ISSN. 2465-8456

2020 March | Vol. 6

03

융합연구리뷰

Convergence Research Review

오명훈 (한국전자통신연구원 데이터중심컴퓨팅시스템 연구실 책임연구원)
김홍연 (한국전자통신연구원 데이터중심컴퓨팅시스템 연구실 책임연구원)
고광원 (한국전자통신연구원 데이터중심컴퓨팅시스템 연구실 책임연구원)

메모리 중심 컴퓨팅 기술 동향

한상욱 (한국과학기술연구원 양자정보연구단 단장)
조영욱 (한국과학기술연구원 양자정보연구단 선임연구원)
임향택 (한국과학기술연구원 양자정보연구단 선임연구원)

양자통신 및 양자컴퓨팅 분야 소개 및 연구동향

융합연구정책센터
Convergence Research Policy Center

CONTENTS

융합연구정책센터
Convergence Research Policy Center

융합연구리뷰 | Convergence Research Review
2020 March vol.6 no.3

발행일 2020년 3월 9일

발행인 김주선

편집인 최수영·권영만

발행처 한국과학기술연구원 융합연구정책센터
02792 서울특별시 성북구 화랑로 14길 5
Tel. 02-958-4980 | http://crpc.kist.re.kr

펴낸곳 주식회사 동진문화사 Tel. 02-2269-4783

메모리 중심 컴퓨팅 기술 동향

다양한 매체를 통하여 매시간 쏟아지는 데이터의 양은 폭발적으로 늘어나고 있다. IDC의 자료에 의하면 2025년 생성 데이터는 163제타바이트(ZB=10^{21}B)에 이르며, 이는 2016년에 비해 처리할 데이터의 규모가 10배 수준으로 증가함을 의미한다. 데이터가 증가함에 따라 이를 분석하고 처리하는 컴퓨팅의 고속처리 능력도 점점 요구되는 상황이다. 하지만 현재의 컴퓨터 구조로는 '데이터 병목 현상'을 극복하는 데 한계가 발생한다.

이에, 본 호 1부에서는 현재 CPU와 저장장치를 사용하며 CPU 내 트랜지스터의 집적도 증가 및 고성능 CPU 아키텍처 적용을 통한 컴퓨팅 능력을 향상하는 소위 "프로세서 중심 컴퓨팅(Processor Centric Computing)"의 한계를 극복할 수 있는 "데이터 중심 컴퓨팅(Data Centric Computing)"에 대해 알아보았다. 데이터 중심 컴퓨팅은 데이터 이동을 최소화하기 위해 데이터가 위치하는 곳에서 데이터를 처리하는 컴퓨팅 모델로 정의되며, 이를 구현하기 위한 핵심 기술은 메모리 중심 컴퓨팅이라고 할 수 있다.

본 호 1부를 통해 메모리 중심 컴퓨팅의 하드웨어(HW), 운영체제(OS) 및 활용 응용프로그램의 기술 동향을 알아보았다. 빅데이터, AI(인공지능), IoT(사물인터넷) 등의 지능화 기술발전에 따른 거대 데이터 처리 요구가 증가함에 따라 병목 현상을 줄여 연산 성능을 높이는 메모리 중심 컴퓨팅이 새로운 차세대 컴퓨팅 패러다임을 제시할 것으로 예상되며, 과거에는 상상하지 못했던 데이터 처리 속도의 혁신을 통해 신기술 및 신시장이 개척되기를 기대해 본다.

양자통신 및 양자컴퓨팅 분야 소개 및 연구동향

인공지능, 기계학습, 초연결 등과 같은 새로운 과학기술을 기반으로 이루어지는 4차 산업혁명시대는 강력한 연산 능력을 필요로 하며, 이러한 연산 능력을 제공하기 위해서는 차세대 통신 및 컴퓨팅이 필수적이다. 그 중, 양자통신과 양자컴퓨팅은 빠른 연산 능력과 강력한 보안성을 바탕으로, 기존의 정보통신기술을 새로운 단계로 이끌어 줄 것으로 기대된다.

이에, 본 호 2부에서는 양자정보통신기술의 개념에 대해 간략하게 서술하고, 이 중에서 양자통신과 양자컴퓨팅에 대한 국내외 기술개발 및 산업계 동향, 향후 기술의 발전에 대해 조망해 보았다. 양자정보통신기술은 기존의 고전정보통신기술에 양자역학적인 원리가 합쳐진 새로운 기술로 양자역학적 특성을 정보통신기술에 적용하기 위하여 양자상태를 생성, 제어, 측정 및 분석하는 기술이다. 양자정보통신기술은 20세기 말, 기초연구가 시작되었고, 21세기에 접어들면서 기초연구 뿐만 아니라 실용적 적용방안 및 가능성을 보여 주는 연구결과들이 나오기 시작했다.

본 호 2부를 통해 양자상태를 활용하여 정보통신의 보안성을 강화하고 양자기기 간의 네트워크를 구성하여 양자기기 간 통신을 지원하는 기술인 양자통신과 기존의 컴퓨터와는 달리 0과 1의 비율과 0과 1 사이의 위상 차이 등을 통해 원리적으로는 무한대의 조합으로 정보 표현이 가능한 양자컴퓨팅에 대해 알아보았다. 아직 상용화보다는 연구단계에 근접한 양자정보통신기술이지만 앞으로의 활용성이 기대되는 기술인 만큼 정부 주도적인 장기투자를 통해 글로벌 경쟁력을 갖출 수 있는 기반이 만들어질 수 있기를 기대해 본다.

융합연구리뷰

Convergence Research Review 2020 March vol.6 no.3

01

메모리 중심 컴퓨팅 기술 동향

오명훈(한국전자통신연구원 데이터중심컴퓨팅시스템 연구실 책임연구원)
김홍연(한국전자통신연구원 데이터중심컴퓨팅시스템 연구실 책임연구원)
고광원(한국전자통신연구원 데이터중심컴퓨팅시스템 연구실 책임연구원)
진기성(한국전자통신연구원 데이터중심컴퓨팅시스템 연구실 책임연구원)
안백송(한국전자통신연구원 데이터중심컴퓨팅시스템 연구실 선임연구원)
김창대(한국전자통신연구원 데이터중심컴퓨팅시스템 연구실 선임연구원)
김강호(한국전자통신연구원 데이터중심컴퓨팅시스템 연구실 책임연구원/실장)
김영균(한국전자통신연구원 초성능컴퓨팅연구본부 책임연구원/본부장)

Ⅰ 서론

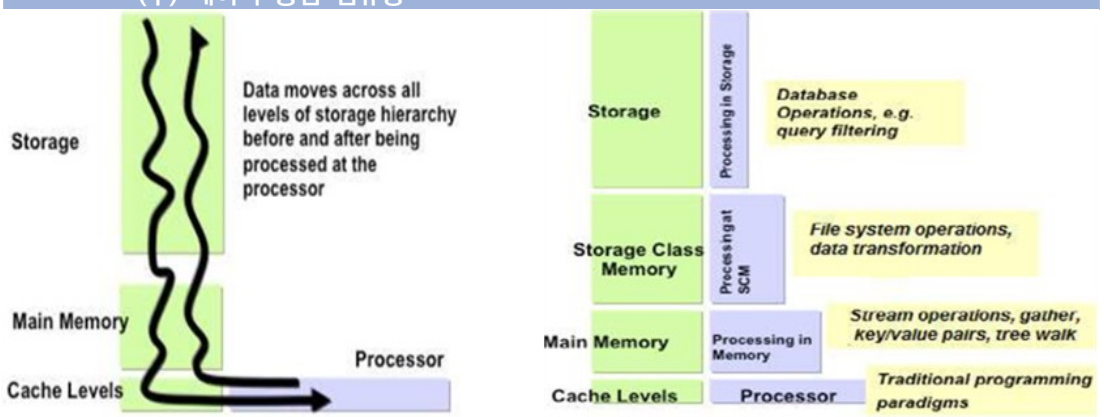
최근까지 빅데이터, AI(인공지능), IoT(사물인터넷) 등의 연관 기술의 발달에 따라, 처리할 데이터가 폭증하고 있다. IDC의 자료에 의하면 2025년 생성 데이터는 163제타바이트(ZB=10^{21}B)로 그중 5.2제타바이트는 실제 데이터 처리 및 분석되어야 하는 것으로 예측된다. 이는 2016년에 비해 처리할 데이터의 규모가 10배 수준으로 증가함을 의미한다(Data Age 2025). 아울러, 자율주행 및 원격 의료 진단과 같은 Life-critical 데이터 처리도 2025년까지 전체 생성 데이터의 20%까지 증가하여 대용량 데이터의 고속처리 또한 요구된다.

현재 컴퓨터는 CPU와 저장장치를 사용하여 컴퓨터를 구동하는 폰 노이만 구조를 사용하고 있으며, CPU 내 트랜지스터의 집적도 증가 및 고성능 CPU 아키텍처 적용을 통한 컴퓨팅 능력을 향상하는 소위 "프로세서 중심 컴퓨팅(Processor Centric Computing)"이 주류를 이루고 있다. 프로세서 중심 컴퓨팅에서 데이터는 저장장치에 저장되고, CPU는 저장장치로부터의 데이터를 순차적으로 처리하고, 저장장치의 계층이 존재함을 가정한다.

이러한 프로세서 중심 컴퓨팅으로 앞서 언급한 데이터 폭증 및 고속처리 요구사항을 만족시키기 쉽지 않다는 의견이 지배적이다. 프로세서 중심 컴퓨팅에서는 병렬처리를 위해서 다수 프로세서의 저장장치 간, 그리고 각 저장장치 계층 간의 데이터 이동이 필수적이다. 그러므로, 대용량 데이터 처리 시에는 이러한 데이터 이동 문제가 더욱 심각해져, CPU의 처리속도가 아닌 데이터 이동속도가 컴퓨팅 성능 및 에너지 소비에 영향을 미치는 데이터 병목 현상(data bottleneck)을 초래한다. 실제로, 순수한 데이터 처리가 아닌 데이터 이동에 필요한 오버헤드가 전체 시스템의 63%를 차지하는 사례도 존재한다(Amirali Boroumand, et. al, 2018).

이에, 프로세서 중심 컴퓨팅의 데이터 병목 현상을 해결할 수 있는 대안으로 "데이터 중심 컴퓨팅(Data Centric Computing)" 개념이 대두되었다. "Data Centric Computing"의 자료("Data Centric Computing", 2011)를 참고하면, 〈그림 1〉과 같이 데이터 중심 컴퓨팅은 데이터 이동을 최소화하기 위해 데이터가 위치하는 곳에서 데이터를 처리하는 컴퓨팅 모델로 정의된다. 특징 요소로는 1) 데이터가 존재하는 곳에서 프로세싱, 2) 데이터 접근 시 낮은 지연시간 및 에너지 소모, 3) 상대적으로 적은 크기의 데이터 저장 및 처리, 4) 효과적인 데이터 관리로 요약할 수 있다(Onur Mutlu, 2019). 데이터 중심 컴퓨팅 개념 자체는 새로운 것이 아니며, 의미상 NDP(Near Data Processing)와 동일하다(Rajeev Balasubramonian et. al, 2014).

그림 1. 프로세서 중심 컴퓨팅과 데이터 중심 컴퓨팅 개념 비교 – (좌) 프로세서 중심 컴퓨팅, (우) 데이터 중심 컴퓨팅

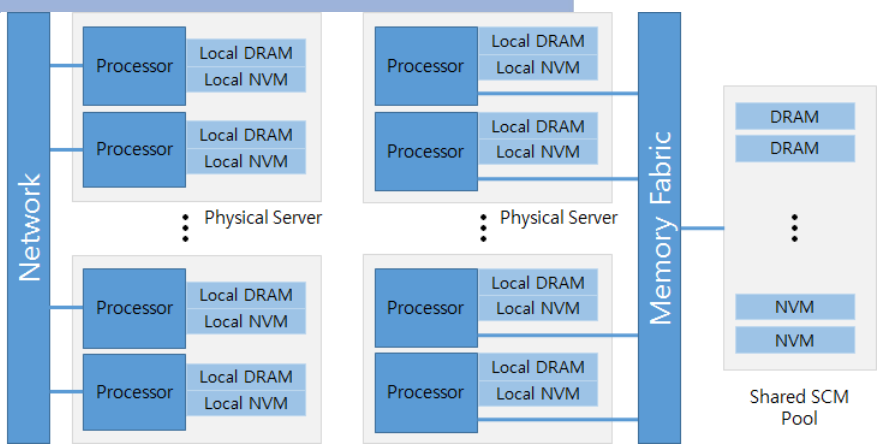


출처: Yoonho Park, 2017

데이터 중심 컴퓨팅의 핵심 기술은 메모리 중심 컴퓨팅이라고 말할 수 있다. 메모리 중심 컴퓨팅은 데이터 중심 컴퓨팅 개념을 계승하여 데이터 이동을 최소화하기 위해, 가능한 메모리에 가까운 곳에서 데이터를 처리하되, 바이트 접근이 가능한 비휘발성의 대용량 메모리를 공유하는 컴퓨팅 모델이다. 용어 자체를 Near-Memory Computing이나 In-Memory Computing을 포괄하는 개념으로 사용하는 경우도 있다 (semiengineering.com). 〈그림 2〉는 기존 프로세서 중심 컴퓨팅 구조와 SCM(Storage Class Memory) 메모리 풀을 사용한 메모리 중심 컴퓨팅의 개념도를 비교해서 나타내고 있다. 기존 컴퓨팅은 각각의 서버에 로컬 메모리를 가지며, 다른 서버의 메모리에 접근하기 위해서는 반드시 네트워크를 통해야만 한다. 그러나, 메모리 중심 컴퓨팅 구조에서는 로컬 메모리는 그대로 사용 가능하며, 대규모 SCM 메모리 풀을 메모리 연결망을 통해 다수의 서버가 공유할 수 있다. 이를 통해 기존 CPU들의 비효율적인 메모리 접근 문제가 해결되며, 스케일아웃(scale-out)이 쉬운 시스템을 구성할 수 있다.

즉, CPU, GPU, FPGA 등 다양한 컴퓨팅 컴포넌트가 연결될 수 있으며, 이들은 메모리 시멘틱 접근, 즉, 읽기/쓰기(load/store) 명령을 통해 대용량 메모리 풀에 접근한다. 이를 통해 대용량 데이터에 대한 인메모리 컴퓨팅이 가능해지고, 더 많은 연산 장치들이 병렬적으로 연산을 수행할 수 있다. 아울러, 대용량 메모리를 적극적으로 활용하면, 연산량 자체를 최적화하여 대규모 데이터 처리를 가속할 수 있는 장점이 있다(김창대 외 6, 2019).

그림 2. 프로세서 중심 컴퓨팅 vs. 메모리 중심 컴퓨팅



메모리 중심 컴퓨팅을 위해서는 크게 다음과 같은 세 가지 기술이 필요하다. 메모리 디바이스 및 메모리 인터커넥트를 포함하는 하드웨어 기술, 대용량 메모리 및 공유 메모리 지원을 위한 OS(운영체제)기술, 메모리 중심 컴퓨팅을 활용한 응용 소프트웨어 기술이다.

본 고에서는 메모리 중심 컴퓨팅을 위한 하드웨어 기술, OS 기술, 응용 소프트웨어 기술의 동향을 살펴보고, 향후 메모리 중심 컴퓨팅 개발의 방향성을 제시한다.

II 메모리 중심 컴퓨팅 하드웨어 기술 동향

1. 대용량 메모리 구성을 위한 메모리 디바이스 기술 동향

메모리 시스템은 현대 컴퓨터 시스템의 성능 및 전력 소모를 결정하는 중요한 부분 중 하나이다. 특히 메모리에서 데이터를 가져오는 속도가 데이터를 계산 및 처리하는 속도보다 느려서 발생하는 메모리 병목 현상(W.A. Wulf, S.A. McKee, 1995)은 잘 알려진 문제이다. 이는 멀티코어, 가속기(accelerator) 등의 계산 성능이 지속해서 발전했지만, 메모리 대역폭은 많이 늘어나지 못한 것이 이유라고 할 수 있다. 또한, 빅데이터로 대표되는 데이터 집약적인 애플리케이션이 점점 더 늘어나고 대용량 데이터를 빠르게 가져올 수 있는 시스템이 필요해지면서 메모리 병목 문제는 더 심각해지고 있다. 아울러, DRAM으로 대표되는 휘발성 데이터의 특성 대신, 비휘발성 데이터의 영속성(persistent)을 극대화(예, 재부팅 속도, 대용량 데이터 로딩속도)할 필요성이 있다.

〈표 1〉은 메모리 소자의 특성을 나타내고 있다. 메모리 소자의 세 가지 주요 기능요소는 집적도, 비휘발성, 속도라고 할 수 있다. DRAM와 NAND 플래쉬(flash)를 비교해 보면, DRAM은 속도는 빠르지만 집적도가 낮고 비휘발성 특성을 갖지 못했으며, NAND 플래쉬는 비휘발성 특성을 보이며 집적도는 매우 우수하지만, 지연시간, 대역폭(bandwidth) 측면에서 속도는 열등하다.

표 1. 메모리 소자 특성 비교

Memory Types	Read Latency (ns)	Write Latency (ns)	Bandwidth (GB/s)	Dynamic Power	Leak Power	Density	Addressability
SRAM, Cache (L1, L2, L3)	2~8	2~8		Low	High	10s MB	byte
DRAM	50~200	50~200	25	Medium	Medium	10s GB	Block/word
Stacked DRAM (HMC, HBM)	40~90	40~90	400	Low	Medium	10s GB	Block or Page
NAND NVRAM	100μs	2~3ms	1GB/s for read, 10MB/s for write	Low for read, High for write	Low	100s GB	Block or Page
3DXPoint	2~3x slower than DRAM	4~6x slower than DRAM				8~10s x of DRAM	

출처: Yonghong Yan, Ron Brightwell and Xian-He Sun, "Principles of Memory-Centric Programming for High Performance Computing," MCHPC'17, November 12-17, 2017.

메모리 병목 현상을 해결하기 위해, 주로 DRAM의 대역폭을 개선하기 위한 기술로, 적층(stacked) DRAM 기술이 연구되었다. 적층 DRAM 기술은 DDR4 또는 GDDR5보다 더 작은 폼팩터를 통해 최대 8개의 DRAM 다이(die)를 적층함으로써 저전력으로 고대역폭을 달성한다. 마이크론을 중심으로 한 HMC(Hybrid Memory Cube)(J. Thomas Pawlowski, 2011)와 삼성과 SK하이닉스를 중심으로 한 HBM(High Bandwidth Memory)(J.C. Lee, et al., 2016)(J. H. Cho, et al., 2018)이 대표적인 사례이다.

데이터 영속성을 유지하면서 저장공간을 확장함으로써 메모리 병목 현상을 해결하는 기술로 2013년도부터 NVDIMM(Non-Volatile Dual In-line Memory Module)이 연구되었다. NVDIMM은 기존의 DRAM 기술과 NAND 플래쉬 기술을 융합하여 기존 DRAM의 폼펙터인 DIMM(Dual In-line Memory Module)을 통해 CPU와 연결하는 기술로, 구성에 따라 메모리 기반의 NVDIMM-N, 스토리지 기반의 NVDIMM-F, 메모리와 스토리지의 하이브리드 형태의 NVDIMM-P 타입으로 나눌 수 있다(Open Server Summit, 2016).

