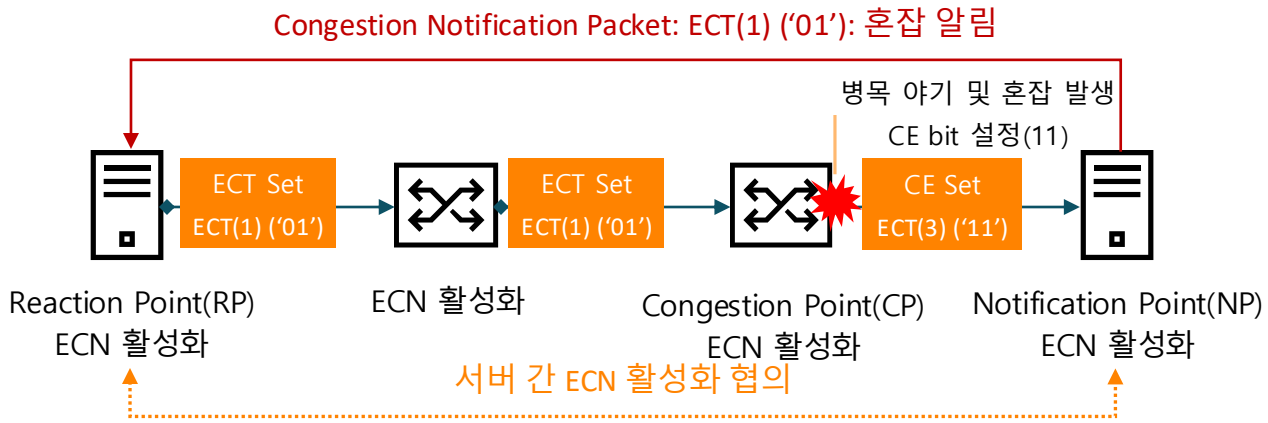
- 트래픽 클래스에 대역폭 할당으로 이더넷 상에 품질 강화
- WRR(Weighted Round Robin) 의 가중치를 규에 적용하여 트래픽 클래스 대역폭 제어

IEEE 802.1Qaz: Data Center Bridging Exchange (DCBx)

- 이웃하는 장치와 기능 수행 여부 및 구성을 교환
- 802.1AB LLDP 의 TLVs(Type-Length-values) 사용
- ETS 가 활성화되는 CNA 연결 포트

Explicit Congestion Notification (ECN)

- 서버 간의 통신에서 발생되는 혼잡을 수신 서버에서 소스 서버로 알림
- 소스 서버에서 트래픽 량 경감



- IP 헤더 **DiffServ** 필드 사용

ECN bits(Code)	내용	설명
00	**Non ECN-Capable Transport, Non-ECT**	
01	ECN Capable Transport, ECT(1)	노드 ECN 활성화
10	ECN Capable Transport, ECT(0)	노드 ECN 활성화
11	Congestion Encountered, CE	혼잡 발생

- 01 과 10은 동일한 의미

Recipes for RoCE Fabrics

출처: https://enterprise-support.nvidia.com/s/article/recommended-network-configuration-examples-for-roce-deployment

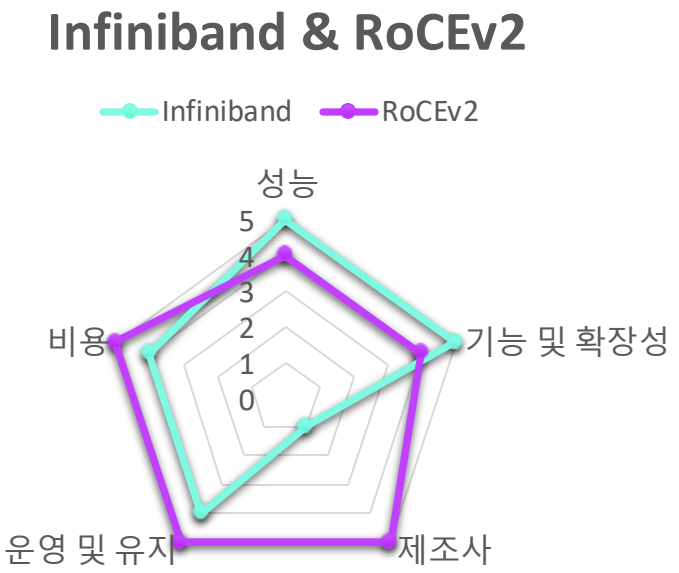
Fabric Attributes			
Fabric Configuration	***Lossy fabric***	***Lossy fabric with QoS***	***Lossless fabric***
Recommended for	Ease of deployment Minimal fabric configuration	Large scale with mixed traffic (TCP/UDP and RoCE)	Uncompromised performance over large scale Storage deployment with heavy back-pressure from PCI*
Trust layer for QoS/PFC**	N/A	L3(DSCP)	L3(DSCP)
VLAN required	No	No	No
ECN (congestion control)	Yes	Yes	Yes
PFC (flow control)	No	No	Yes

* PCI backpressure can occur when working with multiple storage devices or when oversubscribing ports to PCI (2x100G links on 16-lane PCI gen3)

Infiniband vs. RoCEv2 기술 비교

Infiniband 가 뛰어난 성능/솔루션 이나 RoCEv2도 좋은 솔루션

InfiniBand	Ethernet
제조사 종속 기술 – 주요 공급 업체 (Nvidia)	개방형 표준 기반 – 멀티 벤더 지원
IB O&M 전용 시스템	IP 네트워크 O&M 시스템(매뉴얼, SDN, 스크립트…)
• InfiniBand 기반 AI 백 엔드 네트워크와 이더넷 기반 AI 프런트 엔드 네트워크를 관리하려면 서로 다른 운영 기술 필요. - 운영 복잡성 • InfiniBand 네트워크를 운영/관리 제한적 인력	• 백 엔드 및 프런트 엔드 AI 클러스터 네트워크 모두에서 이더넷을 사용하므로 운영상 복잡성 제거 • 풍부한 경험의 이더넷 운영/관리 인력
최대 400Gbps	최대 800Gbps
고 속도의 낮은 대기 시간(End-to-End Delay 2us)	다소 높은 대기 시간 (End-to-End Delay 5us)
RDMA	RoCEV1/V2 프로토콜로 지원되는 RDMA
InfiniBand에 내장된 엔드 투 엔드 흐름 제어로 무손실 네트워크 보장 Credit-Based flow control Mechanism Forwarding based on Local ID Packet-by-packet Adaptive Routing	DCB 기능으로 무손실 이더넷 네트워크 구축 PFC, ECN, DCBX & QOS. IP-Based Forwarding ECMP Routing
IB 스위치 및 IB NIC 필요에 따라 더 높은 비용	비용 효율적이며 표준 이더넷 스위치 및 NIC를 사용

