대규모 GPU 클러스터에서의 Rail Optimized 및 NCCL PXN 기반 통신 최적화

이세하 [1], 신홍일 [2], 구건재 [3]
[1] 고려대학교 빅데이터융합학과, [2] 델 테크놀로지스, [3] 고려대학교 컴퓨터학과

slee585@korea.ac.kr, hongil.shin@dell.com, gunjaekoo@korea.ac.kr

Communication Optimizations on Large-Scale GPU Clusters Using Rail Optimized Networks and NCCL PXN

Seha Lee[1], Hongil Kim[2], Gunjae Koo[3]
[1]Dept. of Big-Data Convergence, College of SW.AI Convergence, Korea University
[2]Dell Technologies
[3]Dept. of Computer Science and Engineering, Korea University

요 약

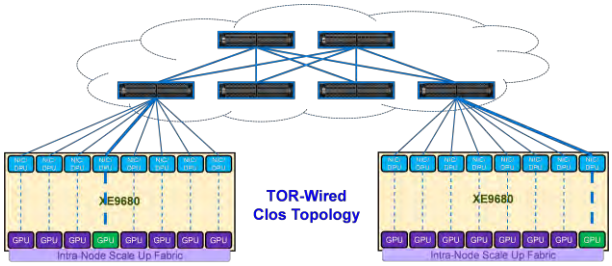
대규모 언어모델을 효율적으로 수행하기 위해서는 여러 대의 GPU 와 CPU 간의 네트워크 통신 최적화가 필요하다. 본 논문에서는 기존의 네트워크 토폴로지 기반 클러스터 구성의 비효율성을 개선하기 위하여 대규모 GPU 클러스터 환경에서 Rail Optimized Network 와 NCCL PXN, NUMA 단순화 및 CPU Affinity 기반의 통신 성능 최적화 기법을 제안한다. 본 연구에서는 제안하는 네트워크 구성 및 라이브러리를 활용하여 Dell XE9680 서버의 4-HCA 구성과 GPU-HCA-CPU 매핑을 통해 집합 통신의 처리량과 지연 시간을 개선하였다. 제안된 방식은 실제 벤치마크를 통해 성능 향상과 확장성을 입증하였으며, 고성능 컴퓨팅 환경에서의 효율적인 네트워크 설계 방향을 제시한다.

1. 서론

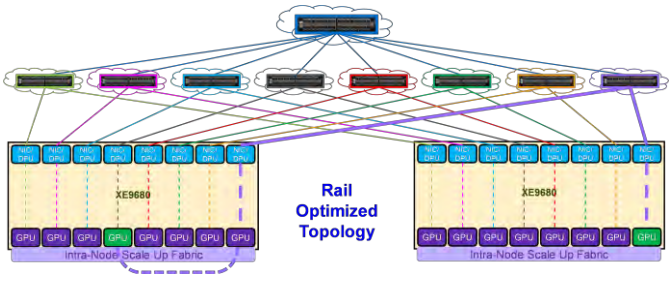
대규모 언어 모델(LLM) 및 생성형 AI 의 확산에 따라, GPU 클러스터의 통신 성능은 모델 학습의 병목을 결정짓는 핵심 요소로 부상하고 있다 [1]. 특히 집합 통신(AllReduce, AlltoAll 등)은 GPU 간 대량의 메시지를 교환하며, 네트워크 혼잡과 지연이 유발될 수 있다 [2]. 이를 해결하기 위해 NVIDIA 는 NCCL 2.12 버전에서 Peer-to-Peer over NVLink (PXN) 기능을 도입하여, GPU 간 통신 경로를 CPU 를 거치지 않고 NVLink 를 통해 직접 연결함으로써 메시지 집약 및 대역폭 최적화를 달성하였다 [2]. 또한 Dell 의 XE9680 서버 기반 설계에서는 GPU 2 개당 1 개의 HCA 구성인 4-HCA 비용 효율적인 구성과 Rail Optimized Network 토폴로지를 적용하여, GPU 통신간 최적의 경로 설정을 통해 통신 효율을 극대화하는 구조를 제시하였다 [3]. 이러한 기술적 진보는 고성능 AI 학습 환경에서의 확장성과 비용 효율성을 동시에 확보할 수 있는 기반을 제공한다 [4].

