Red Hat

실현가능한 미래

Samsung CMM-D를 활용한 Memory Disaggregation

Introduction

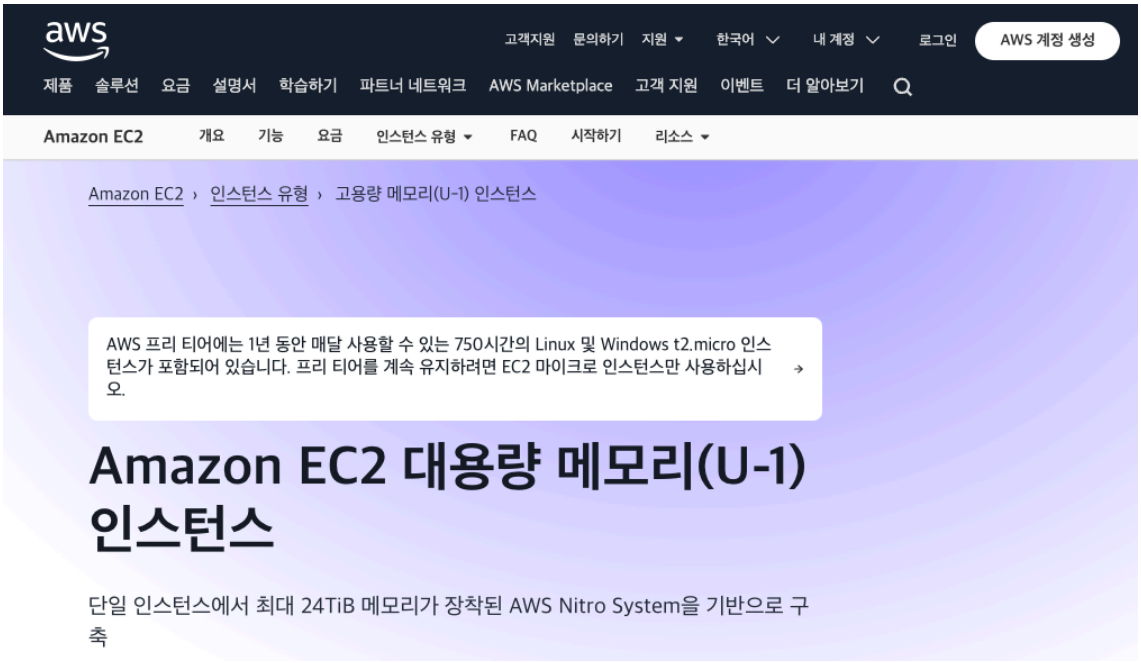
Disaggregated Memory 요구

기존의 컴퓨터 시스템은 프로세서 중심의 컴퓨팅 구조를 기반으로 작동해 왔습니다. 이 시스템에서는 데이터를 처리하는 프로세서가 중심이 되고, 메모리, 저장 장치, 네트워크가 이를 보조하는 역할을 합니다. 그러나 최근 빅데이터와 딥러닝의 부상으로 인해 데이터의 양이 급격히 증가하면서, 기존 구조에서 데이터 이동에 많은 시간이 소요되고 있습니다. 데이터를 프로세서와 저장 장치 간에 이동시키는 작업이 병목 현상을 일으켜 성능 저하가 발생하는 것이 문제입니다.

이러한 문제를 해결하기 위해 Disaggregated Memory 개념이 도입되었습니다. Disaggregated Memory 는 메모리 자원을 독립적으로 관리하고 확장하는 시스템으로, 기존의 프로세서 중심 컴퓨팅 구조에서 벗어나 메모리 중심의 컴퓨팅 시스템을 구현할 수 있습니다. 즉, 데이터를 메모리에 상주시키고, 프로세서가 아닌 메모리에서 직접 데이터를 처리하는 구조입니다. 이를 통해 데이터 이동 시간을 줄이고 처리 속도를 극대화할 수 있습니다.

특히 CXL (Compute Express Link) 같은 기술은 이러한 Disaggregated Memory 구조를 구현하는 데 중요한 역할을 합니다. 이 기술은 여러 프로세서와 메모리, 그리고 저장 장치를 동적으로 연결하고 메모리 풀링을 통해 자원을 공유할 수 있게 합니다. 빅데이터 분석, AI/딥러닝과 같은 대규모 데이터 처리 응용 프로그램에서 효율적으로 메모리 자원을 활용하여 처리 성능을 크게 향상시킬 수 있습니다.

결론적으로, Disaggregation 은 급격하게 증가하는 데이터 처리 요구를 충족시키고 확장성, 유연성, 그리고 성능을 제공하기 위한 필수적인 기술로, 특히 대용량 메모리 활용이 중요한 미래의 컴퓨팅 환경에서 더욱 필수적인 요소로 자리잡고 있습니다.