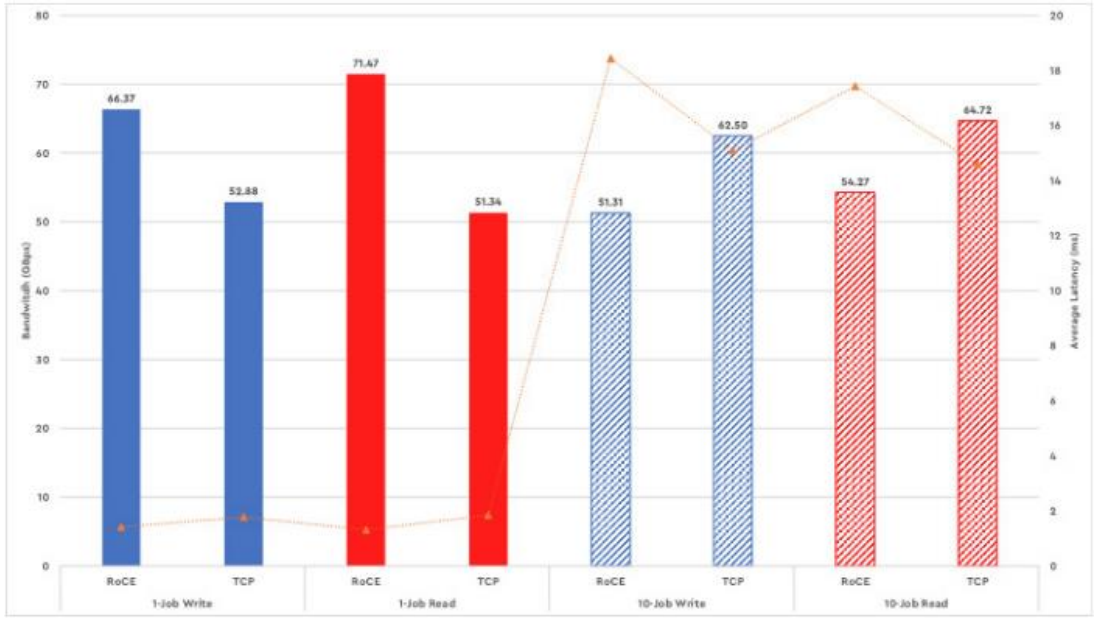


RoCE

스토리지 및 AI/GPU 클러스터 환경에서 고 성능의 컴퓨팅 및 데이터 프로세싱 지원

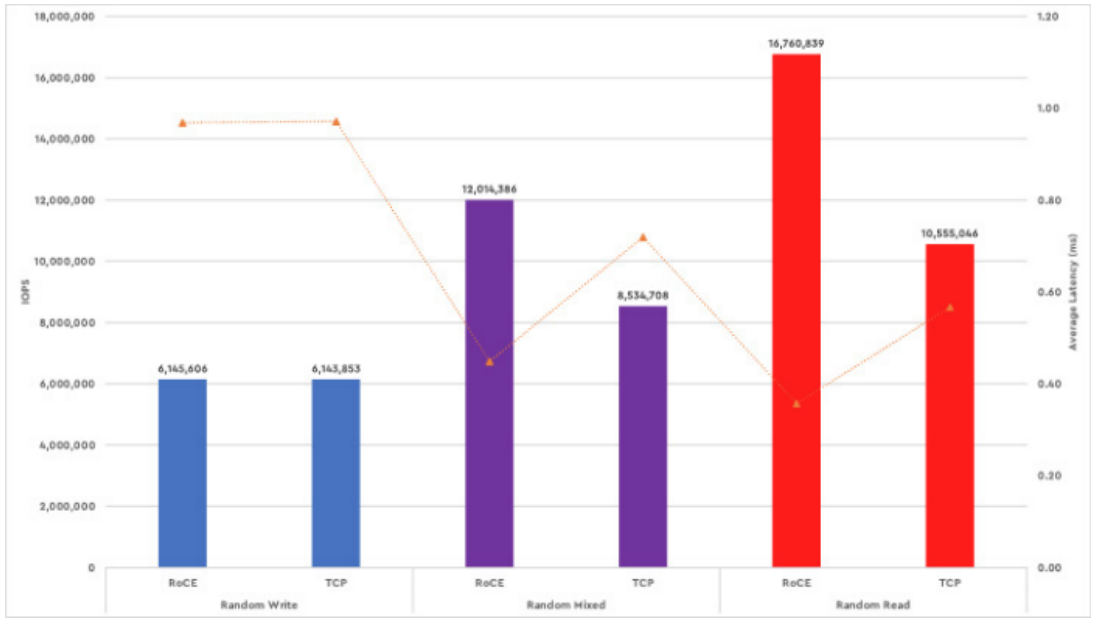
- 일반적인 RoCE 기술 사용은 TCP/IP 보다 프로세서 효율성이 더 높고 특히 활용도가 높은 서버에서 RoCE를 사용하면 RoCE의 CPU 활용률 감소로 인해 TCP/IP에서 나타날 수 있는 성능 제한을 방지하는 데 도움
- ETS 및 RoCEv2 혼잡 관리(RCN)를 추가하면 RoCE 처리량이 향상되고 대기 시간(latency) 단축
- 멀티 벤더 환경 지원

RoCE vs. TCP Sequential Bandwidth vs. Average Latency



일반적으로 RoCE는 더 낮은 대기 시간에 더 나은 대역폭을 제공

RoCE vs. TCP Random IOPS vs. Average Latency



TCP 는 Read IOPS 성능은 RoCE 의 2/3 & Write IOPS 성능은 유사하나 일반적으로 낮은 latency 에서 더 나은 IOPS 제공