2. 연구 동기

최근 대규모 언어 모델(LLM)과 생성형 AI 학습에서는 수백~수천 개의 GPU 를 연결하는 고성능 네트워크 인프라가 필수적이다. 전통적인 Clos 네트워크 토폴로지 (그림 1)는 일반적 구성이지만, 다단계 스위칭으로 인한 지연(Latency)과 네트워크 혼잡이 발생할 수 있다. 이에 대응하여 Rail Optimized Network 구조 (그림 2)가 제안되었으며, 이는 GPU-HCA-Leaf 스위치 간의 직접 경로를 최적화하여 통신 경합을 줄이고 대역폭 활용도를 높인다 [5]. 따라서 본 연구는 실제 Dell XE9680 서버 아키텍처와 NCCL PXN, Rail Optimized Network 를 통합적으로 적용, 분석함으로써, 단순 이론적 설명을 넘어 실질적 성능 향상 수치와 최적화 방법론을 제시한다. 이를 통해 대규모 GPU 학습 인프라에서 발생하는 통신 병목 문제를 해결하고, 차세대 AI 학습 환경의 확장성, 비용 효율성, 안정성을 동시에 확보하는 데 기여하고자 한다.



(그림 1) Clos Topology 구성도 사진



(그림 2) Rail Optimized 구성도

3. 기술 배경 및 관련 연구

3-1 NCCL 과 PXN 기술

NVIDIA Collective Communication Library(NCCL)는 GPU 간 집합 통신(AllReduce, AlltoAll 등)을 최적화하기 위해 설계된 라이브러리로, NVLink, PCIe, InfiniBand 를 포함한 다양한 경로를 활용한다 [3]. NCCL 2.12 에서는 PXN 기능이 도입되어, GPU 간 데이터 전송 시 CPU 를 거치지 않고 NVLink 또는 PCIe 를 통해 직접 통신할 수 있도록 하였다. NVIDIA 공식 블로그에 따르면, PXN 활성화 시 AlltoAll 성능이 최대 2 배 향상되며, 특히 다중 HCA 환경에서 성능 개선 효과가 두드러졌다고 설명하고 있다 [1].

3-2. CPU Affinity 와 NUMA 최적화

대규모 GPU 서버에서는 CPU- GPU- HCA 간의 물리적 연결 관계가 성능에 큰 영향을 미친다. NUMA(Non-Uniform Memory Access) 환경에서 CPU Affinity 를 적절히 설정하면, GPU 통신 경로의 지연을 줄이고 인터럽트 처리 효율을 높일 수 있다. Dell 의 기술 백서에서는 Sub-NUMA Clustering, IRQ Affinity, Core Pinning 등을 활용하여 GPU- HCA 매핑을 최적화하는 방법을 제시하고 있으며, 이를 통해 Throughput 과 Tail Latency 모두 개선됨을 보고하였다 [2][4].

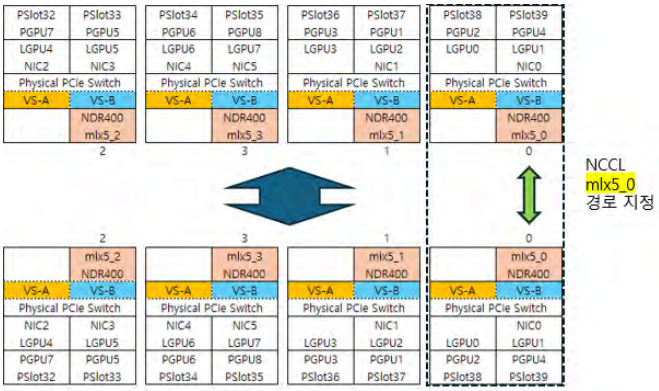
3-3. 기존 연구의 한계

기존 Clos 기반 네트워크는 확장성 측면에서 장점이 있으나, 대규모 집합 통신 시 다단계 스위칭으로 인한 병목이 발생한다. 또한 PXN 미활용 환경에서는 GPU 간 통신이 CPU 를 경유하게 되어 불필요한 지연이 추가된다. Rail Optimized Network 와 PXN, CPU Affinity 를 결합한 최적화 기법은 이러한 한계를 극복하고, 대규모 AI 학습 환경에서의 성능과 확장성을 동시에 확보할 수 있는 대안으로 주목받고 있다.