SAP, Memory Database 와 AI/ML과 같은 대용량의 메모리를 필요로 하는 응용을 Amazon EC2 U-1 인스턴스에 수용할 수 있는 메모리의 최대 용량이 24 TiB로 서비스를 제공함으로써 Disaggregated Memory 의 수요는 이미 증명된 상황입니다.

Disaggregated Memory 의 적용의 핵심 요소 **- Switch**

CXL 3.0 규격은 정의되었지만, 기술적 진전은 CXL 2.0에 머물러 있으며 시장에서 공식적으로 출시되어 사용가능한 CXL Switch는 아직 없는 상황입니다. 이러한 한계로 인해 CXL 기반 Remote Memory Disaggregation 구현에 기술적 어려움이 있습니다. CXL Switch의 부재로 인해 메모리 분산 기술을 대체하는 사용 가능한 솔루션으로 InfiniBand 와 RoCE v2 기술을 기반으로 삼성 CMM-D를 활용한 Remote Memory 아키텍쳐를 구현했습니다.

삼성전자는 세계 최초로 CXL Memory Expander (CMM-D)를 출시[1] 하였으며, Red Hat Enterprise Linux 와의 검증을 통해 CXL 생태계를 주도[2] 하고 있습니다.

두 기술 모두 CXL 생태계가 성숙하기 전까지 성능과 확장성을 고려한 최적의 옵션입니다. InfiniBand 는 특히 고성능 컴퓨팅 및 대규모 데이터 처리에 적합하며, RoCE v2 는 이더넷 인프라를 사용하는 환경에서 활용하기 쉬운 장점이 있습니다. 결과적으로, CXL Switch 가 상용화되기 전까지 삼성 CXL Memory Expander 와 함께 InfiniBand 와 RoCE v2 를 활용한 Remote Memory 솔루션은 현재로서 가장 현실적인 메모리 확장 및 자원 최적화 방안으로 자리 잡을 수 있습니다.

이러한 제안은 AI/ML, 빅데이터 분석, 고성능 컴퓨팅 등 대규모 메모리 요구가 있는 응용 프로그램에서 특히 유용할 것으로 예상되며, 차세대 데이터센터에서의 성능 및 유연성을 크게 향상시킬 수 있는 방안입니다.

On-Premise 환경에서의 Memory Scale-out 솔루션 - CMM-D + InfiniBand

Amazon EC2 U-1 인스턴스는 최대 24 TiB 의 메모리를 지원하여 대규모 데이터베이스와 메모리 집약적 워크로드를 실행할 수 있는 환경을 제공합니다. 이와 같은 메모리 확장은 클라우드에서 대규모 데이터를 처리하기에 적합하지만, On-Premise 환경에서는 이러한 메모리 용량을 구현하는 데 상당한 제약이 따릅니다. 특히, 고비용의 하드웨어 및 제한된 물리적 자원이 문제로 작용합니다.

이러한 제약을 해결하기 위한 Workaround 로 InfiniBand 를 통한 메모리 확장이 중요한 대안이 될 수 있습니다. InfiniBand 는 저지연 및 고대역폭 네트워크 인터페이스를 제공하여, 여러 노드에서 메모리를 연결하고 공유함으로써 단일 서버의 메모리 한계를 극복할 수 있습니다. 예를 들어, 메모리 풀링 또는 리모트 메모리와 같은 기술을 통해 On-Premise 환경에서도 대규모 메모리 수요를 충족할 수 있습니다. 이를 통해 대규모 데이터베이스가 필요로 하는 메모리를 분산 메모리 시스템으로 확장하고, 메모리 집약적 작업을 지원할 수 있습니다.

InfiniBand 를 활용한 이러한 확장은 대규모 데이터 분석 및 인메모리 데이터베이스와 같은 고성능 워크로드를 처리하는 데 있어 매우 유용합니다. 또한, 클라우드 대비 비용 효율적인 방식으로

---

[1] CXL 2.0 device 세계최초 출시 - Press Release, 삼성전자 CMM-D 소개
[2] 업계 최초 삼성 CMM-D 장치 Red Hat Enterprise Linux 인증 - Press Release

On-Premise 인프라를 유지하며 대규모 메모리 활용을 가능하게 합니다. 이는 기업이 클라우드 인프라와 On-Premise 시스템 간의 균형을 유지하면서도, 대규모 데이터를 처리할 수 있는 유연한 확장성을 확보할 수 있음을 의미합니다.

결론적으로, Amazon EC2 U-1 인스턴스에서 제공되는 메모리 확장의 장점을 On-Premise 환경에 적용하는 Workaround 로서 InfiniBand 는 중요한 역할을 하며, 대규모 데이터베이스 환경에서의 성능을 최대화할 수 있는 가능성을 제공합니다.