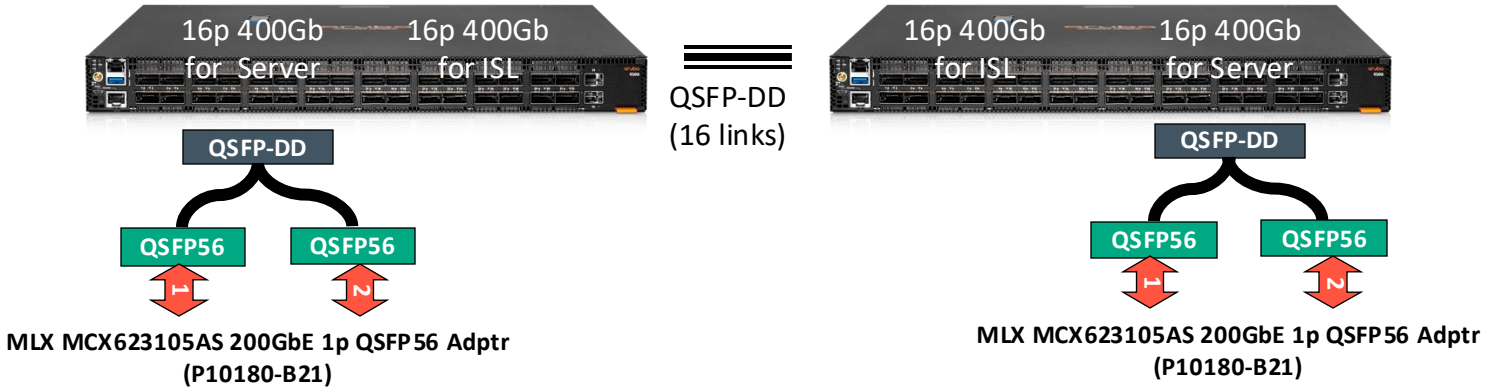
출처: Western Digital

HPC / AI Cluster InfiniBand to Ethernet Migration

HPC AI/Cluster Test

4 x HPC / AI Racks
64 x 200Gb Ports
Currently Infiniband connectivity using Mellanox Switches

2 x CX9300 switches
32 x 400Gb Switch Ports -> 64 x 200Gb NIC Ports
16 x 400Gb ISL (no oversubscription allowed)
2 x 9300-32D switches can interconnect 4 x HPC / AI racks

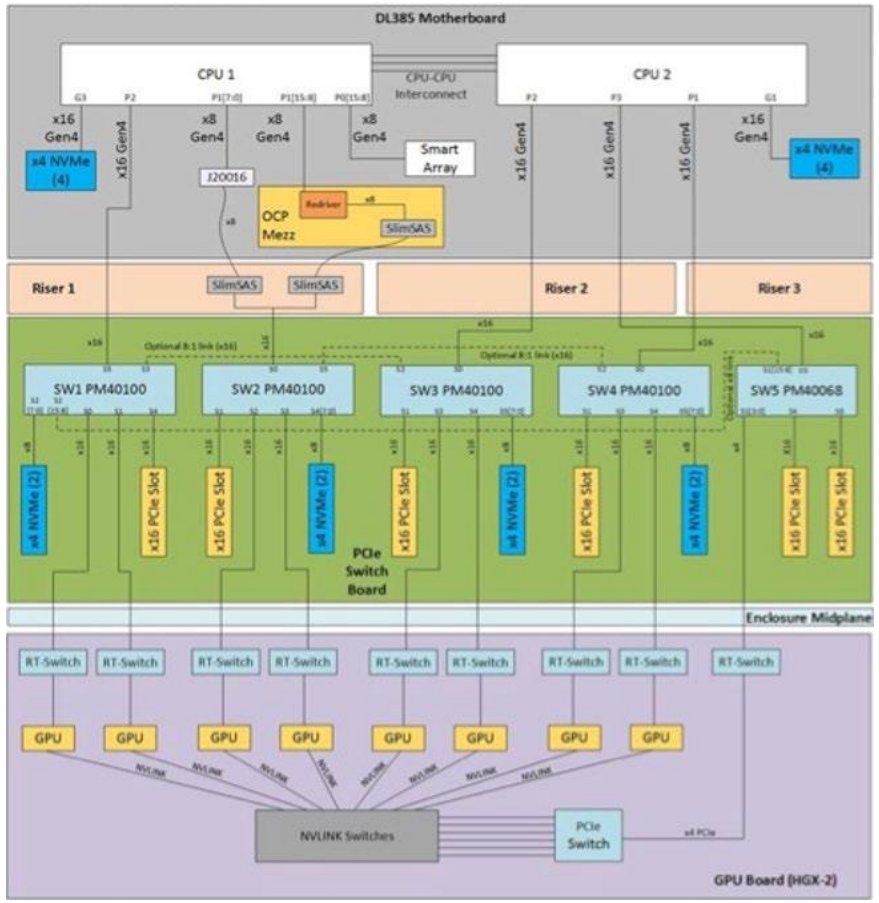


MLX MCX623105AS 200GbE 1p QSFP56 Adptr (P10180-B21)

MLX MCX623105AS 200GbE 1p QSFP56 Adptr (P10180-B21)