NVDIMM-N은 같은 용량의 DRAM과 비휘발성 메모리를 적재한 DIMM이다. DRAM은 시스템 메모리이며, 비휘발성 메모리는 DRAM의 백업 메모리 역할을 수행한다. 따라서, CPU는 NVDIMM-N 내의 DRAM만 접근할 수 있으며, DRAM DIMM과 동일한 기능을 수행한다. NVDIMM-F는 비휘발성 메모리만을 탑재하여, CPU는 이를 스토리지로 취급하므로, DRAM DIMM에서는 불가능한 큰 저장 용량을 확보할 수 있지만, DRAM DIMM보다 액세스 지연시간이 길다. DIMM 폼펙터용 스토리지 장치로 볼 수 있다. NVDIMM-P는 DRAM과 DRAM보다 훨씬 큰 용량의 비휘발성 메모리를 함께 탑재한 형태로, NVDIMM-N 타입과 NVDIMM-F 타입의 두 가지의 동작 모드로 구동된다. NVDIMM-F 타입의 모드로 동작할 때는 DRAM은 저장용 버퍼 메모리로 이용될 수 있다.

NVDIMM 규격과는 별개로 인텔에서는 2019년 2월에 DDR4 DIMM 소켓 폼펙터를 기반으로 비휘발성 메모리를 지원하는 Optane DCPMM(DC Persistent Memory Module)을 출시하였다. 이는 인텔이 독자 개발한 PCM(Phase Change Memory) 기반 3DXpoint(3D 크로스포인트) 메모리 기술이 적용되었고, 일반 DRAM보다는 느리지만, 가격 측면에서 유리하다. 현재 모듈당 128GB~512GB의 용량을 제공할 수 있으며, 최대 읽기, 쓰기의 bandwidth는 256바이트 기준 각각 6.8GB/s, 2.3GB/s 수준이다(Product brief of Intel Optane DC Persistent Memory, Intel). Optane DCPMM에서는 NVDIMM-P의 기능과 유사하게 일반 DRAM을 DCPMM의 캐시 형태로 구동하는 메모리 모드와 DCPMM을 스토리지로 구동하는 AppDirect 모드를 지원하며, 이들을 혼용하여 사용할 수도 있다. 별도의 프로토콜(DDR-T)을 지원하는 Xeon 캐스케이드 레이크(Cascade Lake) 이상급의 프로세서가 장착된 서버에서만 구동할 수 있다.

컴퓨팅 시스템을 고려할 때 저장장치의 계층 구조에서 DRAM과 플래쉬 메모리 사이의 성능 차이가 제일 크고, DRAM의 휘발성 특성으로 인한 시스템 전체의 비효율성이 점차 증가하여 이 두 계층 사이 영역인 SCM 연구가 필요하다. 인텔 3DXpoint가 차세대 비휘발성 메모리인 PCM을 기반으로 개발된 사례와 같이, STT-MRAM(Spin-Transfer Torque Magnetoresistive RAM), RRAM(Resistive RAM) 연구도 활발히 진행 중이다. 〈표 2〉는 차세대 비휘발성 메모리인 SCM을 DRAM과 NAND 플래쉬 메모리와 비교 요약하고 있다.

표 2. SCM 기술 비교

	STT-MRAM	PCM	RRAM	DRAM	NAND Flash
Non-Volatile	Yes	Yes	Yes	No	Yes
저장원리	전자의 스핀 방향 일치에 따른 전류흐름차이	결정과 비결정의 상변화 차이	부도체의 저장 변화에 의한 전류 흐름차이	커패시터 전하 유무 (1, 0) 구분	막층에 전하 유무로 데이터 구분
Power consumption (Ewrite pJ /Bit Iwrite uA)	Medium/low 1/50	Medium 18/100	Low 1/1	Low 〈1pJ	Very high 100pJ
Write/Read Lat. (ns)	High(5/10)	Medium(150/80)	High(50/〈10)	High(5/20-80)	Low ()100,000/ 15,000-50,000)
Program Window	Good	3bit/cell	Variable	Good	4bit/cell
Endurance (Cycles)	Unlimited	Medium (10^8 -10^9)	Low (10^5 -10^{10})	Unlimited	Low (10^5 -10^6)
Retention	Good	R-drift	RTN	64ms	Good
Cell size (cell size in F 2)	Medium 12	Small 4	Medium 4-6	Small 7	Very small 〈4
2014 price ($/Gb)	High ($100-$50/Gb)	Medium (few $/Gb)	Very High ($5,000/Gb)	Low ($1/Gb)	Very low ($0.05/Gb)
Interface	DDR3/4/5	Proprietary	Flash-like	DDR3/4/5	

출처: Emerging Memory Device Technologies," The 49th Annual IEEE/ACM Intl. Symp. on Microarchitecture Tutorial, 2016.10 with modification

RRAM과 PCM은 DRAM보다 더 큰 셀 사이즈를 제공하지만, DRAM보다 훨씬 더 느리고 더 많은 전력을 소비한다. STT-MRAM은 성능 측면에서, DRAM과 비슷한 읽기 및 쓰기 성능을 나타내며, RRAM 및 PCM과 비교해서도 성능 및 내구성 측면에서 훨씬 우수하지만, 구현상 셀 사이즈 측면의 약점 때문에 가격이 증가한다. 따라서, 상대적으로 가격 측면에서 우수한 PCM은 실제 제품으로 상용화되었고, RRAM의 경우는 아직 상용화 사례가 존재하지 않는다. 아울러, STT-MRAM의 경우는 Everspin(https://www.everspin.com/), Avalanche technology(http://www.avalanche-technology.com/) 등의 업체에서 SRAM 대용으로 MRAM을 적용한 제품을 출시하였으며, 영속 메모리 제공을 위해 2021년까지 64Gb급 서버용 DDR4/5 기반 메모리 모듈을 개발할 예정이다.

이렇듯 개별적으로 발전된 메모리 기술들은 대부분 기존의 DRAM 인터페이스인 DDR 기반 DIMM 폼펙터에 기반하고 있다. 그러나, DIMM 폼펙터 기반으로, 메모리 병목 현상을 해결하기 위한 메모리 중심 컴퓨팅의 고대역폭 메모리 풀을 구성하는 데에는 한계가 존재한다. 메모리 중심 컴퓨팅의 메모리 인터커넥트로는 기존의

I/O 인터페이스를 사용하고, CPU 내에 있는 메모리 컨트롤러를 CPU 밖으로 분리하여 별도의 메모리 풀을 구축하는 형태가 바람직하다. 이러한 접근법은 메모리 컨트롤러 기술에서 CPU의 의존성을 제거할 수 있다는 측면에서, 현재 발전하고 있는 다양한 메모리 소자 기술(특히 SCM)들을 적시에 적용할 수 있는 이점도 존재한다. 물론, I/O 인터페이스 위에서 구동되는 프로토콜은 메모리 시멘틱을 지원해야 하며, 저지연 특성도 필요하다.

이외에, 하드웨어 관점에서는 메모리 풀, 메모리 풀 제어기, 메모리 연결망 등이 연구되어야 하고, 소프트웨어 관점에서는 메모리 시맨틱 인터페이스, 공유 메모리 풀 구성 및 관리, 공유 메모리 프로그래밍 도구 등에 관한 연구가 필요하다.

2. 메모리 중심 컴퓨팅을 위한 메모리 인터커넥트 기술 동향

기존의 I/O, 네트워크 프로토콜 대신, 차세대 연결망 프로토콜 개발을 위한 산업계 표준 수요가 증가하여, 2016년에는 Gen-Z(https://genzconsortium.org/), CCIX(https://www.ccixconsortium.com/), Open CAPI(https://opencapi.org/), 2019년에는 인텔을 중심으로 한 CXL(https://www.computeexpresslink.org/) 컨소시엄이 발족되었다.

2016년 10월 HPE, Dell EMC, AMD, ARM, 삼성, SK하이닉스, XILINX 등 세계 유수의 서버, 반도체, 메모리 벤더가 주축을 이루는 Gen-Z 컨소시엄이 구성되었으며 현재 총 50여 개의 업체가 멤버로 활동 중이다. Gen-Z 컨소시엄은 2018년에 Gen-Z 프로토콜의 기능 및 규격을 정의한 코어 규격 문서 1.0을 제정하였고, 이를 개정한 코어 규격 문서 1.1(Gen-Z Core Specification, 2020)을 2020년 2월에 제정하였다. Gen-Z 프로토콜은 DRAM과 NVM을 혼합하여 대용량 메모리 풀을 구성하는 기능을 지원할 뿐만 아니라, 메모리 시멘틱 기반으로 Gen-Z 연결망 내 각 디바이스(CPU, GPU, 가속기 등)가 독립적으로 메모리 풀에 접근할 수 있다. HPE의 경우 2014년부터 문샷(Moon-Shot) 프로젝트를 시작으로 Gen-Z가 추구하는 메모리 중심 컴퓨팅을 구현하고 있으며, 2017년 5월, 40 노드-160 테라바이트 규모의 공유 메모리를 가지는 The MACHINE 프로토타입을 개발하여 발표하였다(https://www.labs.hpe.com/memory-driven-computing).

2016년 5월에 AMD, ARM, 퀄컴, 화웨이, XILINX, Mellanox가 서로 다른 공급 업체의 CPU와 가속기 사이에 메인 메모리를 공유하면서 서로 통신할 수 있는 새로운 인터커넥션 기술 개발을 위해 CCIX(Cache Coherent Interconnect for Accelerators) 컨소시엄을 구성하였다. 궁극적인 목적은 서로 다른 ISA(Instruction Set Architecture)를 기반으로 하는 프로세서들이 FPGA, GPU, 네트워크/스토리지 어댑터, 심지어 ASIC까지 포함하는 수많은 가속기 장치들에 캐시 일관성(cache-coherent)과 피어 프로세싱(peer

processing)을 가능하게 하는 것이다. CCIX 컨소시엄에서는 적용 영역 관점에서, CCIX 프로토콜을 클라우드, 데이터센터, 네트워킹 시스템 내에서 가속(acceleration) 유즈케이스(use-case)를 위한 새로운 수준의 성능, 효율성을 가져오는 캐시 일관성 기반 멀티칩 인터커넥트로 정의하고, 데이터 분석과 검색, 머신러닝, 무선 4G/5G, 메모리 데이터베이스 프로세싱, 비디오 분석, 네트워크 처리 등의 데이터센터 애플리케이션 가속을 위한 오픈 프레임워크에 적용하고자 한다.

2016년 10월에 IBM을 중심으로 AMD, 구글, Mellanox, 마이크론 등 OpenPOWER의 파트너사가 기존 CAPI(Coherent Accelerator Processor Interface) 인터페이스를 기반으로, 이를 공유하기 위해 OpenCAPI 컨소시엄을 설립하였다. CAPI 프로토콜은 CPU와 외부 가속기인 GPU, ASIC, FPGA, Fast Storage 등을 직접 연결하여 I/O 장치 가속기의 고속 구현을 위해 고안된 방식으로, IBM Power8 프로세서에서 구현되었다.

인텔은 고대역폭에서 가속기와 CPU 간의 메모리 공유를 지원하는 인터커넥트 기술을 개발하고자, CXL(Compute Express Link) 컨소시엄에서 CXL 프로토콜 1.0 규격을 2019년 3월 발표하였다.

표 3. Gen-Z, OpenCAPI, CCIX, CXL 비교

Standard	Physical Layer	Topology	Unidirectional Bandwidth	Mechanicals	Coherence	Coverage
CCIX	PCIe PHY	p2p and switched	32-50GB/s x16	PCIe	Full CC between processors and accelerators	Inner Chassis
Gen-Z	IEEE 802.3 Short and Long Haul PHY	p2p and switched	16,25,28,56GT/s per lane, up-to 256 lanes	SFF-TA-1008/9	Does not specify cache coherent agent operations, but does specify protocols that support cache coherent agents	Inter Rack
OpenCAPI	BlueLink 25Gbps PHY	p2p	256GB/s x8	Zaius design	Cache coherence not supported until v4.0	Inner Chassis
CXL	PCIe PHY	p2p	64GB/s x16	PCIe	Coherency Interface	Inner Chassis

출처: CCIX, Gen-Z, OpenCAPI: Overview & Comparison, Brad Benton, AMD Mar. 2017. with modifications

〈표 3〉은 지금까지 설명한 4가지 차세대 연결망 프로토콜을 비교하고 있다. CCIX와 CXL은 캐시 일관성을 지원하며, 기존 PCIe 폼펙터를 물리규격으로 사용한다는 측면에서 유사 기능을 목표로 하고 있으며, 단지 CCIX가 토폴로지에서 좀 더 고확장 특성이 있다. 물론, 캐시 일관성을 위한 시스템 인터페이스를 CCIX, CXL이 각각 대칭형, 비대칭형 구조를 채택하는 것도 다른 점이다. OpenCAPI는 BlueLink를 사용하고, IBM Power 프로세서에 종속되어 x86 계열이나 ARM 프로세서와는 호환성을 제공하지 못하는 것이 제약사항이다.

Gen-Z는 다른 프로토콜에 비해 목적 자체가 확장성을 지향하고 있으므로, 적용 가능한 범위가 가장 넓으며, 확장성이 좋은 것으로 분석된다. 따라서 보드, 섀시를 넘어서는 랙 내 혹은 랙간의 메모리 풀, CPU, 가속기 등 자원에 대한 접근이 가능하다.

3. NDP(Near Data Processing) 기술 동향

본 절에서는 데이터 중심 컴퓨팅의 동일 개념인 NDP 관련 기술 동향을 요약한다.

Patrick Siegl, et. al.에 의하면, 저장장치를 메모리로 한정했을 때, NDP는 NMC(Near-Memory Computing)(혹은 Near-Memory Processing)과 PIM(Processing-In Memory)의 상위 개념으로 간주된다(Patrick Siegl, et. al., 2016). PIM은 프로세싱 유닛이 메모리칩 혹은 메모리 셀과 같이 패키징되어 있지만, NMC는 프로세싱 유닛이 메모리와 근접되어 있되, 같이 패키징되어 있지 않고 분리된 구조를 갖는다.

초기 PIM 영역에 해당하는 연구들이 90년대 말부터 2000년대 초반까지 수행되었으며, 주로 Computational RAM(Duncan G. Elliott, et al., 1999) 등으로 명명되어, RAM 셀 내부에 메모리 어레이(array)와 데이터 라인 사이에 시프트(shift) 등 간단한 연산 기능을 구현하였다. 이후 2000년대 후반부터, 마이크론의 Hybrid Memory Cube, AMD와 하이닉스의 High Bandwidth Memory, 삼성의 Wide I/O, MoSys의 Bandwidth Engine 등의 고대역폭 3D 메모리(stacked DRAM) 기술이 개발되면서, 3D 메모리 기반 PIM 연구가 시작되었다. 아울러, 최근에는 3D 메모리 기술을 확보한 메모리 업체들을 중심으로 PIM 연구가 주도되고 있으므로, 실현성이 높다는 측면에서, 3D 메모리 기반 PIM은 최근 학계 및 산업계의 주목을 받고 있다.

3D 기반 PIM에서는 로직 디바이스와 메모리 스택은 직접 연결되며, 실제로, 로직 디바이스에 프로세싱 기능을 구현할 수 있다. 이 프로세싱 유닛과 메모리의 통합으로 메모리에서 CPU까지 데이터를 가져오는 대신 메모리 안에서 데이터를 처리하기 때문에 기존 시스템과 달리 CPU와 메인 메모리 사이의 대역폭 제한에 영향을 받지 않고 성능을 높일 수 있다. 또한, 메모리 안에서 계산을 수행할 경우 데이터를 CPU까지 가져올 필요가 없으므로 데이터 이동에 소모되는 에너지 또한 크게 줄일 수 있다. 3D 메모리 기술 기반이 아닌 SCM 기반의 PIM 연구 사례도 존재한다(Manuel Le Gallo, et al., 2018).

PIM 보다는 공격적이지 않지만, 메모리와 프로세싱 유닛을 분리하되, 한 보드 내에서 구현하여 성능을 개선하는 Near-Memory Processing 개념의 구현 연구(Mohammad Alian, et al., 2018)가 진행되고 있으며, NMC 기반으로 고성능 Near-Memory Open-CL 기반 가속기 구조를 딥러닝 애플리케이션에 적용하는 연구도 있다(Soroosh Khoram, et al., 2017).

메모리 이외에, 기타 비휘발성 스토리지인 플래쉬, 디스크 내에서 처리하는 메모리처럼 대역폭, 전력, 대기시간 이점 이외에도 밀도가 높아서 훨씬 더 큰 작업을 수행할 수 있다. 이를 위해 Programmable Unit, Fixed-Function Unit, Re-configurable Unit인 3가지 형태로 연구되고 있다(G. Singh, et. al., 2019).

Programmable Unit으로는 GPU를 내장한 SSD에서 MapReduce 프레임워크 기반 API를 제공하는 XSD(B. Y. Cho, et al., 2013), 보안용으로 프로그래밍 가능한 Storage Processor Unit(SPU)을 가진 WILLOW(S. Seshadri, et al., 2014), 호스트 어플리케이션에서 ARM 기반 코어를 내장한 SSD에 CPU 태스크를 오프로딩하는 API를 설계한 SUMMARIZER(G. Koo, et al., 2017), In-storage Processing Subsystem(ISPS) 기반 CompStor(M. Torabzadehkashi, et al., 2018)이 있다.

Fixed-Function Unit으로는 객체기반 통신 프로토콜을 사용하여 SSD 처리 능력을 높이는 Smart SSD (Y. Kang, et al., 2013), NVM이 data comparison write 또는 flip-n-write 모듈 같은 기본 로직과 자연스럽게 협업하는 ProPRAM(Y. Wang, et al., 2015), 프로그래머가 데이터 집약적 애플리케이션을 호스트와 스토리지에 분산할 수 있는 BISCUIT(B. Gu, et al., 2016) 등이 있다.

Re-configurable Unit으로는 플래쉬 기반 플랫폼인 BlueDBM(S. Jun, et al., 2015), 스토리지 노드에 TCP/IP 소켓을 통한 키-값 저장을 위한 인터페이스를 제공하는 CARIBOU(Z. Istvan, D. Sidler, and G. Alonso, 2017)가 있다. 또한, Near-Disk-Storage Processing(G. Davidson, K. Boyack, R. Zacharski, S. Helmreich, and C. J.R., 2006)은 미국 샌디아 국립 연구소(Sandia National Laboratories)에서 Netezza 아키텍처로 제안된 밀결합 서버에 많은 병렬 Snippet Processing Unit(SPU)을 가진 구조로, 각 SPU는 빠른 패턴 매칭을 수행하기 위한 자체 디스크와 스트리밍 데이터베이스 로직 칩을 가지고 있다.

한국전자통신연구원에서는 2018년부터 메모리 중심 컴퓨팅 실현을 위한 메모리 인터커넥트 및 공유 메모리 풀의 하드웨어 개발 과제를 수행하고 있으며, 이를 기반으로, 장기적으로는 NDP로 연구 범위 확장을 계획하고 있다.

