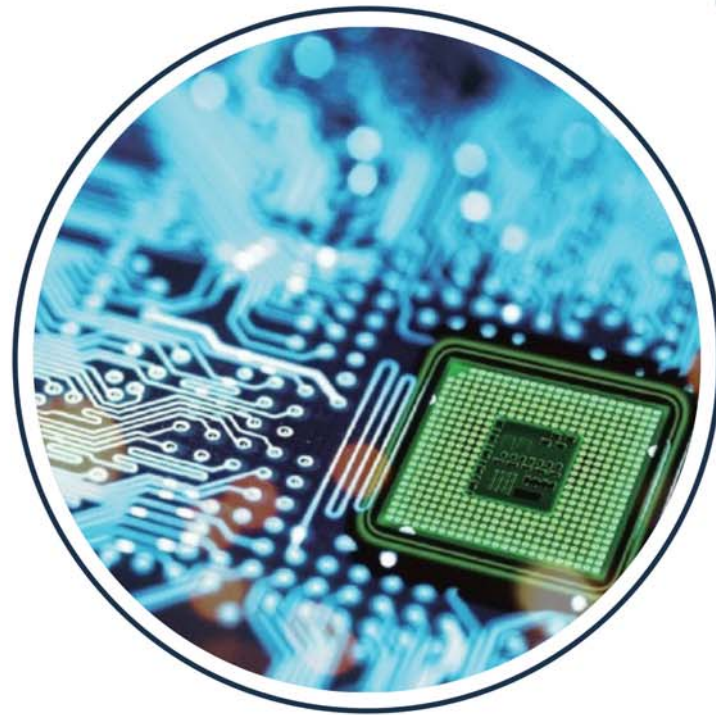




반도체 기술 로드맵

2026

반도체 기술 로드맵 2026

2025년 12월

목차

I. 개요	5
1. 로드맵 수립 동향 및 경과	5
2. 반도체 기술 분류 체계	7
3. 로드맵 개요	8
II. 소자 및 공정 기술 로드맵	15
1. 서론	15
2. 로직 기술 동향 및 로드맵	15
3. 메모리 기술 동향 및 로드맵	19
4. 기술적 극복 과제	24
5. 요약	25
참고문헌	26
III. 인공지능 반도체 로드맵	27
1. 서론	27
2. 응용 분야와 활용 플랫폼	28
3. AI/ML 하드웨어	29
4. 최적화	33
5. 인공지능 반도체 로드맵	34
6. 요약	37
참고문헌	37
IV. 광연결 반도체 로드맵	38
1. 서론	38
2. 광연결 트랜시버 발전 방향	38
3. 응용 분야와 기술 개발 과제	40
4. 광 집적회로 기술	44
5. 광연결 반도체 로드맵	47
6. 요약	52
참고문헌	53
V. 무선연결 반도체 로드맵	56
1. 서론	56
2. 무선연결 반도체 발전 방향	58
3. 반도체 소자 기술 발전 방향	61
4. 무선연결 반도체 로드맵	66
5. 요약	67
참고문헌	68
VI. 센서 기술	69
1. 서론	69
2. 이미지 센서	69
3. 기타 검출기 센서	81
4. AI 시대를 위한 센서 미래 기술 방향	85
5. 요약	87
참고문헌	88
VII. 유선연결 반도체 기술	90
1. 서론	90
2. 유선 연결 기술 동향	91

3. Chiplet 로드맵	97
4. SerDes 로드맵	101
5. 결론	104
참고문헌	104
VIII. PIM(Processing-In-Memory) 기술	106
1. 서론	106
2. 인공지능 하드웨어 관련 최신 기술 동향	106
3. PIM 기술 로드맵	110
4. 해결해야 할 기술	119
5. 요약	120
참고문헌	121
IX. 반도체 패키징 기술	123
1. 서론 및 기술 정의	123
2. 반도체 패키지 핵심 기술 요약	125
3. 반도체 패키지 로드맵 예측	132
4. 요약 및 차후 로드맵 과제	133
참고문헌	134
X. 양자컴퓨팅 반도체 기술	135
1. 서론	135
2. 초전도 큐비트 국내외 기관 기술 동향	138
3. 초전도 큐비트 양자컴퓨터 기술 발전 방향	140
4. 양자컴퓨팅 반도체 로드맵	149
5. 요약	151
참고문헌	152
XI. 결론 및 향후 계획	154
반도체기술로드맵2026 저자목록	155

I. 개요

반도체는 ‘현대 전자 산업의 쌀’이라 불릴 만큼 전자 시스템을 구성하는데 필수적인 핵심 요소이며, 현재 일반 전자 시스템의 구성 부품 중 약 30% 이상을 차지하는 것으로 알려져 있다. 반도체 기술은 인공지능, 클라우드, 빅데이터, 사물인터넷, 로봇, 자율주행차 등 차세대 산업·기술 분야의 구현에 필수적일 뿐 아니라, 다양한 첨단 무기체계 개발에서도 국가 안보의 핵심 기술로서 중요한 역할을 담당한다. 이와 같은 중요성을 고려할 때, 국가 차원에서 반도체 기술 및 산업을 전략적으로 지원하고 경쟁력을 강화하는 것은 필수적이며, 향후 기술 발전 방향을 예측하고 이에 선제적으로 대비하는 노력이 요구된다.

본 반도체기술 로드맵을 통해서 반도체공학회가 전망하는 대한민국 반도체 기술의 중장기 발전 방향과 대응 전략을 제시하고자 한다.

1. 로드맵 수립 동향 및 경과

1958년 잭 킬비에 의해 제안되고 개발된 반도체 집적회로 및 시스템은 반도체 소자 제작 기술의 발달과 혁신에 따라 전자산업에 혁명을 불러왔다. 반도체 소자 및 시스템 기술은 현대 사회를 구성하는 필수 기기인 스마트폰, 컴퓨터, 그리고 데이터센터 등과 같은 전자 시스템을 구현하는 기반 기술이 되었다.

그림 1-1은 글로벌 반도체 로드맵 수립 동향을 요약한 것이다. 1992년 미국은 NTRS(National Technology Roadmap for Semiconductors)를 구성하여 반도체기술 로드맵 수립 활동을 시작하였다. 이는, 1998년부터는 ITRS(International Technology Roadmap for Semiconductors)로 발전하였다. 특히, 2013년 ITRS2.0을 통하여 기존 반도체 소자 및 공정 기술뿐만 아니라 시스템 통합, 외부 시스템 연결, 이종 시스템 집적 기술, CMOS 이후 소자 등 로드맵 수립 기술 분야를 확장하게 된다. ITRS는 이후 2016년부터 반도체 소자와 함께 이를 기반으로 하는 관련 시스템의 발전 로드맵을 포함하도록 IRDS(International Roadmap for Devices and Systems)로 확장된다^[1]. IRDS는 유럽의 SiNANO Institute 및 일본의 SDRJ(The System Device Roadmap Committee of Japan)이 참여하여 활동하고 있다. 우리나라는 2023년 과학기술정보통신부에서 ‘반도체 미래기술 로드맵’을 통하여, 소자, 설계, 공정의 세 분야에서 45개 핵심 기술 주제를 선정 발표한 바 있다^[2].

반도체공학회는 대한민국 반도체 산업의 향후 15년을 전망하는 로드맵의 필요성을 인식하고, 2024년 1월 ‘비전전략 위원회’를 구성하여 로드맵 수립 및 반도체 기술 발전 비전 제시를 위한

활동을 시작하였다. 그 결과, 2024년 12월에는 반도체 소자 및 공정 기술, 인공지능 반도체, 광연결 반도체, 무선연결 반도체 등 4개 기술 분야에 대한 「반도체기술로드맵 2025」를 발표하였다. 이후 2025년 1월부터는 기존 4개 분야에 더해 센서 기술, Processing-In-Memory(PIM) 기술, 유선연결 반도체 기술, 패키징 기술, 양자컴퓨팅 기술 분야까지 로드맵 수립 범위를 확장하였다. 본 보고서는 이러한 일련의 수립 활동을 바탕으로 마련된, 반도체공학회의 두 번째 반도체기술로드맵을 제시하는 문서이다.