8 x Apollo 6500 Gen 10 Plus XL675d Servers
32 x Connect-X 6 200Gb NI

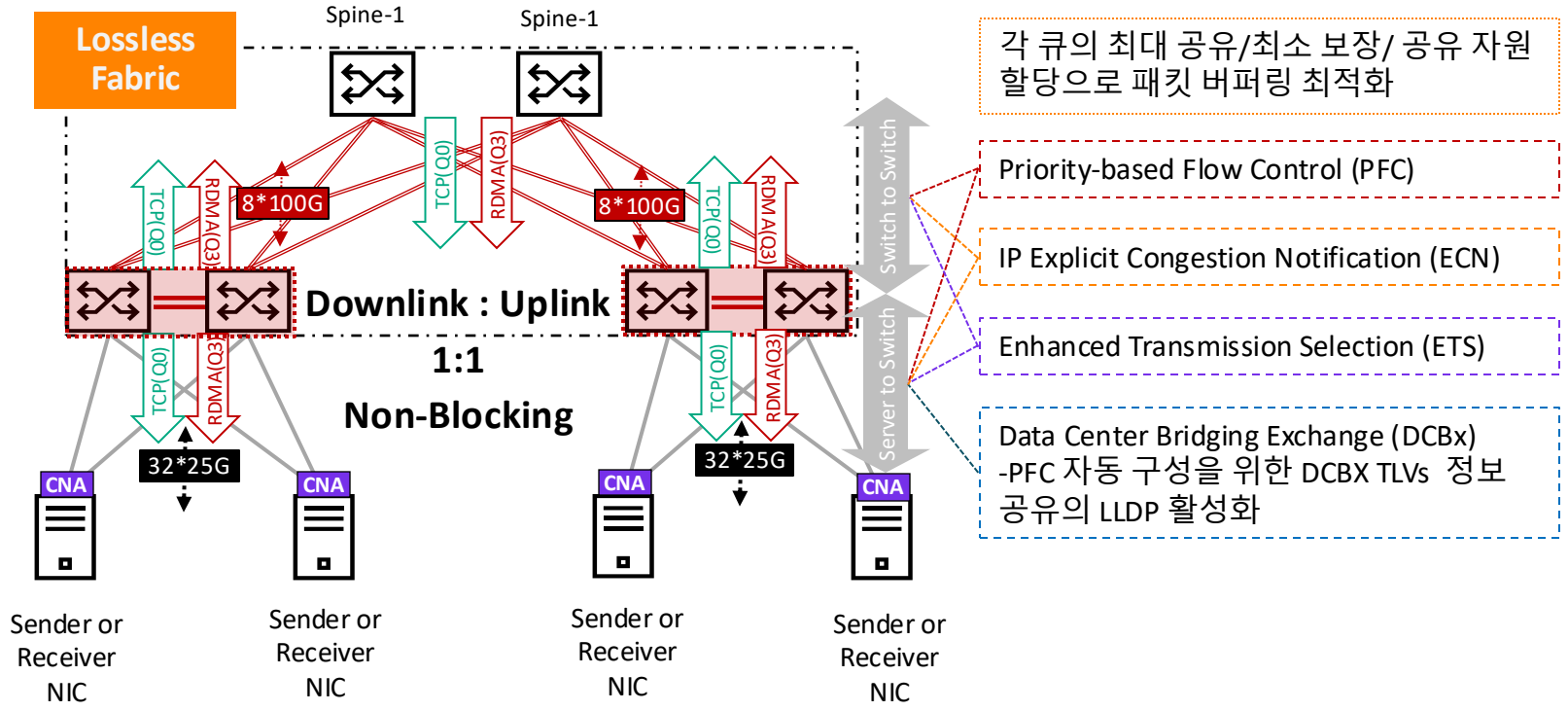
System block diagram – 8 NVLink GPU



AI/GPU 클러스터

Lossless network fabrics

- 고 대역의 속도 (스위치, 워크 로드 등)
  - 100G, 200G, 400G or beyound
- 무 손실 패브릭 (Lossless Fabric)
  - RoCEv1/v2
  - QoS 기능
  - DCB 기능
- Low Latency
  - 일관된 엔드-투-엔드 네트워크 점유 시간
- Non-Blocking
  - No oversubscription
- Load Balancing
  - L2 or L3 연결(Equal Cost Multipathing)
- 네트워크 구조 - 확장이 용이한 구조
  - Single - tier
  - Spine – Leaf Architecture



AI/GPU 클러스터

네트워크 규모

- PCIe 기술
  - Slot Capacity
  - GPU, NIC 등 연결 되는 PCIe 스위치 성능
  - PCIe lane bandwidth 표기는 Giga Bytes per second

PCIe Standard	16 Lane Bandwidth (Single Direction)	Bandwidth in Gigabits/S	PCIe Network Adapter
PCIe 3.0 x16	16GB/s (Max Bandwidth per lane: 1GB/s)	128Gbps	1 x 100G
PCIe 4.0 x16	32GB/s (Max Bandwidth per lane: 2GB/s)	256Gbps	2 x 100G or 1 x/ 200G
PCIe 5.0 x16	64GB/s (Max Bandwidth per lane: 4GB/s)	512Gbps	4 x 100G, 2 x 200G, 1x 400G

- PCIe 장치 수
  - 서버 내의 다수 개의 PCIe 장치 수 (GPU 혹은 NIC)

No. GPU	PCIe 3.0 X16 (e.g Gen 10)	PCIe 4.0 X16 (e.g Gen 10 Plus)	PCIe 5.0 X16 (e.g Gen 11)
1	1 x 100G	2 x 100G or 1 x/ 200G	4 x 100G, 2 x 200G, 1 x 400G
2	2 x 100G	4 x 100G or 2 x/ 200G	8 x 100G, 4 x 200G, 2 x 400G
4	4 x 100G	8 x 100G or 4 x/ 200G	16 x 100G, 8 x 200G, 4 x 400G
8	8 x 100G	16 x 100G or 4 x/ 200G	32 x 100G, 16 x 200G, 8 x 400G

- HPE DL server with PCIe support

Specification	Gen11	Gen10 Plus	Gen10
PCIe	80x PCIe Gen5 lanes /socket	64x PCIe Gen4 lanes /socket	48x PCIe Gen3 lanes /socket

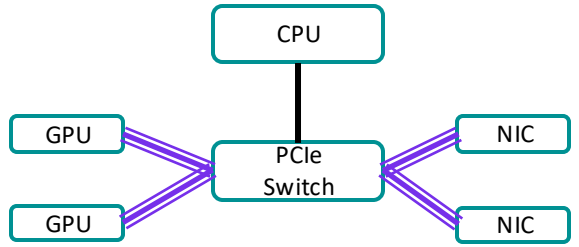
GPU 클러스터 기반의 네트워크 규모

서버 당 8 GPU

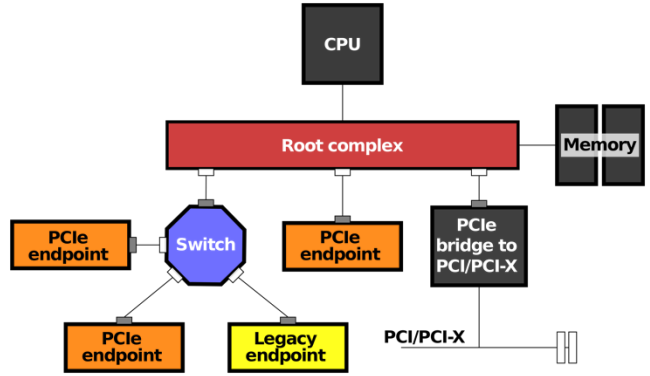
총 서버 수	총 GPU 수	PCIe Gen	총 100G 포트 수
8	64	PCIe 3.0	64
32	256	PCIe 3.0	256
128	1024	PCIe 3.0	1024