III 메모리 중심 컴퓨팅 운영체제 기술 동향

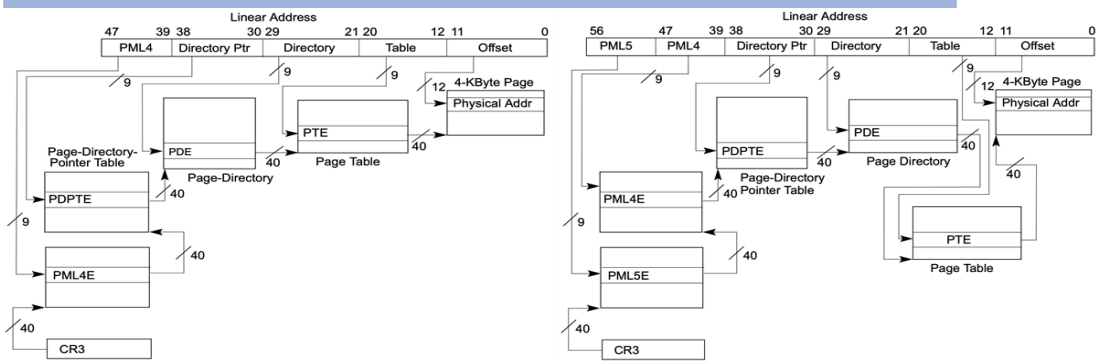
1. 대용량 메모리 지원을 위한 운영체제 기술 동향

최근 시장에서 SAP HANA와 같은 인메모리 데이터베이스 및 인메모리 컴퓨팅이 대중화되고 있다. 시장예측 자료에 따르면 인메모리 컴퓨팅 시장은 2026년까지 약 310억 달러에 이를 것으로 예상되고 있다. 이와 같은 인메모리 컴퓨팅 환경은 고성능/대용량 메모리를 제공하는 하드웨어, 이를 효율적으로 지원하는 시스템소프트웨어, 고성능/대용량 메모리를 적극적으로 활용하는 응용의 융합을 통해 발전하고 있다. 본 장에서는 고성능/대용량 메모리를 지원하기 위한 하드웨어 기술, 운영체제 기술, 가상화 기술의 동향을 기술한다.

1.1 하드웨어 기반 대용량 메모리 주소 공간 지원 기술

최근 대용량 메모리를 제공하기 위한 다양한 하드웨어들이 출시되고 있다. 인텔 DCPMM은 기존의 DIMM에 장착되면서도 대용량과 영속성을 지원하는 메모리다. 하나의 모듈로 현재의 DRAM 기술로는 구현할 수 없는 512GB 크기를 제공한다. 이와 같은 대용량 메모리를 컴퓨터 시스템에서 적극적으로 활용하기 위해서는 해당 메모리를 가상메모리(Virtual Memory) 주소 공간에 매핑할 수 있어야 한다. 하지만 현재 컴퓨터 시스템은 48bit 범위의 가상주소를 지원하기 때문에, 컴퓨터 시스템에서 사용 가능한 메모리의 양이 256TB로 제한된다.

그림 3. 페이징 기반 주소 변환 - (좌) 4-Level 페이징, (우) 5-Level 페이징



출처: "5-Level Paging and 5-Level EPT" (PDF). Intel Corporation. May 2017.

인텔은 이와 같은 문제를 해결하기 위해 5-level 페이징을 개발하여 Sunny Cove(SNC) 구조를 장착한 CPU부터 장착할 것임을 발표했다("5-Level Paging and 5-Level EPT", 20127). 〈그림 3〉에서 나타낸 것처럼, 기존의 4-level 페이징이 단계당 12bit의 주소 공간을 지원하여 총 48bit의 주소 공간을 지원한다. 5-level 페이징은 최상단에 8bit의 주소 공간을 추가하여 총 56bit의 주소 공간을 지원한다. 리눅스 커널에서도 버전 4.14부터 56bit 주소 공간을 지원하기 시작했으며, 이를 통해 SNC 구조 이후의 CPU를 사용할 경우 사용자 프로그램이 최대 64PB의 가상메모리 주소 공간을 사용할 수 있게 되었다.

5-level 페이징은 사용자 프로그램이 더 넓은 공간의 메모리를 매핑하여 사용할 수 있도록 한다. 이는 물리서버 내의 메모리 용량 증가뿐만 아니라, 아래에서 설명할 메모리 확장 기술을 지원하는 데도 필수적이다. 이를 통해 인메모리 데이터베이스나 인메모리 캐시 서버 등 대용량 메모리가 필요한 서비스의 한계를 극복할 수 있게 된다.

1.2 시스템소프트웨어 기반 대용량 메모리 지원 기술

SpaceJMP는 가상주소공간의 제약을 시스템소프트웨어로 해결하기 위한 기술이다(I. El Hajj et. al., 2016). 기존에는 각 프로세스가 실행될 때 해당 프로세스에 종속된 가상주소공간 하나만을 사용할 수 있었다. SpaceJMP는 가상주소공간을 프로세스와는 별도로 운영체제에서 관리한다. 프로세스가 실행될 때, 해당 프로세스가 현재 사용 중인 가상주소공간을 운영체제에서 붙여주게 된다. 더불어 필요에 따라 다른 가상주소공간을 붙이기도 한다. 하나의 프로세스가 여러 가상주소공간을 사용할 수 있게 되면서, 기존보다 훨씬 넓은 가상주소공간을 사용할 수 있게 되는 것이다.

또, SpaceJMP는 특정 가상주소공간을 여러 프로세스가 공유하는 기능도 지원한다. 가상주소공간이 프로세스로부터 독립되어 있으므로, 하나의 가상주소공간을 여러 프로세스에 붙이는 것(attach)이 가능해진 것이다. SpaceJMP는 공통된 가상주소공간에 데이터를 쓰고 읽음으로써 별도의 프로토콜 및 데이터 복제 없이 데이터 공유가 가능하며 공유 데이터 접근 시간도 줄일 수 있다. SpaceJMP 연구진은 이를 여러 프로세스를 사용하는 인메모리 데이터베이스인 Redis에서 활용하여 최대 7배의 성능 향상을 달성했다.

SpaceJMP는 최근 시장에 나타나고 있는 비휘발성 메모리, 대용량 메모리 등에 대응하여 운영체제가 메모리를 다루는 전통적인 방법을 바꾸었다. 프로세스가 활용할 수 있는 가상주소공간이 획기적으로 늘어나고 효율적인 가상주소공간 공유도 가능해지면, 이미 수십 테라바이트에 이르는 스토리지도 가상주소공간을 통해 접근할 수 있게 된다. 특히, SCM도 등장하면서, 메모리와 스토리지로 완전히 구분된 추상화 계층의 경계가 허물어지고 있다. 이에 관해 재고하는 연구가 필요하다.

1.3 가상화 기반 메모리 확장 기술

대용량 메모리 시스템은 특히 클라우드 컴퓨팅 환경에서 주목받고 있다. 대용량 메모리 시스템을 구축하는 비용이 높으므로, 이를 클라우드 컴퓨팅을 통해 공유해서 사용하면 비용이 절감될 수 있기 때문이다. 이 같은 경향에 따라 아마존 EC2 클라우드 서비스는 24,576GB 크기의 메모리 용량을 지원하는 VM instance(인스턴스) 서비스를 제공하기 시작했으며, 마이크로소프트 Azure 클라우드도 11,400GB 크기의 VM instance 서비스를 제공하고 있다(Linux Virtual Machines Pricing). 하지만, 클라우드 컴퓨팅을 위한 데이터센터에 대용량 메모리 시스템의 구축에는 높은 비용이 요구되기 때문에, 전 세계에 걸쳐있는 데이터센터에 적용하기는 쉽지 않다.

클라우드 컴퓨팅을 위한 데이터센터에서 대용량 메모리 시스템의 도입이 어려운 근본적인 이유는 프로세서와 메모리의 독립적인 확장이 불가능한 현대의 컴퓨터 구조 때문이다. 각 프로세서에 연결될 수 있는 메모리 모듈 개수가 제한되어 있고, 각 메모리 모듈이 가질 수 있는 최대용량에도 제약이 존재한다. 이를 메모리 용량 장벽(Memory Capacity Wall)이라고 부른다. 메모리 용량 장벽을 넘기 위해서 산업계에서는 프로세서와 메모리의 독립적인 확장을 요구하기 시작하였으며, 이를 시스템소프트웨어를 통해 극복하는 기법이 연구되고 있다. 대표적인 것이 네트워크를 통해 여러 물리머신의 메모리를 융합해서 쓰는 메모리 분리 시스템 연구(memory disaggregation)이다.

한국전자통신연구원에서는 수행 중인 “패브릭 메모리 컴퓨팅 핵심 기술 연구” 과제를 통해 범용의 대중화된 상용 컴퓨터 시스템(Commodity Computing System)에 기반한 가상화 및 범용 운영체제 기술과 대용량

메모리 기반 응용 알고리즘(4.1장)을 연구하고 있다. 해당 과제에서 연구/개발한 메모리 분리 시스템을 위한 가상화 기술인 Elastic Memory Platform(EMP)은 클라우드 환경에서 같은 랙에 존재하는 고속 네트워크로 연결된 물리서버의 메모리 혹은 고성능 SSD 등을 활용하여 프로세서와 독립적인 메모리 확장을 제공하는 가상화 기술이다. EMP에서는 가상 CPU가 사용하는 메모리의 일부를 고속 네트워크로 연결된 다른 노드에 위치시키거나 고성능 SSD와 같은 I/O에 연결된 저장장치에 위치시킨다. 이때, 빠른 접근이 필요한 메모리를 물리서버에 직접 장착된 지역 메모리인 DRAM에 할당하고, 자주 접근되지 않는 부분을 다른 물리머신의 DRAM 혹은 고성능 SSD를 활용한 원격메모리에 할당하여 가상머신의 성능 저하를 최소화할 수 있다 (K. Koh, K. Kim, S. Jeon and J. Huh, 2019)(K. Koh, K. Kim, C. Kim and J. Huh, 2019).

2. 공유 메모리 지원을 위한 운영체제 기술 동향

메모리 중심 컴퓨팅은 대규모 메모리 풀을 다수의 노드가 공유하는 구조를 기반으로 하며, 이러한 메모리 풀은 대규모 구성에 적합한 고집적 비휘발성 메모리를 포함한다. 따라서, 메모리 풀의 비휘발성 및 노드 간 공유 문제를 효과적으로 해결 가능한 기술이 필요하며, 이와 관련된 운영체제 및 기반 SW 수준 기술 연구가 활발히 진행되고 있다.

메모리의 비휘발성으로 인해 발생하는 대표적인 문제로 휘발성 네임스페이스 문제를 들 수 있다. 기존 DRAM은 전원이 꺼지면 저장된 데이터가 소멸하는 휘발성 메모리이므로 전원이 켜져 있는 동안만 유효한 가상 주소(Virtual Address)를 통해 데이터에 접근하는 것으로 충분했다. 하지만 비휘발성 메모리는 전원 여부와 관계없이 데이터가 유지되며, 다시 전원을 켰을 때 변동성이 있는 가상주소만으로는 어느 데이터가 어느 주소에 있는지 알아내는 것이 불가능하다. 따라서, 비휘발성 메모리에 저장된 데이터에 대한 영속적인 접근성을 보장 가능한 네임스페이스 기능이 필요하며, Mnemosyne, NV-heaps, HEAPO 등 힙(heap) 공간에 대해 영속적인 네임스페이스를 제공하는 연구들이 진행되어 왔다.

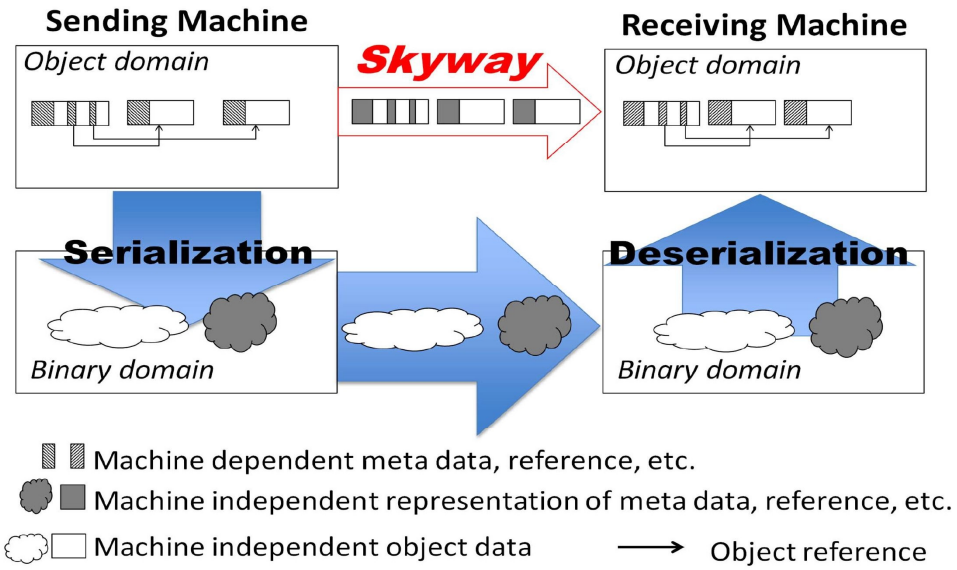
또한, 기존 DRAM은 오류로 인해 저장된 데이터에 문제가 발생하더라도 재부팅을 하면 모든 값이 사라지므로 문제가 없었지만, 비휘발성 메모리를 사용하면 저장된 값이 사라지지 않으므로 재부팅 후에도 문제가 해결되지 않는다. 이러한 문제에 대한 비휘발성 메모리의 일관성 보장을 위해 여러 소프트웨어 기술들이 제안되었는데, 인텔이 공개한 영속 메모리용 개발 라이브러리인 PMDK(Persistent Memory Development Kit)에서 제공하는 트랜잭션 기능 혹은 NV-heap과 같은 라이브러리 수준에서의 일관성 보장 기술 등이 대표적이다.

메모리 중심 컴퓨팅에서 비휘발성 못지않게 중요하게 다루어야 할 기술 이슈는 노드 간 공유되는 메모리 풀에 대한 성능 및 신뢰성 보장 기술이다. 메모리 중심 컴퓨팅의 공유 메모리 기술은 새로운 문제가 아니며 오래전부터 연구된 분산 공유 메모리(Distributed Shared Memory, DSM) 기술과 그 뿌리를 같이 한다. 분산 공유 메모리 기법은 여러 노드에 분산된 메모리를 공유하는 기술로, 공유로 인해 발생 가능한 일관성(Consistency, Coherence) 문제를 해결하는 것이 핵심 기술 이슈다. 그간 분산 공유 메모리 기술은 막대한 성능 부하를 효과적으로 해결하지 못해 활용도가 극히 제한적이었으나, 최근 메모리 및 인터커넥트 성능이 비약적으로 발전함에 따라 메모리 중심 컴퓨팅의 공유 메모리 풀을 구성하는 주요 핵심 기술 중 하나로 재조명되고 있다.

한편, 최근 메모리 공유 기술은 앞서 언급한 비휘발성이 추가됨에 따라 기존 분산 공유 메모리 기술만으로는 해결하기 어려운 새로운 연구 주제로 주목받고 있으며, 대표적인 기술로는 리눅스 운영체제 커널 수준에서 구현된 비휘발성 메모리용 분산 공유 메모리 기술인 Hotpot이 있다. 또한, 기존 분산 공유 메모리의 성능 부하를 줄이기 위해 응용프로그램의 데이터 흐름(Data Flow)이나 그 특성을 최대한 활용함으로써 일관성 보장에 필요한 작업을 최소화하는 메모리 공유 기술 또한 활발히 연구되고 있다. 관련 기술로, 데이터 흐름 의존성에 기반한 분산 공유 메모리 기법인 Givy, 데이터 집약적인 응용프로그램에 최적화된 분산 공유 메모리 기법인 Grappa 등이 있다.

메모리 중심 컴퓨팅 환경에서 공유 메모리 풀을 활용함으로써 기존 노드 간 통신으로 인한 성능 부하를 효과적으로 절감할 수 있다. 기존 노드 간 통신을 위해서는 메모리상에 존재하는 자료 구조 혹은 데이터 오브젝트를 바이트 스트림 형태로 변환하는 직렬화(Serialization, Marshaling) 과정을 거친 이후에 전송할 수 있으며, 이러한 작업이 통신 오버헤드에서 차지하는 비중은 절대적으로 높다. 메모리 중심 컴퓨팅과 같이 노드 간 공유 가능한 메모리가 존재한다면 직렬화 작업 필요 없이 메모리에 저장된 데이터 그 자체로 공유할 수 있으며, 이는 데이터 공유로 인한 통신 부하를 획기적으로 절감할 수 있다. 다만, 기존 소프트웨어 스택은 단절된 노드 간 통신을 위해 의존하는 데이터 직렬화에 기반하여 구현되어 있기에, 메모리 중심 컴퓨팅 환경을 위해 최적화된 메모리 공유 기법이 준비되어야 한다. 한편 메모리 중심 컴퓨팅과 같은 수준은 아니지만, 기존 시스템 환경에서 직렬화 부하를 줄이기 위한 연구는 꾸준히 진행되어 왔다. 네트워크로 연결된 노드 간 JVM 수준 데이터 통신 시 형태 변환 없이 바로 전송함으로써 직렬화 부하를 최소화하는 Skyway(〈그림 4〉 참고)가 발표되었고, 대표적인 인메모리 빅데이터 분석 프레임워크인 Spark의 Worker 간 통신을 메모리 공유를 통해 수행할 수 있도록 개선한 Sparkle 또한 유사한 기술이다.

그림 4. 데이터 통신 방법: 직렬화 기반/직접 전송(Skyway)



출처: Skyway: Connecting Managed Heaps in Distributed Big Data Systems, ASPLOS 2018

다수의 노드가 공유하는 대용량 메모리를 외부 공격 또는 오류 등으로부터 효과적으로 보호하는 것 또한 중요한 이슈 중 하나다. 영속성 공유메모리는 다수의 노드가 동시에 접근 가능하므로 외부 공격에 노출되기 쉽고, 비휘발성 메모리의 특성상 데이터의 오염이나 문제가 발생하면 영속적으로 남는 문제점을 보인다. 페이징 기반 가상메모리 시스템에서 메모리 페이지에 대한 접근 권한은 페이지 테이블에 기록된 관련 정보를 통해 관리되며, 이는 POSIX 인터페이스인 mprotect() 시스템 호출 등에 의해 지원된다. 이때 사용자 공간과 운영체제 커널 간 문맥 교환(Context Switching)을 유발하므로 성능 부하를 초래하게 되며 노드 간 공유되는 메모리의 경우 더욱 증폭되기도 한다. 최신 인텔 프로세서에서 지원하는 MPK(Memory Protection Key)는 메모리 영역별 접근 권한이 명시된 키값을 레지스터 수준에서 관리할 수 있도록 함으로써 문맥 교환 없이 효율적인 쓰레드별 메모리 보호가 가능하다. 한편 MPK의 한계인 제한된 키 개수, 페이지 테이블 및 쓰레드 간 동기화 이슈 등을 SW 라이브러리 수준에서 해결한 libmpk가 최근 발표되기도 하였다.

한국전자통신연구원에서도 메모리 중심 컴퓨팅을 위한 운영체제 수준의 연구를 진행하고 있다. 주요 연구 분야는 자체 개발 중인 고속의 메모리 수준 인터커넥트에 기반한 휘발/비휘발 메모리 융합, 로컬/원격 메모리 융합 연구 등이다. 더 나아가 메모리 종류, 위치와 관계없이 응용 수준에서 일관된 관점으로 이들을 활용할 수 있도록 "Everything is a file"을 "Everything is a memory"라는 관점에서 재조명하는 연구도 추진 중이다.

Ⅳ 메모리 중심 컴퓨팅을 활용한 응용프로그램 기술 동향

메모리 중심 컴퓨팅에서는 대용량·공유 메모리를 제공하기 위해 비휘발성 메모리, 초고속 인터커넥트, 운영체제 기술 등을 사용한다. 이로 인해 기존 DRAM만을 이용했을 때보다는 메모리 접근 시간이 늘어난다. 따라서, 기존 응용을 그대로 실행하면 이득을 볼 수 없기 때문에 메모리 중심 컴퓨팅의 장점을 활용하는 응용프로그램 기술이 필요하다.