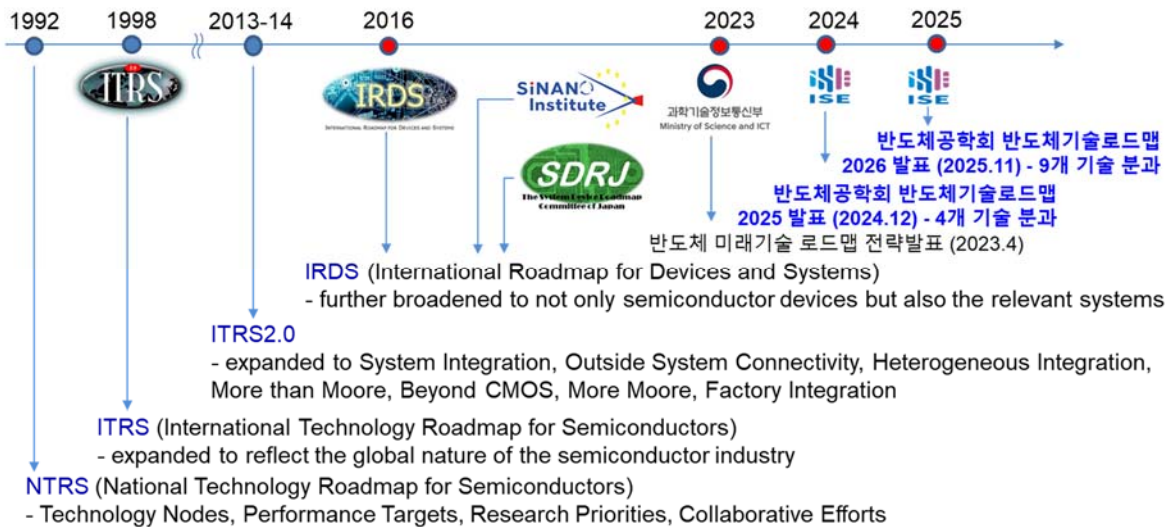


그림 1-1. 글로벌 반도체 로드맵 수립 동향

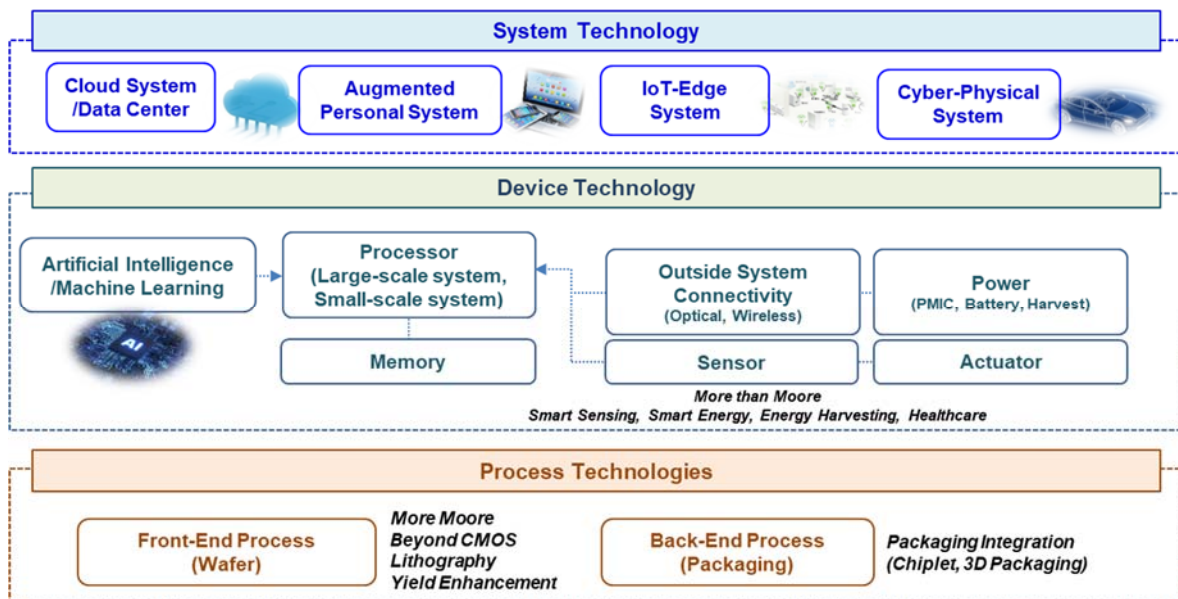


그림 1-2. 반도체 기술 분류 체계

2. 반도체 기술 분류 체계

반도체는 다양한 응용 시스템에 따라 다양한 형태와 용도로 개발되어 적용된다. 따라서, 단순히 반도체 기술 만의 발전을 예측하기 보다는, 이것이 적용되는 전자 부품(Device)과 시스템(System) 기술의 발전을 같이 예측하는 것이 필요하다. 여기서 부품(Device)은 전자 시스템을 구성하는데 필요한 프로세서, 유무선 통신 송수신기, 전력관리, 메모리, 센서 등의 구성 요소가 되는 반도체 칩, 소자, 집적회로 등을 포함하는 반도체 부품을 의미한다.

그림 1-2는 반도체 기술 분류체계를 보여주고 있다. 시스템 기술(System Technology)은 데이터 센터와 같이 대규모 컴퓨팅 및 처리를 목적으로 하는 클라우드 시스템(Cloud System), 스마트폰 같은 다양한 기능을 하는 증강 개인 시스템(Augmented Personal System), 사물인터넷의 센서 노드와 같은 IoT-E 시스템(IoT-Edge System), 무선센서네트워크(Wireless Sensor Network), 커넥티드 카(Connected Car), 지능형 교통정보 시스템(Intelligent Transportation System)과 같은 사이버 공간에 연결되어 동작하게 되는 사이버-물리공간 연결 시스템(Cyber-Physical System)의 네 가지로 분류할 수 있다.

시스템을 구현하기 위한 전자 부품 기술(Device Technology)로는, 데이터 처리를 위한 프로세서(Processor)가 중심에 있다. 클라우드 시스템 등 전력 소모에 제한 받지 않는 대규모 시스템용 프로세서(Large-Scale system Processor)와 개인기기 또는 IoT-E 시스템 등 소규모 시스템 용 프로세서(Small-Scale system Processor)로 분류할 수 있다. 또한, 데이터의 저장을 위한 메모리 부품이 필수적이다. 프로세서에 센서(Sensor)와 구동기(Actuator)가 연결되어 기존 반도체 칩 만이 아닌 다양한 부가 기능이 통합되는 모어댄무어(More-than-Moore)로 발전할 수 있다.

부품 기술의 또 한가지 핵심요소는 외부 시스템으로 데이터를 주고 받을 수 있는 연결 및 통신(Outside System Connectivity) 기능이 있어야 한다. 이는 무선 연결, 유선 연결, 광 연결 기술을 통해 이루어진다. 또한, 모든 기기의 전력(Power) 공급 및 관리를 담당하는 부품이 필요하다. 마지막으로, 최근에는 인공지능 및 기계학습(Artificial Intelligence(AI)/Machine Learning(ML)) 기능이 다양한 형태로 프로세서와 부품에 적용되고 있다. 이러한 부품 기술을 이용하여 앞서 언급한 네 가지 전자 시스템을 구현할 수 있는 것이다.

이러한 부품과 시스템을 구현하는 반도체 공정 기술(Process Technology)은 전공정(Front-End Process)과 후공정(Back-End Process)으로 구분된다. 전공정 기술은 무어 법칙에 따른 집적화 기술(More Moore), 신소자 기술(Beyond CMOS), 리소그래피 기술(Lithograph), 수율 개선 기술(Yield Enhancement)등을 포함하며, 후공정 기술은 칩렛(Chiplet), 3차원 패키징 기술(3D Packaging) 등을

포함한다.

반도체 기술 로드맵에서는 이와 같이 공정기술, 부품기술, 시스템기술을 아우르는 기술 발전 동향 및 방향을 통합적으로 제시하는 것이 필요하다. 본 로드맵에서는 소자 및 공정 기술(More Moore), 인공지능 반도체 기술(AI Semiconductor), 광연결(Optical Connectivity) 반도체 기술, 무선연결(Wireless Connectivity) 반도체 기술, 센서 기술, 유선연결 반도체(Wireline Connectivity) 기술, PIM(Processing-In Memory) 기술, 패키징 기술, 양자컴퓨팅 기술 분야를 9가지 핵심 분야로 선정하여 이에 대한 로드맵을 제시하고자 한다.

3. 로드맵 개요

본 로드맵 보고서에서 선정한 '소자 및 공정 기술', '인공지능 반도체 기술', '광연결 반도체 기술', '무선연결 반도체 기술', '센서 기술', '유선연결 반도체 기술', 'PIM 기술', '패키징 기술', '양자컴퓨팅 기술'의 9가지 기술 분야에 대한 상세한 기술 발전 로드맵을 논의하기 전에, 여기서는 간단한 요약을 제시한다.

(1) 소자 및 공정 기술

로직 소자 및 공정 기술은 2025년 2nm급 GAA 소자에서 2040년 0.2nm급 CFET 소자로 발전할 것으로 예측된다. 극자외선(EUV) 기술은 공정 비용 상승을 억제하면서 공정 복잡성을 줄이는 동시에 미세화의 기본 원칙을 가능하게 할 것이다. 반도체 소자의 물리적 미세화는 2031년경, 0.75NA EUV 도입되면서 포화되고, 2031년 이후에는 3D 집적으로의 전환과 확장적 SoC(시스템 온 칩) 구현을 위한 Beyond-CMOS 소자의 사용이 예측된다.

3D 집적의 주요 과제는 소자, 금속 배선, 메모리, 아날로그, 입력/출력(I/O)과 같은 하위 시스템을 더 효과적으로 활용하기 위해 시스템을 분할하는 방법이다. 기생 캐패시터 문제 해결을 위해 Low-k 소자 스페이서(Spacer)가 도입되어 2025년 이후 소자 성능 개선을 위한 주요 기술이 될 것이다. 금속 배선 저항, EM(electromigration) 및 TDDDB(time-dependent-dielectric-breakdown) 한계를 제어하는 것이 점점 더 어려워지고 있다. Cu 확산 방지막의 비이상적 스케일링으로 인해 금속화 부피가 줄어들고 표면 및 결정립 경계면에서 산란이 증가하기 때문에 금속 배선 저항은 기하급수적으로 커질 수 있다. 따라서 새로운 확산 방지막 물질, 원자층 증착(ALD) 기반 확산 방지막 증착 그리고 non-Cu 금속 배선화와 같은 해결 방안이 필요하다. 저항 문제 외에도 TDDDB는 주어진 Low-k 유전체에 대해 인접한 라인 사이의 최소 공간이 한정되어 유전율 스케일링을 더디게

만든다.

2025년 11nm급 DRAM이 상용화되었고, 2040년에는 최소 반선폭이 6nm까지 작아질 것으로 보이며, 비트(bit) 밀도 증가, 셀 정전용량 증가, 기생 용량 감소를 위한 기술개발과 함께 3D DRAM과 4F2 셀과 같은 새로운 아키텍처 개발도 수행되고 있다. 또한 AI 반도체 수요 급증으로 고성능 고대역폭 메모리(HBM, High Bandwidth Memory) 시장이 급증하고 있고, 2025년 12단의 2TB/s, 2031년 20단의 8TB/s, 2040년 30단 이상의 128TB/s 대역폭으로 발전할 것으로 예측된다. 이를 위해서 Bump-less Cu-Cu Direct Bonding 기술과 임베디드 냉각기술이 개발되어야 한다.