4. 제안 아키텍처 및 최적화 기법

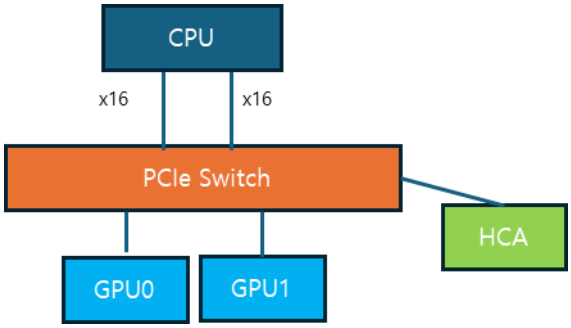
본 연구의 제안 구성은 다음과 같다.

1. NCCL PXN 활성화: NVLink 기반 GPU 간 데이터 전송 경로를 검증하고, 특정 GPU 를 선택하는 방식으로 NCCL 기동하는 조건으로 최적화 경로를 명확히 검증하는 방식으로 Direct HCA 경로와 전체 GPU 경로의 성능을 비교 분석한다 (그림 3).



(그림 3) NCCL PXN 활성화 구성도

2. CPU Affinity 지정 및 활성화: GPU 통신 성능은 CPU 스케줄링 정책과 밀접한 관련이 있으며, 특히 NUMA 기반 시스템에서는 CPU 코어와 GPU 간의 물리적 근접성이 중요하다. 본 연구에서는 GPU 순서와 CPU 코어 번호를 매핑하여 스케줄링 지연을 최소화하고, taskset 을 활용해 특정 코어 범위(예: 0- 143)를 고정한다. NCCL 의 자동 최적화 기동으로 CPU ↔ GPU ↔ NIC 간 데이터 경로에서 발생하는 불필요한 컨텍스트 스위칭을 줄이고, 통신 처리 효율을 높인다 (그림 4).



<그림 4> CPU↔GPU↔NIC 경로 관계

3. NUMA 최적화: NUMA 환경에서 메모리 접근 지연은 성능 저하의 주요 원인 중 하나이다. 본 연구에서는 BIOS 에서 Sub-NUMA Cluster 기능을 비활성화하고, GPU 와 물리적으로 가까운 CPU 코어를 고정하여 데이터 경로의 지연을 최소화한다.
세 가지 기법을 종합적으로 적용함으로써, 본 연구는 NCCL 기반 집합 통신에서 처리량과 지연 시간을 동시에 개선하는 것을 목표로 한다.

5. 실험 환경 및 결과

5-1. 실험 환경

- 서버: Dell XE9680 2 대(서버 당: 8×NVIDIA H100 GPU, 4×NVIDIA ConnectX-7 HCA, 2×Intel Xeon CPU)
- 네트워크: 400GbE 스위치, Rail Optimized Network
- 소프트웨어: Ubuntu 22.04 LTS, CUDA 12.8, NVIDIA Driver 570.xx, NCCL 2.26+

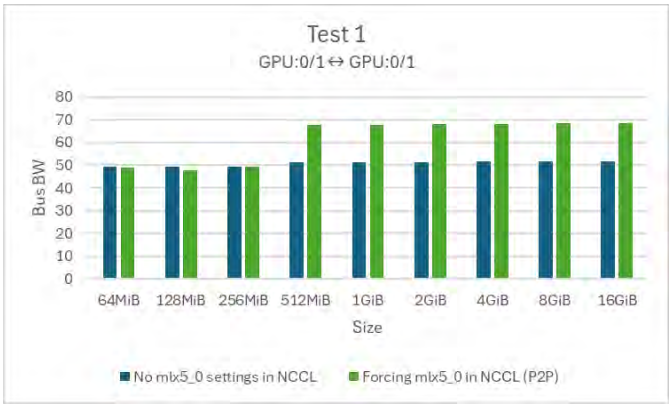
5-2. 밴치마크

- 측정 지표: NCCL bus bandwidth (GB/s), 선택된 GPU 쌍/그룹 및 전체 집합 경로.
- GPU Size: 64 MiB ~ 16 GiB

5-3. 실험 결과

Test 1 (PXN 지정, NUMA 기본)