총 서버 수	총 GPU 수	PCIe Gen	총 200G 포트 수
8	64	PCIe 4.0	64
32	256	PCIe 4.0	256
128	1024	PCIe 4.0	1024

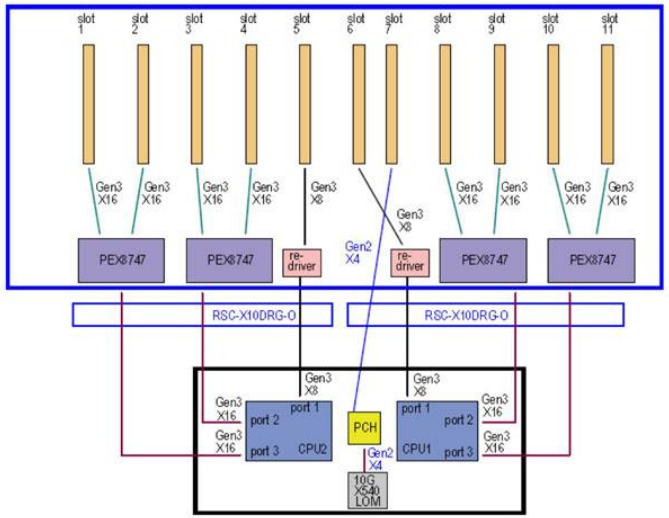
총 서버 수	총 GPU 수	PCIe Gen	총 400G 포트 수
8	64	PCIe 5.0	64
32	256	PCIe 5.0	256
128	1024	PCIe 5.0	1024



PCI Express topology



자료 출처: 위키피디아



NVIDIA(1/2)

ConnectX-7 400G Adapters

Product Sepccification	
Supported network protocols	InfiniBand, Ethernet
InfiniBand speeds	NDR 400Gb/s, HDR 200Gb/s, EDR 100Gb/s
Ethernet speeds	400GbE, 200GbE, 100GbE, 50GbE, 25GbE, 10GbE
Number of network ports	1/2/4
Host interface	PCIe Gen5, up to x32 lanes
Form factors	PCIe HHHL, FHHL, OCP3.0 TSFF, SFF
Interface technologies	NRZ (10G, 25G), PAM4 (50G, 100G)

출처: https://nvdam.widen.net/s/csf8rmnqwl/infiniband-ethernet-datasheet-connectx-7-ds-nv-us-2544471

NVIDIA(2/2)

MELLANOX ADAPTERS - Comparison Table

Class	Feature	ConnectX-3	ConnectX-3 Pro	ConnectX-4	ConnectX-4 Lx	ConnectX-5	ConnectX-6	References and Notes
Interface	Port/Speed options	2 ports of 10/40/56GbE	2 ports of 10/40/56GbE	2 ports of 100/56/50/40/25/10GbE	2 ports of 10/25GbE 1 port of 40/50GbE	2 ports of 100/56/50/40/25/10GbE	2 ports of 200/100/56/50/40/25/10GbE	**Note**: There various of adapter cards. Some OPNs for example could be speed limited, e.g.ConnectX-4 FDR (that won't link up EDR) or Connect-IB that is x8 only and not x16.
Interface	PCIe	x8 Gen3	x8 Gen3	x8, x16 Gen3	x8 Gen3	x16 Gen3 x16 Gen4	2x Gen3 x16 Gen4 x16	
Interface	PCIe Switch					PCIe x16	PCIe x32	
Interface	Coherent Accelerator Processor Interface (CAPI)	-	-	-	-	Supported (v2)	Supported (v2)	IBM - Coherent Accelerator Processor Interface (CAPI) http://openpowerfoundation.org/blogs/interconnect-your-future-mellanox-100gb-edr-capi-infiniband-and-interconnects/
IB RDMA / RoCE	IB RDMA / RoCE	IB RDMA, RoCE*	IB RDMA, RoCE	IB RDMA, RoCE	RoCE	IB RDMA, RoCE	IB RDMA, RoCE	RDMA/RoCE Solutions
RoCE	RoCE Congestion Control	-	Supported	Supported	Supported	Supported	Supported	RDMA/RoCE Solutions Understanding RoCEv2 Congestion Management RoCEv2 CNP Packet Format Example HowTo Configure RoCE Congestion Control for Windows 2012
HPC	CORE-Direct®	Supported	Supported	Supported	Supported	Supported	Supported	Exploited by FCA and extended by Mellanox SHARP technology. http://www.mellanox.com/related-docs/whitepapers/TB_CORE-Direct.pdf
HPC	PeerDirect®	Supported	Supported	Supported	Supported	Supported	Supported	GPU Direct Uses PeerDirect. http://www.mellanox.com/page/products_dyn?product_family=116
HPC	Dynamically Connected Transport (DCT)	-	-	Supported	Supported	Supported	Supported	http://www.mellanox.com/related-docs/applications/SB_Connect-IB.pdf
CPU Offloads	Stateless Ethernet offload	Supported	Supported	Supported + LRO, LSOv2	Supported + LRO, LSOv2	Supported + LRO, LSOv2	Supported + LRO, LSOv2	**LRO** = Large Receive Offload. **LSO** = Large Send offload. See MLNX_OFED User Manual for more info.
CPU Offloads	RSS (MAC, VLAN 5 Tuple)	Supported	Supported	Supported +	Supported +	Supported +	Supported +	
Virtualization	SR-IOV	Supported	Supported	Supported	Supported	Supported	Supported	Virtualization Solutions
Virtualization	Multi Host	-	-	4 hosts	4 hosts	4 hosts	8 hosts	http://www.mellanox.com/page/products_dyn?product_family=210&mtag=multihost
Open V-Switch (OVS)	OVS Offload	-	-	-	Supported	Supported +	Supported +	Describes whether the adapters can run the OVS data-plane in HW. Supported means that at least matching for a specific flow is performed in HW. Supported+ means that HW Encap/Decap is also included. See http://www.mellanox.com/page/asap2.
Overlay Network	Stateless offload	-	Supported	VXLAN / NVGRE / GENEVE	VXLAN / NVGRE / GENEVE	VXLAN / NVGRE / GENEVE and FlexParse	VXLAN / NVGRE / GENEVE and FlexParse	Virtualization Solutions
Overlay Network (VXLAN/NVGRE)	Encap/Decap (in HW)	-	-	-	Supported	Supported	Supported	
Storage	Erasure Coding Offload	-	-	Supported	Supported	Supported	Supported	Reed Solomon Erasure Coding hardware offload is supported by the adapters. Verbs API are available. Understanding Erasure Coding Offload
Storage	T10/DIF Signature Handover	-	-	Supported	-	Supported	Supported	HowTo Enable T10-PI (T10-DIF) Data Integrity Protection in iSER with LIO Target
Storage	NVMe oF Target Offload (also for Burst Buffer)	-	-	-	-	Supported	Supported +	
Storage	Host Chaining	-	-	-	-	Supported	Supported	
Cloud Integration	Mirantis Fuel	Supported Fuel 7.0/8.0	Supported Fuel 7.0/8.0	Supported Fuel 8.0	Supported Fuel 8.0	Supported	Supported	Cloud Solutions
Media & Entertainment	Packet Pacing	-	-	Supported	Supported	Supported	Supported	
Security	Secure Firmware update	-	-	Supported	Supported	Supported	Supported	
Security	Secure Boot	-	-	-	-	-	Supported	