플래시 메모리 기술은 2025년 321단의 제품 양산이 시작되었으며, 2031년에는 1,000단이 양산될 것으로 예측된다. 이를 위해서 고종횡비 식각, 공정 미세화, 다단 적층 및 고품질 Wafer Bonding 기술이 개발되어야 한다.

(2) 인공지능 반도체

인공지능 반도체 로드맵은 AI/ML 영역의 애플리케이션 벤치마킹에 초점을 두어, GPU, FPGA, ASIC 등 다양한 프로세서에 기반한 컴퓨팅 접근법에 대한 성능 목표를 제공하고자 한다. 특히 성능 메트릭 (처리량, 지연 시간, 정확도, 에너지 효율)의 최근 추세를 정확하게 보고함으로써, 가까운 미래까지 이러한 추세의 미래 진화를 추정한다.

학습을 위한 하드웨어 연산 효율은 2040년까지 GPU (FP64)의 경우 100 TOPS/W 이상, ASIC (FP16)의 경우 10 TOPS/W 이상으로 예측하고, 추론을 위한 하드웨어 연산 효율은 ASIC (INT16)의 경우 10 TOPS/W 이상, INT8의 경우 100 TOPS/W 이상으로 예측한다.

(3) 광연결 반도체

광연결 반도체 로드맵은 데이터 센터를 중심으로 각각의 시스템 내(Intra-System)의 전송과 시스템 간(Inter-System) 전송을 위한 광 연결과 해당 재료, 부품, 중간 모듈 및 시스템에 한정하기로 한다. 응용 영역에 따라서 시스템 내(Intra-System)를 표현하는 Inside-of-Rack 과 시스템 간(Inter-System)을 표현하는 Outside-of-Rack 관점의 분류와 전송 거리에 따라 XSR (Extreme Short Reach, < 1m), SR (Short Reach, < 100m), DR (Data Center Reach, < 500m), 그리고 FR (Fiber Reach, < 2km) 등으로 구분할 수 있다.

2025년부터 PAM4 방식 기준의 200G/lane 급의 도입을 시작으로 400G/lane까지 증가할 것으로 예상된다. 또한 시스템 내에서 Gen 3의 NPO(Near-Packaged Optics) 구현까지 시도되며, 기술 및 산업기반으로는 실리콘 포토닉스(Silicon Photonics: SiPh) 산업의 표준 제정이 예상된다. 실리콘 포

토닉스(SiPh)는 SOI 기반의 도파관(Waveguide)이나 변조기(Modulator) 구현이라는 기술적인 관점으로 제시되었으나, 현재는 Silicon (Semiconductor)-based Optical Device 를 총괄하는 산업적인 개념으로 진화하고 있다. 향후에는 FOWLP를 활용하여 반도체 공정/소자들과 광학부품과의 이중 패키징을 포함한 광융합반도체 산업으로 확장될 것으로 예상된다.

2040년까지는 고급 Modulation 방식을 채택한 800G/lane 이상의 성능 확보, Gen 4의 CPO(Co-Packaged Optics) 적용 및 초고속 데이터전송을 위하여 E/O 변환을 최소화한 Hybrid E/O 변환 및 광로직(Optical logic)이 도입될 것으로 예상된다. 또한, 광로직을 위한 관련 기술·재료·소자·시스템과 양자컴퓨팅과의 결합에 의한 시너지로 초고속 연산 및 데이터 전송에서 획기적인 성과가 창출될 것으로 예상된다.

(4) 무선연결 반도체

사물 인터넷(Internet of Things: IoT)을 넘어 만물 인터넷(Internet of Everything: IoE) 시대에는 고성능 무선 통신을 요구하는 애플리케이션들이 계속 증가하고 있다. 대부분의 애플리케이션은 RF/마이크로파 무선 통신과 고성능 백홀을 사용하여 인터넷과 클라우드 데이터 센터에 연결된다. 현재 모바일 장치는 4G LTE 및 5G 네트워크에서 초당 최대 10Gbps의 다운로드 데이터 속도를 지원한다.

무선 기술의 발전으로 모바일 데이터 전송 속도는 계속 증가하고, 안테나 수는 MIMO(Multi-Input Multi-Output) 기술과 함께 계속 늘어나게 된다. 데이터 속도 증가를 지원하기 위해 전이중(Full Duplex) 통신 방식이 채택되어야 하고, 다수의 안테나를 위한 빔포밍 기술도 발전해야 한다. 전력 증폭기 및 A/D 컨버터를 포함하여 무선통신 회로는 높은 데이터 속도에서 휴대폰의 배터리 사용 시간을 개선하기 위해 밀리미터파(Millimeter-Wave: mmWave) 주파수 및 광대역에서 향상된 에너지 효율성을 제공해야 한다. 레이더와 같이 전파를 활용하는 센싱과 무선통신은 상호 협력하여 운영되어 센싱과 통신 결합 서비스(Integrated Sensing and Communication: ISAC)로 발전할 것으로 기대된다. 향후 우주 인터넷 망을 위한 위성 통신용 통신 반도체 기술도 연구되어야 한다. 이러한 무선 통신에서의 마이크로파 회로 기술은 향후 양자컴퓨터의 양자 상태 검출(Readout) 회로 개발에도 활용될 것이다.

2040년까지 무선 데이터 전송률은 1000Gbps를 달성할 것으로 예측되고 이를 달성하기 위한 고주파 반도체 소자 개발, 밀리미터파 및 서브테라헤르츠(Sub-THz) 대역 활용 회로 및 시스템 기술, 새로운 무선 네트워크 기술 등이 필요할 것으로 예상된다.

(5) 센서 기술

센서 기술은 반도체 공정 고도화와 AI 융합을 중심으로 고해상도·저전력·지능형 구조로 발전하고 있다. 이미지 센서는 픽셀 미세화와 메타옵틱스, 고성능 HDR·LFM 기술에 힘입어 2030년경 0.5 μm 이하, 초고감도·초고해상도 구조를 갖춘 형태로 진화할 전망이다. 글로벌 셔터와 DPS는 모바일과 산업용 영역에서 확산되며, 단일광자 센서(PCS)는 초저조도 환경의 핵심 기술로 자리잡을 것으로 예측된다.

비가시광 센서는 UV·NIR·SWIR·LWIR 영역에서 소재 혁신이 주도한다. UV는 PQD·GaN 기반 고감도 센서 개발이 가속화되고, NIR은 SPAD·RGB+IR 센서가 보안·LiDAR 중심으로 고도화된다. SWIR에서는 InGaAs를 대체하는 QD·Ge 기술이 2040년까지 상용화될 가능성이 높으며, LWIR은 VOx-a-Si 기반 마이크로볼로미터의 소형화와 내구성 개선이 진행될 것이다. 이벤트 기반 비전 센서는 초고속·저전력 동작을 기반으로 2040년 8K급 해상도, 120 dB 이상의 DR을 지닌 센서로 진화하며, CIS-EVS 하이브리드 구조가 로보틱스·모빌리티 분야에서 확대될 것으로 전망된다. MEMS 기반 초음파 센서, 나노소재 기반 이온·바이오 센서, SAW 기반 VOC 센서는 정밀도·선택성을 높이며 환경·헬스케어·스마트 제조 영역에서 핵심 역할을 강화한다.

AI 시대 센서는 In-Sensor DNN, 초저전력 AON 기술, 멀티센서 퓨전이 중심이 된다. 2040년에는 센서 내부에서 100fps 이상의 지능형 처리와 대역폭 절감이 가능해지고, 카메라·라이다·레이다 등이기종 센서가 융합된 거대 센서 네트워크가 자율주행·스마트시티의 기반 기술로 자리잡을 것이다.

(6) 유선연결 반도체

유선 연결 반도체 기술은 고속·저지연 데이터 전송을 위해 패키징, 칩렛, SerDes를 중심으로 통합적 진화를 거듭하고 있다. 데이터 전송 속도가 지속적으로 증가하면서 기존 보드·패키지 기반 전송선의 기생 성분이 성능 한계로 부상하였고, 이에 따라 고집적 패키징 기반의 이종 집적, 칩렛 기반 단거리 인터커넥트, 고속 SerDes 회로가 핵심 기술축을 형성하고 있다. 이종 집적 기술에서는 실리콘·유리·유기 기판 인터포저가 활용되며, TSV/TGV 기반 수직 연결을 통해 신호/전원 무결성을 향상시키는 방향으로 발전하고 있다. 패키지 내 인터커넥트 기술은 wire bonding에서 hybrid bonding까지 진화하며, 신호 속도는 1 GHz 미만에서 100 GHz 이상까지 확장되었다. 한편 칩렛 아키텍처는 UCle를 중심으로 표준화가 빠르게 진행되고 있으며, 32~64 Gbps급 단거리 직렬 인터커넥트를 통해 전력·비용 효율적 시스템 통합을 가능하게 한다. SerDes 기술은 PAM4·PAM6 기반의 고차 변조와 AI 기반 적응형 equalization을 통합하며 100~400 Gbps급 회로로 진화하고 있다.

기술 전망 측면에서 2030년 이전까지는 UCle 기반 칩렛 상용화와 2.5D·3D 패키징의 확산을 통해 64~96 Gbps급 인터커넥트와 높은 I/O 집적도가 구현될 것으로 예상된다. 2035년 이후에는 하이브리드 본딩, 실리콘 브릿지, 패널 레벨 공정이 성숙하여 패키지 내 완전 3D 수직 통합 구조가 등장하고, 전기·광 인터커넥트가 병행되는 혼합형 I/O 구조가 주류가 될 전망이다. 2040년경에는 200 Gbps 이상급 신호 전송, 500 TB/s/mm²급 I/O 집적도, 0.5 ns 수준의 지연 특성을 갖는 시스템이 가능해지고, 전기·광·양자 신호를 하나의 칩 패키지 내에서 처리하는 통합 인터페이스 구조가 상용화될 것으로 예측된다. 결국 유선 연결 기술은 칩 간 물리적 경계를 제거하는 완전 집적형 Polyolithic SoC 및 Photon-First SerDes로 진화하며, 차세대 AI-HPC 시스템의 핵심 기반 기술로 자리매김할 것이다.