1. 대용량 메모리를 활용한 대규모 메모이제이션 기법

메모이제이션(memoization) 기법이란 반복되는 연산을 메모리에 저장해두고 연산을 수행하는 대신 메모리에 저장해둔 값을 읽는 것으로 대체하는 기법으로, 금융 관련 응용을 대상으로 많이 연구되어 왔다(G. Agosta, M. Bessi, E. Capra and C. Francalanci, 2011)(A. Moreno and T. Balch, 2014). 일반적으로 메모이제이션은 대상 함수에 대해 입력값에 대한 해시테이블을 만들어 결과값을 저장하는 방식이 사용된다. 함수에 입력값이 주어졌을 때, 먼저 해시테이블에 저장된 결과값이 있는지 찾아보고, 없는 경우에만 실제 연산을 수행하는 것이다. 직접 연산을 수행하는 것보다 해시테이블을 통해 결과값을 찾는 것이 더 빠르다면 성능상의 이득을 볼 수 있다. 단, 매 연산을 수행할 때마다 해시테이블을 찾아보는 과정도 거쳐야 하므로, 입력값이 자주 반복되어 결과값이 해시테이블에 저장되어 있을 확률이 높아야만 성능상의 이득을 볼 수 있다.

메모리 중심 컴퓨팅은 대용량 메모리를 제공하므로, 메모이제이션 기법에 사용되는 해시테이블을 훨씬 크게 만들 수 있다. 결과값이 해시테이블에 저장되어 있을 확률이 더 높아지므로 성능 이득을 볼 확률이 높아질 수 있다. 입력값의 범위와 메모리의 용량에 따라 모든 결과값을 미리 저장해두고 활용하는 것도 가능할 것이다. 그러면 입력값의 반복 여부와 상관없이 성능상의 이득을 볼 수 있게 된다.

HPE에서는 몬테카를로 시뮬레이션(Monte Carlo Simulation)에 다른 형태의 대규모 메모이제이션 기법을 적용한 바 있다(K. Keeton, 2017). 몬테카를로 시뮬레이션은 랜덤하게 생성된 여러 포인트에 대해 시뮬레이션을 수행하고 그 결과값을 모아 분석하여 전체 영역에 대한 통계적 근사치를 계산하는 기법이다. HPE에서는 대용량 메모리를 활용해서 대표적인 포인트들에 대한 시뮬레이션 결과값을 저장해둔 테이블을 구축했다. 그리고 랜덤하게 생성된 포인트들에 대해 시뮬레이션을 수행하는 대신, 저장된 결과값을 변형해서 결과값을 계산하는 기법을 개발했다. 이를 통해 기존 몬테카를로 시뮬레이션 기법 대비 8,600배에서 10,200배에 달하는 성능 향상을 이루었다.

메모리 중심 컴퓨팅이 대용량 메모리를 제공하지만, 무한한 크기의 메모리를 제공하는 것은 아니다. 그러므로 HPE에서 개발한 것처럼 응용의 특성을 활용해 메모이제이션의 효과를 극대화하는 기법을 개발하는 것이 필요하다. 다양한 응용에 대해 이와 같은 기술이 개발된다면 메모리 중심 컴퓨팅의 효용성도 증명될 수 있을 것이다.

2. 비휘발성 메모리를 활용한 응용 성능 향상 가능성

비휘발성 메모리는 메모리와 스토리지의 경계 지점에 있다. 굳이 따지자면 비휘발성 측면에서는 스토리지나 데이터베이스 관련 기술과 맥이 닿아 있으면서, 접근성 측면에서는 블록 장치로도, 그리고 메모리처럼도 접근할 수 있는 특징을 가지고 있다. 비휘발성 메모리의 특성을 이용한 전용 자료구조와 알고리즘은 개발과 적용 부담이 있기는 하지만 이를 통해 훨씬 큰 폭의 성능 향상이 가능하다.

전용 자료구조와 알고리즘을 연구하는 사례에서 비휘발성 메모리의 바이트/워드 단위 접근성(byte addressability), 워드 단위 원자적 업데이트 명령어(atomic update instruction), 비휘발성 도메인 명령어(persistent domain instruction) 등 새롭게 추가된 전용 CPU 연산을 최대한 활용하여 기존 블록장치/파일시스템 계층 통과 비용을 원천 제거하고, 최소잠금/로깅/플러시(minimal lock/log/flush)로 데이터의 일관성 유지에 따른 성능 저하를 최소화한다.

CCEH(Cacheline-Conscious Extendible Hashing)(M. Nam, H. Cha, Y. Choi, S. H. Noh and B. Nam, 2019)는 비휘발성 메모리의 특성을 고려한 전용의 해싱 기법이다. CCEH는 CPU 캐시 친화성을 위해 3단계 해시테이블과 디렉토리 더블링(directory doubling) 기법을 제안하였다. 이 과정에서 데이터의 일관성 유지를 위해 8바이트 원자적 접근으로 결함원자성(failure-atomicity)을 보장하도록 하였다.

wB+Tree(Shimin Chen and Qin Jin, 2015)는 B+ Tree 인덱스를 비휘발성 메모리 특성에 맞춘 전용의 인덱싱 기법이다. wB+ Tree에서는 언두/리두(undo/redo) 로깅 및 쉐도잉 등 전통적인 기법들이 과도한

비휘발성 메모리 기록과 CPU 캐쉬 플러시를 유발하는 문제를 제기하였고, 이를 해결하기 위해 무정렬 리프노드 구조, 원자적 변경 연산, 리두로깅 기법을 활용하여 약 3배의 성능 향상이 가능함을 보였다.

PMDK(https://pmem.io/pmdk/)는 인텔에서 출시한 DCPMM 전용 라이브러리로 low-level lib(libpmem)을 기본으로 하며, 이를 기반으로 다양한 자료 구조를 지원하는 부가 라이브러리(object, log, block, malloc 등)를 API로 제공한다. 이를 이용하여 전용의 자료 구조와 알고리즘을 개발하거나, 제공되는 자료 구조를 활용할 수 있다. 제공되는 자료 구조의 경우 소프트웨어 트랜잭션과 장애 일관성을 지원하기 위해 로깅, 쉐도잉 등의 방법을 내부에서 사용하여 응용 개발자의 관련 부담을 줄였다.

MOD(Swapnil Haria et. al.)는 CCEH, wB+ Tree와 같은 전용의 자료구조와 PMDK같은 범용의 라이브러리 사이에서 각각의 장점을 취하기 위한 새로운 접근법이다. 전용의 자료구조는 성능은 좋지만, 개발 난도가 매우 높고, 범용의 라이브러리는 편의성은 좋으나 범용 트랜잭션 로직으로 인한 성능 저하 문제를 지적하였다. 이에 MOD에서는 PMDK와 같이 맵(map), 셋(set), 스택(stack), 큐(queue), 벡터(vector) 등과 같은 범용 자료 구조를 제공하지만, 데이터의 기록 순서를 완화하고 캐시 플러시를 겹치도록 하여 성능을 끌어 올린 구조를 제안하였다.

Malicevic(J. Malicevic et. al.) 등은 대규모 그래프 분석 응용을 대상으로 비휘발성 메모리의 성능 저하 영향을 최소화하는 연구를 했다. 비휘발성 메모리를 장착한 시스템에서 CPU의 데이터 프리페치(prefetch)와 메모리 접근에 대한 병렬성(Memory-level Parallelism)을 최적화하는 방법을 제안했다. 그리고 작은 용량의 DRAM을 추가하고 DRAM에 중요 데이터를 배치함으로써 DRAM만을 사용했을 경우와 비교했을 때 성능 부하를 20% 수준으로 낮추었다.

3. 대용량 공유 메모리를 활용한 성능 향상 가능성

메모리 중심 컴퓨팅 구조의 또 다른 특징인 공유 메모리 구조에 관한 연구도 진행 중이다. 이 구조는 다수의 컴퓨터들 중심에 공유 메모리를 두고, 매우 빠른 인터커넥트를 이용하여 공유 접근하는 것이 핵심이다. SMP(Symmetric Multi Processing)나 NUMA(Non-Uniform Memory Access) 구조가 다수의 CPU가 노드 안에서 메모리를 공유하는 구조라면 이 구조는 다수의 노드가 메모리를 공유하는 구조이다. 이러한 구조 연구는 HPE의 The Machine으로 대표되며, 이 구조에서 전통적인 데이터베이스, 인메모리 컴퓨팅, 거대 그래프 분석과 같은 응용의 성능 향상 가능성이 연구되었다.

FOEDUS(H. Kimura, 2015)는 메모리 중심 컴퓨팅을 위한 키-밸류(key-value) 저장소다. FOEDUS는 데이터 읽기/쓰기의 많은 부분을 노드 별 DRAM에서 독립적으로 하도록 만들어서 확장성을 높였다. 읽기 작업을 위해서는 노드 별 DRAM에 비휘발성 메모리상의 데이터에 대한 캐시를 만들었다. 읽기 작업이 캐시에 많이 적중될수록 성능이 높아진다. 쓰기 동작을 위해서는 노드 별 DRAM에 로그를 만들었다. 쓰기는 일차적으로 노드별 로그에 쓰는 것으로 끝나고, 이후에 별도의 쓰레드가 해당 로그를 비휘발성 메모리로 옮긴다. 이때 낙관적(optimistic) 동시성 제어를 도입하여 병렬성을 높였다. 이를 통해 메모리 중심 컴퓨팅의 많은 연산장치가 동시에 빠르게 데이터에 접근할 수 있게 했다.

Sparkle(M. Kim et. al., 2017)은 메모리 중심 컴퓨팅을 위한 데이터 분석 플랫폼이다. 기존의 Spark를 개선하여 대용량 공유 메모리를 효과적으로 활용할 수 있도록 만들었다. 데이터 통신을 할 때 공유 메모리를 활용함으로써 가용 대역폭이 늘어났고, 네트워크를 통해 데이터를 주고받을 때 필요한 데이터 마샬링(mashalling) 단계도 제거했다. 그리고 공유 데이터를 수정할 때 공유 메모리 상에 직접 수정하게 하여, 여러 버전의 데이터를 생성하고 관리하던 Spark의 메모리 압박 문제를 해결했다. 그 결과, 가비지 컬렉션(Garbage Collection)에 대한 성능 부하도 줄어 성능이 개선됐다.

Chen 등은 반복 그래프 처리 기법(iterative graph processing)의 특성을 활용해 메모리 중심 컴퓨팅의 효과를 극대화하는 방법을 제안했다(F. Chen et. al., 2016). 반복 그래프 처리 기법을 수행할 때 각 꼭짓점에 대한 데이터는 여러 쓰레드가 사용하는 경우가 많지만 한 꼭짓점의 데이터가 두 개 이상의 쓰레드에 의해 업데이트되는 경우는 드물다. 게다가 동기화가 엄격하지 않아도, 반복 처리 과정 중에 오류가 수정될 수 있다. 메모리 중심 컴퓨팅에서 빠른 데이터 공유를 위해, 데이터에 대한 락(Lock)을 없애고 비동기적 업데이트를 허용함으로써 성능 확장성을 극대화했다. 이를 통해 기존 방식 대비 5배의 성능 향상을 이루었다.

V 결론

빅데이터, AI(인공지능), IoT(사물인터넷) 등의 지능화 기술 발전에 따른 거대 데이터 처리 요구가 증가함에 따라 중앙처리장치(CPU)와 메모리 간 병목 현상을 줄여 연산 성능을 높이는 이른바 데이터 중심 컴퓨팅 시대가 열리고 있다. 데이터 중심 컴퓨팅에서 가장 중요한 개념은 메모리 중심 컴퓨팅으로, 컴퓨팅의 중심을 CPU에서 메모리로 옮기는 것이 그 핵심이다. 메모리 중심 컴퓨팅은 대용량 비휘발성 메모리 풀과 많은 연산 장치를 고속 연결하는 인터커넥트 기술, 이를 효율적으로 제어하기 위한 OS 핵심 기술과 응용프로그램 기술의 혁신을 통해 컴퓨팅의 새로운 패러다임을 제시한다.

본 고에서는 컴퓨팅 방식을 근본적으로 바꿀 수 있는 메모리 중심 컴퓨팅 시대의 기술 트렌드 및 연구 동향을 소개하였다. 그 기술적 변화의 시작은 메모리 디바이스와 이들을 상호 연결하기 위한 고속 인터커넥트 및 하드웨어 기술의 급속한 발전에 기인한다. DRAM의 용량과 대역폭의 한계를 극복하고 데이터 영속성 보장을 위한 비휘발성 메모리 기술 연구가 활발히 진행되고 있고, 다수의 비휘발성 메모리를 연결하여 거대 주소 공간을 효율적으로 관리하기 위한 기법들이 하드웨어, 운영체제(OS) 각각의 영역에서 빠르게 진화하고 있다. 특히, 이를 활용한 다양한 응용 사례들의 등장은 메모리를 중심으로 컴퓨팅의 관점이 변화하는 기술적 패러다임의 흐름을 방증하고 있다. 메모리 중심 컴퓨팅을 위한 메모리 하드웨어와 핵심 소프트웨어의 융합, 그리고 이를 기반으로 한 응용 서비스의 진화는 상호 보완적이고 의존적이다. 하드웨어는 거대 데이터 처리의 고속화를, 소프트웨어는 메모리 중심 하드웨어의 효율적 제어 방법을, 응용 서비스는 거대 데이터를 이용한 신산업 도출에 초점을 맞춘다. 하지만, 이들 중 하나만 살아남는 것이 아니라, 궁극적으로는 상호 공존하는 방향으로 진화하게 될 것이다.

메모리 중심 컴퓨팅의 가장 큰 장점은 '속도의 혁신'이다. 단순히 과거에 수행했던 업무의 처리속도가 약간 빨라진다는 것을 의미하는 것이 아니다. 속도의 혁신을 통해 과거에는 상상하지 못했던 일들을 할 수 있게 만든다는 점이 매력적이다. 전혀 새로운 차원의 서비스 창출이 가능하다는 것이다. 기존 컴퓨팅 방식의 한계가 이미 다양한 분야에서 나타나고 있고, 이를 해결하기 위한 요소 기술들이 점차 성숙해가는 지금이 컴퓨팅 기술 및 시장의 주도권을 확보하기 위한 가장 적절한 시기이다.

저자 _ 오명훈(Myeong-Hoon Oh)

• **학 력**

광주과학기술원 정보통신공학 박사
전남대학교 컴퓨터공학 석사
전남대학교 컴퓨터공학 학사

• **경 력**

現) 한국정보통신기술협회
(클라우드컴퓨팅 프로젝트 그룹) 간사
現) 과학기술연합대학원대학교 겸임교수
現) 한국전자통신연구원 데이터중심컴퓨팅시스템 연구실
책임연구원

저자 _ 김홍연(Hong Yeon Kim)

• **학 력**

인하대학교 전자계산학 박사
인하대학교 전자계산학 석사
인하대학교 통계학과 학사

• **경 력**

現) 한국전자통신연구원 데이터중심컴퓨팅시스템 연구실
책임연구원

저자 _ 고광원(Kwangwon Koh)

• **학 력**

한국과학기술원 전산학 박사

• **경 력**

現) 한국전자통신연구원 데이터중심컴퓨팅시스템 연구실
책임연구원

저자 _

진기성, 한국전자통신연구원 데이터중심컴퓨팅시스템 연구실, 책임연구원
안백송, 한국전자통신연구원 데이터중심컴퓨팅시스템 연구실, 선임연구원
김창대, 한국전자통신연구원 데이터중심컴퓨팅시스템 연구실, 선임연구원
김강호, 한국전자통신연구원 데이터중심컴퓨팅시스템 연구실, 책임연구원/실장
김영균, 한국전자통신연구원 초성능컴퓨팅연구본부, 책임연구원/본부장

참고문헌

국내문헌

1) Data Age 2025, "The Evolution of Data to Life-Critical Don't Focus on Big Data; Focus on the Data that's Big," IDC, 2017.
2) Amirali Boroumand, et. al, "Google Workloads for Consumer Devices: Mitigating Data Movement Bottlenecks", Proceedings of the 23rd International Conference on Architectural Support for Programming Languages and Operating Systems(ASPLOS), Williamsburg, VA, USA, March 2018.
3) "Data Centric Computing," SPXXL/SCICOMP - Summer 2011.
4) Onur Mutlu, "Memory-Centric Computing in the Big Data Era," FMS Special Session Invited Talk, August 8, 2019.
5) Rajeev Balasubramonian et. al, "NEAR-DATA PROCESSING: INSIGHTS FROM A MICRO-46 WORKSHOP," IEEE Micro, Vol 34, Issue 4, July-Aug., 2014.
6) Yoonho Park, "IBM Data Centric Systems & OpenPOWER," HPC User Forum in Santa Fe, May 6, 2017.
7) https://semiengineering.com/in-memory-vs-near-memory-computing/
8) 김창대 외 6, "메모리 드리븐 컴퓨팅 연구동향", 정보과학회지, 제37권 제6호, pp. 43-51, 2019년 6월
9) W.A. Wulf, S.A. McKee, "Hitting the memory wall: implications of the obvious," SIGARCH Comput. Archit. News, 23 (1) (1995), pp. 20-24
10) J. Thomas Pawlowski, "Hybrid Memory Cube," HotChips 23, 2011
11) J.C. Lee, et al., "A 1.2V 64Gb 8-channel 256GB/s HBM DRAM with peripheral-base-die architecture and small-swing technique on heavy load interface," IEEE ISSCC Dig. Tech. Papers, pp. 318-319, Feb, 2016
12) J. H. Cho, et al., "A 1.2V 64Gb 341GB/S HBM2 stacked DRAM with spiral point-to-point TSV structure and improved bank group data control," IEEE ISSCC, Feb. 2018
13) Providing Storage at Memory Speed Using NVDIMMs Sponsored by the SNIA NVDIMM SIG, Open Server Summit, April 14, 2016.
14) "The Challenge of Keeping Up With Data," Product brief of Intel Optane DC Persistent Memory, Intel.
15) https://www.everspin.com/
16) http://www.avalanche-technology.com/
17) https://genzconsortium.org/

18) https://www.ccixconsortium.com/
19) https://opencapi.org/
20) https://www.computeexpresslink.org/
21) Gen-Z Core Specification version1.1,. 2020.02.
22) https://www.labs.hpe.com/memory-driven-computing
23) Patrick Siegl, et. al, "Data-Centric Computing Frontiers: A Survey on Processing-In-Memory," MEMSYS 2016, Oct. 2016.
24) Duncan G. Ellott, et al., "Computational RAM: Implementing Processors in Memory," IEEE Design& Test of Computers, 1999.
25) Manuel Le Gallo, et al., "Mixed-Precision In-Memory Computing," Nature Electronics, 2018.10
26) Mohammad Alian, et al., "Application-Transparent Near-Memory Processing Architecture with Memory Channel Network", 51st Annual IEEE/ACM International Symposium on Microarchitecture, 2018
27) Soroosh Khoram, et al., "Challenges and Opportunities: From Near-memory Computing to In-memory Computing," ISPD '17, March 19-22, 2017.
28) G. Singh, et. al, "Near-Memory computing, "Near-Memory Computing: Past, Present, and Future," Journal of microprocessors and microsystems, Vol. 71, Nov. 1, 2019.
29) B. Y. Cho, et al., "XSD: Accelerating MapReduce by Harnessing the GPU inside an SSD," in Proceedings of the 1st Workshop on Near-Data Processing, 2013.
30) S. Seshadri, et al., "Willow: A User-programmable SSD,"in Proceedings of the 11th USENIX Conference on Operating Systems Design and Implementation, 2014, pp. 67-80.
31) G. Koo, et al., "Summarizer: Trading Communication with Computing Near Storage," in Proceedings of the 50th Annual IEEE/ACM International Symposium on Microarchitecture, 2017, pp. 219-231.
32) M. Torabzadehkashi, et al., "Comp-Stor: An In-storage Computation Platform for Scalable Distributed Processing," in 2018 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)2018, pp. 1260-1267.
33) Y. Kang, et al., "Enabling Cost-effective Data Processing with Smart SSD," in Mass Storage Systems and Technologies (MSST), 2013 IEEE 29th Symposium on., 2013, pp. 1-12.
34) Y. Wang, et al., "ProPRAM: Exploiting the Transparent Logic Resources in Non-Volatile Memory for Near Data Computing," in Proceedings of the 52nd Annual Design Automation Conference. ACM, 2015.
35) B. Gu, et al., "Biscuit: A Framework for Near-data Processing of Big Data Workloads," SIGARCH Comput. Archit. News, vol. 44, no. 3, pp. 153-165, Jun. 2016.
36) S. Jun, et al., "BlueDBM: An Appliance for Big Data analytics," in 2015 ACM/IEEE 42nd Annual International Symposium on Computer Architecture (ISCA), June 2015, pp. 1-13.