출처: https://enterprise-support.nvidia.com/s/article/mellanox-adapters---comparison-table

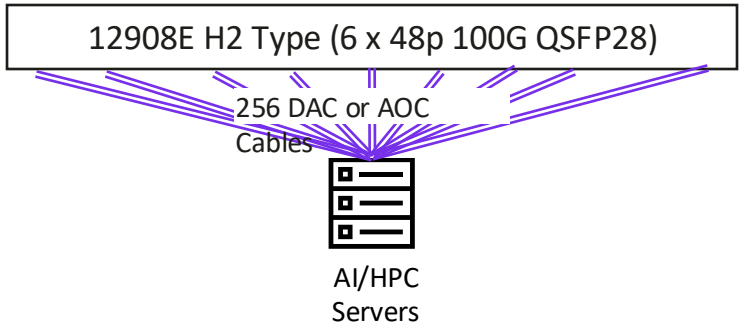
AI/GPU 클러스터

네트워크 규모와 비용

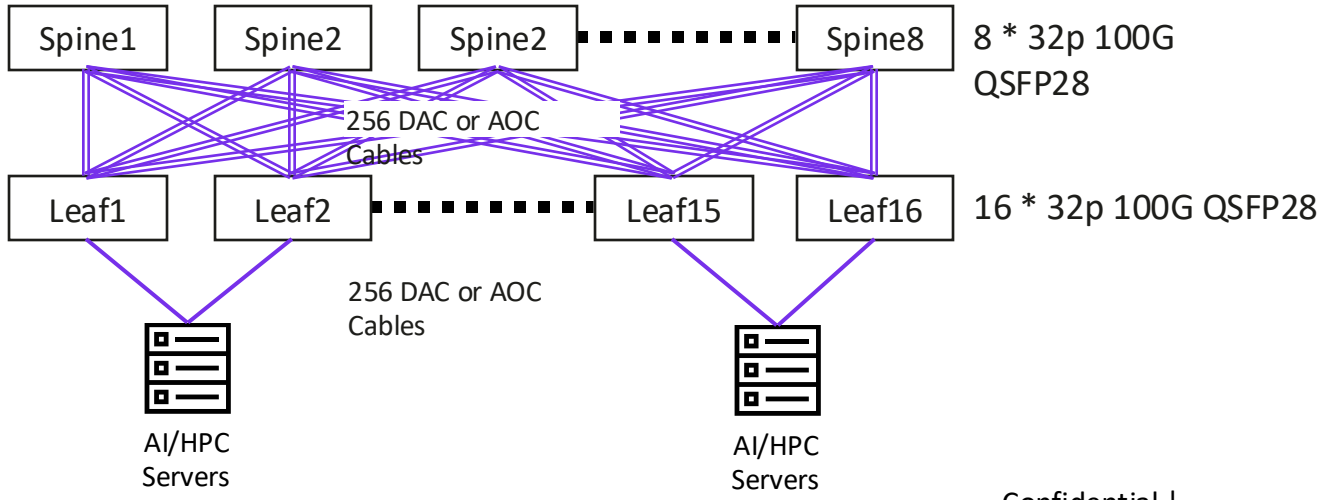
- 네트워크 패브릭
  - GPU 당 비용
  - 장치 및 링크 수량
- 에너지 효율성
  - GPU 당 BTU/hr

구조	장비 수량	케이블 수	에너지 효율	Cost/GPU
1-tier	1	256	118 BTU/hr	1 time
2tiers	24	512	120 BTU/hr	1.68 time

