GPU 클러스터 시스템의 계산노드 간 인터커넥트 네트워크 통신 성능 비교 분석 연구

권민우*, 안도식*, 홍태영*
*한국과학기술정보연구원 슈퍼컴퓨팅인프라센터
mwkwon81@kisti.re.kr

A study on comparison and analysis of interconnect network communication performance between computing nodes in GPU cluster system

Min-Woo Kwon*, Do-Sik An*, TaeYoung Hong*
*Dept. of Supercomputing Infrastructure Center, KISTI

요 약

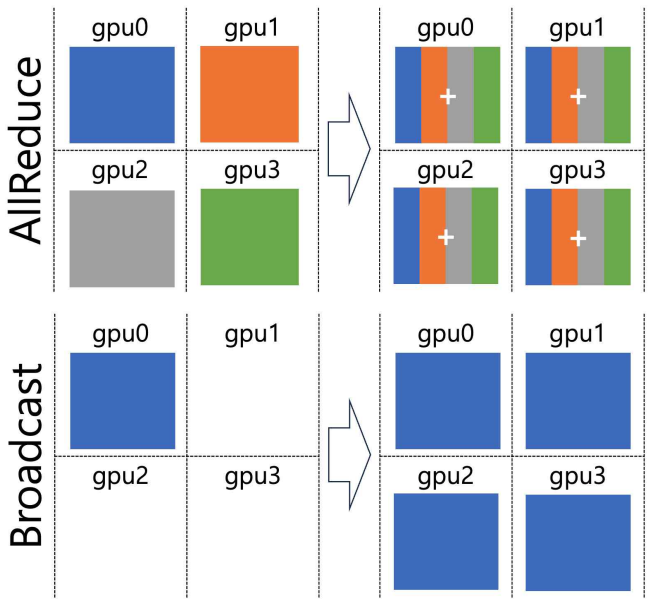
KISTI의 GPU 클러스터 시스템인 뉴론은 NVIDIA의 A100과 V100 GPU가 총 260개 탑재되어 있는 클러스터 시스템이다. 뉴론의 계산노드들은 고성능의 인터커넥트인 Infiniband(IB) 케이블로 연결되어 있어 멀티 노드 작업 수행 시에 고대역 병렬통신이 가능하다. 본 논문에서는 NVIDIA사에서 제공하는 NCCL의 벤치마크 코드를 이용하여 인터커넥트 네트워크의 통신 성능을 비교분석하는 방안에 대해서 소개한다.