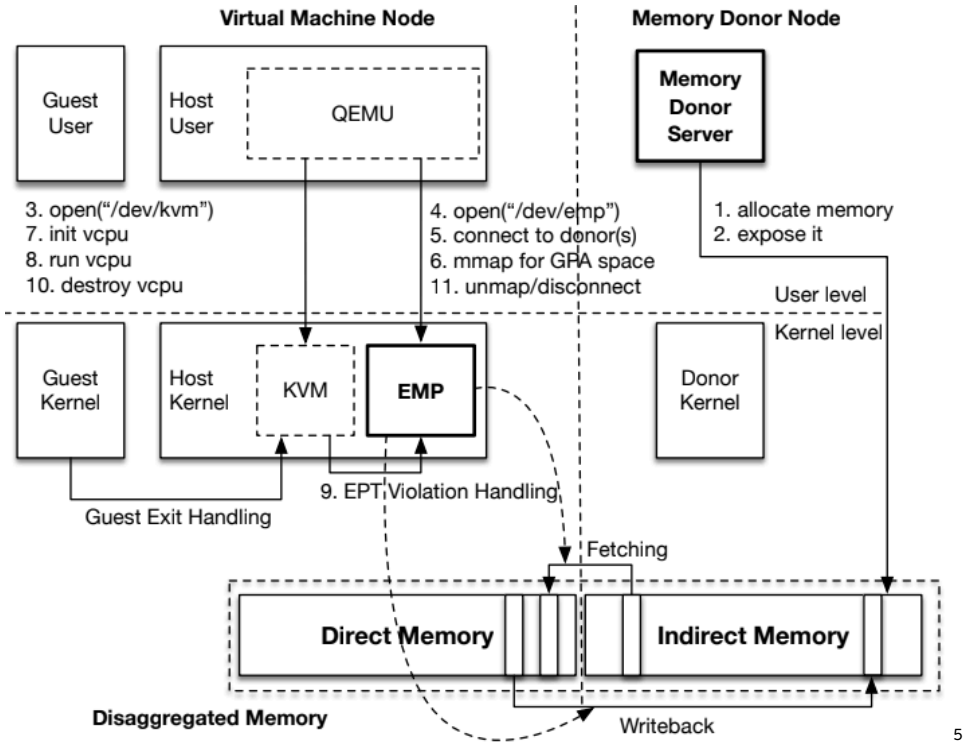
ETRI[3] EMP - Disaggregated Memory 를 가능하게 하다.

지금까지는 CXL 장치, 특히 CMM-D 를 Remote 호스트가 사용할 수 있게 하는 연결에 목표가 있었다면, Local Memory 와 Remote Memory로 계층화된 메모리를 효율적으로 사용할 수 있도록 하는 Memory 관리 방법이 필요합니다.

ETRI EMP[4] - Remote Memory 의 Enabler

기존 메모리 관리 시스템에서는 비활성 메모리 (Inactive Memory)가 DIMM 을 사용하게 되며, 이로 인해 신규 메모리 (Free Page) 요청시 비용 증가와 성능 저하가 발생합니다. EMP 는 이러한 문제를 해결하기 위해 고안된 기술로, 메모리 리소스를 효율적으로 관리하고 최적화합니다. ETRI EMP 의 기본 아이디어는 메모리를 계층화하고, 가장 비싸고 빠른 Local Host 의 System Memory 사용률 을 극대화 하는데 있습니다. 활성 메모리 (Active Memory)를 DIMM으로 적극 활용하고, 비활성 메모리 (Inactive Memory)를 평가하여 'Stranded Memory' 가 DIMM 에 상주하는 것을 최소화함으로써 메인 메모리의 활용을 극대화 할 수 있습니다.

[3] ETRI (Electronics and Telecommunications Research Institute) - 한국전자통신연구원
[4] EMP (Elastic Memory Platform)



[5]

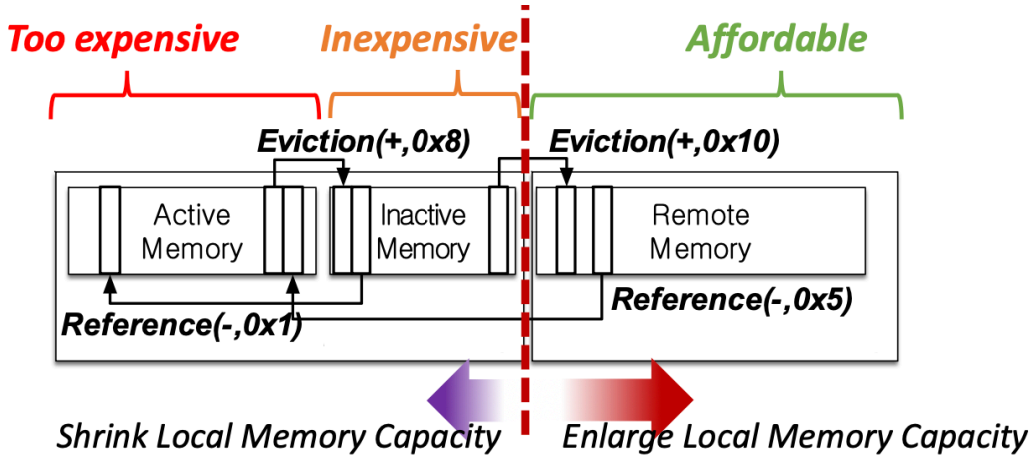
상기 다이어그램은 Disaggregated Memory 시스템에서 가상 머신 노드와 메모리 노드 간의 상호작용을 도식화하였습니다.