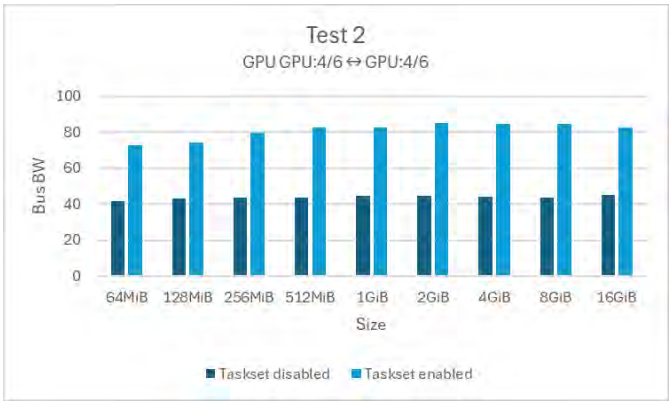
- 구성: 8×GPU - 4×HCA, Rail Optimized 배선, 각각의 GPU 선택후 Direct HCA 경로 설정, PXN 지정.
- 실험 방법: 강제로 mlx5_0 이라는 Direct HCA 경로 사용하여 성능 향상되었는지 확인한다. (NCCL_IB_HCA=mlx5_0)
- 결과: 지정하기 전, 평균 bus bandwidth 은 50.77 GB/s 로 확인됐지만, 선택된 GPU 와 가까운 HCA 를 강제로 지정하여 구성했을 때, 61.77 GB/s 로 약 21.7% 의 성능 향상을 확인했다 (표 1).



<표 1> Forcing mlx5_0 (P2P) 사용 전 후 Bandwidth 비교

Test 2 (CPU Affinity 지정 및 활성화)

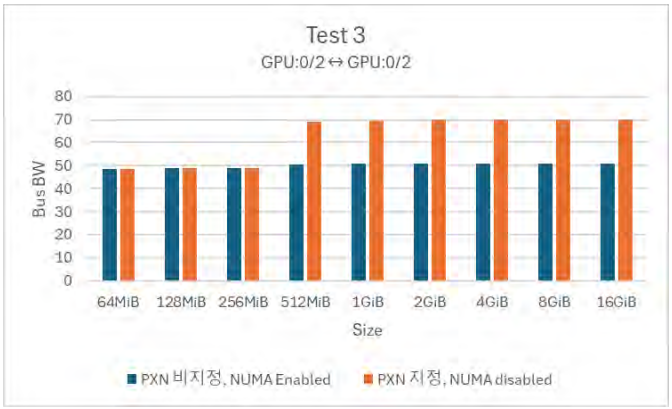
- 구성: PXN ON + Rail Optimized, CPU Affinity 미적용과 PXN ON + Rail Optimized + CPU Affinity 지정 비교.
- 실험 방법: Taskset 으로 특정 CPU# 지정한 후 성능 향상 비교 (eg: taskset -c 1,145)
- 결과: taskset 지정하는 기능 이용한 성능 값이 평균 80.66 GB/s 로 지정 전 값 43.75 GB/s 보다 약 47% 더 향상된 값이 나왔다 (표 2).



<표 2> CPU Affinity 사용 전, 후 Bandwidth 비교

Test 3 (NUMA 최적화 확장 BIOS + 코어 매핑)

- 구성: PXN ON + Rail Optimized + BIOS 최적화 설정(Sub-NUMA Cluster Disabled) + CPU Affinity 적용
- 실험 방법: Sub-NUMA Cluster 설정을 기본으로 한 값과, disabled 설정한 값을 비교 분석한다.
- 결과 값: BIOS 최적화 전 및 PXN 경로 미지정은 약 50.05 GB/s, BIOS 최적화 및 PXN 경로 지정한 값은 약 62.52 GB/s 로 약 24.8%의 성능 향상을 확인할 수 있었다 (표 3). 특히 512MB 사이즈때 부터 급격히 두 모델의 차이가 커지는 것 또한 눈여겨 볼 점이다.



<표 3> BIOS 최적화 전, 후 Bandwidth 비교

6. 결론 및 향후 연구

본 연구에서는 Dell XE9680 서버(8×NVIDIA H100 GPU, 4×ConnectX-7 HCA) 기반 환경에서 NCCL 통신 성능 최적화를 위해 PXN 지정, CPU Affinity 적용, NUMA 및 BIOS 설정 변경의 효과를 분석하였다. 주요 결과는 다음과 같다.

Direct HCA 경로 지정 (PXN 활용)

mlx5_0 HCA 를 강제로 지정한 경우 평균 NCCL bus bandwidth 가 50.77 GB/s → 61.77 GB/s 로 약 20% 향상됨을 확인하였다.

