The bridge to possible

Cloud + AI Infrastructure

# Cisco Nexus로 구현하는 AI-Ready 네트워크 아키텍처

시스코 시스템즈, 클라우드 & AI 인프라팀
장 희성이사, Account Executive Specialist

2025. 6. 18

# 목 차



# AI 시장 및 인프라 혁신

## 전 산업군의 업무 효율화를 위해 발전 중

지식기반 Copilot

AI 비서

컨텐츠와 코드 생성

텍스트 | 이미지 | 비디오 | 코드

리포팅과 데이터 분석

텍스트 요약 | 시각화 생성

언어 번역

다국어 실시간 커뮤니케이션

가상 에이전트 & Chatbot

도메인별 특화된 Chatbot

탐지 & 예측

예측 | 이상 징후 | 인사이트

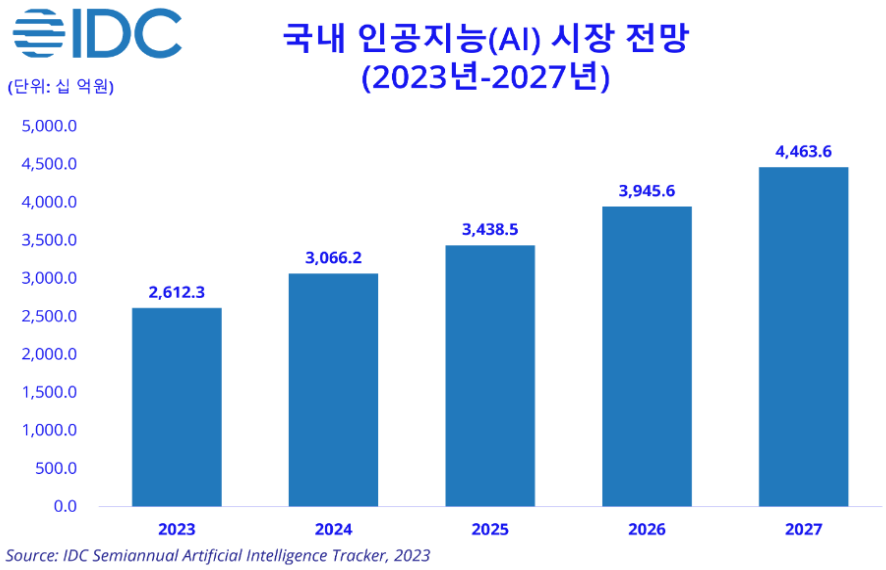
**Model Building/Training**

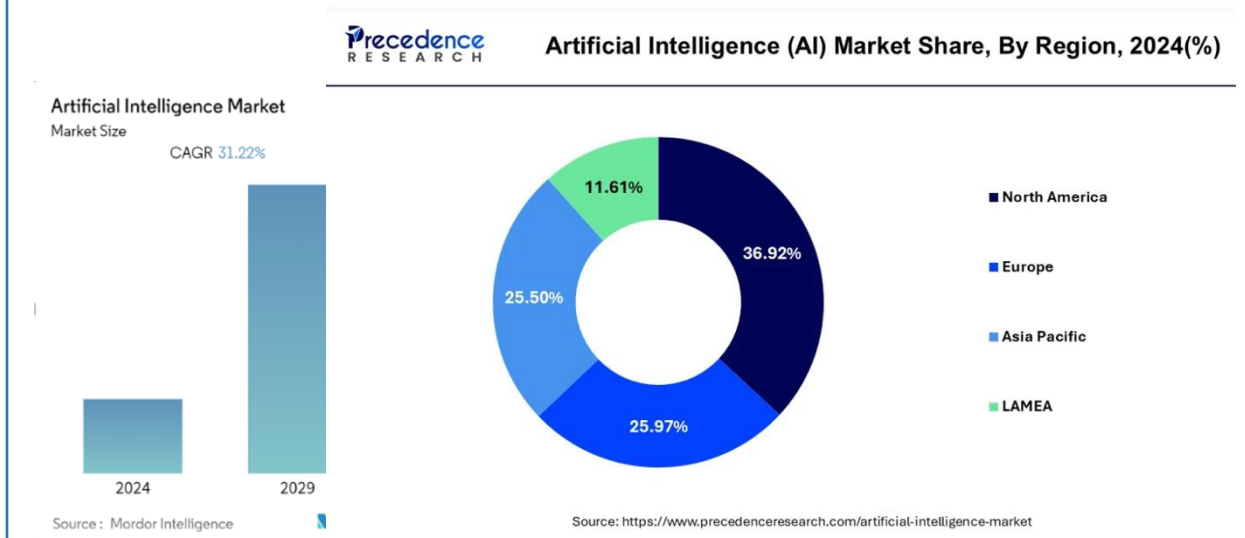
**Model Fine-tuning/RAG**

**Model Inferencing**

# AI 시장 및 인프라 혁신

연평균 30% 이상 성장




## 국내 인공지능 시장 전망

- 2027년 전체 시장규모는 4조 4천억, 2배 이상 성장 전망
- 하드웨어 인프라가 절반 차지, AI 패브릭은 20%인 4천억 예상
- 클라우드 사업자, 통신, 디지털 네이티브 고객 위주 시장 발전 예측