1-tier network design for 256 GPU on PCIe 3.0 HW cluster

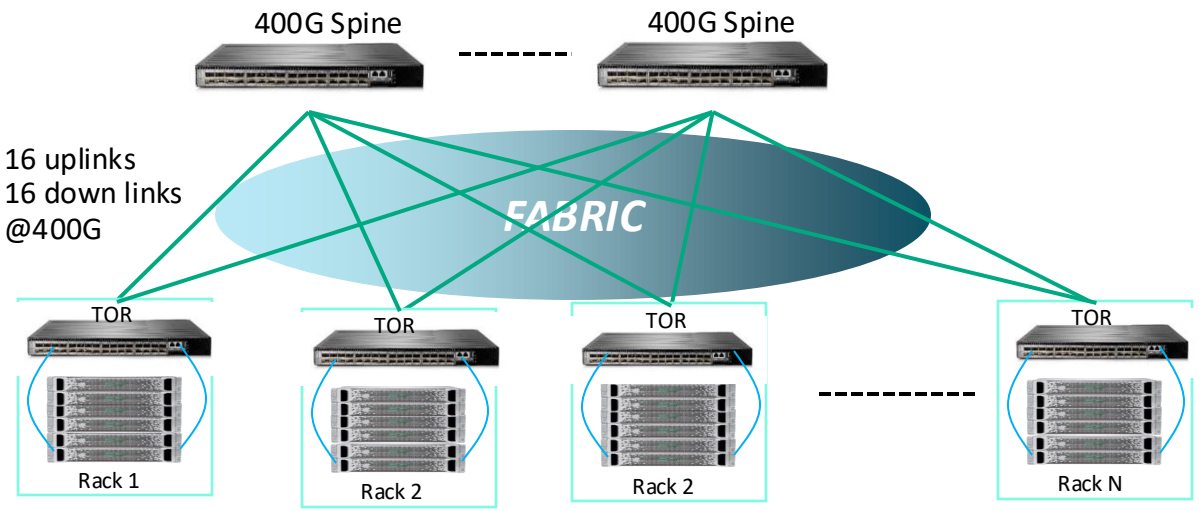


2-tier network design for 256 GPU on PCIe 3.0 HW cluster



AI/GPU DCN DESIGN OPTION 1: TWO TIER SPINE/LEAF (400G)

➢ **Small to medium-sized data center**



- 128 to 512 physical leaf ports, **16~64 Nodes** with 8 x GPUs
- 256 to 1024 physical leaf ports, **32~128 Nodes** with 8 x GPUs

- **16-64 Nodes :** Spine: 5960, 32*400G
  Leaf: 5960, 32*400G

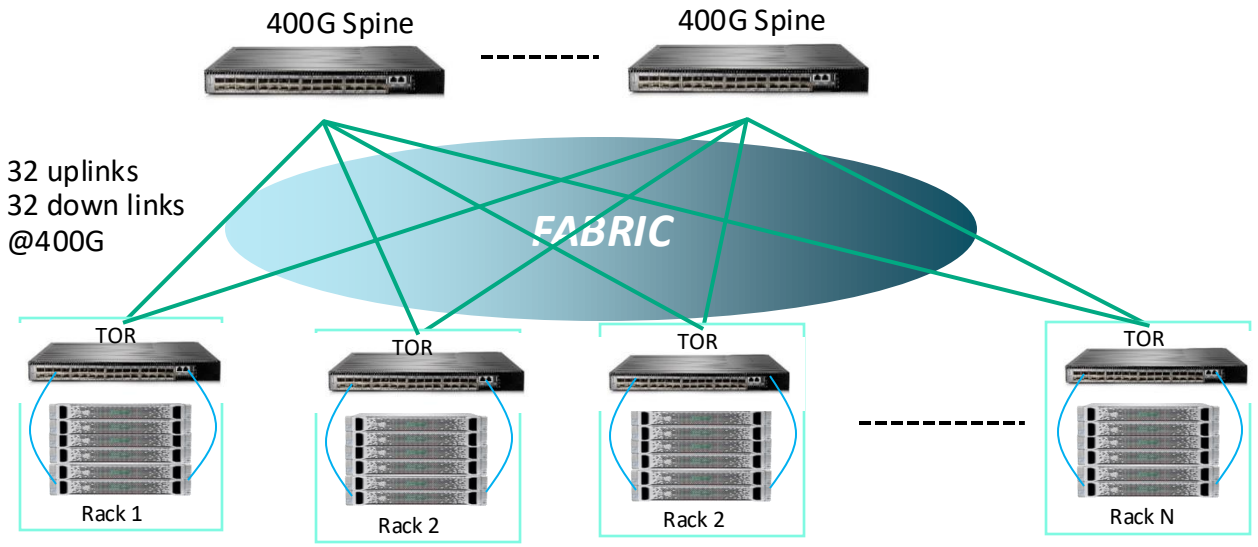
Leaf ports	Leaf Speed	Inter-switch links	Leaf	Spine
128	400G	128	8	4
256	400G	256	16	8
512	400G	512	32	16

- **32-128 Nodes:** Spine: 5965, 64*400G
  Leaf: 5960, 32*400G

Leaf ports	Leaf Speed	Inter-switch links	Leaf	Spine
256	400G	256	16	4
512	400G	512	32	8
1024	400G	1024	64	16

AI/GPU DCN DESIGN OPTION 2: TWO TIER SPINE/LEAF (400G)

➢ **Medium to large-sized data center**



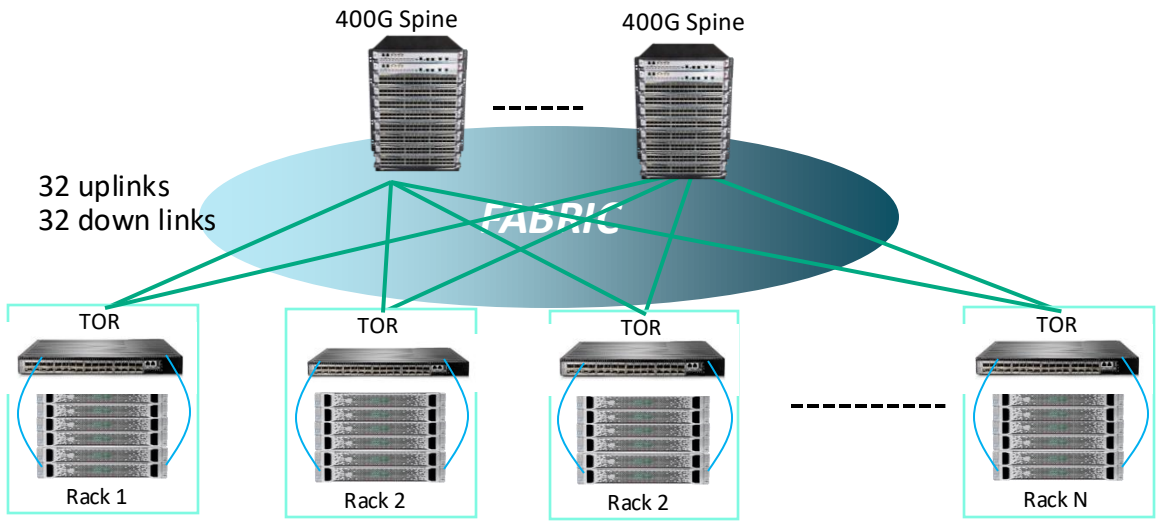
- **32-256 Nodes :** Spine: 5965, 64*400G
  Leaf: 5965, 64 *400G