1. Memory Donor Node : Memory Donor 서버는 가상 머신이 사용할 수 있도록 메모리를 할당하고 노출하는 역할을 수행합니다. Donor에서 동작하는 유저레벨 프로세스를 통해 메모리를 관리하며, RDMA(InfiniBand/RoCE) 등의 프로토콜을 통해 원격으로 메모리를 제공합니다.
2. Virtual Machine Node
    a. QEMU 는 가상 머신을 실행을 관리하는 소프트웨어로, Host 유저 레벨에서 동작을 합니다. QEMU 는 KVM 을 통해 Local Memory 와 Remote Memory 를 가상 머신에 연결합니다.
    b. ETRI EMP는 메모리 페이징과 EPT(Extended Page Table) Violation 을 처리합니다.
    c. EMP Module 은 KVM 과 통합되어, 원격 메모리 노드와의 연결을 처리하고 가상 주소 공간을 매핑합니다. 이를 통해 원격 메모리의 페이지를 가져와 (Fetching) 메모리 관리 효율성을 높입니다.

[5] Koh, Kwangwon, et al. "Disaggregated cloud memory with elastic block management." IEEE Transactions on Computers 68.1 (2018): 39-52

3. Memory Interaction : 가상 머신이 메모리에 액세스할 때, 직접 메모리 (Direct Memory)와 원격 메모리(Indirect Memory) 간의 전환이 이루어지며, 필요시 Fetching 과 Writeback 이 수행됩니다.

ETRI EMP 는 Application 에 할당된 메모리를 지속적으로 자체 알고리즘으로 평가하고, 계층화 메모리의 값에 따라, 비활성메모리를 Remote Memory에 이동함으로써 활성메모리의 활용율을 극대화 합니다.



삼성 CMM-D 를 Remote Memory 로 활용하기 위한 주요 요소를 정리하면 아래와 같습니다.

- CXL Switch 를 대체하는 InfiniBand over IB 혹은 RoCE v2 (Ethernet)
- Remote Memory 를 공유가능하게 하는 ETRI EMP Donor 기능
- Remote Memory 를 효율적으로 사용하게하는 계층화 메모리 활용 솔루션인 ETRI EMP

Red Hat, 삼성전자, ETRI는 사용성 검증을 위해 삼성전자내 SMRC (Samsung Memory Research Center) Lab에서 공동으로 시험을 실시하였고 이에 대한 구성을 Reference Architecture 문서로 배포함으로써 일반사용자들이 삼성 CMM-D 를 활용한 Remote Memory - Scale-out 모델을 구성할 수 있는 기틀을 마련하였습니다.

STREAM[6] 으로 수행한 성능 시험 시, System Memory + Swap 을 기준으로 하였을 때, IPoIB 로 구성된 Remote Memory 시스템은 21 배, RoCE v2 로 구성한 Remote Memory 는 20 배의 성능 개선을 확인하였습니다.

(3. Key performance Indicator of the remote memory 성능 테스트 30% 결과 기준)