37) Z. Istvan, D. Sidler, and G. Alonso, "Caribou: Intelligent Distributed Storage," Proceedings of the VLDB Endowment, vol. 10, no. 11, pp.1202-1213, 2017.
38) G. Davidson, K. Boyack, R. Zacharski, S. Helmreich, and C. J.R., "Data-centric computing with the netezza architecture," Technical report sand2006-3640, Sandia National Laboratories, April 2006.
39) "5-Level Paging and 5-Level EPT" (PDF). Intel Corporation. May 2017.
40) I. El Hajj, A. Merritt, G. Zellweger, D. Milojicic, R. Achermann, P. Faraboschi, W.-m. Hwu, T. Roscoe, and K. Schwan. SpaceJMP: Programming with Multiple Virtual Address Spaces. ASPLOS, 2016.
41) Linux Virtual Machines Pricing, https://azure.microsoft.com/en-us/pricing/details/virtual-machines/linux/
42) K. Koh, K. Kim, S. Jeon and J. Huh, "Disaggregated Cloud Memory with Elastic Block Management," IEEE Transactions on Computers, vol. 68, no. 1, 2019.
43) K. Koh, K. Kim, C. Kim and J. Huh, "End Performance SLA Support for Disaggregated Memory," WORD, 2019.
44) G. Agosta, M. Bessi, E. Capra and C. Francalanci, "Dynamic memoization for energy efficiency in financial applications," 2011 International Green Computing Conference and Workshops, Orlando, FL, 2011, pp. 1-8.
45) A. Moreno and T. Balch, "Speeding up Large-Scale Financial Recomputation with Memoization," 2014 Seventh Workshop on High Performance Computational Finance, New Orleans, LA, 2014, pp. 17-22.
46) K. Keeton, "Memory-Driven Computing," Keynote of FAST, 2017.
47) M. Nam, H. Cha, Y. Choi, S. H. Noh and B. Nam, "Write-Optimized Dynamic Hashing for Persistent Memory," FAST, 2019.
48) Shimin Chen and Qin Jin. Persistent b+-trees in non-volatile main memory. Proceedings of the VLDB Endowment, 8, February 2015.
49) https://pmem.io/pmdk/
50) Swapnil Haria and Mark D. Hill, and Michael M. Swift, "MOD: Minimally Ordered Durable Datastructures forPersistent Memory," ASPLOS 20
51) J. Malicevic, S. Dulloor, N. Sundaram, N. Satish, J. Jackson and W. Zwaenepoel, "Exploiting NVM in Large-scale Graph Analytics," INFLOW, 2015.
52) H. Kimura, "FOEDUS: OLTP Engine for a Thousand Cores and NVRAM," SIGMOD, 2015.
53) M. Kim, J. Li, H. Volos, M. Marwah, A. V. Ulanov, K. Keeton, J. Tucek, L. Cherkasova, L. Xu and P. Fernando, "Sparkle: optimizing spark for large memory machines an d analytics," SoCC, 2017.
54) F. Chen, M. T. Gonzalez, K. Viswanathan, Q. Cai, H. Laffite, J. Rivera, A. Mitchell and S. Singhal, "Billion node graph inference: iterative processing on The Machine," HPE Tech Report, 2016.

융합연구리뷰

Convergence Research Review 2020 March vol.6 no.3

02

양자통신 및 양자컴퓨팅 분야 소개 및 연구동향

한상욱(한국과학기술연구원 양자정보연구단 단장)
조영욱(한국과학기술연구원 양자정보연구단 선임연구원)
임향택(한국과학기술연구원 양자정보연구단 선임연구원)

Ⅰ 양자정보통신기술의 개요

2016년 3월 인류 역사에서 손에 꼽을만한 큰 사건이 한국에서 일어났다. 인류를 대표하여 한국의 바둑기사 이세돌 9단이 구글의 자회사인 딥마인드의 인공지능 바둑기사 알파고와 자존심을 걸고 바둑 대국을 가졌다. 대국 시작 전, 이세돌 9단은 본인이 4:1 또는 5:0으로 승리할 거라 말했고, 많은 사람들이 이세돌 9단의 승리를 예상했다. 그러나 모두가 알다시피 결과는 이세돌 9단이 1승 4패로 알파고에게 패하였다. 이는 인류에게 있어서 바둑만큼은 컴퓨터가 인간을 이길 수 없다는 자존심을 철저하게 깨버린 사건이었으며, 오히려 알파고가 보여준 바둑은 지금까지 인류가 쌓아온 바둑의 정석과는 전혀 다른 새로운 방식이었다는 점에서 바둑계에 큰 혼란을 가지고 왔다.

현재 우리는 4차 산업혁명 시대를 준비하고 있다고 한다. 처음 4차 산업혁명에 대한 이야기가 나왔을 때, 실체가 없는 뜬구름을 잡는 이야기 같았지만, 위의 알파고 사례와 같이 4차 산업혁명은 이미 우리 주위에서 진행 중이다. 4차 산업혁명은 인공지능, 기계학습, 초연결 등과 같은 새로운 과학기술에 의해 주도될 것으로 예상되고 있는데, 인공지능 및 기계학습과 같이 강력한 연산 능력을 필요로 하는 분야에 빠른 연산 능력을 제공할 수 있는 양자컴퓨팅과 초연결 사회에 필수적인 강력한 보안성을 제공할 수 있는 양자통신을 포함하는 양자정보기술 또한 4차 산업혁명에서 핵심적인 역할을 할 기술로 여겨지고 있다.

본 고에서는 양자정보통신기술의 개념에 대해 간략하게 서술하고, 이 중에서 양자통신과 양자컴퓨팅에 대한 국내외 기술개발 동향과 산업계 동향에 대해 소개한 후, 향후 기술의 발전에 대해 조망해 보고자 한다.

1. 양자정보통신기술의 개념

양자정보통신기술은 기존의 고전정보통신기술에 양자역학적인 원리가 합쳐진 새로운 기술로 양자역학적 특성을 정보통신기술에 적용하기 위하여 양자상태를 생성, 제어, 측정 및 분석하는 기술이다. “양자”는 쉽게 말하면 불연속적으로 표현되는 물리량을 이야기하는데, 일반적으로 양자역학적인 현상은 빛의 최소 단위인 광자, 물질의 최소 단위인 원자 및 전자와 같은 미시적인 세계에서 일어나기 때문에 일상생활에서 관찰할

수 없는 신기한 현상들로 이루어져 있다. 이러한 양자현상 중에서 양자정보통신에 사용되는 중요한 양자적인 특성들에는 양자중첩(quantum superposition), 양자얽힘(quantum entanglement), 복제불가원리(no-cloning theorem) 등이 있다.

양자상태는 고전상태와 달리 서로 다른 두 상태의 중첩이 가능하다. 즉, 전자스핀의 경우 스핀 업(Up)과 스핀 다운(Down)의 서로 다른 두 상태가 동시에 존재할 수 있는데, 이 중첩의 원리 때문에 양자정보의 기본 단위인 큐비트(Qubit, Quantum bit의 줄임말)는 0과 1의 상태를 동시에 가질 수 있다. 반면에, 고전정보의 기본 단위인 비트는 0 또는 1의 두 가지 상태 중 하나만을 표현한다. 그리고 양자역학의 가장 특이한 현상 중 하나인 양자얽힘은 양자상태 간의 특수한 상관관계를 나타내는데, 두 상태가 시공간적으로 무한히 멀리 떨어져 있더라도 양자얽힘 관계를 가지고 있다면, 하나의 입자에 양자적 측정을 했을 때, 그 즉시 다른 양자상태에도 영향을 주게 된다. 이처럼 양자얽힘은 비국소적인 상관관계이며, 고전물리로는 설명되지 않는 양자역학적인 상관관계를 말한다. 그리고 임의의 양자상태를 완벽하게 복제하는 것이 불가능하다는 것이 잘 알려져 있는데, 이를 '복제불가원리'라고 한다. 즉, 고전정보의 경우 컴퓨터에서 하나의 파일을 여러 곳의 저장장치로 복사하는 것처럼 정보를 무한히 복제할 수 있지만, 양자정보의 경우 양자역학적으로 복제가 불가능하므로 뒤에서 소개하는 양자암호통신의 안전성을 보장하는 중요한 원리가 된다.

양자정보통신기술은 크게 양자통신, 양자컴퓨팅, 양자센서/이미징 등이 있으며, 본 기고문에서는 이 중에서 양자통신과 양자컴퓨팅 분야의 동향에 대해서 주로 소개하고자 한다.

2. 양자통신

영화 '이미테이션 게임'은 영국의 수학자인 앨런 튜링(Alan Turing)이 나치 독일의 암호체계인 에니그마(Enigma)를 해독하는 과정을 그린 실화를 바탕으로 한 영화이다. 독일군의 공격에 대해 속수무책으로 당하던 연합국은 앨런 튜링의 에니그마 암호 해독을 통해 전세를 뒤집고 제2차 세계대전에서 승기를 잡을 수 있었다. 이 에니그마 암호 해독으로 인하여 독일군의 통신 보안은 무너지게 되었고, 이는 국가 안보와 직결되는 전쟁의 승패에 통신의 보안이 얼마나 중요한 역할을 하는지 잘 보여주고 있다. 그뿐만 아니라 현대 사회에서 온라인 뱅킹, 신용카드와 같은 금융정보에서부터 주민등록번호나 전화번호 등 개인 신상정보 등도 정보의 보안이 깨진다면, 개개인의 삶에도 큰 악영향을 끼칠 수 있다. 특히, 계산의 복잡성을 이용하여 보안의 안정성을 추구하는 기존의 방식은 양자컴퓨팅 기술이 개발되어 고전컴퓨터보다 더 빠른 계산이 가능하게 되면 더 이상 안전성을 보장할 수 없기에 이를 방지할 수 있는 새로운 암호체계가 필요하다.

이런 위협에 대한 대응기술로써 등장한 것이 바로 양자통신이라고 할 수 있다. 양자통신은 양자상태를 활용하여 정보통신의 보안성을 강화하고 양자기기 간의 네트워크를 구성함으로써 양자기기 간 통신을 지원하는 기술을 포함하고 있다. 이러한 양자통신은 크게 양자암호통신과 양자전송, 그리고 양자네트워크 세 분야로 나눌 수 있다. 양자암호통신은 암호통신에서 사용되는 비밀키를 통신에 참여하는 사용자들 사이에서 절대적으로 안전하게 나누어 가질 수 있게 만드는 혁신적인 차세대 보안통신이다. 양자암호는 양자상태를 이용하여 비밀키 정보를 정당한 사용자들끼리 나누어 갖는데, 이때 임의의 양자상태는 완벽하게 복제할 수 없다는 '복제불가원리'에 의해 양자역학적으로 그 안전성을 보장한다. 도청자가 중간에서 양자측정을 통해 양자상태에 대한 정보를 가져가면, 양자상태에 변화를 주게 되고 이러한 변화는 송신자와 수신자의 키 분배의 오류 통계에 변화를 가져오게 되므로 송신자와 수신자는 도청자의 존재를 확인할 수 있다. 양자암호통신은 양자역학적인 특성을 이용하여 송수신자 간 비밀키를 나누는 것이 핵심이기 때문에 양자 키 분배(QKD, Quantum key Distribution)로 명명하기도 한다.

양자전송(Quantum teleportation)은 송신자와 수신자 사이에 공유하고 있는 양자얽힘을 이용하여 송신자가 가지고 있는 양자정보를 수신자에게 전달하는 기술이다. 미국의 SF 드라마 스타트랙을 보면, 원격 전송(teleportation)을 통해 사람을 멀리 떨어진 공간으로 전송을 시키는 모습을 볼 수 있는데, 이러한 원격 전송은 다양한 SF 영화 등에서 상상의 기술로 등장하지만, 양자역학의 세계에서는 양자얽힘을 이용하여 양자정보를 멀리 떨어진 곳으로 전송할 수 있다. 이때 SF 영화와 다른 것은, 양자전송에서는 원자와 같은 물질을 직접 원격 전송하는 것은 불가능하며 원자가 갖는 양자상태에 대한 정보(양자정보)를 전송할 수 있다는 것이다.

양자네트워크는 좀 더 넓은 범위의 일반적인 양자통신 네트워크를 의미한다. 단순히 송신자와 수신자 두 사용자가 참여하는 통신이 아닌 통신상에 퍼져 있는 많은 수의 노드(node) 간의 양자통신체계를 구축하는 것이라고 할 수 있다. 실제 고전정보통신에서도 일대일 통신뿐만이 아닌 일대다(多), 다대다 등의 많은 수의 참여자 사이의 통신이 이루어지고 있다. 이러한 양자네트워크를 구축하기 위해서는 각 노드들에서 양자상태 및 양자정보의 생성 및 제어, 측정이 가능해야 하며, 정보 전송, 처리 및 공유를 할 수 있어야 한다. 특히, 실제로 멀리 떨어진 두 노드 간에 양자통신을 하기 위해서는 두 노드 사이에 있는 중간 노드를 거쳐 양자통신이 이루어져야 한다. 일반적으로 양자네트워크는 더 큰 개념으로 양자암호통신과 양자전송 등을 포함한다고 볼 수 있으나, 양자네트워크는 아직 기초연구단계에 해당하며 각 노드 간 연결을 위해 필요한 양자기술이 앞의 두 분야와 다르므로 여기서는 따로 구분하여 분류하기로 한다.

양자통신에 필요한 다양한 요소기술을 아래 〈표 1〉에 정리하여 나타내었다.

표 1. 양자통신 기술 세부 분류

소분류	세분류	요소기술
양자통신	양자암호통신	양자광원 생성, 단일 광자 검출, 양자 난수 발생 기술, QKD 프로토콜, QKD 후처리 기술, QKD 시스템 기술, 양자 해킹 방지, 양자 서명/인증, QKD 안전성 연구, QKD 프로토콜 이론 등
	양자전송	양자얽힘 생성, 비국소적 양자측정, 양자얽힘 정제, 양자통신 오류정정, 양자전송 프로토콜 등
	양자네트워크	양자 중계기/메모리, 양자얽힘 교환, 양자 신호 파장 변환, 양자얽힘 기반 양자네트워크, 양자 스위치/라우터, QKD 네트워킹 기술 등

3. 양자컴퓨팅

20세기 중반에 현대적인 컴퓨터가 등장한 이후 컴퓨터의 연산 속도 및 연산 능력은 비약적으로 발전하였다. 20세기 말 개인용 컴퓨터(PC)가 등장하면서 가정에서도 컴퓨터를 사용할 수 있게 되었으며, 현재는 사람들이 사용하고 있는 휴대폰마저도 상당한 컴퓨팅 파워를 가지고 있다. 전자산업 및 정보기술의 비약적인 발전으로, 집적 회로의 소형화를 통해 더 많은 소자를 집적화 하여 컴퓨팅 파워를 증가시켜 왔으나 이러한 방법은 이미 10nm 이하의 구조를 갖는 회로 소자를 제작해야 하는 단계에 이르렀기에 동일한 방법으로 더 이상 집적화를 증가시킬 수 없는 수준까지 도달했다. 우리가 물질로 회로를 만드는 이상, 원자의 사이즈보다 더 작은 구조를 만들 수는 없기 때문이다. 또한, 이렇게 미시적인 세계로 들어오면, 양자역학적인 효과가 두드러지게 나타나게 되는데 이는 큰 구조에서는 나타나지 않았던 새로운 현상이 나타날 수 있음을 의미한다. 따라서 새로운 컴퓨팅 패러다임의 전환을 통한 연산 능력의 비약적인 향상을 도모해야 하는 상황이 되었다. 4차 산업혁명으로의 진입은 전보다 훨씬 더 빠른 속도의 연산 능력을 필요로 할 것이므로 고전컴퓨팅을 넘어 양자역학적인 현상을 이용함으로써 새로운 방식으로 연산 속도를 향상시키려는 연구가 활발하게 진행되고 있다.

양자컴퓨팅이 고전컴퓨팅에 비해 가장 다른 점은 바로 새로운 정보처리의 기본 단위를 사용한다는 것이다. 고전컴퓨터에서는 0과 1로 표현되는 비트가 기본 정보 단위이지만, 양자컴퓨터는 위에서 소개한 것처럼 0과 1의 중첩 상태로 표현되는 큐비트를 사용한다. 즉, 비트는 두 가지의 경우만 존재하는 반면에 큐비트는 0과 1의 비율과 0과 1 사이의 위상 차이 등 원리적으로는 무한대의 조합으로 정보를 표현할 수 있다. 그리고 다수의 큐비트들을 준비하여 이 큐비트들이 양자얽힘을 공유하게 만들면, 하나의 큐비트에 대한 조작을 통해 다른 큐비트에 영향을 줄 수 있다. 고전컴퓨터에서는 하나의 연산에서 하나의 결과만을 얻을 수 있지만, 양자중첩

및 양자얽힘을 이용하는 양자컴퓨터에서는 모든 조합의 결과값을 동시에 연산할 수 있다. 이러한 특성을 이용함으로써 특정 문제에 대해서는 양자컴퓨터가 고전컴퓨터에 비해 연산 속도가 기하급수적으로 빠를 수 있다는 것이 잘 알려져 있다.

양자컴퓨팅을 통해 연산 속도를 비약적으로 향상시킬 수 있는 연산의 대표적인 것으로는 소인수분해가 있다. 1994년에 피터 쇼어(Peter Shor)는 소인수분해를 효율적으로 할 수 있는 양자컴퓨팅 알고리즘을 제안했는데, 기존의 고전컴퓨팅을 기반으로 한 소인수분해 알고리즘이 숫자가 커짐에 따라 지수적으로 연산이 증가하는 반면, 쇼어의 알고리즘은 연산이 polynomial하게 비례하기 때문에 양자컴퓨터를 이용하여 소인수분해를 하면 연산속도를 기하급수적으로 빠르게 향상시킬 수 있다. 또한, 1996년 로프 그로버(Lov Grover)가 제안한 그로버 양자알고리즘은 정렬되지 않은 데이터베이스에서 원하는 특정 데이터를 찾는 알고리즘으로, 기존의 고전컴퓨팅 알고리즘의 경우 N개의 데이터 중에서 특정 데이터를 찾기 위해서 O(N)번의 시도가 필요하지만, Grover의 양자알고리즘을 사용하면 $O(\sqrt{N})$번의 시도로 찾을 수 있다. 이 외에도 다양한 양자알고리즘이 있으며, 특정 문제들에 대해서 고전컴퓨터로는 연산시간이 천문학적으로 오랜 시간이 필요하여 현실적으로 해결할 수 없는 문제들을 빠르게 계산 할 수 있다.