(7) Processing-In-Memory (PIM) 반도체

PIM(Processing-In-Memory) 기술은 메모리 내에서 직접 연산을 수행함으로써 기존 프로세서 중심 구조에서 발생하는 '메모리 월' 문제를 근본적으로 해결하려는 접근으로, AI-HPC 환경에서 요구되는 대규모 병렬 연산의 지연과 전력 소모를 크게 감소시키는 데 목적이 있다. DRAM 기반 PIM은 기존 D램 아키텍처를 확장하여 내부 뱅크·로우 동작을 이용해 벡터 연산을 수행하는 구조가 핵심이며, HBM-PIM은 TSV·스택형 구조를 활용해 높은 대역폭을 유지하면서 병렬 연산을 통합하는 방향으로 발전하고 있다. 특히 국내 기업이 제안한 HBM-PIM 구조는 AI 가속기와의 병렬 처리 효율을 극대화하면서 표준 HBM 인터페이스를 유지해 상용화 가능성을 크게 높였다. 한편 비휘발성 메모리 기반 PIM(NVM-PIM)은 신경망 연산에 특화된 아날로그 매트릭스-벡터 곱 연산(MVM)을 초저전력으로 수행할 수 있어 차세대 뉴로모픽 연산기술의 유력한 후보로 평가된다.

기술 전망을 보면, 2025~2030년에는 HBM-PIM의 본격적인 상용 사용이 확대되며 AI-데이터센터에서 GPU와 결합된 다중타일 구조가 도입될 것으로 예상된다. 2030년대 중반에는 PIM-전용 명령어셋 및 실행 런타임이 표준화되고, DRAM·NVM·3D 패키지 기반 혼합형 PIM 구조가 등장해 작업 부하에 따른 동적 연산 분배가 가능해진다. 같은 시기 NVM 기반 아날로그 PIM 기술은 고정밀 ADC·보정 회로의 성숙을 통해 실제 제품 수준의 신뢰도를 확보할 전망이다. 2040년경에는 PIM이 메모리·연산·네트워크를 통합한 'Near/In-Memory Computing Fabric'으로 확장되며, 초대형 AI 모델 학습·추론을 위한 전력·대역폭 최적 구조의 핵심 기술로 자리하게 될 것이다.

(8) 패키징 기술

반도체 패키징은 전통적으로 칩 보호와 기판 연결을 담당하는 후공정 기술로 인식되어 왔으나, 공정 미세화에 따른 비용 증가와 시스템 복잡도 확대에 따라 성능·전력·열·신뢰성에 영향을 미치

는 시스템 통합 기술로 그 역할이 점차 확대되고 있다. 단일 대형 SoC 구조는 수율 및 비용 측면에서 제약이 커지고 있으며, 이에 따라 기능별·공정별로 분리된 다수의 칩을 하나의 패키지로 통합하는 Multi-Chip Integration과 칩렛 기반 구조가 주요 집적 방식 중 하나로 활용되고 있다.

패키지 구조 측면에서는 전통적인 2D 패키지가 여전히 폭넓게 사용되고 있으나, 고대역·저지연 요구가 증가함에 따라 실리콘 또는 유리 인터포저 기반 2.5D 구조, RDL 및 브리지 기반 2.xD 구조, TSV 및 μ -bump 또는 하이브리드 본딩을 활용한 3D 적층 구조에 대한 연구와 적용 사례가 점진적으로 확대되고 있다. 특히 로직 칩과 HBM의 근접 집적, 미세 피치 die-to-die 인터커넥트, Fan-out 및 Panel-Level 패키징 기술은 AI 및 HPC 응용을 중심으로 활용 가능성을 지속적으로 검토하고 있다.

패키지 집적도의 증가와 함께 전력 소모 및 발열 밀도 또한 증가하고 있어, 냉각 구조, 전력 전달 경로, 열·기계적 응력 요소를 포함한 시스템 수준의 패키지 설계 중요성이 커지고 있다. 이에 따라 전기·열·기계 특성을 통합적으로 고려하는 모델링과 시뮬레이션, 그리고 Co-Design 접근이 패키지 설계 과정에 점진적으로 도입될 것으로 예상된다.

기술 전망을 보면, 2030년 전후 시점에는 하이브리드 본딩과 die-to-die interconnection의 미세화가 주요 패키징 기술 옵션으로 논의되고 있으며, 전기·광혼합 interconnection 역시 HPC 시스템을 중심으로 적용 가능성이 논의되고 있다. Fan-out RDL 및 Panel-Level 패키징은 미세화와 생산성 향상을 목표로 공정 고도화와 표준화 논의가 진행 중이다. 중장기적으로는 3D 적층 구조와 칩렛 기반 패키지에 대한 연구와 적용이 확대되면서, 패키징 기술이 시스템 구조 설계에서 차지하는 비중이 점차 증가할 것으로 전망된다.

(9) 양자컴퓨팅 기술

초전도 큐비트 기반 양자컴퓨팅은 조셉슨 접합을 이용한 비선형 초전도 회로를 통해 양자 2준위 시스템을 구현하며, CMOS 공정과의 높은 호환성과 빠른 게이트 시간을 바탕으로 현재 가장 산업화에 근접한 양자 하드웨어 플랫폼으로 평가된다. 트랜스몬 및 플럭소늄 큐비트 구조는 잡음 민감도를 낮추고 결맞음 시간과 제어 안정성을 향상시키면서 현대 양자 프로세서의 사실상 표준이 되었다. 큐비트 제어는 정밀한 마이크로파 펄스 합성에 의해 구현되며, 고속 DAC 기반 직접 합성, I/Q 믹서 기반 SSB-DSB 구조, FDM 기반 다중화 방식 등이 대규모 확장을 위해 활용되고 있다. 기존 상온 제어 구조의 지연·열잡음 문제를 해결하기 위해 3~4 K에서 동작하는 Cryo-CMOS 기반 제어·리드아웃 회로가 필수 기술로 부상하고 있으며, 이는 큐비트 인접 환경에서 펄스를 직접 생성해 신호 품질과 확장성을 크게 향상시킨다. 리드아웃은 공진기 기반 Dispersive 방식이 표준으로, 양자

한계 증폭기(JPA-TWPA), 저잡음 LNA, Mixer, ADC 및 DSP를 조합해 반사 신호의 위상·진폭 변화를 분석하여 큐비트 상태를 판별한다. 확장성을 높이기 위해 볼로미터 기반 에너지 감지형 리드아웃과 극저온에서 ADC·DSP 없이 동작하는 Homodyne Cryo-CMOS 리드아웃 SoC 등이 제안되고 있다.

기술 전망 측면에서, 2024~2030년에는 Cryo-CMOS 통합과 FDM 기반 제어 기술의 고도화를 통해 수백~수천 물리 큐비트 및 10~50 논리 큐비트의 안정적 운용이 가능해질 것으로 예상된다. 2030년대 중반에는 칩렛 기반 Cryogenic 패키징과 양자 네트워킹이 결합되면서 수십만 물리 큐비트의 초대형 초전도 양자 시스템이 등장하고, 화학·재료·AI 최적화 분야에서 실질적 양자 우위가 실현될 것으로 전망된다. 2040년경에는 제어·리드아웃·QEC·인터커넥트가 통합된 Quantum System-on-Chip과 Quantum-Centric Computing 아키텍처가 상용화되어 데이터센터의 표준 연산 자원으로 자리잡을 것으로 예측된다.

II. 소자 및 공정 기술 로드맵

1. 서론

반도체 산업의 목표는 지속적인 소자의 미세화를 통해 전력과 비용을 낮추고 성능을 향상시키는 것이다. 최종 칩의 성능은 고속화, 고밀도화, 저전력화, 폼팩터 소형화, 재료비 절감, 기능성 향상 등 다양한 방식으로 측정될 수 있다. 전통적 기법에 기반한 칩 미세화는 물리적 한계에 이르렀고 칩 미세화를 지속하기 위해서는 공정 모듈, 툴, 재료 특성 등을 통해 해결해야 할 과제들을 갖고 있다.

본 로드맵에서는 IRDS의 More Moore IFT(More Moore International Focus Team)^[1]과 IMEC의 ITF World 2023 및 2024 자료에서 예측한 로드맵 등을 참조하여 빅 데이터, 모빌리티, 클라우드 서버 애플리케이션의 전력·성능·면적·가격 [Power-Performance-Area-Cost: PPAC] 고도화를 위한 로직 및 메모리 기술에 대한 물리적, 전기적, 신뢰성 요구 사항을 기술하였다. 본 로드맵에서는 로직과 메모리 반도체의 소자 및 공정 기술에 대하여 양산수준 기술의 주요 방향을 2025년부터 3년 마다 노드 발전에 따라 15년간, 2040년까지 로드맵을 제시하고자 한다.

2. 로직 기술 동향 및 로드맵

소자공정 기술 발전은 공정 피치 미세화와 낮은 공급 전압에서 성능과 전력 스케일링을 유지하기 위한 효과적인 솔루션 개발에 초점을 맞추고 있다. 미세화의 기본 원칙은 성능 향상 및 칩 면적 감소를 추구하는 것이다. 그러나, 미세화가 진행됨에 따라 기생 성분으로 인한 부하 비중이 증가하고, 성능과 전력소모 측면에서 역효과가 생길 수 있다. 따라서 디바이스 및 금속 배선의 기생 성분들을 고려하면서 미세화 기술을 개발할 필요가 있다. 미세화의 기본 원칙은 칩 내부 디바이스·회로 수준에서 공정 미세화로 인한 성능 저하나 제조 한계를 설계 기술로 보완하는 DTCO(Design-Technology Co-Optimization)에서, 더 나아가 여러 칩렛(chiplet), 패키지, 메모리, 인터커넥트 구조 등을 포함하여 전체 시스템 아키텍처 차원에서 최적화를 수행하는 STCO(System-Technology Co-Optimization)를 필요로 한다.