STREAM Performance

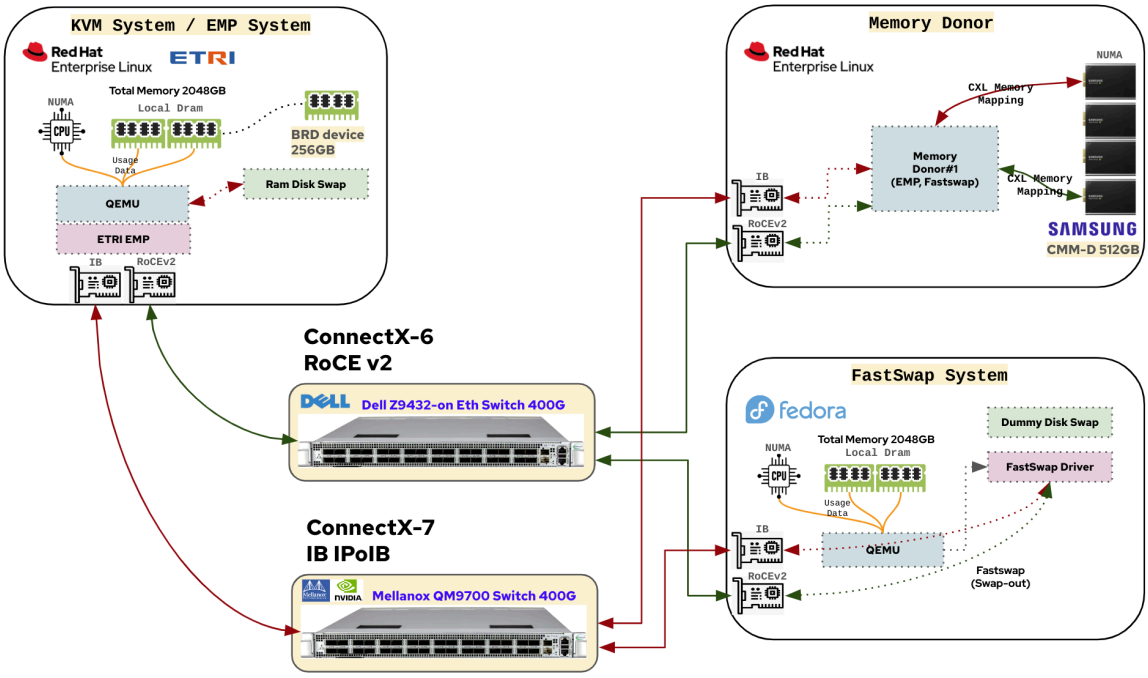
Scenario	System Memory + Swap	System Memory + Remote Memory (IB)	System Memory + Remote Memory (RoCE v2)
Result	**1**	**21**	**20**

[6] STREAM - Sustainable Memory Bandwidth in High Performance Computers

Performance

1. Environment (성능측정 환경)

RDMA (IPoIB, RoCE v2) 환경에서 EMP 원격 메모리를 활용한 성능측정을 위한 환경은 아래와 같습니다. 공정한 성능측정을 위해 동일한 하드웨어 환경에서 RHEL 9, Fedora 38 운영체제를 기반으로 ETRI EMP (Elastic Memory Platform) 에 대한 원격메모리 성능측정을 진행하였습니다.



KVM 을 사용한 메모리 성능을 비교하기 위해, Control Group v2 (cgroup2) 로 메모리 자원을 제한하여 Swap out 상황을 만들었으며, 최신 커널 기반의 ETRI EMP 분산 메모리 시스템과의 비교를 위해 Local Ram Disk를 Backend로 사용하는 이상적인 Swap 환경과 Remote RDMA FastSwap 을 대상으로 비교 검증을 진행하였습니다.

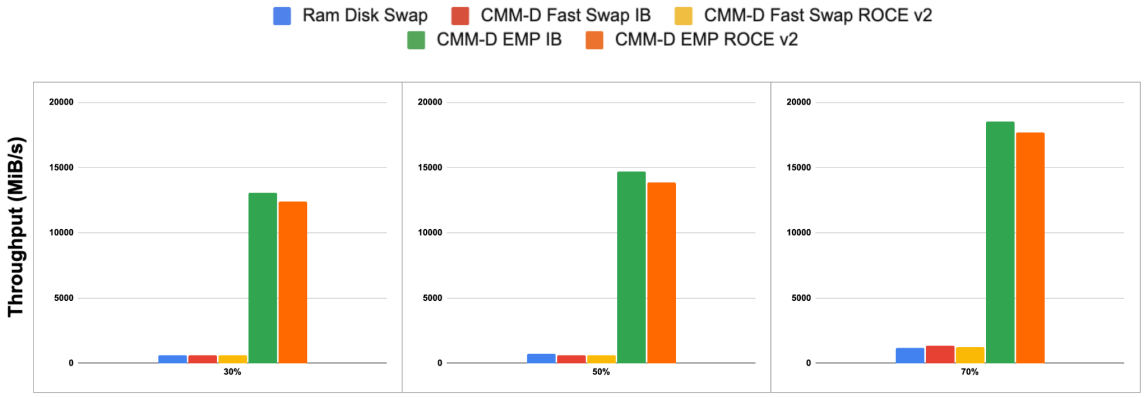
- Remote RDMA FastSwap 의 경우 별도의 커널 Build 가 필요함에 따라 Fedora 38 (Kernel 6.1) 환경에서 설치하였고, 이 외 나머지 구성은 동일하게 구성하였습니다.
- Local Ram Disk Swap 의 경우 4 KiB 블록 크기를 사용하며, 128 KiB 블록 크기를 사용하는 Remote RDMA EMP 와 성능을 비교합니다.