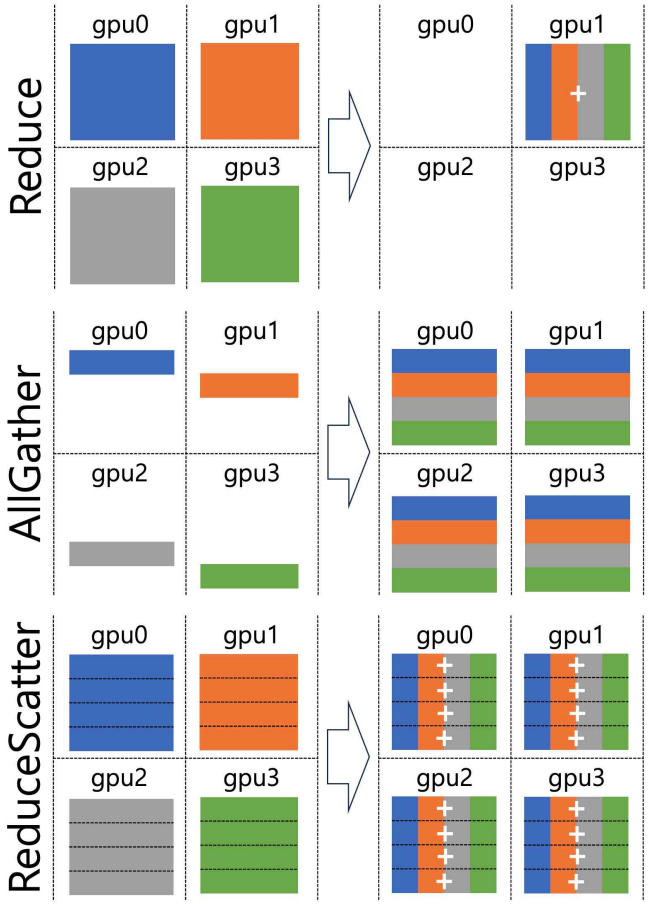
1. 서론

한국과학기술정보연구원(KISTI)에서는 슈퍼컴퓨터인 누리온의 보조시스템인 뉴론을 운영하고 있다. 누리온 시스템이 CPU 기반 시스템으로 HPC 어플리케이션 수행에 최적화되어 있는 반면에 뉴론 시스템은 NVIDIA A100, V100 GPU를 탑재한 계산노드로 구성되어 있어 AI 어플리케이션 수행에 최적화되어 있다. KISTI는 슈퍼컴퓨터 사용자가 누리온에서 수행한 HPC 어플리케이션 입출력 데이터를 뉴론에서 즉시 AI 연구에 활용할 수 있도록 HPC+AI +Data Analitics 통합 서비스를 구축하였다. 이를 위해 누리온과 뉴론 시스템은 동일한 계정관리(LDAP) 서버를 사용하며, 라우터(Lnet)를 이용하여 인터커넥트 네트워크가 상호 연결되어 있다. 최근 AI 모델의 사이즈가 커짐에 따라 멀티 GPU를 넘어 멀티 노드를 활용한 작업 수행의 필요성이 증대되고 있다. 이러한 멀티 노드 병렬 작업 성능을 끌어올리기 위해서는 고성능의 인터커넥트를 적절한 네트워크 토폴로지로 구성해야 한다. 본 논문에서는 NVIDIA사에서 제공하는 NCCL의 벤치마크 코드를 활용하여 뉴론과 같은 클러스터 시스템의 인터커넥트 네트워크 통신 성능을 비교분석하는 방안에 대해 소개한다.

2. NCCL (NVIDIA Collective Communication Library)

NCCL은 NVIDIA사에서 제공하는 멀티 GPU 및 멀티 노드 집합 통신 라이브러리이다. GPU간에 PCIe, NVLink, 인터커넥트로 연결되어 있는 상황에서 고대역폭과 낮은 Latency를 제공하기 위한 최적화된 집합 통신 함수를 제공한다[2]. NCCL은 그림 1과 같이 다양한 집합 통신 함수를 제공한다.





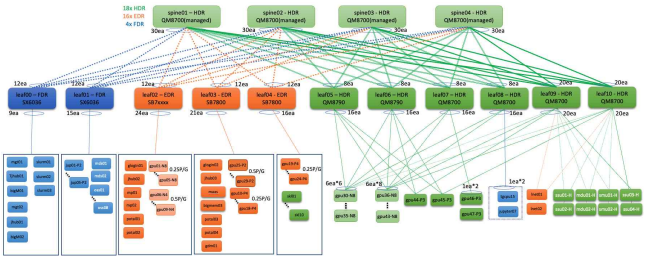
(그림 1) NCCL 집합 통신

또한 ncclSend/Recv함수를 이용하여 AllToAll과 같은 Point-to-point(이하 P2P) 통신도 구현이 가능하다[2]. NVIDIA사는 NCCL 함수의 기본적인 성능을 측정할 수 있는 벤치마크 코드를 GitHub 사이트에서 제공하고 있다[3]. 본 논문에서는 NCCL 벤치마크 코드를 뉴론의 계산노드에서 수행하고 AllGater, AllReduce, AllToAll 성능에 대한 결과를 분석하였다.

3. 뉴론 인터커넥트 네트워크 토폴로지

뉴론은 그림 2와 같이 인터커넥트(Infiniband, 이하 IB) 네트워크가 Fat-tree 토폴로지로 구성되어 있다. 2014년 슈퍼컴퓨터의 성능 벤치마크 시스템으로 처음 구축된 뉴론은 연차별로 계산노드, 인터커넥트, 스토리지가 증설되어 다양한 종류의 시스템으로 구성되어 있다. 인터커넥트의 경우 FDR(56Gbps), EDR(100Gbps), HDR(200Gbps)로 구성되어 있으며, FDR 스위치(그림 2의 파란색 영역)에는 고성능의 대역폭을 필요로 하지 않는 인프라 노드가 연결되어 있다.