양자컴퓨팅 기술을 세부기술로 나누는 방식은 기준에 따라 다양할 수 있으나, 본 기고문에서는 크게 범용 양자컴퓨팅 기술과 양자시뮬레이팅 기술, 그리고 양자소프트웨어 기술로 나누어 설명하고자 한다. 범용 양자컴퓨팅은 양자알고리즘에 사용되는 범용 양자게이트인 Controlled-NOT 게이트, Hadamard 게이트, 위상 게이트 등을 이용하여 양자알고리즘을 모두 수행할 수 있는 완전한 양자컴퓨팅을 의미한다. 즉, 우리가 생각하는 일반적인 양자컴퓨팅을 의미한다고 볼 수 있다.

반면에, 양자시뮬레이팅 기술은 범용 양자게이트를 이용하지 않고 고전컴퓨팅으로 시뮬레이션하기 힘든 양자역학적인 현상을 양자시스템을 이용하여 시뮬레이션 하는 것을 말한다. 이러한 양자시뮬레이팅 기술은 일반적으로 범용 양자게이트를 만들기 어려운 물리 시스템에서 제한된 양자연산을 이용하여 특정 문제에 대하여 제한된 알고리즘만 효율적으로 수행하는 기술을 말한다. 그렇기 때문에, 범용 양자컴퓨팅과 달리 양자시뮬레이팅에서는 큐비트 기반의 디지털 컴퓨팅뿐만 아니라 아날로그 방식도 사용할 수 있다.

양자소프트웨어 기술은 범용 양자컴퓨팅 및 양자시뮬레이팅에서 문제를 해결하기 위해 필요한 방법론이 고전컴퓨팅과 다르기 때문에, 양자컴퓨팅 및 양자시뮬레이팅을 구동하기 위한 소프트웨어 및 이론 연구가 필요하다. 이러한 기술을 양자소프트웨어 기술이라 부른다. 이 분류에는 양자컴퓨팅을 이용하여 고전컴퓨팅에 비해 더 빠른 계산을 할 수 있는 양자알고리즘을 개발하는 연구도 포함된다.

양자컴퓨팅, 양자시뮬레이팅과 양자소프트웨어 분야에 필요한 요소기술을 〈표 2〉에 정리하여 나타내었다.

표 2. 양자컴퓨팅 기술 세부 분류

소분류	세분류	요소기술
양자컴퓨팅	범용 양자컴퓨팅	범용 게이트용 큐비트 기술, 오류 검출 기술, 오류 정정 기술, 양자 논리게이트 구현 기술, 논리 큐비트 구현 기술, 양자 특성측정 검증평가 기술, 개별제어 가능한 다중 큐비트 기술 등
	양자시뮬레이팅	양자 어닐링 기술, 양자 제어 기술, 다중 큐비트 기술, 시뮬레이션 기술 등
	양자소프트웨어	양자알고리즘, 양자컴퓨팅 소프트웨어, 양자컴퓨팅 이론 등

II 각국의 정책 동향

양자정보통신기술은 20세기 말에 기초연구가 시작되었고, 21세기에 접어들면서 기초연구 뿐만 아니라 실용적으로 인류 사회에 양자정보통신기술을 적용할 수 있는 가능성을 보여주는 연구결과들이 나오기 시작하고 있다. 이에 각국에서는 새롭게 등장한 양자정보통신 분야에서의 기술을 선점하기 위하여 다양한 정책을 도입함으로써 양자정보통신기술의 발전을 지원하고 있다. 본 챕터에서는 간략히 여러 국가의 양자정보과학 정책을 시행하고 있는지 소개하고, 우리나라의 정책 동향에 대해서도 간략하게 소개하고자 한다.

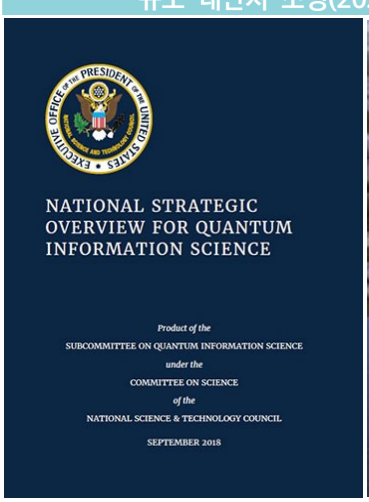
1. 해외 정책 동향

표 3. 양자정보분야 해외 각국의 정책 동향

국가	연구현황
미국	• 2008년 12월 양자정보과학비전을 수립하여 양자정보과학의 연구개발을 통합함으로써 미국의 양자정보기술개발을 위한 정책 마련 • 양자정보과학 분야를 미국이 선도하기 위하여 새로운 세대의 연구자에 대한 교육, 연구결과의 공유 및 협력을 위한 기반을 구축하는 등 10년 이상의 장기간 집중 지원 체계를 구축 • 2016년 7월에 '발전하는 양자정보과학'이라는 백악관 보고서에서 양자정보과학 분야의 기술발전을 저해하는 요인을 분석하였고, 이를 토대로 하여 미국의 향후 추진방안을 제시함 • 최근에는 상용화 수준의 양자정보통신 기술을 개발하기 위하여 5년간 8천억 원 이상의 예산을 집중투자하는 내용을 담은 '양자컴퓨팅 연구법 2018' 법률안 추진 중 • 양자기술 심화, 연구인력 양성 및 대규모 연구시설 마련 등을 통해 양자 산업을 육성하고자 하는 '국가 양자 이니셔티브' 법률안도 동시 추진 중
유럽연합(EU)	• 제2차 양자혁명에 대비하기 위하여 2016년 5월 유럽연합회원국 간에 합의문 체결하고, 양자정보통신 분야의 R&D 투자를 강화함으로써 2035년 이후까지 확보하고자 하는 양자정보통신 기술 로드맵 제시 • 2017년 2월, 유럽연합이 제2차 양자혁명을 선도하는 그룹이 되기 위하여 양자기술을 향후 10년간의 전략적 연구 의제로 설명(양자기술 플래그쉽 중간 보고서) • 2018년부터 10년간 약 10억 유로 규모의 양자기술 개발 프로그램 개시 • QuantERA 프로그램을 통하여 유럽연합 소속 25개국, 32개 연구기관 네트워크를 구성하여 양자기술과 관련된 공동연구 지원 • 양자정보기술 분야의 특성상 국가 간, 연구기관 간 공동 및 협동 연구에 대한 수요가 증가하고 있기 때문에 국가 또는 지역의 한계를 넘어 공동 펀드 프로그램을 구성하여 지원

국가	연구현황
영국	• 2013년부터 2017년까지 약 2억 7천만 파운드를 투자하는 국가 양자기술 프로그램 발족 • 영국 내 17개 대학과 50개 이상의 파트너 기관이 참여하는 양자기술 허브 구성 • 2015년 2월, '양자기술 국가 전략'에서 영국이 보유하고 있는 선두 양자기술 연구결과를 시장성이 있는 상품으로 발전시키기 위한 국가 전략을 제시하고, 이를 통해 양자정보기술 분야의 시장 진입을 위한 제품 로드맵 함께 제시
중국	• 5개년 국가 과학기술혁신 계획(2016-2020년, 13차 5개년 계획)에서 '국가 전략적 요구와 연관된 기본연구'에 '양자 제어' 분야를 포함하여 양자정보과학에 대한 우선순위 강화 • 2020년까지 기존 컴퓨터에 비해 연산 능력이 100만 배 이상 빠른 양자컴퓨터를 개발하는 것을 목표로 '국립 양자정보과학연구소' 설립 준비 중
일본	• 정보통신연구소(NICT) 내의 Quantum ICT Advanced Development Center에서 2035년까지의 중장기 기술로드맵을 제시하였고, 주로 광 기반의 양자정보이론 및 측정 등에 관한 연구를 추진 중 • 2017년 2월, 문부과학성의 '양자과학기술의 새로운 전개를 위한 추진방안'을 통하여 '광, 양자기술'을 새로운 가치창출의 기반기술로 선정하고 연구 지원 및 국제협력, 인재육성 등이 필요함을 강조
호주	• 2015년 12월, 5년간 2천 6백만 호주 달러를 투입하여 양자컴퓨팅 기술을 개발할 계획 발표 (국립 혁신 및 과학 의제) • 정부의 지원 하에 New South Wales 대학에 양자컴퓨팅 및 통신기술센터를 설립하고 전 세계의 파트너 기관과 공동연구를 통해 양자기술 연구 진행 중 • 이 센터는 2017년에 우수센터로 지정되었고, 3천 3백만 호주 달러를 추가로 지원받았으며, 참여 파트너 기관을 38개로 확대

그림 1. (좌) 미국의 양자정보 분야 전략 보고서. 미국 백악관 양자 회담(Quantum summit) (우) 중국의 양자정보과학국가연구소 조감도. 안후이성 허페이시에 10조 원, 37헥타르 규모 대단지 조성(2020년 완공 예정)



2. 국내 정책 동향

우리나라는 2014년 12월에 미래창조과학부에서 '양자정보통신 중장기 추진전략'을 수립하여 양자정보통신 분야의 기술개발, 인력양성 및 기반 조성의 필요성을 강조하였으며, 2016년 하반기부터는 '양자정보통신 중장기 기술개발' 정책을 기획함으로써 투자 확대를 추진하고 있다. 2019년에는 양자정보통신 기술발전을 위한 '정보통신 진흥 및 융합 활성화 등에 관한 기술특별법(ICT특별법)' 개정안이 마련되어 국회 통과를 기다리고 있다.

III 국내외 연구 동향

1. 양자통신

1.1. 해외 기술개발 동향

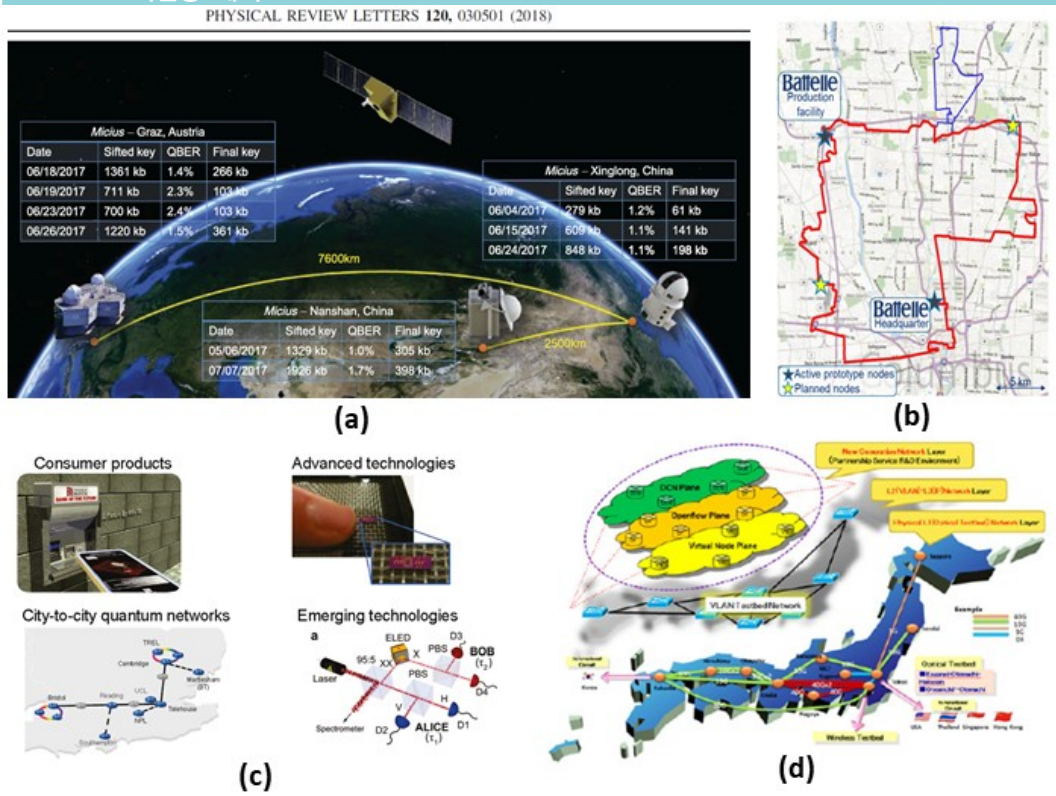
양자통신은 주로 양자암호통신 기술을 중심으로 활발한 연구개발 활동이 이루어지고 있다. 1984년 IBM의 Charles H. Bennett과 몬트리올 대학의 Gilles Brassard가 최초의 양자암호 프로토콜인 BB84를 제안한 이후 양자 응용기술 중 가장 비약적인 기술발전을 보여주고 있다. 1980년대와 90년대를 거치면서 이미 실험실 수준의 검증을 마쳤고, 2000년대에 들어서는 시스템 구현 및 시험망 테스트까지 미국, 중국, 일본, 스위스 등 여러 기술 선진국에서 성공적으로 수행했다. 2010년부터는 본격적인 상용화를 위한 시스템 고도화 및 응용 서비스 개발을 진행하여 몇몇 시범 서비스들이 제공되는 수준까지 이르렀다. 특히, 최근에는 중국의 연구개발 활동이 두드러지는데, 2016년 베이징-상해 간 2,000km 양자백본망 구축과 2018년 양자 위성 발사를 통한 중국-오스트리아 대륙 간 7,600km 양자통신 실험은 기술개발 측면에서 큰 이정표가 되는 사건이었다.

세계 각국의 주요한 기술개발 현황을 정리하면 아래 〈표 4〉와 같다.

표 4. 양자암호통신 해외 주요연구현황

국가	연구현황
중국	• 2007년 안후이성과 베이징을 연결한 양자암호 시험통신망 구축 • 2010년대에 Quantum Ctek, Qasky 등의 회사 설립을 통해 QKD 시스템 구현 • 2016년 베이징-상해 2,000km 구간을 32개의 신뢰연계점을 이용하여 양자백본망을 구축하고 시범서비스 제공 • 2016년 양자위성을 발사하고, 이를 이용하여 2018년에 중국-오스트리아 7,600km 대륙 간 양자암호통신 실현
미국	• 1984년 IBM의 Charles H. Bennett이 몬트리올 대학의 Gilles Brassard와 함께 BB84 프로토콜 제안 • 2004년 BBN tech는 DARPA 프로젝트를 통해 보스턴 및 하버드 대학을 연결한 최초의 QKD 네트워크 구축 • 2000년대 이후 LANL(Los Alamos National Laboratory)는 위성기반 양자암호통신, 초소형 양자암호통신 시스템 개발 • 2013년 Battelle는 오하이오州 콜럼버스와 워싱턴DC를 연결하는 770km 정도의 자체망에 양자암호통신 네트워크 구축
유럽	• 2002년 스위스는 세계 최초로 양자암호 시스템 장비 회사인 ID Quantique를 설립하고 상용시스템 출시 • 유럽 12개국, 40여 개의 기관, 대학 및 기업의 참여로 이루어진 SECOQC project를 통해 2008년 주요국 QKD 네트워크 보유 및 국가 간 연결 • 2009년 Toshiba 유럽연구소(영국)에서 1GHz급 QKD 시스템을 구현하여 100km 거리에서 10.1kbps의 암호키 생성 성공 • 영국은 2015년부터 양자통신허브를 구축하여 4개의 대형 업무 패키지를(Work Package) 분류하여 WP1: Short Range Consumer QKD, WP2: Chip Scale QKD, WP3: Quantum Networks, WP4: Next Generation Quantum Communication 등을 연구하고 있으며 본 프로그램을 통해 City-to-City 퀀텀 네트워크 구축
일본	• 2009년에 미국 스탠퍼드 대학교, Hamamatsu, 일본 NII 및 NTT가 공동으로 DPSK 프로토콜 기반의 시스템을 구현하여 10km 거리에서 1.3Mbps의 키생성 • 2010년에는 일본 NICT 주관으로 NEC, NTT, Mitsubishi, Toshiba, All-Vienna, ID Quantique 등이 참여하여 도쿄에 6개 노드로 이루어진 QKD 네트워크(Tokyo QKD network) 구축 • NICT, NTT, NEC 등은 Tokyo QKD 네트워크의 후속 연구로 장시간 운용, DPS 프로토콜의 안전성 검증, CV-QKD 연구를 진행하고 있으며, 2차 QKD 시범 네트워크 구축 중

그림 2. 양자암호통신 해외 주요연구현황 (a) 중국 인공위성 기반 양자암호통신 (b) 미국 오하이오 州 바텔 양자암호 시험망 (c) 영국 양자정보통신 연구 허브 현황 (d) 일본 양자암호통신 시험망 계획



1.2. 국내 기술개발 동향

양자정보 분야 기초 이론 연구가 대학을 중심으로 1990년대 이후 이루어졌고, 이를 바탕으로 본격적인 양자암호통신 연구는 2000년대 중후반부터 한국과학기술연구원(KIST), 한국전자통신연구원(ETRI) 등 국내 정부출연연구소가 중심이 되어 시작되었다. 다른 기술 선진국에 비해 연구개발을 뒤늦게 시작했고, 정책적으로 주목을 받지 못한 연구였기 때문에 소규모 산발적인 연구 활동에 머무르는 수준이었다. 그러다가 2010년대부터 전 세계적으로 양자기술이 주목받기 시작하면서 국내에서도 본격적인 연구 활동이 시작되었다. 한국과학기술연구원은 2013년 플러그앤플레이(plug-and-play) 방식의 유선 QKD 시스템을 세계양자암호 학회에서 발표 및 시연에 성공했다. 그리고 2018년 단대단 방식을 뛰어넘어 일대다 양자암호 네트워크 시스템을 개발하고, 이를 서울

지역에 포설된 광케이블망에서 시험 검증하였다. ETRI는 2000년대 후반에 실험실 수준에서 양자키분배 시스템을 구현하는데 성공한 것을 시작으로, 2017년 무선 QKD 핵심 부품을 소형 칩으로 개발하고, 이를 이용한 무선 QKD 야외 실험에 성공했다. 특히 야간뿐 아니라 주간에도 QKD 동작을 보이고 계속해서 초소형 QKD 시스템 구현 연구를 수행 중이다. 국가보안기술연구소는 QKD 시스템의 안전성 평가 기준을 만들기 위한 노력을 진행 중이며, 그 일환으로 국내 표준화를 위한 활동을 활발히 진행하고 있다. 국내에서는 통신 대기업에서 관련 기술에 대해 활발한 연구개발 활동을 진행하고 있다. SKT는 2011년 퀀텀랩을 중앙연구소 내에 설치하고, 2014년 QKD 시스템을 개발, 2016년 분당사옥과 용인집중국 간 68km 구간(왕복) 등 총 5개 구간에 양자암호통신 국가시험망을 구축하였으며, 4G LTE망에 QKD를 적용하는 등, 상용화 기술개발을 선도하고 있다. KT는 한국과학기술연구원과의 협력을 바탕으로 일대다 QKD 시험망을 구축했고, 특히 국제적인 표준화 활동을 활발히 진행하고 있다.

국내 주요한 기술개발 현황을 정리하면 다음 〈표 5〉와 같다.