2. Environment (하드웨어 사양)

Specification	
Server Platform	Supermicro Super Server X13DSF-A
CPU	Intel(R) Xeon(R) Gold 6442Y CPU @ 2.6GHz (24C) * 2EA
Local Main Memory	Samsung DDR5 4800 MHz 64GB * 32EA (Total 2048GB)
RDMA HCA	Mellanox Technologies MT28908 Family [ConnectX-6]
RDMA HCA	Mellanox Technologies MT2910 Family [ConnectX-7]
IPoIB Switch	Mellanox QM9700 Switch 400G
RoCE Switch	Dell Z9432-on Eth Switch 400G
CXL Memory Expander	**Samsung CMM-D 128GB * 4EA (Total 512GB)**
Operating System	Red Hat Enterprise Linux 9.2
Operating System	Fedora38

3. Key performance Indicator of the remote memory (1)

Memory Usage					
Benchmark	Description	Memory Usage	Local Memory		
			30%	50%	70%
STREAM	vector multiply & add	112G	34G	56G	78G



Performance Result

위 STREAM 벤치마크는 KVM Ram Disk Swap, CMM-D 기반 FastSwap, CMM-D 기반 EMP 과 비교한 결과입니다. EMP 에 대한 검증은 128 KiB 블록 크기를 사용하며, 삼성 CMM-D 기반에서 수행되었습니다. 성능 결과는 cgroup2를 사용하여 메모리 자원을 제한한 KVM Ram Disk Swap의 성능을 기준으로 정규화 되었고, 각 애플리케이션의 작업 세트 크기의 30%, 50%, 70%의 로컬 메모리 크기를 설정하여 실험을 수행했습니다.

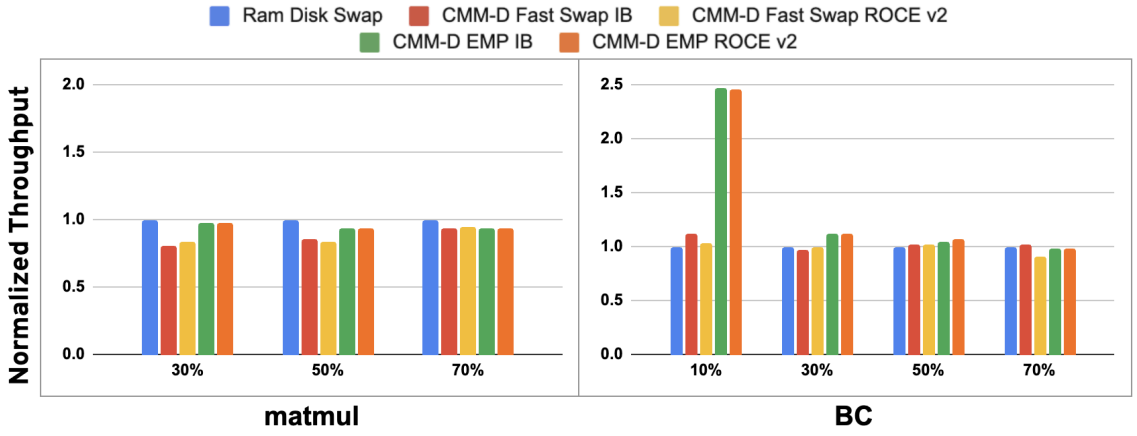
EMP 에 설정된 128K Block는 프리패칭 효과로 인해 KVM Ram Disk Swap, FastSwap 보다 우수한 성능을 보입니다. 이는 FastSwap 이 비동기 I/O 작업으로 성능을 향상시키고 EMP 128K 가 분산 메모리 시스템을 위한 별도의 LRU 체인을 사용하여 성능을 개선한 결과입니다. 또한 Mellanox ConnectX-6 100GbE 환경의 RoCE v2 와 Mellanox ConnectX-7 400GbE 환경의 IPoIB 와 비교시에도 비슷한 성능을 보여 줍니다.

Red Hat

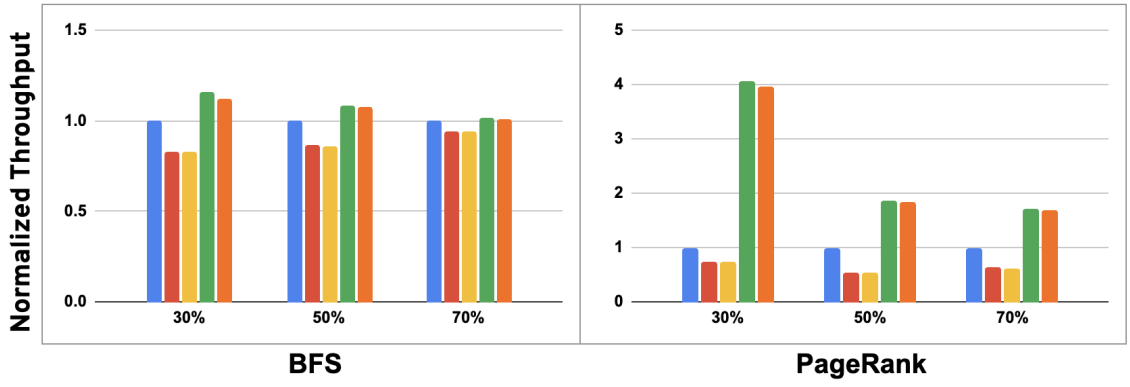
4. Key performance Indicator of the remote memory (2)