소자 구조 및 공정 기술의 핵심 변수에 대한 로드맵을 표 2-1에 제시하였다. 반도체 소자성능을 기준으로 2025년 2nm급 소자에서 2031년 1nm급 소자, 2040년 0.2nm급 소자까지 발전할 것으로 예측된다. 3차원 게이트 구조 및 피치, 금속 배선 피치, 게이트 길이 및 3D Tier 적층 기능은 미세화에서 중요한 인자이다. 표 2-1에서 알 수 있는 것처럼 2031년경 0.75NA EUV가 도입되면서 현재의 0.33NA EUV보다 2.3배 선폭을 줄일 수 있지만, 미세공정에 의한 물리적 미세화는 포화될 것으로 예측된다. PMOS, NMOS의 3차원 집적화, CFET라고 하는 3D

VLSI를 통하여 소자성능을 향상시키는 방향으로 발전할 것으로 예측되는데, 기존 하부 칩의 손상을 방지하고 후속공정을 진행하기 위해서는 저온공정도 소자 제작이 기술적 과제가 될 것이다.

표 2-1. 로직 소자에 대한 소자 구조 및 공정 기술 핵심 변수

물리적 미세화의 한계를 3차원 집적화로 해결						
Year of production	2025	2028	2031	2034	2037	2040
Logic industry "Node Range" labeling	"2.0 nm eq"	"1.4 nm eq"	"1.0 nm eq"	"0.5 nm eq"	"0.3 nm eq"	"0.2 nm eq"
Fine-pitch 3D integration scheme	Stacking	Stacking	3DVLSI	3DVLSI	3DVLSI	3DVLSI
Logic device structure options	GAA	GAA	FSFET (IMEC) CFET	CFET	CFET	CFET (Atomic)
Gate pitch (nm)	45	42	40	38	36	34
Gate length (nm)	14	12	12	10	10	8
M0 pitch (nm)	20	18	16	16	16-12	16-12
Device technology inflection	GAA	CFET-SRAM	Low-Temp Device	Low-Temp Device	Low-Temp Device	Low-Temp Device
Patterning technology inflection for Mx interconnect	193i, 0.33NA EUV	193i, 0.33NA EUV (TSMC) 193i, 0.55NA EUV	193i, 0.75NA EUV	193i, 0.75NA EUV	193i, 0.75NA EUV	193i, 0.75NA EUV
Channel material technology inflection	SiGe60%	SiGe70%	SiGe70%, Ge	2D Mat	2D Mat	2D Mat
Process technology inflection	Channel, RMG, Lateral/Atomic Etch	P-over-N, N-over-P	3DVLSI	3DVLSI	3DVLSI	3DVLSI
Stacking generation inflection	3D-stacking, Mem-on-Logic	3D-stacking, CFET, Mem-on-Logic	3D-stacking, CFET, 3DVLSI	3D-stacking, CFET, 3DVLSI	3D-stacking, CFET, 3DVLSI	3D-stacking, CFET, 3DVLSI

※ GAA (gate-all-around-device), FSFET (forksheet FET), CFET (complementary FET), 3DVLSI (fine-pitch 3D logic sequential integration)

저전압 소자에서 미세화를 유지하기 위해 최근 몇 년간은 채널 격자 변형(Mobility 촉진), HKMG(High-k Metal Gate), Contact 저항 낮추기, 정전기 개선 등 기술 개발을 통한 성능 향상에 집중하고 있었다. 이는 휴대용 배터리를 통한 동작이 보편화 되는 상황에서 공급 전압을 낮춰야만 하기 때문에 게이트 구동 손실을 최소화하기 위한 것이었다.

2025년 이후에는 로직 소자 트랜지스터의 주력이 FinFET에서 4면을 채널로 사용하는 GAA(Gate All Around)로 변화되고 있다. GAA는 향상된 정전기 특성에 의해 게이트 구동을 유지하는 수평형 나노 시트(Lateral Nano Sheet)로서 공급 전압을 작게 할 수 있다^[3]. GAA 소자제작을 위해서는 기생 Capacitance 문제, 유효 채널 폭(W_{eff}) 제어, RMG(replacement metal gate) 집적이 숙제이다^[4].

Monolithic 3D(M3D) 집적 기술을 통해 소자를 수직적으로 쌓을 수 있다^[5]. 미세화의 초점은 Single-Thread 성능 이득, 전력 감소에 있고, 이후에는 낮은 동작 전압(Vdd) 및 단위 공간에 내장된 더 많은 기능을 허용하는 고병렬 3D 아키텍처로 진화할 것이다.

3D 하이브리드 메모리-온-로직(3D Hybrid Memory-on-Logic) 기술은 로직 칩과 메모리를 수직으로 적층하고 하이브리드 본딩 기술로 직접 연결하여 성능과 집적도를 극대화하는 기술로서 인공지능(AI), 고성능 컴퓨팅(HPC) 등 고사양 애플리케이션의 병목 현상을 해소하고 온디바이스 AI 구동에 필수적인 고용량 메모리 통합에 기여할 수 있다.

이것은 향후 3D IC 기술로 진화할 것으로 예측된다. 수평으로 연결했던 칩을 직접 붙이는 만큼 데이터 전송속도나 레이턴시는 개선되겠지만, 더욱 어려워진 칩의 이종결합, 즉 본딩(Bonding) 난이도와 HBM(High Bandwidth Memory, 고대역폭 메모리 반도체)와 GPU(Graphics Processing Unit, 그래픽 처리 장치)에서 발생하는 발열이 전가되는 등의 문제를 해결해야 한다.

소자 구조 변화와 함께 공정 요소 기술에서도 발전이 이루어진다. 기존 실리콘에서 SiGe, Ge 등의 2차원(2-Dimensional: 2D) 재료로 진화하는 채널 물질, 실리콘에서 더 낮은 쇼트키 장벽 높이(SBH)를 제공하는 새로운 재료로 진화하는 콘택트 모듈, 콘택트 표면적을 늘리기 위한 wrap-around contact 집적 방식이 포함될 수 있다.

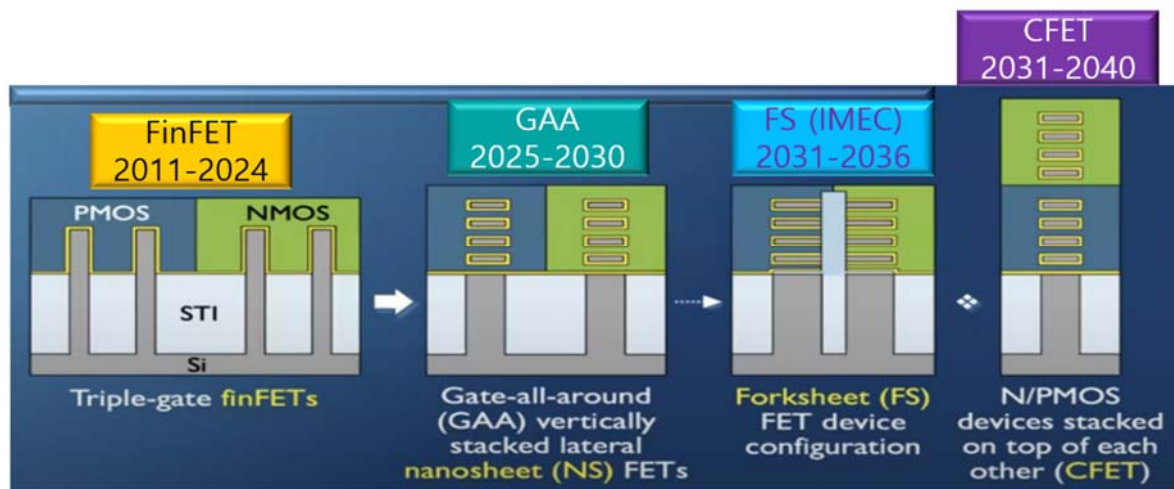


그림 2-1. 트랜지스터 구조의 발전 (출처: ITF World 2023 (IMEC))

그림 2-1은 소자구조 발전 동향을 정리한 것이다. FinFET 및 GAA 아키텍처는 완전 공핍 채널과 완전 반전 채널(볼륨 반전)을 이용한다. 더 나아가 FS-FET(Fork-Sheet FET)는 나노 시트 사이에 절연층으로 분리하여 N-소자와 P-소자를 분리하는 구조로 디바이스의 크기를 획기적으로 줄일 수 있다^[6]. CFET(Complementary FET)는 P-소자가 N-소자 위에 적층되는 3D 형태로 진화할 것이 예측된다^[7].

2025년부터 2040년까지 로드맵에서 예측된 6개 Technology Node에서 로직 코어 소자의 전기적 사양은 표 2-2에서 보는 바와 같이 같이 2nm급 이하의 로직 소자에서는 미세화가 진행됨에 따라 기생 성분예 의한 부하 비중이 증가하고, 성능과 전력소모 측면에서 역효과로 인해서 공급전압은 0.5V~0.4V로 크게 개선되지 않지만, Transconductance 등 아날로그 특성은 유지될 것이다.