표 5. 양자암호통신 국내 주요연구현황

기관	연구현황
한국과학기술연구원 (KIST)	• 2012년 양자정보연구단 설립 • 2013년 세계양자암호 학회에 플러그앤플레이 방식의 QKD 시스템 발표 및 시연 • 2017년 무선 양자암호 시스템 구현 • 2018년 유선 1xN 양자암호 네트워크 시스템 개발 및 시험망 검증 • 현재 LiNbO3 기반 QKD 칩 기술 개발 및 양자인증/서명 연구 중
한국전자통신연구원 (ETRI)	• 2005년 25km 유선 양자암호 시스템 실험실 검증 • 2017년 초소형 칩 기반 무선 양자암호 시스템 구축 및 주야간 동작 검증 (@100m) • 현재 Si, SiO2 기반 QKD 칩 및 양자광원/검출 소자 기술 개발 중
국가보안기술연구소	• 양자암호 시스템 안전성 평가 기준안 개발 • 양자암호 국내 표준화 연구
SKT	• 2011년 퀀텀랩 설립 • 2014년 현대암호 장비와 연결한 QKD 시스템 구현 • 2016년 양자암호 테스트베드 구축 • LTE 망에 양자암호 적용 및 서비스 실현 • 초소형 QRNG 개발 및 상용화 • 2018년 스위스 IDQ 기업 인수 • 2019년 5G 망에 양자암호 적용
KT	• 2018년 양자암호 네트워크 시험망 구축 • 2019년 ITU 양자암호 표준화 활동

2. 양자컴퓨팅

2.1. 해외 기술개발 동향

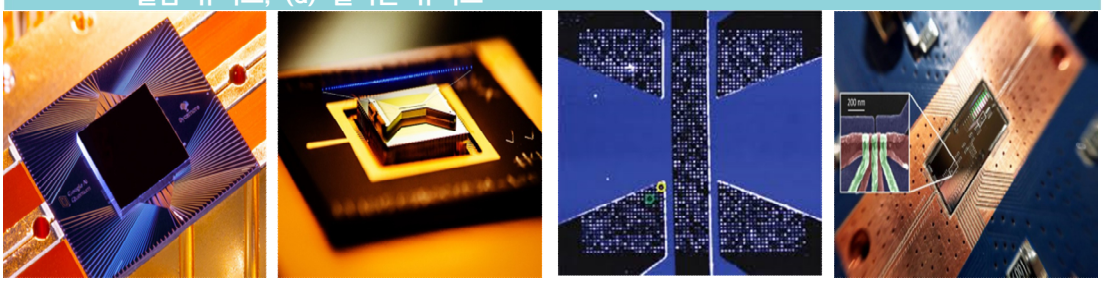
2.1.1. 범용 양자컴퓨팅 연구현황

양자컴퓨팅 기술은 미국, 유럽, 중국, 일본 등 세계 각국에서 전략기술로 인식하고 범정부 차원의 계획을 수립하여 기술선점 경쟁이 가속화되고 있다. 다만, 미국을 비롯한 어느 나라도 현재 양자컴퓨터를 완성할 기반 기술을 가지고 있지 못하며 앞으로 개발해야 할 이론 및 원천 구현 기술이 더 많이 남아 있는 것으로 여겨지고 있다. 이러한 기술 상황에 따라 산업계 연구도 포함하여 특정 기관 중심의 폐쇄적 연구보다는 외부 연구자들의 아이디어를 적극 수용하고, 대규모 공동연구를 수행하는 추세이다. 또한, 현재 여러 가지 플랫폼 중 어떠한 시스템이 최종적인 양자컴퓨터의 형태가 될 것인지는 아무도 예단하기 어려운 상황이다. 따라서, 여러 가지 물리적 큐비트 시스템이 경쟁적으로 연구되고 있으며 주요 플랫폼별 연구 동향은 아래와 〈표 6〉과 같다.

표 6. 범용 양자컴퓨터를 위한 물리 큐비트 해외 주요연구현황

물리 큐비트	연구현황
초전도 큐비트	• 중국 USTC에서는 12개 초전도체 큐비트 기반 12개 큐비트의 진성 양자얽힘 상태를 신뢰도 70%로 구현 • 스타트업 기업인 Rigetti에서는 8큐비트 양자프로세서 칩에서 2큐비트 양자얽힘을 99.2%의 높은 신뢰도를 달성 • 최근 구글은 양자우월성을 달성했다고 보고
이온덫 큐비트	• 미국의 양자컴퓨팅 스타트업 기업인 IonQ는 11개의 이온 큐비트를 이용한 범용 양자프로세서 개발을 발표하고, 단일 큐비트 게이트 연산 평균 신뢰도 99.5%, 그리고 2큐비트 게이트 연산 평균 신뢰도 97.5% 달성
고체점결함 큐비트	• 2019년 다이아몬드 내 탄소 동위원소 핵스핀 큐비트 7개를 NV 센터 보조 큐비트를 이용하여 진성 양자얽힘 상태를 생성하고 신뢰도 60% 달성 • 네덜란드 TU Delft 그룹은 서로 떨어진 다중 큐비트 양자노드를 광자 큐비트를 활용하여 연결할 수 있음을 보여줌
실리콘 큐비트	• 호주 UNSW에서 실리콘 큐비트 역시 2큐비트 양자게이트의 구현 가능함을 보여줌

그림 3. 범용 양자컴퓨팅을 위한 주요 플랫폼: (a) 초전도 큐비트, (b) 이온덫 큐비트, (c) 고체점 결함 큐비트, (d) 실리콘 큐비트



(a) (b) (c) (d)

2.1.2. NISQ 양자컴퓨팅 기술과 양자시뮬레이터 연구현황

앞서 살펴본 바와 같이 양자컴퓨팅 연구는 지난 10년간 괄목할 만한 성과를 이루었음에도 실용적인 활용에 적용하지는 못하는 상황이다. 양자컴퓨터는 큐비트의 개수를 확장 시키더라도 양자 에러에 의해 그 성능을 제한받게 된다. 따라서 연산 과정에서 필연적으로 발생하는 양자 에러를 최소화하고 이를 정정하는 과정이 필요하다. 이러한 양자오류정정이 구현되어 있는 에러 내성형 양자컴퓨터(Fault-tolerant quantum computer)가 궁극적으로 지향해야 할 양자컴퓨터의 형태이다. 다만, 아직까지 기존의 디지털 기술 기반의 슈퍼컴퓨터로 모사할 수 없는 양자우월성을 보여줄 수 있는 50큐비트급 이상의 양자컴퓨터에 양자오류정정 기술이 적용된 적은 없다. 현재의 기술 수준은 양자 우월성 기준을 충족할 수 있는 50큐비트를 생성할 수는 있으나 이러한 대규모 시스템에서 높은 신뢰성을 가진 연산을 구현하기는 어려운 실정이다. 이러한 현재의 양자컴퓨터 기술 수준을 연구자들은 Noisy Intermediate-Scale Quantum(NISQ) 양자기술이라 부르고 있다. 즉, 큐비트의 개수가 많아 기존의 슈퍼컴퓨터로는 모사할 수는 없지만 아직은 에러율이 높아 신뢰성이 높은 연산을 수행할 수 없는 중간 단계의 기술 수준으로 보고 있다.

따라서 많은 연구자들이 현 기술 수준인 NISQ 양자컴퓨팅 기술 수준에서 필요한 다양한 연구를 진행하고 있다. 먼저, 꾸준히 큐비트 개수를 확장하기 위한 다양한 노력을 기울이고 있다. 이는 양자컴퓨터의 성능지표를 결정짓는 가장 큰 요인이기 때문이다. 그렇지만, 큐비트의 개수만으로 양자컴퓨터의 성능을 결정할 수 없기 때문에 양자 에러를 줄이기 위한 연구 노력 역시 중요한 연구 분야이다. 이러한 전통적인 연구목표 이외에도, NISQ 기술단계에서 실용적인 문제를 해결해 보고자 하는 시도들이 생겨나고 있다. 에러 내성을 가진 범용 양자컴퓨터는 우리가 알고 있는 이론적인 양자알고리즘들을 모두 구현할 수 있는 궁극적인 기술이지만, 현재의 수준인 에러를 가지고 있는 NISQ 양자컴퓨팅 기술 역시 제한적이지만 실용적인 쓰임새가 있을 것으로 많은

연구자들이 기대하고 있다. 따라서, 제한적인 NISQ 양자컴퓨팅 플랫폼을 활용하여 실질적인 문제에 활용하고자 하는 연구 노력이 최근 활발히 이루어지고 있다.

2.2. 국내 기술개발 동향

국내의 양자컴퓨팅에 관한 연구는 주로 광자 큐비트를 기반으로 한 양자정보 기초연구를 중심으로 진행되어 왔다. 2010년대 이후 신진 인력이 국내에 들어오면서 기초연구 위주에서 물리 큐비트 구현기술에 대한 연구가 진행되고 있으며, 일부 양자컴퓨팅 이론 및 아키텍쳐 연구를 진행하는 그룹이 생기기 시작했다. 또한, 범용 양자컴퓨팅을 위한 물리 큐비트 구현기술에 대한 연구뿐만 아니라, 현재의 양자기술을 실용적인 문제에 적용하기 위한 양자시뮬레이팅 연구의 중요성 및 관심 역시 증대되고 있는 추세이다.

2.2.1. 범용 양자컴퓨팅을 위한 물리 큐비트 구현

현재 국내의 경우에도 양자컴퓨팅 구현을 위한 다양한 주요 물리 큐비트 구현기술 연구가 진행 중이다. 범용 양자컴퓨팅 기술을 구현하기 위해서는 양자컴퓨터를 구성하는 물리 큐비트의 성능지표 향상이 필수적이며, 이를 위해 국내에서는 소규모 5큐비트 정도의 물리 큐비트를 구현하고 성능지표 향상을 위한 개발 연구가 진행 중이다. 이와 관련된 연구현황을 정리하면 〈표 7〉과 같다.

표 7. 범용 양자컴퓨터를 위한 물리 큐비트 국내 연구현황

물리 큐비트	연구현황
초전도 큐비트	• 해외에서는 주로 IT 기업 중심으로 연구되고 있는 초전도 큐비트 기반 양자컴퓨팅 기술 연구는 국내에서는 한국표준과학연구원 중심으로 이루어지고 있음 • 최근 두 개의 Transmon 초전도 큐비트를 구현하였으며, 단일 큐비트 게이트 신뢰도 99.5%, 두 개의 큐비트 사이의 다중 큐비트 양자 게이트의 신뢰도 65% 수준 달성
이온덫 큐비트	• 서울대학교에서는 Yb+ 이온을 이용한 양자컴퓨터 및 양자정보 연구를 수행 • 이온덫 큐비트의 확장을 위해 MEMS 기반 이온덫 칩을 개발하여 Yb+ 이온 포획 성공 후 단일 큐비트 게이트 구현 • 이온 큐비트 사이 2큐비트 양자게이트는 연구개발 수행 중
고체점결함 큐비트	• 한국과학기술연구원은 다이아몬드 Nitrogen-Vacancy(NV) 센터와 같은 고체 내 점결함 큐비트를 활용한 양자컴퓨팅 연구 수행 중 • NV 전자스핀과 주변의 핵스핀이 클러스터 형태로 집적된 큐비트 소자 제작, 제어, 측정 기술 보유 • 단일 큐비트 초기화 신뢰도 97% 수준 달성한 바 있으며 단일 큐비트 게이트 구동 시연 • 확장 가능한 양자컴퓨팅 시스템을 위한 광자기반의 양자인터페이스 역량 보유 • 효율적인 양자컴퓨팅 프로세스 분석 원천기술 보유
실리콘 큐비트	• 서울대학교 및 한국전자통신연구원 연구진은 최근 실리콘 큐비트 기술개발 시작

2.2.2. 양자시뮬레이팅 연구현황

범용 양자컴퓨팅 기술 전 단계 기술로써, 양자시뮬레이팅 기술은 양자기술을 적용하여 실용적인 문제에 활용하는 기술을 지칭한다. 대표적인 예로 기존 머신러닝에 필요한 계산을 효율적으로 할 수 있는 양자머신러닝, 그리고 분자 구조 및 전자 구조를 계산할 수 있는 양자계산화학 분야가 있다. 또한, 대규모 양자시스템을 구현하고 이에 대한 양자물리 현상을 모사하는 분야 역시 전통적인 양자시뮬레이션 연구로 진행되어 왔다.

국내에서 연구되고 있는 주요 양자시뮬레이터 플랫폼으로는 광자 기반 그리고 중성원자 시스템을 기반으로 하는 연구가 있다. 광자 큐비트의 경우 선형광학계가 가지는 확률적 연산의 어려움 때문에 범용 양자컴퓨팅으로써의 단점이 있지만, 하나의 광자에 고차원의 양자상태를 인코딩할 수 있어 양자시뮬레이터로 활용 가능성이 높다고 할 수 있으며, 이러한 특성을 이용하여 확장 가능한 양자시뮬레이터를 개발하고 실용적 문제에 활용하고자 하는 연구가 진행 중이다. 중성원자 시스템의 경우 연산 신뢰도가 낮은 단점이 있지만, 수백 개에 이르는 다량의 큐비트를 쉽게 확보할 수 있는 뚜렷한 장점이 있어 다체계 물리 현상과 같은 기존의 디지털 컴퓨터로는 모사할 수 없는 양자현상을 시뮬레이션하여 그 특성을 분석하는 연구가 진행 중이다.

표 8. 국내 양자시뮬레이터 연구현황

양자 시뮬레이터 플랫폼	연구현황
광자 기반 양자시뮬레이터	• 양자광집적회로에 필요한 간단한 양자간섭계에 대한 연구는 간간이 학회에서 보고된 바 있으나 아직 국내에서는 양자게이트를 포함한 복잡한 양자광집적회로를 이용한 양자시뮬레이터에 대한 연구결과는 없음 • 한국과학기술연구원, 포항공과대학교, 아주대 연구팀은 고차원 양자상태를 활용한 확장 가능한 광자 기반의 집적 양자시뮬레이터 플랫폼 연구를 진행 중 • 또한, KIAS, 한양대, 성균관대 등의 이론 연구진은 양자머신러닝 및 양자계산화학 양자시뮬레이션 문제와 같은 실용적 문제를 광자기반의 양자시뮬레이터를 활용하기 위한 알고리즘 연구를 수행 중
중성원자 기반 양자시뮬레이터	• 한국표준과학연구소, 서울대학교, KAIST 연구팀은 초저온 보스아인슈타인 응축체 (BEC: Bose-Eistein condensate)를 광격자에 넣어 다체계 응집물리현상을 시뮬레이션 하는 연구를 수행 중 • 또한, KAIST 연구팀은 광 트위저 트랩을 이용한 기술로 중성원자의 결정적 로딩 및 양자얽힘 상태 생성 성공

IV 산업계 동향

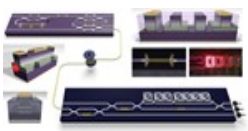
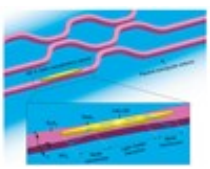
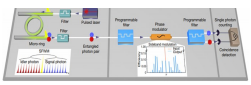
1. 양자통신

1.1. 해외 동향

양자통신은 현재 부분적으로 상용화 시장이 형성되어 있다. 양자암호 시스템 장치는 이미 2000년대부터 상용화가 시작되어 현재는 시스템 고도화 및 저가격화를 통한 장비 시장 확대를 도모하고 있다. 아울러 차별화된 새로운 보안 서비스를 개발하여 시장을 넓혀나가고자 하는 노력이 이루어지고 있다. 본격적인 상용화를 위해서는 기술적으로 거리 제한 문제를 해결할 수 있어야 하고, 시스템의 저가격화를 위한 추가 개발도 필요하며, 동시에 시스템 및 서비스에 대한 표준화가 선결되어야 한다. 따라서, 산업계에서는 거리 문제를 단기간에 해결하기 위해 임시방편으로 사용할 수 있는 신뢰연계점 기술을 개발하고 있으며, 시스템의 저가격화를 위해 양자집적화 칩 기반의 시스템 연구가 활발히 진행되고 있다. 그리고 지속적인 표준화 활동을 진행하고 있는데 대표적인 사례들을 소개하면 다음과 같다.

2019년 스위스 IDQ(SKT)는 스위스 제네바, 독일 베를린, 스페인 마드리드, 오스트리아 비엔나 등 유럽 주요국의 14개 구간 (1구간에 100km)에 양자암호 시험망을 구축하여 유럽 전역을 연결하는 프로젝트에 착수했다. 이 프로젝트에는 도이치텔레콤, 오렌지, 노키아, 애드바 등 이동통신사와 통신 장비사는 물론 정부, 대학의 연구기관까지 총 38개의 파트너가 참여하는 대규모 프로젝트로 본격적인 상용화의 가능성을 실증할 것으로 기대된다. 시스템 저가격화, 소형화 그리고 안정적인 동작을 위한 양자집적화 칩 연구는 현재 학계를 중심으로 이루어지고 있지만 산업계에서 이를 활용하고자 하는 움직임이 나타나고 있다. 주요 연구 결과를 살펴보면 아래 표와 같다.

표 9. 양자집적소자 기반 양자통신 주요 해외 연구

발표 논문	Nat. nano. 13, 835 (2018)	Nat. comm. 8, 13984 (2017)	Nat. comm. 8, 889 (2017)	Nat. 546, 622 (2017)
연구 개념	칩 기반 싱글포톤 소스	칩 기반 양자암호분배	on-chip 퀀텀 닷	on-chip 양자얽힘광원
대표 그림				
연구 그룹	영국 셰필드 대학	영국 브리스톨 대학	미국 표준연구소	캐나다 국립과학연구기관

표준화와 관련해서는 ETSI, ISO, ITU 등의 표준화 기구를 통해 주요국들이 주도권을 잡기 위해 치열한 논의를 진행하고 있다.

1.2. 국내 동향

국내에서는 통신 사업자 중심으로 산업계가 활발한 활동을 보여 주고 있다. 특히 SKT는 공격적인 연구개발 투자를 통해 LTE망, 5G 망에 양자암호 기술을 적용하여 차별화된 보안 서비스를 마케팅 포인트로 활용하고 있다. IDQ 인수를 통한 QKD 시스템뿐 아니라 핵심 부품인 QRNG도 상용 제품을 출시하는 등 공격적인 모습을 보여 주고 있다. KT도 한국과학기술연구원과의 협력을 통해 양자암호 네트워크 시험망을 설치하고, 선제적인 표준화 활동을 통해 기술을 리딩하기 위한 노력을 지속하고 있다. 또한 우리로광통신, EYL, 아이에이네트웍스, 우리넷, 텔레필드 등 여러 중소기업들이 양자암호 핵심 소자 개발, 양자암호 호환 전송장비 개발 등을 진행하고 있다. 양자암호 관련 제품의 인증을 위해 표준화 활동도 이루어지고 있는데 그 성과로 TTA에서 양자키분배 시스템 관련 표준화가 속속 이루어지고 있다.