Memory Usage

Benchmark	Description	Memory Usage	Local Memory		
			30%	50%	70%
MatMul	matrix multiplication	40G	12G	20G	28G
BC	betweenness centrality	71G	21G	35G	50G
PageRank	ranking web pages	71G	21G	35G	50G
BFS	breadth-first search	26G	8G	13G	18G



Random memory access patterns - graph processing



Sequential memory access patterns - graph loading

Performance Result

IPoIB/RoCE v2 환경에서 구동되는 CMM-D 기반 EMP 는 무작위 메모리 접근 패턴을 가지는 데이터 분석의 경우 로컬 메모리 (DDR5 4800 MHz 64GB * 32EA) 에서 구성된 KVM Ram Disk Swap 과 비교 시 비슷하거나 좀 더 나은 성능을 보여줍니다. 특히, BC 의 경우 로컬 메모리 크기가 10% 정도 일때, 원격 메모리 접근이 많아져서 성능 변화 파악에 용이합니다. 또한, 데이터를 메모리에 적재하는 데이터 로딩 부분에서는 EMP 가 여러 페이지 (128KiB 블록 크기) 단위로 처리하기 때문에 성능을 향상시킬 수 있습니다.

Conclusion

데이터 집약적인 응용의 증가로 인해 Memory Disaggregation의 중요성이 커지고 있습니다. 이를 활용하기 위해 CXL(Compute Express Link) 기술이 도입되었지만, CXL Switch 미출시로 인해 완전한 Memory Disaggregation 구현이 어려운 상태입니다.

이에 삼성전자, ETRI, Red Hat 은 RDMA (Infiniband, RoCE v2) 기술을 활용하여 원격 메모리 솔루션을 제안, 기능 검증을 완료하였고 이에 대한 구성을 Reference Architecture 문서로 배포함으로써 일반사용자들이 삼성 CMM-D 를 활용한 Remote Memory-Scale-out 모델을 구성할 수 있는 기틀을 마련하였습니다.[7]

STREAM 으로 수행한 성능 시험 시, System Memory + Swap 을 기준으로 하였을 때, IPoIB 로 구성된 Remote Memory 시스템은 21배, RoCE v2 로 구성한 Remote Memory 는 20배의 성능 개선을 확인하였습니다.

STREAM Performance

Scenario	System Memory + Swap	System Memory + Remote Memory (IB)	System Memory + Remote Memory (RoCE v2)
Result (정규화)	1	21	20

KVM 기반의 가상화를 사용하는 경우, CMM-D & EMP 환경에서는 Swap-out 으로 인한 성능문제 발생 시 Remote 영역으로 효과적인 메모리 교환이 이루어지고, 일반적으로 사용할 수 있는 Ethernet 기반의 RoCE v2 프로토콜을 이용하여 기존 인프라 환경에서 메모리 분산구조를 가져갈 수 있습니다.

삼성전자와 Red Hat의 협력을 통해 Type 3 Memory Expander(CMM-D)를 활용한 Remote Memory Disaggregation 아키텍쳐를 활용가능한 수준으로 발전시켰으며, CXL 생태계가 성숙해지기 전까지 최적의 솔루션으로 평가됩니다. 향 후 CXL 2.0 기술 및 IMDB/AI 등 다양한 산업분야에서 삼성과 지속적 협업을 통해 CXL 생태계 확대에 기여하기를 기대합니다.