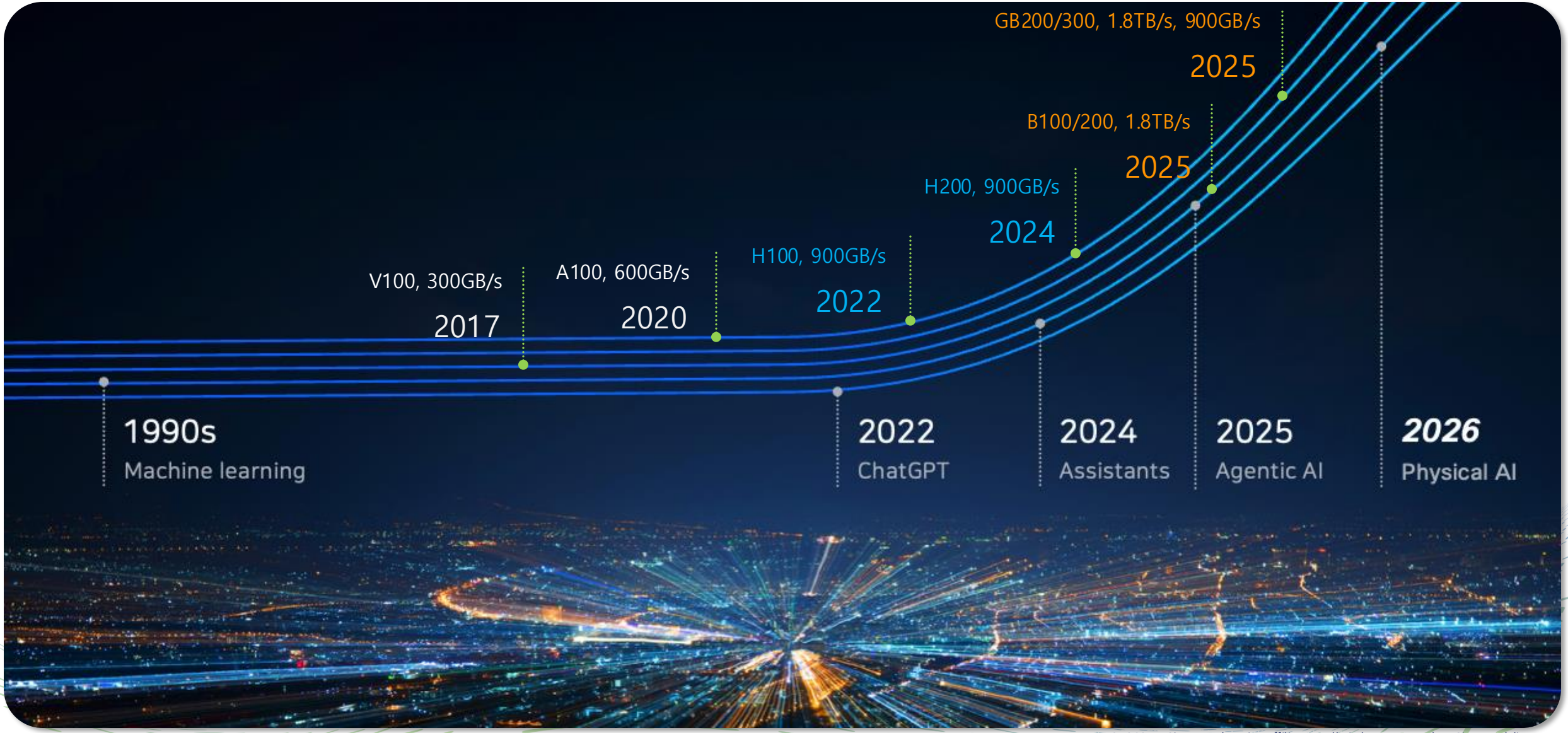
Sources: PWC, IDC

## 국외 인공지능 시장 전망

- 2024년 ~ 2029년 평균 31.22 % 성장 예상
- 가장 빠른 성장 시장은 Asia Pacific
- 가장 규모가 큰 시장은 North America