표 2-2. 로직소자기술 로드맵 (출처 [1] IRDS의 MM-9에서 발췌하고, 수정함)

YEAR OF PRODUCTION	2025	2028	2031	2034	2037	2040
Logic industry "Node Range" labeling	"2 nm eq"	"1.4 nm eq"	"1.0 nm eq"	"0.5 nm eq"	"0.3 nm eq"	"0.2 nm eq"
Fine-pitch 3D integration scheme	Stacking	Stacking	3DVLSI	3DVLSI	3DVLSI	3DVLSI
Logic device structure options	GAA	GAA	FSFET (IMEC) CFET	CFET	CFET	CFET (Atomic)
LOGIC DEVICE ELECTRICAL SPECS						
Power Supply Voltage -V _{dd} (V)	0.5 ~ 0.6	0.5 ~ 0.6	0.5 ~ 0.6	0.4 ~ 0.5	0.4 ~ 0.5	0.4 ~ 0.5
Subthreshold slope (mV/dec) - HP (mV/dec)	72	70	70	70	70	70
Capacitive equivalent thickness (CET) (nm)	1.00	0.90	0.90	0.90	0.90	0.90
Effective mobility (cm ² /V.s)	100	80	60	40	40	40
Ion (uA/um) at I _{off} =10nA/um - HP	787	851	753	737	753	753
C _{ch, total} (fF/um ²) - HP/HD	34.52	38.35	38.35	38.35	38.35	38.35
ANALOG SPECIFICATIONS OF LOGIC DEVICE						
Transconductance - gm (μS/μm)	1621	1751	1725	1653	1684	1684
High-current gain cut-off frequency - f _T (GHz)	304	358	411	403	411	411
1/f-noise (μV ² ·μm ² /Hz)	16	13	13	13	13	13
Analog gain (dB)	40	36	38	39	39	39

금속 배선에 대한 가장 어려운 과제는 금속 배선 전도도, 유전체 유전율 감소, 신뢰성을 충족하는 새로운 재료를 개발하는 것이다. 식각 공정에서 해결해야 할 금속 배선 기술의 과제는 기생 저항 및 커패시턴스 변동성을 줄이기 위해 Low-κ 유전체 재료에 정밀한 트렌치 및 비아 구조를 형성하는 것이다. 배선 구조 집적에 사용되는 다마신¹ 방식은 패턴, 식각 및 평탄화 공정의 정밀한 제어가 필요하다. 미세 금속 배선 구조에서 발생하기 쉬운 RC 열화 문제를 극복하기 위해서 배선 구조를 정밀하게 제어해야 한다. 이를 위해서 고종횡비(high aspect ratio) 구조의 측정을 위한 고속 이미징 계측법이 요구되고, 접착 특성 및 결함의 인라인 모니터링을 위한 새로운 계측 기술도 필요하다. 웨이퍼 대구경화와 테스트 웨이퍼 사용 제한을 위해서는 다양한 In-situ 공정 제어 기술 개발이 필요하다.

기존의 반도체 설계에서는 전력과 신호가 모두 트랜지스터의 앞쪽(Front-side)으로 전달된다. 이 방식에서는 전력 전달과 신호 전송이 동일한 경로를 공유하기 때문에 신호 간섭이나 저항 증가 등의 문제로 인해 성능이 제한될 수 있다. 백사이드 전력 공급은 이러한 문제를 해결하기 위해 도입된 새로운 방법으로, 전력 공급을 반도체 칩의 뒷면(Back-side)에서 전달하는 방식이다. 이를 통해 신호 경로와 전력 경로를 분리하여 설계할 수 있어, 신호 간섭을 줄이고 전력 효율을 극대화할 수 있다. 또한 전력 공급 구조가 칩 전면에서 후면으로 전환되어 로직(Logic) 회로와 전력 공급 회로를 분리함으로써 칩 집적도 및 면적 효율도 개선할 수 있다.

표 2-3의 금속배선의 미세화 로드맵에서 제시된 것처럼 2028년을 기점으로 후면 전력공급 기술

¹ Cu 등의 배선 공정에서 휘발성 반응 부산물을 만들지 못해 식각이 어려운 경우, 이를 해결하기 위해 식각하는 방법이 아닌 절연막에 먼저 Trench를 만든 후 여기에 배선소재를 채우고 나서 CMP 공정으로 상부의 배선막을 연마함으로써 배선을 형성하는 공정

(Backside Power Delivery Network, BSPDN)이 적용될 것으로 예측된다. 2031년 이후에는 후면 전력공급에서 파워 비아 기술이 적용되고, 전력 레일 피치(Pitch)가 40nm 수준으로 급격히 미세화될 것이다.

표 2-3. 금속배선 미세화 로드맵(출처 [1] IRDS의 MM-13에서 발췌하고 수정함)

YEAR OF PRODUCTION	2025	2028	2031	2034	2037	2040
Logic industry "Node Range" labeling	"2 nm eq"	"1.4 nm eq"	"1.0 nm eq"	"0.5 nm eq"	"0.3 nm eq"	"0.2 nm eq"
Fine-pitch 3D integration scheme	Stacking	Stacking	3DVLSI	3DVLSI	3DVLSI	3DVLSI
Logic device structure options	GAA	GAA	FSFET (IMEC) CFET	CFET	CFET	CFET (Atomic)
INTERCONNECT TECHNOLOGY						
Metallization – M0	Co, W, Ru	Co, W, Ru	Co, W, Ru	Co, W, Ru	Co, W, Ru	Co, W, Ru
Metallization – M1, Mx	Cu, Co, Ru	Cu, Co, Ru	Cu, Co, Ru	Cu, Co, Ru	Cu, Co, Ru	Cu, Co, Ru
Barrier Metal – M0, M1, Mx	0.5nm, TiN+WC	0.5nm, TiN+WC	0.5nm, TiN+WC	0.5nm, TiN+WC	0.5nm, TiN+WC	0.5nm, TiN+WC
Dielectrics k value – M0, M1, Mx	SiCOH (2.70-3.20)	SiCOH (2.70-3.20)	SiCOH (2.70-3.20)	SiCOH (2.70-3.20)	SiCOH (2.70-3.20)	SiCOH (2.70-3.20)
POWER DELIVERY TECHNOLOGY						
	Without BS-PDN	BSPDN (Backside Power Delivery Network)				
Power rail structure	Frontside Buried Rail	Backside Buried Rail	Backside Power Via	Backside Power Via	Backside Power Via	Backside Power Via
Power rail material	W, Ru	W, Ru	W, Ru	W, Ru	W, Ru	W, Ru
Power rail pitch	160~65 nm	115~42 nm	40~34 nm	40~28 nm	40~28 nm	40~28 nm

3. 메모리 기술 동향 및 로드맵

인공지능(AI)의 활용이 확대되면서 데이터센터를 비롯한 각종 AI 서버 등에서 차세대 메모리 제품 수요가 늘어나고 있다. 차세대 메모리 제품은 고용량, 고속연산, 고성능, 저전력의 특성이 요구된다. DRAM과 비휘발성 메모리(NVM)은 메모리 기술 발전을 주도해 왔고 독립형 표준 제품으로 개발되고 있기 때문에, 여기서는 DRAM과 NVM의 기술 발전 로드맵을 제시하고자 한다. 임베디드 메모리는 이와는 일정한 시간적 차이를 두고 같은 경향성을 갖고 발전할 것이다.

(1) DRAM

40여 년의 역사가 있는 DRAM 기술은 다양한 응용시스템에서 사용되고 있다. 예를 들어, 컴퓨터 및 DIMM(Dual In-line Memory Module) 제품용 DDR 메모리, 모바일용 저전력 LPDDR 메모리, 그래픽 전용 GDDR 메모리, AI 프로세서용 고대역폭 HBM 메모리, 고속 동작을 위한 LLDRAM(Low Latency DRAM) 메모리, 캐시 메모리용 eDRAM(embedded DRAM) 메모리 등이 있다.

그림 2-2처럼 DRAM 셀 구조는 현재의 트랜지스터에서 수직 채널 트랜지스터(Vertical Channel Transistor), 적층형 DRAM(stacked DRAM)으로 발전하고, 셀 레이아웃도 6F²에서 4F²로 미세화될 것으로 예측되지만, CMOS 회로를 메모리 어레이에 직접 결합하는 하이브리드 본딩 방식인 CBA(CMOS Directly Bonded to Array)도 셀 구조 발전에서 혁신적인 기술로 적용될 수 있다.

표 2-4는 DRAM 메모리 기술 로드맵이다. 여러 기술 중에서 Dual Work Function Word Line, One-Sided 캐패시터 공정, 그리고 Recessed 채널 S/A Transistor가 DRAM 제품에 적용되고

있으며^[8], EUV Lithography 기술도 본격적으로 적용되기 시작했다.

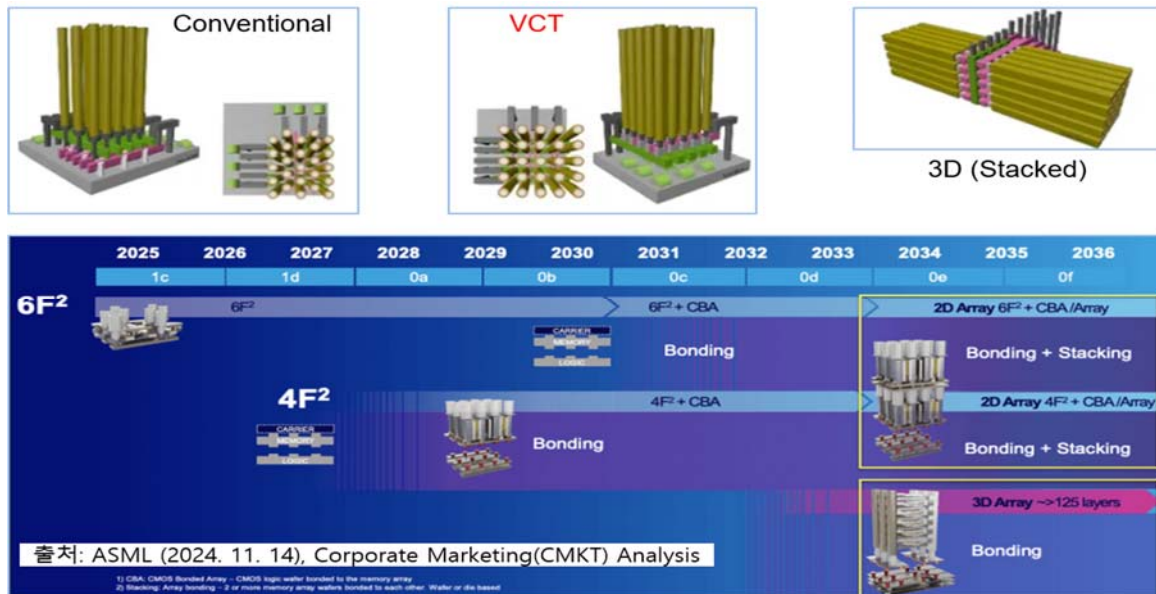


그림 2-2. DRAM 셀 구조 발전 방향

※ VCT (vertical channel transistor), Stacked (stacked DRAM), CBA (CMOS Bonded Array)