계산노드의 경우, 병렬 통신에서 고성능의 대역폭이 필요하기에 EDR 스위치(그림 2의 주황색 영역)와 HDR 스위치(그림 2의 초록색 영역)에 주로 연결되어 있다. Fat-tree 토폴로지의 상단과 하단사이에는 2to1으로 IB 케이블이 연결되어 있으며, 계산노드 내에 GPU카드와 IB 포트의 연결 비율은 아래와 표 1과 같다.



(그림 2) 뉴론 인터커넥트 네트워크 구성도

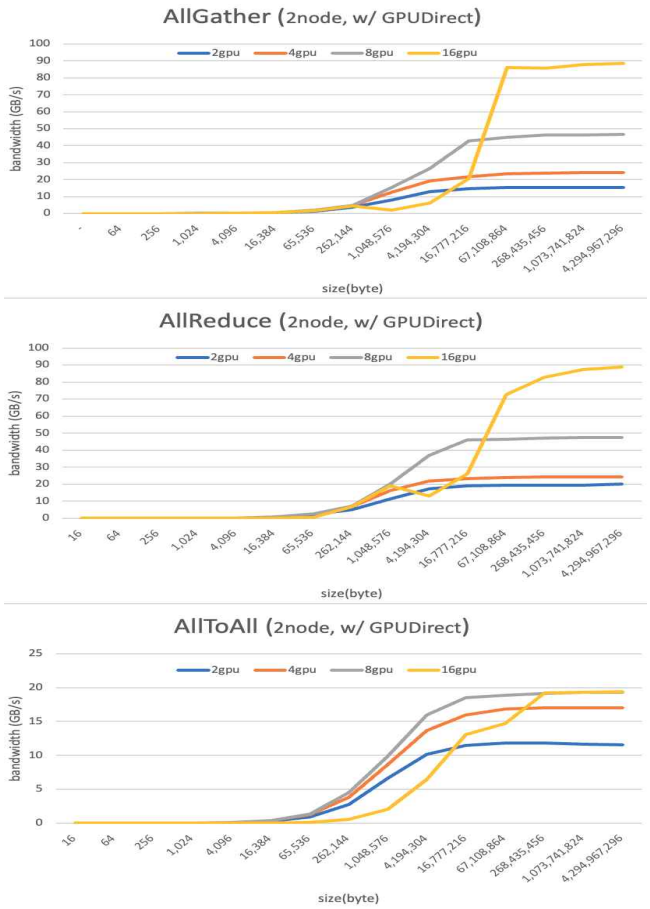
<표 1> GPU 노드별 GPU:IB 비율

노드종류	GPU개수	IB포트수	GPU:IB
A100 8GPU 장착	8	4	2:1
V100 8GPU 장착	8	2	4:1
V100 4GPU 장착	4	1	4:1
V100 2GPU 장착	2	1	2:1

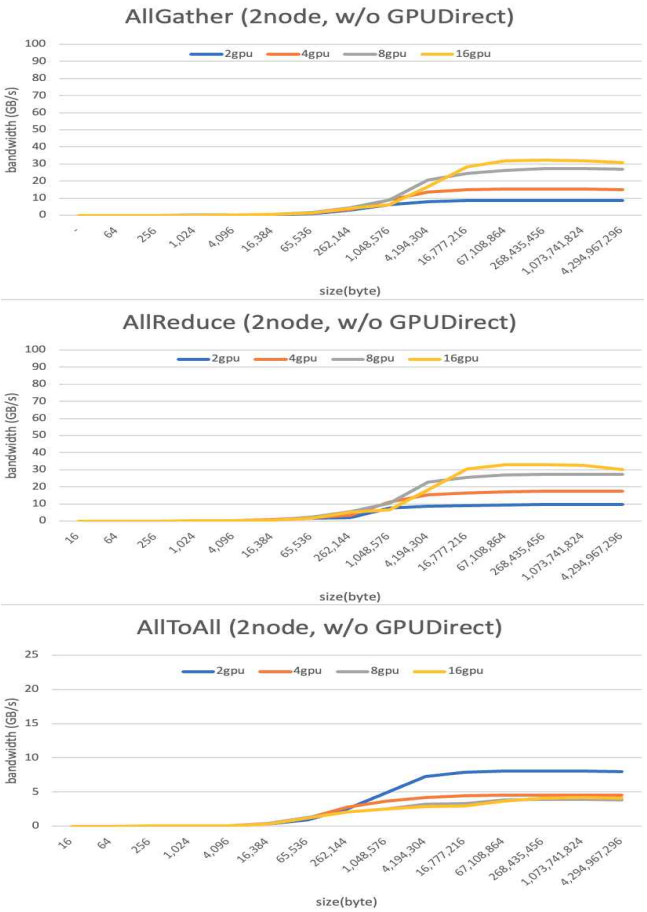
본 논문에서는 A100 8GPU를 장착한 노드를 이용하여 NCCL 벤치마크 코드의 성능을 분석하였다.