# AI 시장 및 인프라 혁신

## GPU 진화에 따른 AI 서비스 혁신

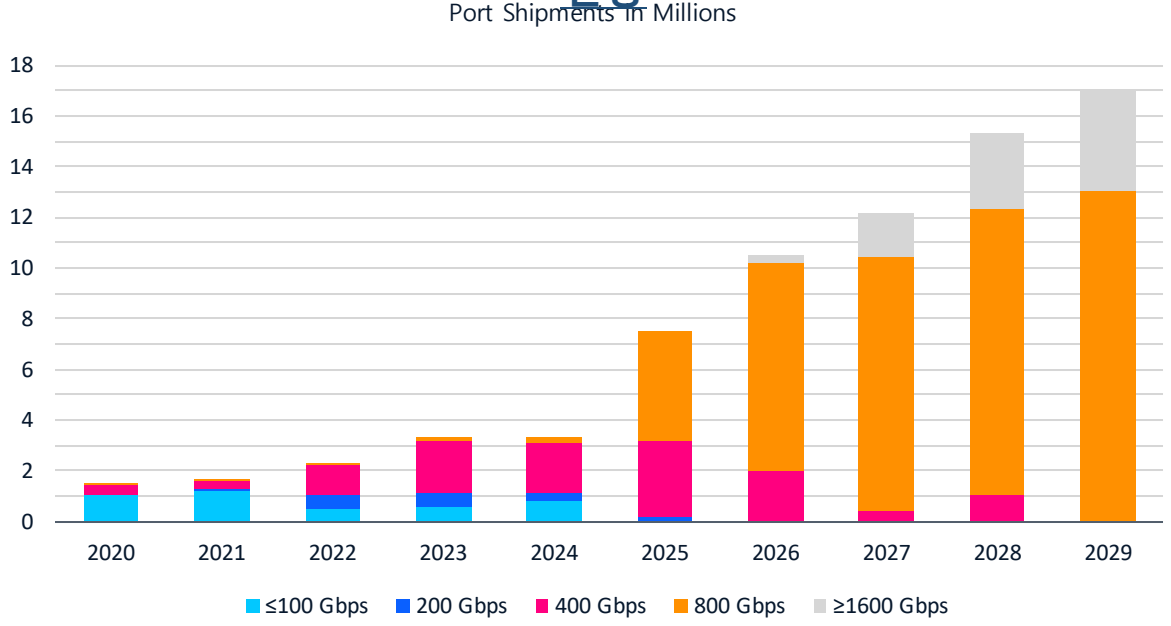


# AI 시장 및 인프라 혁신

## 데이터센터 네트워크 대역폭 증가

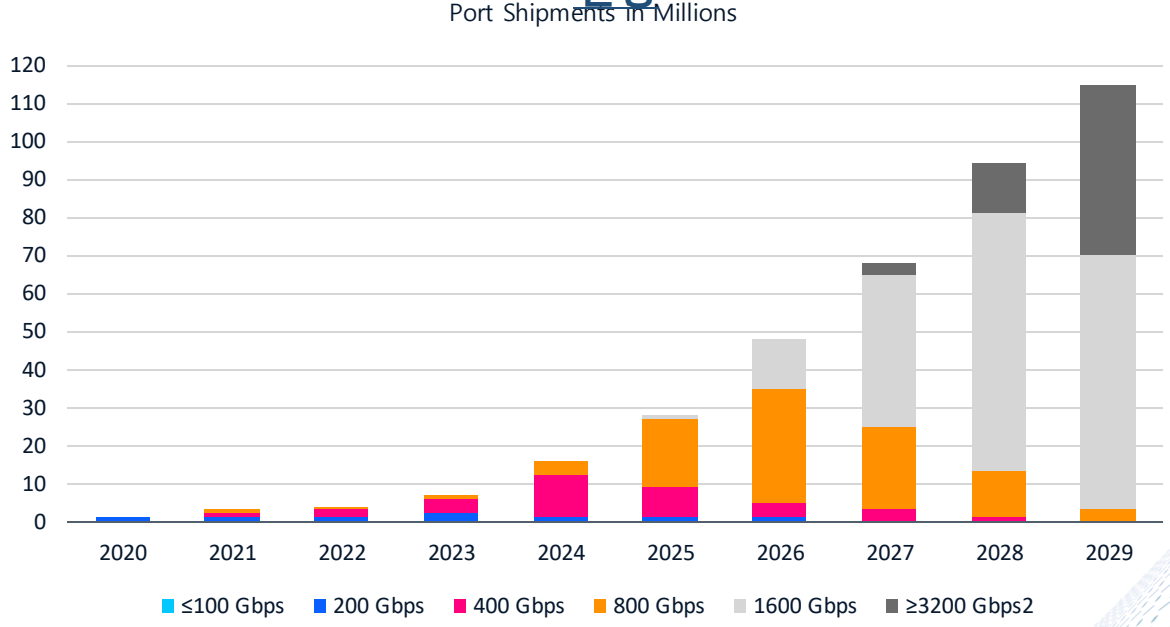
### AI Front-End 와 Back-End 네트워크 대역폭 차이는 GPU 간 데이터 통신 유무에 따라 차이 발생

**AI Front-End 네트워크 대역폭 전망**




**Speed Migration in AI Front-End Connectivity of Accelerated Servers, Dell'Oro, 2025**

**AI Back-End 네트워크 대역폭 전망**



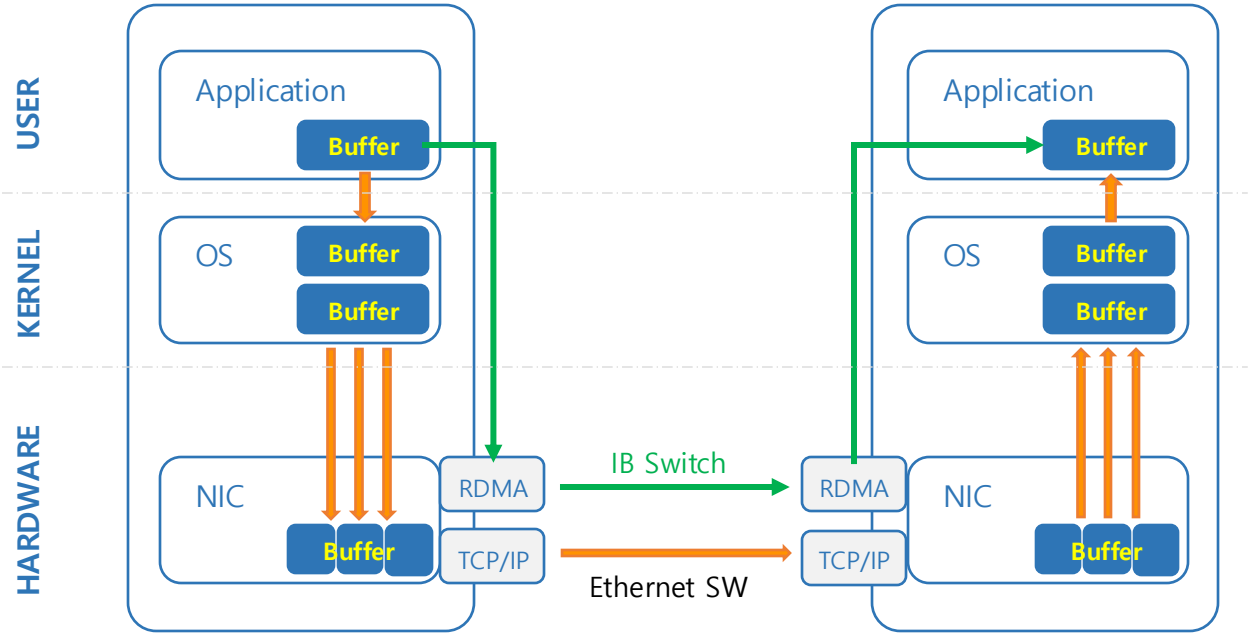

**Speed Migration (Ethernet & InfiniBand) in AI Back-End Networks, Dell'Oro, 2025**

Source: Dell'Oro Group and 650 Group

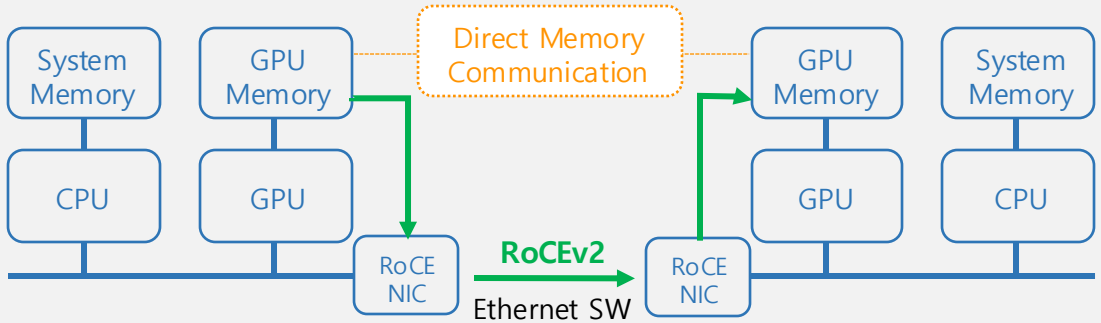
# AI 워크로드를 위한 네트워크 요구사항

## Low Latency - RDMA, iWARP, RoCEv2

AI/HPC 환경에 Low latency 구현을 위해 개선된 원격 직접 메모리 접근 기술이 적용되어 왔습니다.