표 2-4. DRAM 기술 로드맵

YEAR OF PRODUCTION	2025	2028	2031	2034	2037	2040
Logic industry "Node Range" labeling	"2 nm eq"	"1.4 nm eq"	"1.0 nm eq"	"0.5 nm eq"	"0.3 nm eq"	"0.2 nm eq"
Fine-pitch 3D integration scheme	Stacking	Stacking	3DVLSI	3DVLSI	3DVLSI	3DVLSI
Logic device structure options	GAA	GAA	FSFET (IMEC) CFET	CFET	CFET	CFET (Atomic)
INTERCONNECT TECHNOLOGY						
Metallization – M0	Co, W, Ru	Co, W, Ru	Co, W, Ru	Co, W, Ru	Co, W, Ru	Co, W, Ru
Metallization – M1, Mx	Cu, Co, Ru	Cu, Co, Ru	Cu, Co, Ru	Cu, Co, Ru	Cu, Co, Ru	Cu, Co, Ru
Barrier Metal – M0, M1, Mx	0.5nm, TiN+WC	0.5nm, TiN+WC	0.5nm, TiN+WC	0.5nm, TiN+WC	0.5nm, TiN+WC	0.5nm, TiN+WC
Dielectrics k value – M0, M1, Mx	SiCOH (2.70-3.20)	SiCOH (2.70-3.20)	SiCOH (2.70-3.20)	SiCOH (2.70-3.20)	SiCOH (2.70-3.20)	SiCOH (2.70-3.20)
POWER DELIVERY TECHNOLOGY						
	Without BS-PDN	BSPDN (Backside Power Delivery Network)				
Power rail structure	Frontside Buried Rail	Backside Buried Rail	Backside Power Via	Backside Power Via	Backside Power Via	Backside Power Via
Power rail material	W, Ru	W, Ru	W, Ru	W, Ru	W, Ru	W, Ru
Power rail pitch	160~65 nm	115~42 nm	40~34 nm	40~28 nm	40~28 nm	40~28 nm

※ VCT (vertical channel transistor), Stacked (stacked DRAM), MR (Mass Reflow), MUF(Molded Underfill)

워드라인(Word Line)과 비트라인(Bit Line)의 저항 감소와 공정 개선을 위해 Ru, Mo 등을 포함한 여러 신재료의 개발이 진행되고 있다. 하지만 이러한 노력에도 불구하고 현재의 BCAT(Buried Channel Array Transistor) 셀에 대한 미세화는 7~8nm급 기술 노드에서 한계에 이를 것으로 예측된다.

이를 극복하기 위한 새로운 DRAM 아키텍처가 절실하게 필요한 상황이다. 이를 위해, High-NA

EUV 도입, X-DRAM을 포함한 여러 종류의 3D DRAM, 4F2 Cell, Capacitorless DRAM 등의 구조 평가와 공정 통합이 개발되고 있다. 물론, 새로운 구조 평가 이외에도 제품의 성능 향상과 전력 개선을 위해 추가적인 클로킹 기능 개선 등 회로 설계 및 동작 개선도 병행하고 있다.

DRAM 공정 개발의 핵심 이슈로는, 지속적인 셀 크기 감소, HKMG를 포함한 페리 영역의 게이트 공정, 셀 워드라인 및 비트라인의 신재료 적용, EUV 패터닝 적용 확대, 캐패시터 고유전율 막질 개선 및 신물질 개발, 차세대 3D DRAM 또는 4F2 DRAM 셀의 안정적 공정 확보 등이 중요하다. 또한, 중장기적으로는 고용량 하이브리드 본딩 DRAM 칩 개발과 고층 HBM DRAM 다이 패키지 기술 개발 등이 더욱 필요한 상황이다.

최근 AI 반도체 수요 급증으로 고대역폭 메모리(High Bandwidth Memory, HBM) 시장이 확대되고 있다. HBM은 여러 개의 DRAM 다이를 수직으로 쌓아 올려 동작 시킴으로써 저전력과 고대역폭을 갖도록 하는 구조를 갖는다. HBM은 메모리 반도체에 속하며, AI 학습을 위해 필요한 많은 양의 데이터를 GPU와 같은 시스템 반도체에 전달 및 저장하는 역할을 한다. 즉, 초거대 생성형 AI와 같은 대규모 데이터를 학습한 인공지능이 원활하게 연산할 수 있도록 도와주는 것이다. AI가 발전하며 학습량이 대폭 늘어나고 있기 때문에, 메모리 반도체의 중요성은 더욱 커지고 있다. 표 2-4에 HBM의 로드맵에 대해서 예측되어 있다.

HBM 주요 기술은 TSV 공정, 균일한 전력 배분, 발열 제어 기술, 마이크로 뱀프 인터페이스 공정, I/O 수 증가 등이 있으며, HBM 메모리 패키지 측면에서 보면 메모리와 프로세서 코어 사이의 신호 이동 거리를 줄이는 기술과 전력 사용 감소, 사이클 당 소요 시간 감소 역시 지속해서 개발해야 하는 과제들이다. 특히, TSV 기반의 TC-NCF(Thermal Compression Non Conductive Film)와 MR-MUF(Mass Reflow Molded Underfill) 기술에 하이브리드 본딩 기술을 접목할 수 있다면 앞으로 한층 뛰어난 성능의 제품을 개발할 수 있을 것으로 전망된다. 또한, 기존 컴퓨팅의 근간이 되는 폰 노이만 구조의 한계를 뛰어넘을 수 있도록 하는 PIM(Processing-In-Memory), CIM(Computing-In-Memory), AIM(Accelerator-In-Memory)과 같은 기술 개발이 필요하다. HBM 및 GDDR 기반의 AI 향 메모리 패키지 기술과 CXL 메모리 등의 고속 메모리 기술도 함께 개발돼야 할 필수 기술로 꼽힌다.

(2) NVM(비휘발성메모리): FLASH

비휘발성 메모리에는 다양한 형태와 구조가 있다. Kilo-Bit 수준의 저용량에서 수 Tera-Bit 수준의 고용량까지 응용 분야에 따라 요구 사항과 도전 과제가 다르다. 플래시 메모리는 한 개의 트랜지스터(1T) 셀을 기반으로 하며, 트랜지스터는 액세스 소자와 저장 노드의 역할을 모두

수행한다. 2D 플래시 메모리 밀도는 전하 기반 소자의 지속적 미세화에 의해 무한정으로 증가할 수 없고, 메모리 층을 수직으로 쌓음으로써 밀도 증가가 지속될 수 있다.

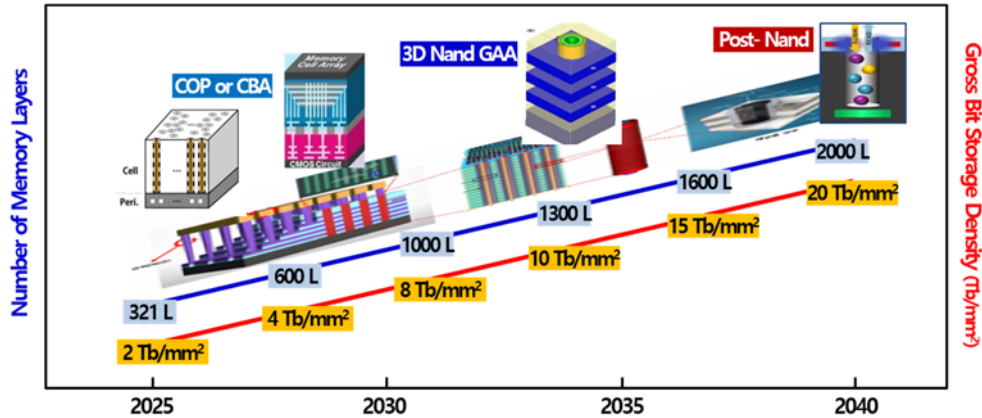


그림 2-2. 플래시메모리 기술 발전 방향

표 2-5. 플래시 메모리 로드맵

Year of production	2025	2028	2031	2034	2037	2040
Flash memory cell structure	3D-NAND	3D-NAND	3D-NAND	3D-NAND Post-NAND	3D-NAND Post-NAND	3D-NAND Post-NAND
Key technology for capacity increase	Multi-tiering	Multi-tiering	Multi-tiering	Multi-wafer bonding	Multi-wafer bonding	Multi-wafer bonding
	COP CBA	COP CBA	COP CBA			
Flash 2D NAND Flash uncontacted poly ½ pitch – F (nm) 2D	15	15	15	15	15	15
Product highest density (Tb/mm²) = $L \times N_{\text{cells}} \times B / A_{\text{chip}}$	2	4	8	10	15	20
Flash product maximum bits/cell	4	4	5	5	6	6
Flash 3D NAND maximum number of memory layers	321	600	1000	1300	1600	2000
Number of stacked tiers @ 3D-NAND	2	3~4	5~7	8~9	10~11	12~14
Maximum chip size (mm²)	140	140	140	140	140	140
High Bandwidth Flash Memory (HBF)		1st HBF				

설계기술과 함께 Read/Write 수명 혁신이 필요

표 2-5는 플래시 메모리 로드맵이다. 3D-NAND 메모리의 제조 공정에서 핵심은 Deep Hole 식각이고, 유전체 또는 폴리실리콘의 다층 박막을 증착하는 것 뿐만 아니라, 다층 막질과 Deep Hole에 대한 계측 기술도 필요하다. 많은 층들을 적층하는 것은 웨이퍼 뒤틀림(Warping) 문제가 발생할 수 있으며 공정기술로서 극복해야 한다. 3D-NAND 기술이 2025년에 321단의 플래시 메모리가 양산 시작되었으며, 2028년 이후에 600단이 가능하고 2031년경에는 1,000단의 플래시 메모리가 양산되는 것으로 예측되고 있다. 후술하는 공정 미세화, 3D 하이브리드 본딩 기술 등이 적용된다면 2040년에는 2000단도 가능할 것으로 기대된다.