4. 뉴론 인터커넥트 네트워크 통신 성능 비교 분석

본 논문에서는 IB를 이용한 GPU 통신 시에 성능 향상을 위해 원격 GPU간에 직접 통신이 가능하게 해주는 GPUDirect RDMA 기능을 활성화, 비활성화 시키며 성능 변화를 분석하였다[4]. NVIDIA에서 HDR 8포트를 탑재한 DGX A100에서 테스트한 결과를 보면 집합통신과 P2P통신의 경우 대역폭이 각각 192GB/s, 24GB/s로 측정되었다[5]. 그림 3과 같이 GPUDirect RDMA 기능이 활성화 되어있는 경우, 2개의 노드에서 16개의 GPU 간의 AllGather, AllReduce 통신 테스트를 수행한 경우 대역폭이 90GB/s 까지 측정되었다. 본 논문에서 사용한 노드는 HDR 4포트를 사용하므로, NVIDIA에서 발표한 집합통신 대역폭(192GB/s)의 절반정도의 성능이 나온 것을 확인할 수 있었다. AllToAll의 경우는 P2P 통신이므로, NVIDIA에서 발표한 P2P통신 대역폭(24GB/s)에 근접한 20GB/s까지 측정되었다. 최상의 통신 성능을 갖는 시스템을 구성하기 위해서는 Fat-tree 토폴로지와 GPU:IB의 비율을 Non-blocking(1:1)으로 구성해야하는데, 이는 인터커넥트



(그림 3) NCCL 테스트 결과 (w/ GPUDirect)



(그림 4) NCCL 테스트 결과 (w/o GPUDirect)

네트워크 구성에 상당한 비용이 발생하게 된다. 추가적으로 GPU Direct RDMA를 비활성화했을 때는 2노드에서 16GPU를 사용하여 수행한 결과를 살펴보면, 활성화했을 때의 결과에 비해 AllGather과 AllReduce는 2.7배, AllToAll은 4.6배 정도의 성능하락이 있는 것을 확인할 수 있었다.

4. 결론 및 향후 연구 방향

본 논문에서는 뉴론의 A100 8GPU를 장착한 계산노드에서 NVIDIA NCCL 벤치마크 코드를 활용하여 인터커넥트 네트워크 통신 성능을 비교 분석하는 기법을 소개하였다. 멀티 GPU를 장착한 멀티 노드 병렬 통신에서 최대한의 성능을 얻기 위해서는 인터커넥트 네트워크 구성과 같은 하드웨어적인 구성과 동시에 GPUDirect RDMA와 같은 소프트웨어적인 구성 역시 중요함을 확인할 수 있었다.

향후에 구축될 KISTI의 슈퍼컴퓨터 6호기에서도 인터커넥트 네트워크 통신 성능을 극대화하기 위한 구성을 적용하고 다양한 벤치마크 성능 테스트를 수행하여 비교 분석하는 연구를 수행할 예정이다.

사 사

이 논문은 2023년도 한국과학기술정보연구원(KISTI)의 기본사업(과제번호:K-23-L02-C01-S01) 및 자체사업(과제번호:J-23-NB-C03-S01)으로 수행된 연구입니다.

참고문헌

[1] KISTI 국가슈퍼컴퓨팅센터 홈페이지, 뉴론소개, https://www.ksc.re.kr/byjw/neuron

[2] NVIDIA NCCL Manual, Collective Operations, https://docs.nvidia.com/deeplearning/nccl/user-guide/docs/usage/collectives.html

[3] GitHub, NVIDIA NCCL Tests, https://github.com/NVIDIA/nccl-tests

[4] NVIDIA 개발자 매뉴얼, GPUDirect RDMA, https://developer.nvidia.com/gpudirect

[5] NVIDIA On-Demand, Scaling Deep Learning Training: Fast Inter-GPU Communication with NCCL, https://www.nvidia.com/en-us/on-demand/session/gtcspring23-s51111/