| 항목 | TCP/IP 방식 | RDMA |
|---|---|---|
| 데이터 흐름 | App - OS - NIC - 네트워크 -NIC - OS - App | App - NIC - Network - NIC - App |
| 메모리 복사 단계 | 최소 3단계 복사 | Zero-Copy |
| CPU 개입 | 필수 (커널 스택 사용) | 거의 없음 (CPU Offload) |
| 지연 시간 (Latency) | 수십 ~ 수백 μs | 수 μs |
| 처리 속도 | 일반 이더넷 수준 | 높은 대역폭 |
| OS 커널 개입 | 필요 (커널 Bypass 불가) | 없음 (Kernel Bypass) |
| I/O 효율성 | 낮음 (CPU 자원 소모) | 높음 (네트워크 오프로드) |
| 복잡성 | 단순 (표준화) | 설정 복잡 (메모리 등록, QP 설정 등) |



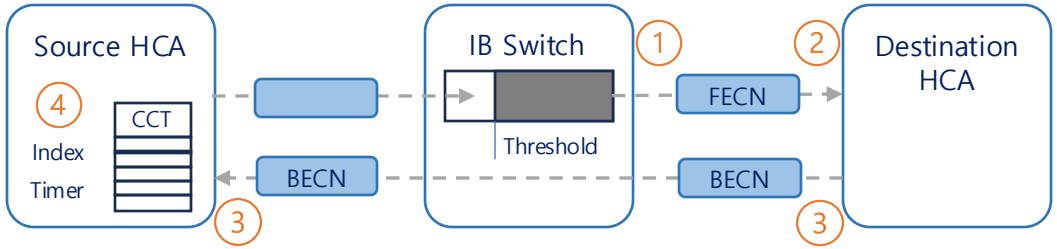

### RoCEv2

- RDMA over Converged Ethernet (UDP/IP, UDP 포트 4791)
- GPU Direct RDMA 기능 - GPU 간 통신 환경에서 Low latency 구현
- 혼잡 관리 (Congestion Management) - ECN, PFC
- L3 라우팅 지원으로 확장성 뛰어남

# AI 워크로드를 위한 네트워크 요구사항

## Congestion Management - ECN (IB), ECN & PFC (RoCEv2)

### IBA - InfiniBand Architecture




**① 혼잡 발생 감지**

- IB 스위치는 트래픽 버퍼가 임계치를 넘을 경우 혼잡 상태로 판단
- 이때 전방향 패킷에 FECN (Forward ECN) 비트 마킹

**② FECN 마킹된 패킷 전달**

- 혼잡 상태인 스위치를 지나가는 패킷은 목적지 HCA 도달
- 목적지 HCA는 패킷의 FECN 마크 인식

**③ BECN 전송**

- 목적지 HCA는 FECN 비트가 마킹된 패킷 수신 시, 송신지 HCA로 BECN 메시지 전송
- 이 과정은 혼잡이 발생했음을 송신 측에 알리는 역할

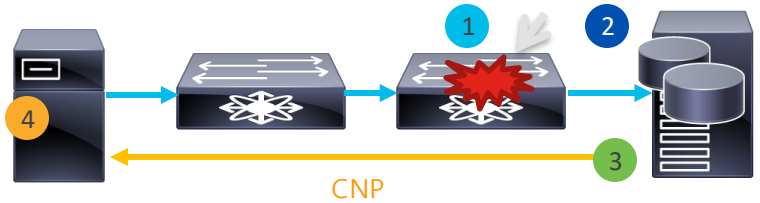
**④ 전송 속도 조절**

- 송신지 HCA는 BECN 수신 시 CCT 를 참조하여 해당 연결에 대한 전송 속도 낮춤
- 이후 타이머에 따라 점진적으로 속도 회복

**HCA**: Host Channel Adapter (NIC 역할)
**FECN**: Forward Explicit Congestion Notification
**CE**: Congestion Experience
**BECN**: Backward Explicit Congestion Notification
**CCT**: Congestion Control Table (혼잡 제어 테이블)
**CNP**: Congestion Notification Packet

### DCQCN - Data Center Quantized Congestion Notification

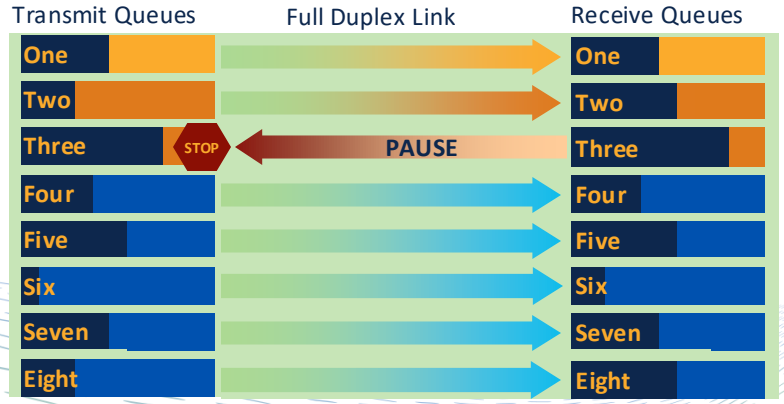
#### ECN (Explicit Congestion Notification)