CPU Affinity 적용

PXN ON + Rail Optimized 환경에서 taskset 을 통한 CPU 코어 고정 시, 평균 성능이 43.75 GB/s → 80.66 GB/s 로 약 47% 향상되었다. 이는 CPU taskset 지정이 NCCL 성능에 큰 영향을 미침을 시사한다.

BIOS NUMA 최적화(Sub-NUMA Cluster Disabled)

BIOS 최적화 및 PXN 경로 지정 시 평균 성능이 50.77 GB/s → 62.42 GB/s 로 약 20% 향상되었으며, 특히 512MB 이상 메시지 크기에서 성능 격차가 크게 확대되는 경향을 보였다.

이상의 결과는 GPU-NIC 매핑, CPU 코어 바인딩, NUMA 정책이 고대역폭 GPU 통신에서 핵심적인 역할을 한다는 점을 명확히 보여준다.

NVIDIA 는 PXN, NVLink, SHARP 와 같은 기능을 지속적으로 개선하고 있으므로, 향후 연구에서는 이러한 기술적 변화가 통신 및 연산 성능에 미치는 영향을 정량적으로 분석할 필요가 있다. 특히 Collective 알고리즘의 구조적 변화가 PXN 경로 및 NUMA 최적화와 어떤 상호작용을 보이는지에 대한 심층적인 고찰이 요구된다.

현재 PXN 경로 지정과 CPU Affinity 설정은 모두 수동으로 수행되고 있어, 운영 환경의 효율성을 높이기 위해 자동화가 필요하다. GPU 와 NIC 간의 최적 경로를 계산하고 NUMA 코어를 자동으로 매핑하는 Topology-aware Scheduler 를 개발한다면, 관리 부담을 줄이는 동시에 안정적이고 일관된 성능을 확보할 수 있을 것이다.

본 연구에서는 두 개의 노드에서만 실험이 진행되었으나, 실제 대규모 학습 환경에서는 멀티 GPU 및 멀티 노드 구성이 일반적이다. 따라서 향후에는 두 대 이상의 XE9680 서버를 연결하여 400 GbE 기반 NCCL 성능을 측정하고, SHARP In-Network Aggregation 이나 NVLink Switch 와 같은 고급 네트워크 기능이 전체 시스템 성능 및 확장성에 미치는 영향을 분석해야 한다. 특히 AllReduce, AllGather 등 Collective 연산에서의 성능 확장성 검증이 핵심 과제가 될 것이다. 또한 CPU 아키텍처에 따른 성능 차이 역시 중요한 변수로 작용할 수 있다. Intel 과 AMD 프로세서는 NUMA 구조와 메모리 접근 방식이 상이하므로, nvidia-smi topo 및 lstopo 등의 도구를 활용해 시스템 아키텍처와 NUMA 토폴로지를 정확히 파악하는 것이 중요하다. 이를 통해 각 아키텍처별 자원 배치 최적화 및 통신 효율 개선 방향을 도출할 수 있을 것이다.

참고문헌

[1] Mandakolathur, K., & Jeaugey, S. (2022, February 28). Doubling all2all performance with Nvidia Collective Communication Library 2.12. NVIDIA Technical Blog. https://developer.nvidia.com/blog/doubling-all2all-performance-with-nvidia-collective-communication-library-2-12/

[2] Dell Technologies info hub. (2025, July). Generative AI in the enterprise with Nvidia GPUs, Dell Networking, and Nvidia Software Stack. [White paper]. https://infohub.delltechnologies.com/en-us/t/generative-ai-in-the-enterprise-with-nvidia-gpus-dell-networking-and-nvidia-software-stack/

[3] NVIDIA. (2020). Overview of NCCL - NCCL 2.28.3 documentation. [White paper] NVIDIA. https://docs.nvidia.com/deeplearning/nccl/user-guide/docs/overview.html

[4] Dell Technologies info hub. (2024, April). Generative AI in the Enterprise – Model Training. [White paper]. https://infohub.delltechnologies.com/en-au/t/white-papers/

[5] B. Kitor and K. Rygol, "Rail Optimized PCIe Topologies for LLMs," ISC High Performance 2025 Research Paper Proceedings (40th International Conference), Hamburg, Germany, 2025, pp. 1-11.