[7] Remote Memory Reference Architecture

Red Hat

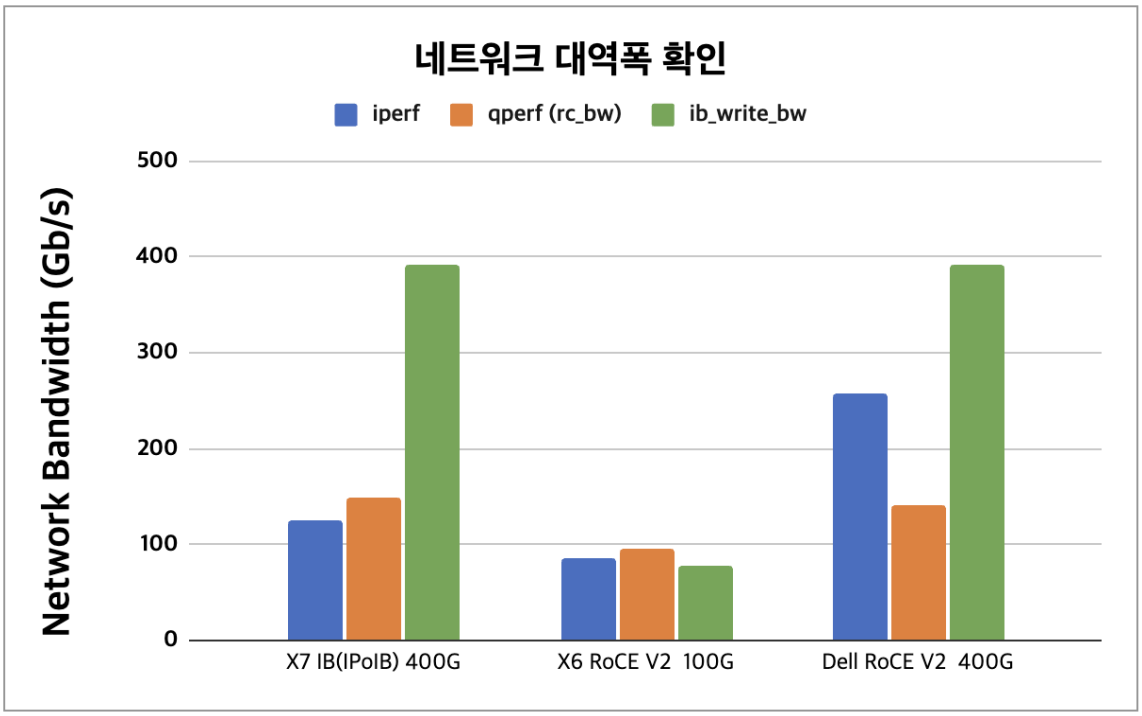
Appendix

InfiniBand (IB) 400G Bandwidth	
Model	**NVIDIA Mellanox QM9700 Switch**
Networking Type	InfiniBand Switch
Networking / Ports Qty	64 x NDR InfiniBand
Data Transfer Rate	OSFP cable or connector for 40/56/100/200/400 Gb/s
GBIC	NVIDIA MMA4Z00
Connected Adapter	**MT2910 Family [ConnectX-7]**

Ethernet (RoCE v2) 100G Bandwidth	
Model	**Dell Z9432-on Eth Switch**
Networking Type	Ethernet Switch
Networking / Ports Qty	Multi-rate 400GbE ports support 10/25/40/50/100GbE. 40GbE ports support 10/40GbE
Data Transfer Rate	100/400GbE
GBIC	Dell QSFP28 100GBASE-SR4-AOC
Connected Adapter	**MT28908 Family [ConnectX-6]**

Ethernet (RoCE v2) 400G Bandwidth	
Model	**Dell PowerSwitch Z9864F-ON (800G) Switch**
Networking Type	Ethernet Switch
Networking / Ports Qty	Multi-rate 100/200/400/800GbE Max 64 Ports 800GbE Max 128 ports 400GbE via Breakout
Data Transfer Rate	100/200/400/800GbE
Connected Adapter	**BCM57608 NetXtreme-E (400G)**

Ethernet 기반 RoCE v2 (RDMA) 환경 구성 시 **Broadcom 400G BCM57608 Adapter** 와 **Dell Z9864F-ON 800G** 스위치를 이용하면, Remote Memory 와 같은 미션 크리티컬한 워크로드에서 최고 수준의 네트워킹 성능과 안정성을 발휘 할 수 있습니다. 아래는 고성능 Broadcom BCM57608 과 Dell Z9864F-ON 스위치에 대한 네트워크 성능 측정 결과입니다.



최신 네트워킹 기술의 발전을 반영하는 **Broadcom 400G BCM57608** 어댑터와 **Dell Z9864F-ON 800G** 스위치 기반의 RoCE v2 구성은 RDMA 벤치마크 검증에서 뛰어난 성능을 보여주고 있습니다. 이러한 첨단 기술의 조합은 기존 백서에서 보고된 성능 측정 환경과 비교하여 상당한 성능 향상을 제공할 것으로 예상되며, 특히 AI/ML, 클라우드 컴퓨팅, 고성능 컴퓨팅 및 대규모 데이터 센터와 같은 고성능을 요구하는 환경에서 더욱 뛰어난 효과를 발휘할 것으로 기대 됩니다. 이는 데이터 처리 속도 향상, 리소스 관리 효율성 증대, 복잡한 계산 처리 능력 개선 등을 통해 현대의 데이터 중심 애플리케이션의 성능을 크게 향상시킬 수 있는 잠재력을 보여주는 것으로, 향후 네트워킹 기술의 발전 방향을 제시하는 중요한 지표가 될 것 입니다.