**① 혼잡 발생 감지** **② ECN CE 마킹된 패킷 전달**
**③ CE 마크된 패킷 수신 시 CNP 패킷 전달** **④ 전송 속도 조절**

#### PFC (Priority-based Flow Control)

- 특정 우선순위별로 링크 레벨에서 정지 신호 (Pause Frame) 전송
- 일시적으로 흐름을 멈추어 트래픽 손실 방지

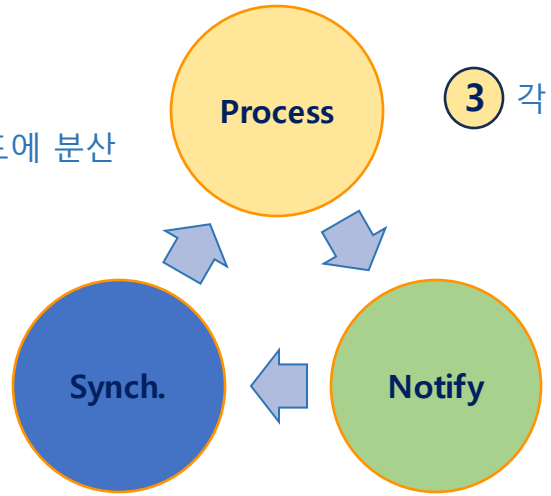


# AI 워크로드를 위한 네트워크 요구사항

GPU 병렬 통신에 따른 네트워크 토폴로지 및 패킷 분산 방식이 중요

## 각 GPU의 JCT(Job Completion Time) 동일 유지가 GPU 클러스터 성능 좌우

1. 학습데이터를 작은 청크 단위로 분할
2. 데이터 청크를 클러스터의 모든 GPU 노드에 분산

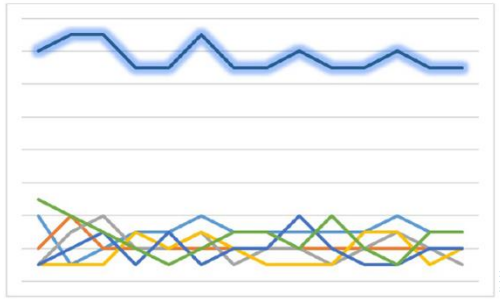



3. 각 GPU 노드는 할당된 데이터 청크를 학습
4. 각 GPU 노드는 할당된 데이트 청크 학습을 완료
5. 완료되면 결과가 모든 GPU 노드에 공유
6. 모든 노드의 결과를 결합하여 최종 AI 모델을 생성

### 일반적인 DC 트래픽 패턴

Cumulative Traffic

Individual Traffic

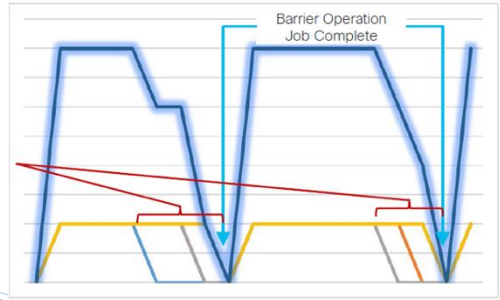


다량의 비동기식 small BW 플로우 패턴

GPU Stalled:
다른 GPU의 작업이 완료될 때까지 Waiting

- 데이터 불일치
- Deadlock 발생
- 병렬처리 장점 감소
- 성능 저하

### AI (All-to-All) 트래픽 패턴




Cumulative Traffic

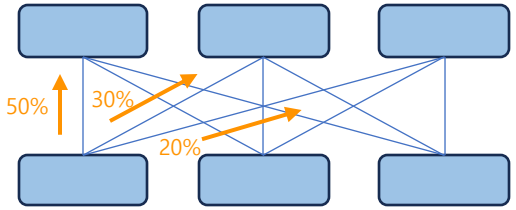
Individual Traffic

소량의 동기식 high BW 플로우 패턴

# AI 워크로드를 위한 네트워크 요구사항

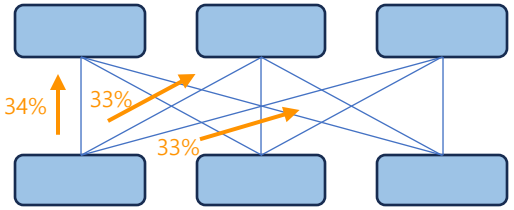
GPU 병렬 통신에 따른 네트워크 토폴로지 및 패킷 분산 방식이 중요

## ECMP (Equal-Cost Multi-Path)



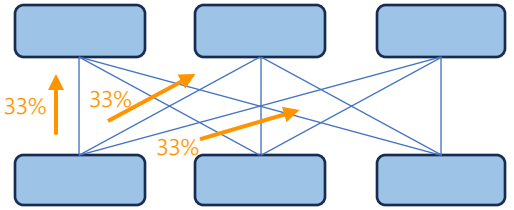

- Hash CRC16 (SIP, DIP, SPort, DPost, Protocol) / 경로 수
- 같은 플로우는 항상 같은 경로 사용
- 일반 L3 스위치 / 라우터에서 사용

## DLP (Dynamic Load Balancing)




- 스위치의 큐 상태, RTT 등 실시간 혼합 정보 기반 동적 분산
- 주로 AI, HPC, RoCE 환경에 적합

## Packet Spray (Per-packet Load Balancing)




- 같은 플로우의 모든 패킷을 경로별로 나눠서 분산 전송
- Packet 단위로 분산하므로 가장 세밀한 로드밸런싱
- 가장 이상적인 분산 방식

| 항목 | ECMP | DLB | Packet Spray |
|---|---|---|---|
| 분산 단위 | Flow 단위 (고정 해시) | Flowlet 단위 (혼잡도 기반) | Packet 단위 |
| 경로 선택 기준 | 해시값 기반 고정 경로 | 실시간 혼잡 상태 (큐/RTT/버퍼 상태 등) | 라운드로빈 등 경로 순회 |
| 순서 보장 | O (보장) | O (flowlet 기준으로 보장) | ✕ (패킷 순서 꼬임 가능성 높음) |
| 구현 복잡도 | 낮음 | 중~높음 (ASIC 지원 필요) | 낮음 (그러나 적용 제한 많음) |
| 장점 | 간단하고 빠름 | 지연 최소화, 트래픽 균형화 | 가장 세밀한 분산 가능 (실험적) |
| 단점 | 해시 편향 가능성 (정적) | 디버깅 어려움, 구현 복잡 (동적) | 패킷 순서 문제, RDMA와 호환 안됨 |
| 적합 환경 | 일반 L3 트래픽 | **RoCEv2, AI 트래픽, 대용량 통신** | 단순 트래픽, 비-RDMA 환경 |