표 10. 양자암호통신 국내 산업계 현황

기관	연구현황
SKT	• 2011년 퀀텀랩 설립 • 2014년 현대암호 장비와 연결한 QKD 시스템 구현 • 2016년 양자암호 테스트베드 구축 • LTE 망에 양자암호 적용 및 서비스 실현 • 초소형 QRNG 개발 및 상용화 • 2018년 스위스 IDQ 기업 인수 • 2019년 5G 망에 양자암호 적용
KT	• 2018년 양자암호 네트워크 시험망 구축 • 2019년 ITU 양자암호 표준화 활동
우리로광통신	• 양자암호 시스템용 단일광자 검출 소자 및 검출기 개발
EYL	• QRNG 개발
아이에이네트웍스	• QRNG 패키징 기술 개발
우리넷	• 양자암호 호환 전송 장비 개발
텔레필드	• 양자암호 호환 전송 장비 개발

2. 양자컴퓨팅

2.1. 해외 동향

범용 양자컴퓨팅 기술은 IBM, 구글, 마이크로소프트와 같은 전통적인 IT 기업의 적극적인 투자로 인해 견인되어 왔다. 이러한 IT 기업들은 학계 위주의 기초 연구에서 탈피하고 실제 양자컴퓨팅 기술을 활용할 수 있는 대규모 큐비트 구현기술에 중점을 두고 있다. 이 중 선두 주자라고 할 수 있는 IBM과 구글은 초전도 방식의 큐비트 기술을 기반하고 있다. 다른 대표적인 물리큐비트 구현기술은 미국의 IonQ가 중점적으로 연구를 진행하고 있는 이온덫 큐비트 방식이다. 다만 양자컴퓨터는 큐비트 개수를 확장하는데 있어 양자프로세서 내부에서 양자역학적 성질을 유지할 수 있어야만 양자컴퓨팅을 실질적 활용이 가능할 전망이다. 따라서 양자컴퓨팅 프로세서의 큐비트 개수뿐만 아니라 연산정확도를 함께 고려해야만 양자컴퓨팅 시스템의 종합적 평가가 가능하다. 초전도 방식의 큐비트 기술은 기존 반도체 기술을 활용하여 큐비트 소자를 만들 수 있는 장점이 있어 큐비트 개수를 확장하는데 용이한 것으로 알려져 있는 반면, 이온덫 방식의 큐비트는 가장 높은 수준의 연산 정확도를 가지고 있어 두 시스템간의 경쟁이 치열하다고 볼 수 있다.

현재 산업계 위주의 양자컴퓨팅 연구는 양자우월성(Quantum supremacy) 달성을 보여주기 위한 경쟁이 뜨겁다. 양자우월성은 특수 조건에서 현존하는 슈퍼컴퓨팅 기술을 뛰어넘는 것을 보여주는 이정표이다. 최근 구글은 최초로 샘플링 문제에서 최초로 양자우월성을 입증했다는 논문을 발표하였다. 이에 대해 IBM은 구글이 보여 준 문제에 대해서 고전 알고리즘을 최적화함으로써 기존 슈퍼컴퓨팅 기술로 풀 수 있는 문제라고 지적하는 등 양자우월성 달성 이정표를 먼저 달성하기 위한 산업계 간의 경쟁 역시 치열하다고 볼 수 있다.

표 11. 양자컴퓨팅 해외 산업계 현황

미국	
IBM	초전도 큐비트을 기반으로한 양자정보처리 프로세서를 개발하고 있으며, 16큐비트 양자컴퓨팅 디바이스를 클라우드 서비스로 오픈
구글	UCSB의 John Martinis 초전도 큐비트 기반 양자컴퓨팅 연구그룹을 합류시켜 연구개발을 수행중이며, 72큐비트 양자컴퓨터 개발에 성공했다고 홍보하고 있음
마이크로소프트	양자오류에 강하다고 알려진 토폴로지컬 양자컴퓨터를 개발 중이나 아직 뚜렷한 성과를 말하기는 어려운 단계
IonQ	메릴랜드 대학, 듀크 대학을 중심으로 창업한 양자컴퓨팅/양자시뮬레이션 스타트업 기업으로 최근 11큐비트 범용 양자컴퓨팅 플랫폼을 구현하였으며 이를 양자화학 시뮬레이션에 적용
PsiQuantum	영국 브리스톨 대학의 우수한 연구를 바탕으로 집적 양자광학계 기반 양자시뮬레이션/양자컴퓨터 하드웨어 개발 중
캐나다	
D-Wave	초전도 기반 양자 어닐링 머신을 개발하였으며, 이를 양자컴퓨팅/양자시뮬레이션 응용에 사용하기 위한 연구개발을 진행 중
Xanadu	신생 스타트업으로 광자기반 양자컴퓨터 및 응용기술 연구개발 중

2.2. 국내 동향

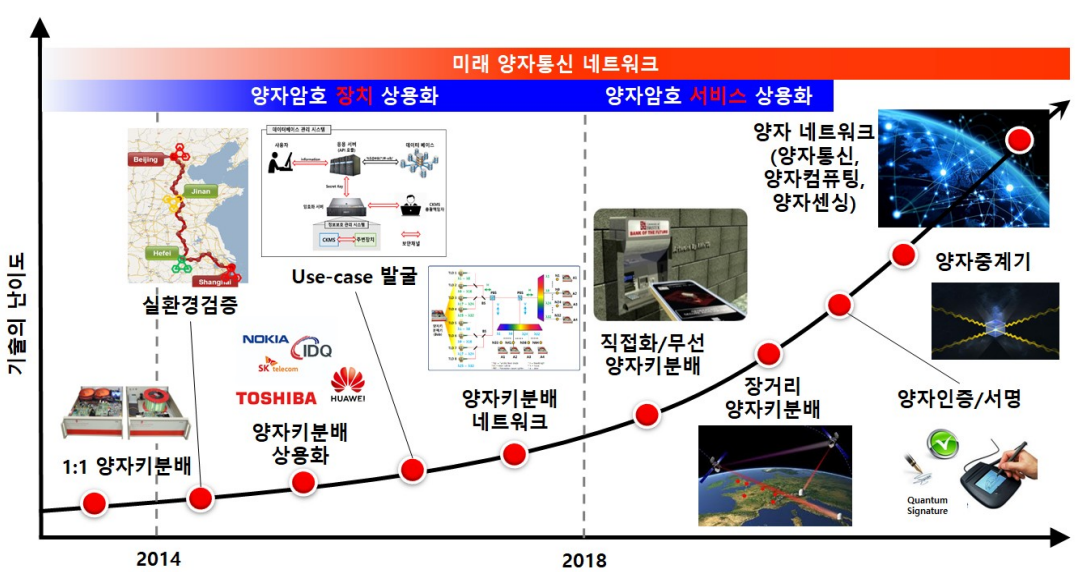
국내에서는 양자컴퓨팅에 산업계 기반의 연구 환경은 아직은 미흡하다고 할 수 있으나, 최근 미래 기술인 양자컴퓨팅 연구의 중요성을 많은 대기업이 인식하기 시작하였다. 특히 삼성전자는 IBM과 파트너십을 맺고 초전도체 기반의 큐비트 소자 개발을 착수한 것으로 알려져 있다. 또한, 현대자동차, LG화학, LG전자 등의 국내 주요 대기업 역시 양자컴퓨팅 기술의 활용 가능성을 염두해 두고 관련 연구팀이 신설될 것으로 알려져 있다. 이러한 현황을 비추어 보아 국내 산업계의 구체적인 연구개발 수준은 아직 알려진 바는 없으나 국내 산업계 역시 미래 양자컴퓨팅 기술의 미래가치에 대해 인식한 것으로 볼 수 있다.

Ⅴ 양자정보통신 분야의 향후 해결과제 및 전망

지금까지 양자정보통신 기술의 개념과 각국의 정책 동향, 국내외 연구 동향 및 산업계 동향에 대해서 알아보았다. 본 챕터에서는 양자통신 및 양자컴퓨팅 분야의 향후 해결과제 및 전망에 대해서 간략하게 서술하고 본 기고문을 마무리하고자 한다.

1. 양자통신 분야의 향후 해결과제 및 전망

그림 4. 양자통신 기술발전 전망



양자통신에서 가장 활발하게 개발이 진행되고 있는 양자암호통신 분야는 주로 실용적인 1 x 1 양자키분배 시스템 개발이 집중적으로 진행되었다. 특히, 포설된 광케이블 상에서 실제 통신이 이루어지는 다양한 환경에서 양자암호통신을 구현하고 실증하는 개발 연구가 활발히 진행되고 있다. 점차 자유 공간에서 인공위성을 통해

양자암호통신을 하는 방법에 대한 연구가 활성화되고 있으며, 광섬유를 기반으로 한 양자암호통신에서는 양자암호를 사용할 수 있는 거리를 증가시킴으로써 국가 내에서 주변 국가 간, 그리고 멀리 떨어진 국가 간에 양자암호통신망을 구축하는 연구가 추진될 것으로 예상된다. 최근에는 다양한 극한 상황에서도 문제없이 양자암호통신망을 구성하는 연구도 진행되고 있다 (해저 양자암호통신 등).

이러한 연구개발들의 결과로 단기적으로는 하나의 도시 내에서는 실용적인 양자암호망이 도입될 수 있을 것으로 보이며, 중장기적으로는 현대암호기술과 결합하여 실제로 양자암호를 서비스하는 시장이 생길 것으로 예상된다. 이를 위해 새로운 Use case를 발굴하는 노력이 더 활발히 이루어져야 하며 특히 양자키 분배 거리의 증가뿐만 아니라 양자암호시스템의 소형화 및 저가격화를 위한 기술 개발을 가속화해야 한다.

장기적인 연구로는 실용적인 양자네트워크망을 구축하기 위해 필요한 양자중계기를 개발하기 위한 핵심 원천기술 확보를 위한 연구가 활발하게 진행될 것으로 전망된다. 양자중계기가 개발된다면, 현재 약 100km 정도가 통신거리의 한계인 광섬유 기반의 양자통신이 거리의 한계를 넘어 중거리 통신에서 장거리 양자통신을 가능하게 할 것으로 예상된다. 그리고 개발된 양자중계기를 적절하게 배치한다면, 장기적으로 도시 내의 양자네트워크망을 넘어 우리나라 내의 양자네트워크망을 완성할 수 있을 것이다. 이처럼 양자중계기 기반의 양자네트워크망이 구축된다면 양자정보통신 분야에서 큰 패러다임의 변화가 올 것으로 예상되며, 다양한 상업적 양자암호통신 서비스를 제공하는 시장이 생길 것으로 기대된다.

2. 양자컴퓨팅 분야의 향후 해결과제 및 전망

양자컴퓨팅 분야에서는 지금까지 주로 기초연구분야에서 많은 연구가 이루어져 왔으나 최근 들어 해외 대기업 (구글, IBM)과 다양한 해외의 스타트업 회사들이 일부 상용 서비스도 시작하고 있는 단계이다. 그렇지만 현재 상용 서비스의 수준을 감안하면 의미 있는 수준에서 실용적인 상용 서비스를 제공하기까지는 최소 5~10년 정도의 시간이 필요할 것으로 예상된다. 국내의 경우 아직까지 경쟁력 있는 상용 서비스를 제공하거나 국제적으로 양자컴퓨팅 기술을 선도하고 있는 회사는 아직 없는 상황이나, 아직은 회사들이 수익성을 내고 있는 단계가 아닌 기초연구 단계이므로 정부 주도적인 장기 투자가 이루어진다면 약 10여 년 후에는 해외 기업들과 경쟁할 수 있는 기반이 만들어질 수 있을 것으로 기대하고 있다.

현재 진행되고 있는 양자컴퓨팅 연구는 아직 불완전한 부분이 많다. 실제 양자컴퓨터가 제대로 동작하기 위해서는 오류가 발생했을 때 정정할 수 있는 양자오류정정이 가능한 논리가 필요한데, 이를 아직 사용하고

있지 않기 때문이다. 즉, 현재의 양자컴퓨팅은 물리 큐비트 기반의 양자컴퓨팅이며, 완전한 의미의 양자컴퓨터로 거듭나기 위해서는 논리 큐비트 기반의 양자컴퓨팅 기술 개발이 필수적이다. 이는 앞으로 약 10년 이상의 시간이 걸릴 것으로 예상된다. 그렇기 때문에 양자컴퓨팅 분야에서 현재 연구자들이 논리 큐비트의 개발과 함께 중요하게 생각하는 연구방향이 있는데, 이것은 향후 5년 정도의 단기간에는 불완전한 물리 큐비트을 기반으로 하는 양자컴퓨터를 이용하여 양자시뮬레이터로써 활용하는 것이다. 이것이 위에서 소개한 NISQ 양자컴퓨팅이라 하는데, 현재 가지고 있는 양자컴퓨터의 불완전함을 감안하면서도 고전컴퓨터가 할 수 있는 계산보다 더 나은 계산능력을 보이는 문제를 찾고, 이를 양자컴퓨터를 이용하여 계산하려는 것이다. NISQ 양자컴퓨팅에서 다루고자 하는 문제의 사이즈는 큐비트 약 10~100개 정도의 중규모 양자컴퓨팅이며, 현재 가지고 있는 양자컴퓨터가 갖는 특성에 맞는 좋은 연구 주제를 발굴하는 방향도 논리 큐비트 개발과 함께 연구가 진행될 것으로 전망된다.

또한, 현재 초전도체 기반과 포획 이온을 기반으로 한 양자컴퓨팅 시스템이 가장 앞서있는 기술로 평가받고 있는데, 현재 사용되는 수십 개의 큐비트 기반의 양자컴퓨팅에 알맞은 물리계가 초전도체 또는 포획 이온일 수 있지만, 향후 수백, 수천 개의 큐비트로 구성된 양자컴퓨터가 필요할 때는 적합한 물리계가 다른 물리계일 수 있다. 그러므로 앞으로도 다양한 물리계에서 범용 양자컴퓨팅 게이트를 개발하고, 확장성 있는 양자컴퓨팅을 구성하려는 노력이 필요할 것으로 예상되고 있다.

마지막으로 대규모 양자컴퓨팅으로 나아가기 위해서는 앞서 이야기한 논리 큐비트를 이용한 양자오류정정 기술이 필수적이므로, 논리큐비트 기반의 양자컴퓨팅 기술개발뿐만 아니라 이에 적용하기 위한 다양한 이론연구 및 소프트웨어 요소개발도 함께 요구되고 있다. 특히, 양자기술의 특성상 큐비트의 개수가 늘어남에 따라 필요한 자원 및 노력이 기하급수적으로 늘어나는데 이 과정에서 나타나는 다양한 문제를 발견하고 그에 맞는 해결책을 찾는 이론 및 실험적 연구가 활발하게 진행될 것이다.

논리 큐비트를 기반으로 하여 양자오류정정이 가능한 대규모의 범용 양자컴퓨터가 나오기까지는 앞으로도 많은 시간과 노력 및 자원이 필요할 것으로 예상된다. 그렇지만, NISQ 기반의 중규모 양자컴퓨팅에서도 충분히 양자컴퓨터를 이용한 계산이 고전컴퓨터보다 더 나은 계산능력을 보여주는 사례가 나오기 시작할 것이다. 최종적으로 완성될 범용 양자컴퓨터가 어떤 물리계를 기반으로, 어떤 방식으로 동작할지는 아직 예상할 수는 없지만 양자컴퓨터를 이용하여 의미 있는 계산 결과를 얻을 수 있는 시기는 우리 예상보다 빨리 올 수 있을 것이라 기대한다.

저자 _ 한상욱(Sang Wook Han)

• **학 력**
KAIST 전기 및 전자공학과 박사
KAIST 전기 및 전자공학과 석사
KAIST 전기 및 전자공학과 학사

• **경 력**
現) 한국과학기술원 책임연구원
前) 삼성종합기술원 전문연구원
前) 픽셀플러스 선임연구원

저자 _ 조영욱(Young-Wook Cho)

• **학 력**
포항공과대학교 물리학 박사
포항공과대학교 물리학 학사

• **경 력**
現) 한국과학기술연구원 선임연구원
前) The Australian National University
박사후연구원

저자 _ 임향택(Hyang Tag Lim)

• **학 력**
포항공과대학교 물리학 박사
포항공과대학교 물리학 학사

• **경 력**
現) 한국과학기술연구원 선임연구원
前) 스위스 ETH Zurich 물리학과 박사후 연구원

참고문헌

국내문헌

1) 2020 차세대반도체연구소 백서(KIST).
2) 임향택, 라영식, 홍강희, 송영선, 김윤호, 광자 기반의 양자정보 연구, Optical Science And Technology, April, 2014.

국외문헌

3) Acin, A. et. al. "The quantum technologies roadmap: a European communicty view", New J. Phys. 20, 080201 (2018).
4) Arute, F. et al. "Quantum supremacy using a programmable superconducting processor," Nature 574, 505 (2019).
5) Bennett C. H. and Brassard G., "Quantum cryptography: public key distribution and coin tossing," Theor. Comput. Sci. 560, 7 (2014).
6) Biamonte, J. et. al. "Quantum machine learning," Nature 549, 195 (2017).
7) Bruzewicz, C. D. et. al. "Trapped-ion quantum computing: Progress and challenges," Appl. Phys. Rev. 6, 021314 (2019).
8) Humphreys, P. C. et. al. "Deterministic delivery of remote entanglement on a quantum network," Nature 558, 268 (2018).
9) Kandala, A. et al. "Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets," Nature 549, 242 (2017).
10) Ko, H. et. al. "High-speed and high-performance polarization based quantum key distribution system without side channel effects caused by multiple lasers," Photon. Res. 6, 214 (2018).
11) Liao, S.-K. et. al. "Satellite-relayed intercontinental quantum network," Phys. Rev. Lett. 120, 030501 (2018).
12) Monroe, C., Raymer, M. G., and Taylor J. "The U.S. National Quantum Initiative: From Act to action", Science 364, 440 (2019).
13) Park, B. K. et. al. "User-independent optical path length compensation scheme with sub-nanosecond timing resolution for a 1 × N quantum key distribution network systemQKD system with fast active optical path length compensation," Photon. Res. 8, 296 (2020).
14) Patel, K. A. et. al. "Quantum key distribution for 10 Gb/s dense wavelength division multiplexing networks," Appl. Phys. Lett. 104, 051123 (2014).
15) Raymer, M. G. and Monroe, C. "The US National Quantum Initiative", Quantum Sci. Technol. 4, 020504 (2019).

16) Sasaki, M. et. al. "Field test of quantum key distribution in the Tokyo QKD Network," Opt. Express 19, 10387-10409 (2011).
17) Sibson, P. et. al. "Chip-based quantum key distribution," Nature Commun. 8, 13984 (2017).
18) Yamamoto, Y., Sasaki, M., and Takesue, H. "Quantum information science and technology in Japan", Quantum Sci. Technol. 4, 020502 (2019).
19) Wang, Y. , Li, Y., Yin Z.-Q., and Zeng, B. "16-qubit IBM universal quantum computer can be fully entangled," npj Quantum Information 4, 46 (2018)
20) Wehner, S., Elkouss, D., and Hanson R. "Quantum internet: A vision for the road ahead," Science 362, eaam9288 (2018)
21) Zhang, Q., Xu, F., Li, L., Liu, N.-L., and Pan, J.-W. "Quantum information research in China", Quantum Sci. Technol. 4, 040503 (2019).

기타문헌

22) http://biz.newdaily.co.kr/site/data/html/2019/10/27/2019102700049.html
23) https://epsrc.ukri.org/newsevents/pubs/quantumtechroadmap/
24) https://qist.lanl.gov/pdfs/rm_intro.pdf
25) https://qt.eu/app/uploads/2018/04/QT-Roadmap-2016.pdf
26) http://www.epnc.co.kr/news/articleView.html?idxno=92501
27) https://www.etnews.com/20200110000136
28) https://www.itu.int/en/ITU-T/Workshops-and-Seminars/2019060507/Documents/_Seong%20Su%20Park3.pdf
29) https://www.whitehouse.gov/wp-content/uploads/2018/09/National-Strategic-Overview-for-Quantum-Information-Science.pdf

융합연구리뷰

Convergence Research Review 2020 March vol.6 no.3