Leaf ports	Leaf Speed	Inter-switch links	Leaf	Spine
256	400G	256	8	4
512	400G	512	16	8
1024	400G	1024	32	16
2048	400G	2048	64	32

- 256 to 2048 physical leaf ports, **32~256 Nodes** with 8 x GPUs
- Lesser hardware required at Spine tier, low hardware failure risk

AI/GPU DCN Design OPTION 3: TWO TIER, HIGHER-DENSITY SPINE (100/400G)

➢ **Medium to large-sized data center**



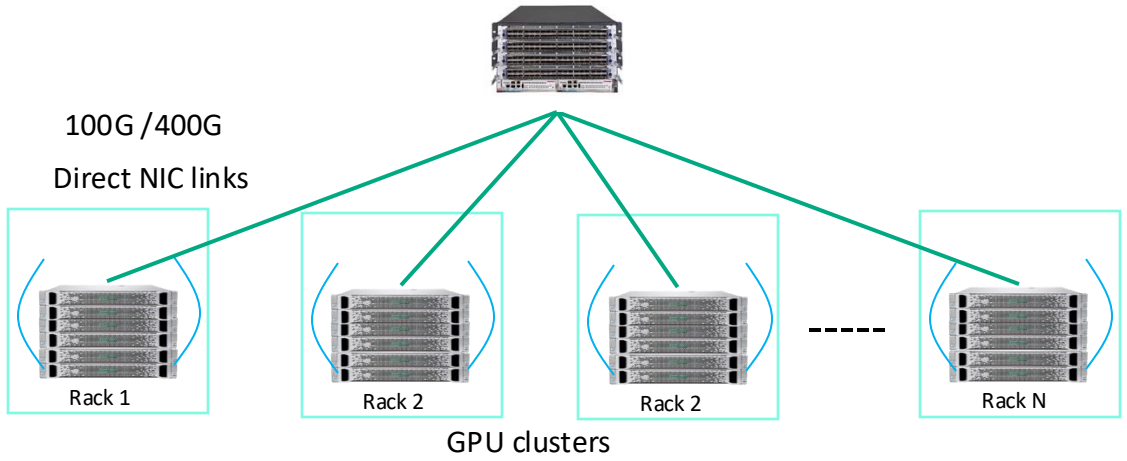
- **32-256 Nodes :** Spine: 12908E+ 6*24*400G H2 Modules
  Leaf: 5965, 64 *400G

Leaf ports	Leaf Speed	Inter-switch links	Leaf	Spine
256	400G	256	8	2
512	400G	512	16	4
1024	400G	1024	32	8
2048	400G	2048	64	15

- 256 to 2048 physical leaf ports, **32~256 Nodes** with 8 x GPUs
- Even lesser hardware required as compared to box-box design
- Flexible connectivity option at the Spine tier i.e 100G/400G on same chassis
- Fully redundant hardware design at spine tier

AI/GPU DCN Design Option 4 : Single Tier (100/400G)

➢ **1 tier architecture with 12900E chassis switch**



- More than **700 * 100G** Port for single tier
- Fully redundant lossless design
- More than 70% interconnect cables reduced
- Flexible scalability options:
  --Move to two tier architecture (recommended)
  --Use another 12900E chassis & connect the two in DRNI

Leaf ports	Leaf Speed	HH2 Modules	Switch
144	100G	4*36*100G	12904E X
288	100G	8*36*100G	12908E X
384	100G	8*48*100G	12908E H2
768	400G Breakout 100G	8*24*4*100G	12908E H2

Leaf ports	Leaf Speed	H2 Modules	Switch
192	400G	8*24*400G	12908E H2
96	400G	4*24*400G	12904E H2

HPE Aruba Data Center Use Cases

데이터센터 네트워크 패브릭

- Aruba CX 10000 with AMD/Pensando
- 서버 및 스토리지를 위한 Server ToR 교체
- 고성능 100 - 400G Leaf-Spine
- L2/3 VxLAN, EVPN, DCI, DCB, 클라우드
- 수십 개의 Rack/POD, 수천대의 서버로 확장 가능

Aruba CX 9300-32D
100/200/400G Spine

aruba Fabric Composer

규모에 맞게 간소화되고
자동화된 패브릭 오케스트레이션

Aruba CX 10000
10/25G 분산형 서비스

보안성 중요
ZTNA, PCI, 공공/정부, 규제

ZERO TRUST

Aruba CX 8100/8360/8325
10/25G, 유연한 모델 선택

핵심 엔터프라이즈 워크로드
ERP, CRM, MSFT, VMW …

Microsoft vmware

Aruba CX 8325
10/25/100G 스토리지 최적화

메인/보조 스토리지
NAS, SDS, Flash, HCI …

nimble NUTANIX

Aruba CX 9300S-32C
100G 고성능

HPC, Spine-Leaf 스케일 아웃
CSP, NFV, ML/AI, NVMe

SUPERDOME HPE CRAY

HPE ARUBA를 통한 차세대 데이터센터 패브릭 구현의 이점

01

광범위한 스위치 포트폴리오와 안정적인 네트워크 운영 체제

- 최적화된 폼팩터
- 일관된 운영체제 라인업
- 마이크로서비스 기반 운영 시스템

02

SDN 자동화 및 오케스트레이션

- 패브릭 구축 간소화
- 클라우드 경험을 제공하는 써드파티 통합 모듈
- 엔드-투-엔드 물리 및 가상 네트워크 가시성

03

분산형 보안 서비스

- 가장 필요한 곳에 분산된 보안 서비스 제공
- 네트워크 기능 + 800G Stateful 방화벽
- 실시간 원격 분석을 통한 세분화된 가시성

04

HPE와 OEM 솔루션과의 검증된 통합 연계

- HPE 연구소에서 검증한 솔루션 가이드
- IP 스토리지 패브릭 구축에 대한 인증
- 25개 이상의 통합 솔루션

Thank you

정 수 현 (soohyun.cheong@hpe.com), 010-3318-7466

Hewlett Packard
Enterprise