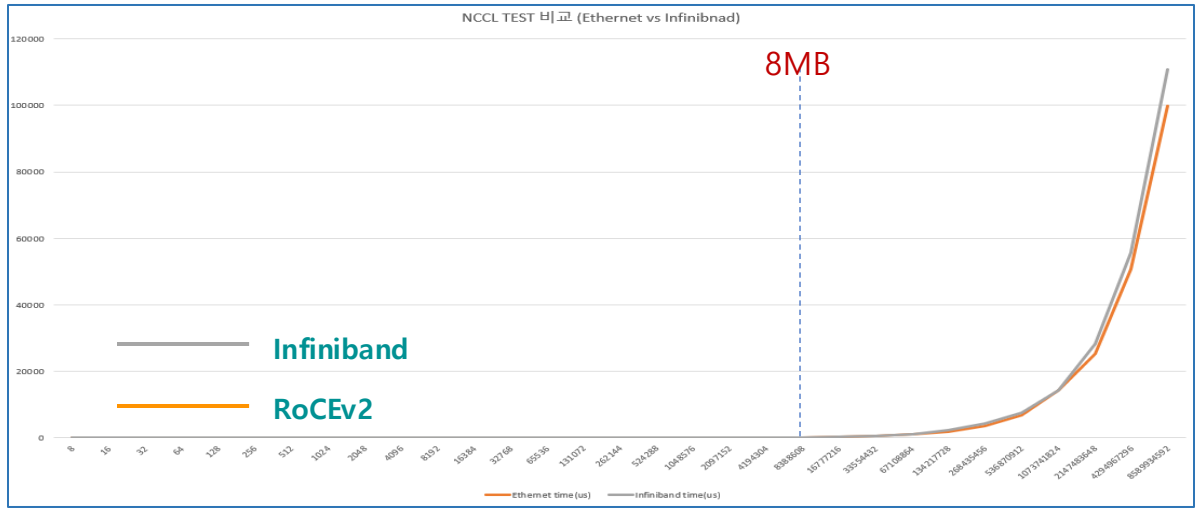
# AI 네트워킹 성능 테스트 결과

## Ethernet RoCEv2 vs. Infiniband - Cisco Korea Test

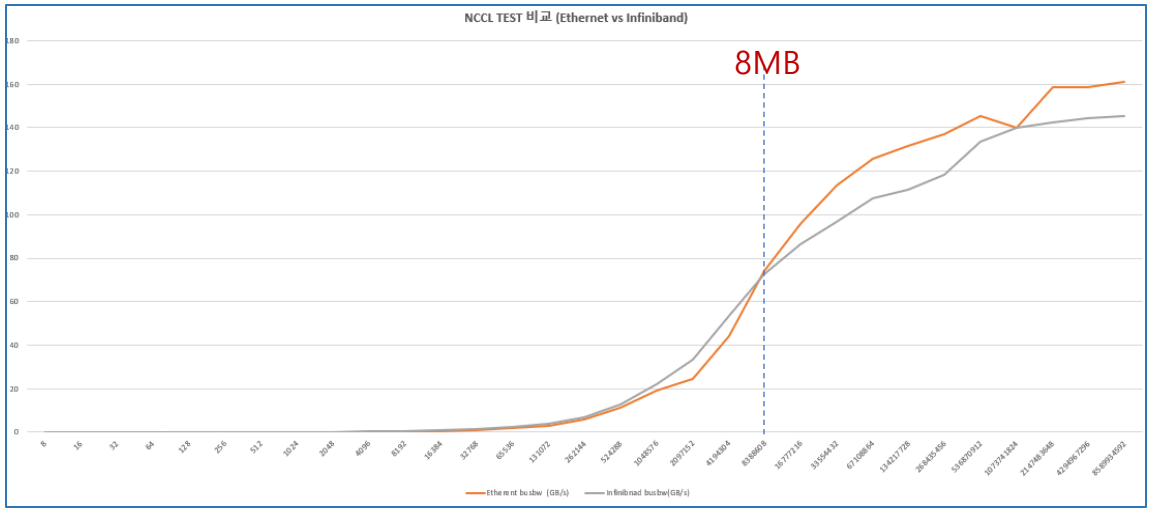
RoCEv2 : 8 x 100G vs. Infiniband : 4 x 200G

성능 측정 툴 : NCCL (NVIDIA Collective Communication Library)

Latency 비교




BW 비교

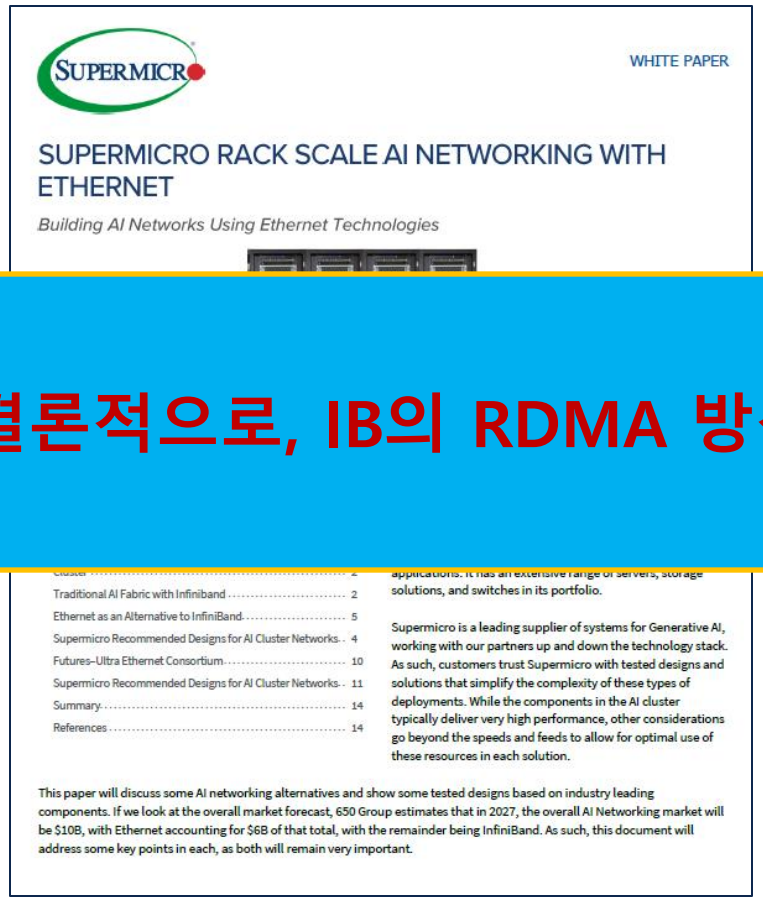



테스트 결과 특징 – 8MB 기점으로 데이터 크기가 클수록 RoCEv2에서 지연 시간과 성능에서 우수

멀티모달 GPT-4 환경에서는 데이터 크기가 커지므로 이더넷 기반의 RoCEv2 패브릭이 상대적으로 우수할 것으로 (

# AI 네트워킹 성능 테스트 결과

## Ethernet RoCEv2 vs. Infiniband - Supermicro



SUPERMICRO

WHITE PAPER

SUPERMICRO RACK SCALE AI NETWORKING WITH ETHERNET

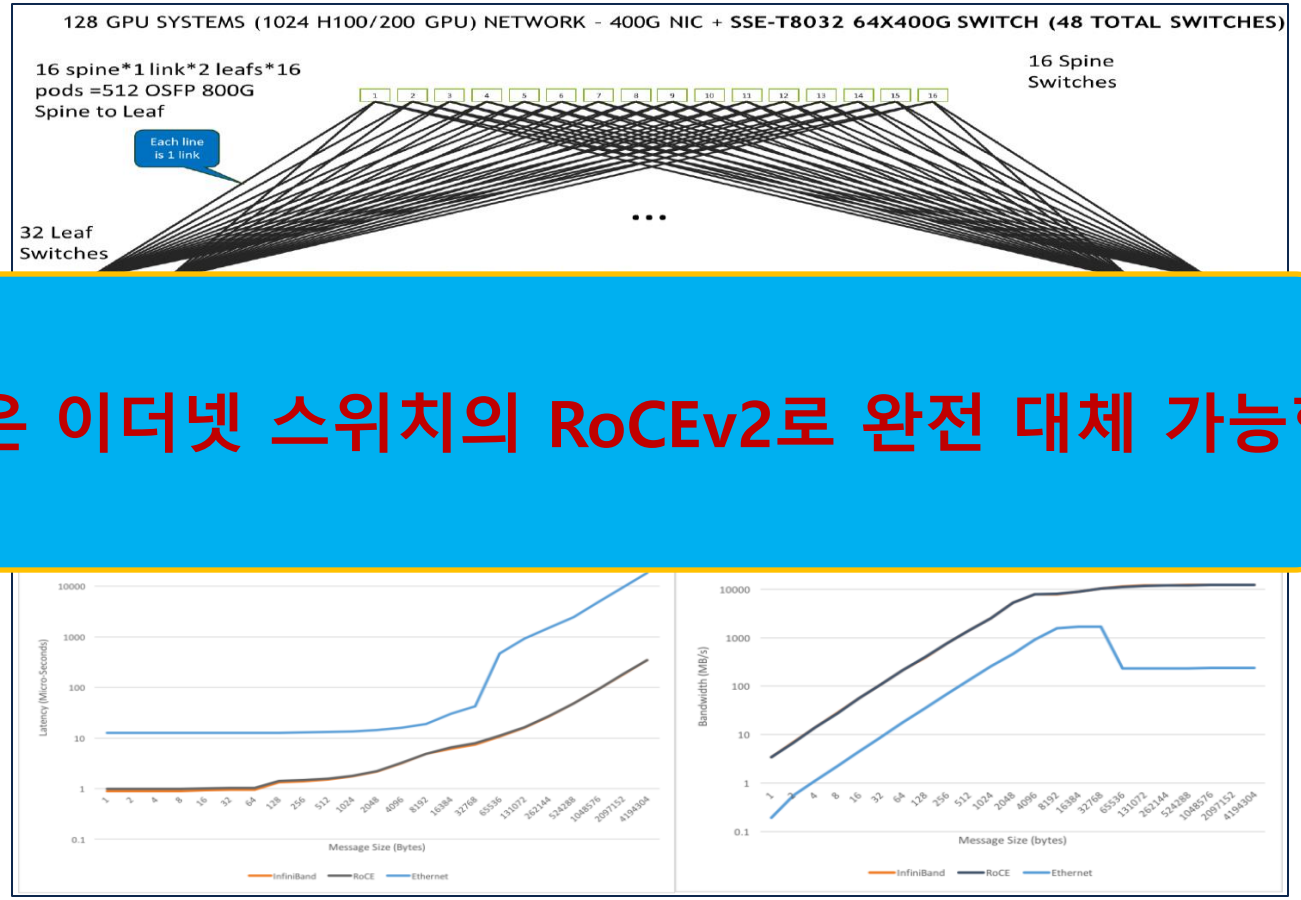
*Building AI Networks Using Ethernet Technologies*



solutions, and switches in its portfolio.

Supermicro is a leading supplier of systems for Generative AI, working with our partners up and down the technology stack. As such, customers trust Supermicro with tested designs and solutions that simplify the complexity of these types of deployments. While the components in the AI cluster typically deliver very high performance, other considerations go beyond the speeds and feeds to allow for optimal use of these resources in each solution.

This paper will discuss some AI networking alternatives and show some tested designs based on industry leading components. If we look at the overall market forecast, 650 Group estimates that in 2027, the overall AI Networking market will be $10B, with Ethernet accounting for $6B of that total, with the remainder being InfiniBand. As such, this document will address some key points in each, as both will remain very important.




**결론적으로, IB의 RDMA 방식은 이더넷 스위치의 RoCEv2로 완전 대체 가능합니다.**

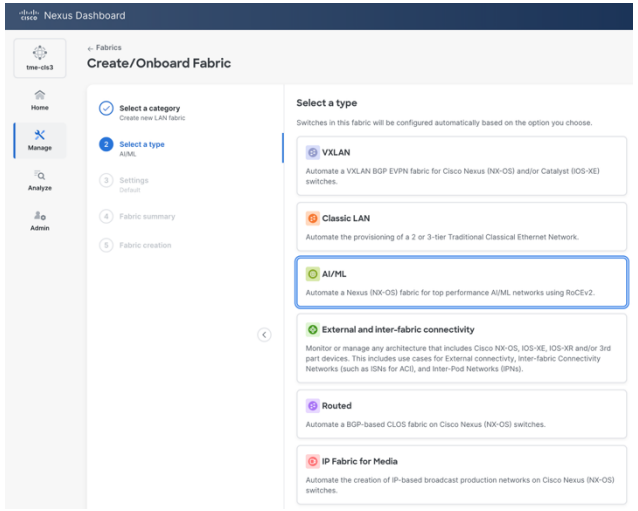
supermicro.com/white_paper/White_Paper_AI_Networking_With_Ethernet.pdf

# Cisco AI 인프라의 차별성 및 통합 관리

## 시스코만의 차별화 포인트

### 자동화

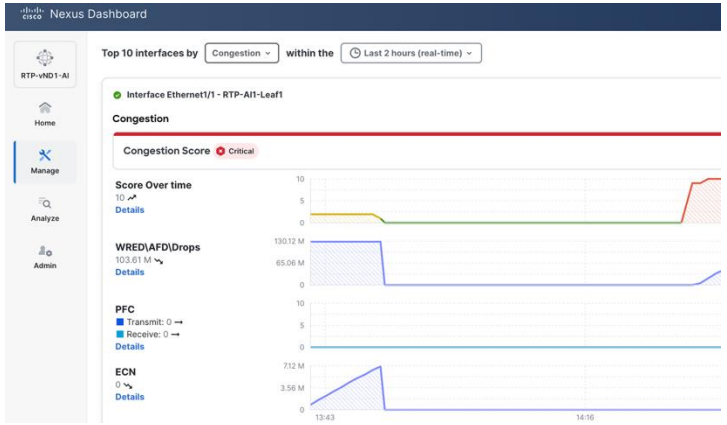
NDFC의 AI/ML 패브릭 QoS 정책 템플릿이 내장되어 오류 없이 확장 가능한 AI 네트워크 구축 가능




NDFC - Nexus Dashboard Fabric Controller

### 텔레메트리와 선제적 분석

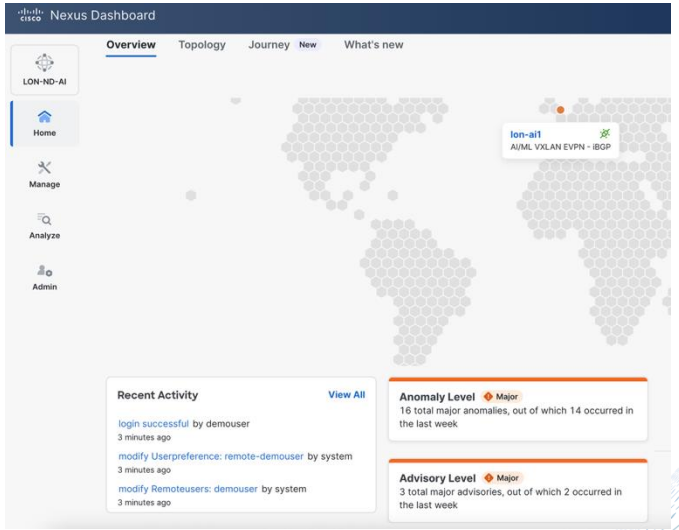
실시간 텔레메트리와 혼잡 점수를 통해 네트워크 문제를 사전에 감지 및 해결




NDI - Nexus Dashboard Insights

### 단일 운영 모델

프런트엔드와 백엔드 네트워크 운영을 통합하여 일관되고 간소화된 네트워크 관리 제공

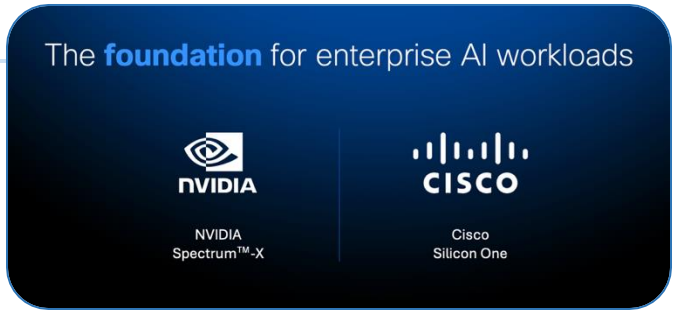



NDFC - Nexus Dashboard Fabric Controller

시스코 Advanced Service를 이용한 AI 네트워킹과 컴퓨팅 기술 지원 모두 제공 가능

# Cisco + NVIDIA 전략적 파트너쉽

## AI Innovation Journey Together




| | | |
|---|---|---|
| 1 | **Platform**<br>800G and 1.6T switches | SiliconOne and CloudScale ASICs<br>NVIDIA Spectrum ASICs |
| 2 | **Software**<br>Intelligent Packet Flow | Advanced load balancing techniques<br>Adaptive routing |
| 3 | **Operations**<br>Provisioning and Visibility | Nexus Dashboard AI fabric enhancements<br>Visibility into GPU / NIC<br>Monitoring AI jobs |
| 4 | **Partnerships**<br>Vendor agnostic functioning | GPU: NVIDIA, AMD reference architecture<br>NIC: Compatibility w/ major NICs<br>Server: Reference designs with UCS, Lenovo, Supermicro |
| 5 | **Market readiness**<br>Benchmark performance testing | ML Commons, IBPerf, NCCL<br>Customer success stories |

# Nexus AI Customer Wins

## Q3 FY2025 Major Wins

- XiaoAI 및 스마트폰, 스마트 하드웨어, 스마트 홈, 특히 XIAOMI EV 전기차의 첨단 운전자 보조 시스템(ADAS) 설계 지원
- 32개의 H20 GPU로 성능 테스트, 딥시크의 극단적 트래픽 패턴 시험
- 8,000개의 GPU로 디자인된 AI 네트워킹 패브릭 수주

Xiaomi
Training & Inferencing

El Puerto de Liverpool
Inferencing

XAI
Inferencing

Emirates Investment
Inferencing

Qatar Ports
Inferencing

Hong Kong University
Inferencing

Optiver US
Inferencing

Singapore Home Affairs
Inferencing

G-Research
Inferencing

Groq
Inferencing

Group42
Training & Inferencing

Quadrature
Inferencing

Largest AI bookings quarter

200+
new customers

800G
customer wins