지속적으로 증가하는 적층은 그에 대응하는 많은 워드 라인에 대한 contact 구성이 필요하며, 그에 따른 면적 오버헤드도 증가하므로 공정 비용을 줄이기 위해서 기술적 극복이 필요하고

워드라인 수직 방향 선평(Pitch)은 지속적으로 감소하여 가까운 시기에 40nm 이하가 요구될 것이다.

아울러 메모리 셀당 저장 비트 수를 늘리는 것은 기술적으로 어렵지만 진전이 있는 것으로 보인다. TLC(셀 당 3비트)를 넘어 현재 QLC(셀 당 4비트)가 대량 생산 중이며, 가까운 미래에 PLC(셀 당 5비트)등이 가능할 것으로 보인다. 셀의 저장 비트 수가 많으면 프로그래밍과 읽기에 시간이 더 오래 걸리고 논리 레벨을 서로 가깝게 압축할 때 신뢰성이 손상되므로 성능 측면에서는 트레이드오프가 필요하다.

3차원 구조 낸드 플래시 메모리의 고용량, 고집적, 고성능, 저비용화를 위해서 CuA(CMOS-under-Array) 구조, 하이브리드 본딩 기술 등이 제안되어 일부는 상용화까지 이어지고 있다. 현재 기술적인 도전 이슈로는, Cell string 전류 확보, HARC(High Aspect Ratio Contact) 장비/공정 개발, 극저온 식각기술 적용 확대, Carrier Mobility 증가, Stress/Warping 보상 공정, 고품질 하이브리드 본딩 기술, XY-셀 미세화, Multi-Bonded 낸드 기술, CMOS/Peripheral Area Scaling 등이 있으며 이러한 문제의 해결을 위한 노력이 지속되어야 한다.

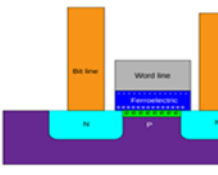
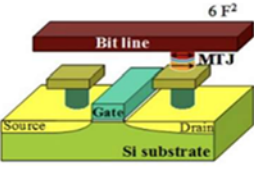
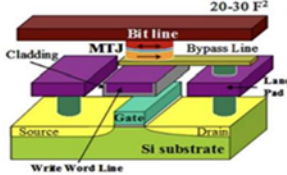
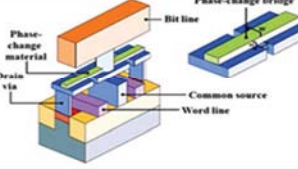
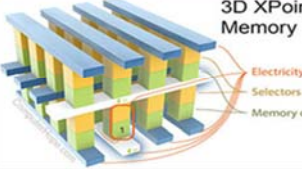
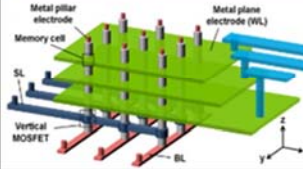
(3) 차세대 비휘발성 메모리

전하 저장에 기반하지 않는 몇몇 새로운 방식의 비휘발성 메모리(FeRAM, MRAM, PCRAM, ReRAM 등)가 개발되고 있는데, 이는 차세대 비휘발성 메모리 범주로 구분된다. 표 2-6은 차세대 비휘발성 메모리 기술 전망을 요약하였다.

FeRAM은 고속, 저전력, 저전압 비휘발성 메모리(NVM)의 대안이므로, 무선 주파수 식별(RFID), 스마트 카드, ID 카드 및 기타 임베디드 애플리케이션에 적합하다. 하지만, 공정 난이도로 인해 광범위한 채택에는 한계가 있다. 최근에는 강유전성이 FET의 문턱전압(V_t)을 변화시키는 역할을 하여 플래시와 유사한 1T 셀을 형성할 수 있는 HfO_2 기반의 강유전 전계효과 트랜지스터(FET)가 제안되었다. 개발이 성숙되면 저전력 및 매우 빠른 플래시와 같은 메모리 역할을 할 수 있다.

독립형 낸드 플래시를 대체할 STT-MRAM(spin-transfer torque-magnetic RAM)은 시간이 걸릴 것으로 보인다. STT-MRAM은 임베디드 플래시(NOR) 대체를 위한 잠재적인 솔루션이 될 것이다. 반면에 더 높은 메모리 밀도를 사용하는 다른 임베디드 시스템 애플리케이션의 경우, NOR 플래시가 여전히 훨씬 더 비용측면에서 효율적이고 미리 입력된 코드를 잃지 않고 인쇄 회로 기판(PCB) 솔더링 프로세스($\sim 250^\circ\text{C}$)에 견딜 수 있도록 잘 확립되어 있기 때문에 계속해서 우세할 것으로 예상된다.

표 2-6. 차세대 비휘발성 메모리 기술 전망

구 분	강유전체 메모리(FeRAM)	자기 메모리(MRAM)	
		MRAM	STT-MRAM
아키텍처			
특 징	- 빠르고 저전력이며 저전압 동작 - 안정한 강유전체 특성 물질 및 전극 재료를 찾는 것이 어려움	빠르고 무한 순환이 가능하여 SRAM 및 DRAM뿐만 아니라 NVM으로도 사용 가능	- 낮은 전력 소모로 고속 스위칭 구현 - 내장형 SRAM 대체 뿐만 아니라 내장형 플래시(NOR) 대체에도 가능
기술 수준	최근 강유전체 특성이 FET의 V_t를 변경할 수 있어 플래시 메모리와 유사한 1T 셀을 형성할 수 있는 HfO_2 기반 강유전체 FET가 제안됨	- Field Switching MTJ MRAM은 성공적으로 제품화 - FS-MTJ MRAM이 65nm 노드 이상 미세화 어려움	- 열 잡음이 발생하므로 32nm 이하에서 수직 자화물질이 필요할 것으로 예상 - Switching 전류가 커서 Shrink가 어렵고 신뢰성 해결이 필요
구 분	상변화 메모리(PCRAM)		저항 메모리(ReRAM)
	PCRAM	Cross Point	
아키텍처			
특 징	1T-1R(또는 1D-1R) 구조로 NOR Flash 대비 셀크기 장점 없음 - 임베디드 플래시, 뉴로모픽 소자 등에서 적용 가능성 있음	- BEOL 공정 제작으로 적층 공정이 용이하고 저렴 - 시스템에서 DRAM 기능을 강화하는 데 적합하여 광범위한 채택 가능성 존재	<10nm 미세화 낮은 전력 소비 및 높은 동작주파수(<ns) 특성 - 대형 어레이의 신뢰성 특성 충족 어려움으로 진입 장벽 높음
기술 수준	- 탄소 나노튜브 전극을 적용하여 5nm 이하의 미세화 구현 - 얇은 상변화 필라멘트를 생성하여 초저전력 동작이 가능한 기술 개발	- Chalcogenide 기반 상변화 물질의 OTS 특성을 갖는 X-point 셀의 Selector 구현 - 최초로 상업적 SCM을 실현함	- 저항성 메모리의 대부분은 아직 연구 개발 단계에 있음 - PCRAM의 Selector 기술 발전으로 연구 개발 가속화

I/O 처리량 향상, 전력 및 비용을 절감을 위해 스토리지 클래스 메모리(Storage Class Memory, SCM)의 3D Cross 메모리가 입증되었다^[9, 10]. Selector 소자를 포함하는 3D Cross Point 메모리는 BEOL(Back-end-of-line) 공정으로만 구현되므로 비트 비용을 줄이기 위해 다층을 적층 하는 것이 효율적이다. 다이오드의 작동 범위의 한계 때문에 고밀도 저항성 RAM (ReRAM) 개발이 제한되어 왔다. 그러나 최근 3D Cross 포인트 메모리의 발전으로 이러한 병목 현상이 해결될 것으로 보이며 불규칙한 비트와 같은 다른 기술적 문제가 해결되면 ReRAM이 빠르게 발전할 수 있다. PCM은 RRAM의 높은 가변성과 MRAM의 낮은 온-오프저항 비(R_{on}/R_{off})와 같은 부담을 극복하는 우수한 스케일링 방식으로 보인다.

4. 기술적 극복 과제

표 2-7은 소자 및 공정 기술의 미세화를 해결하기 위한 방안 및 차세대 소자 개발 수요에 대해서 단기(2025-2031) 및 장기(2032-2040)로 분류하여 제시하였다.

표 2-7. 기술적 극복과제

구분	단기(2025~2031) 해결 방안	장기(2032~2040) 해결 방안
성능	• 2nm급 GAA 소자 양산 수율 확보(@로직소자) • High-K 물질 도입(@DRAM 커패시터) • 초고종횡비 에칭(저온공정) 기술과 웨이퍼 휨 제어 기술 (@플래시메모리) • 하이브리드 본딩 기술(@로직 소자, DRAM, 플래시메모리) • 3D 적층을 통해 wire 길이와 단위 공간 당 구동 수 최소화 • 병렬 처리 방식 및 이중 적층의 수직 적층 로직-인-메모리 (logic-in-memory; LIM) 채용	• 수직 적층 N-FET, P-FET 공정(@로직소자 CFET) • 인공지능반도체에 HBF 적용을 위한 셀 수명 혁신 기술 • 유니버설 메모리: MRAM(자기저항), PRAM(상변화), ReRAM(저항변화) 등 • 뉴로모픽 & 양자 컴퓨팅 소자 개발
전력	• 소자 기생 효과 감소 • 고대역폭 메모리 도입에 의한 병렬 처리 소자 증가 • 세분화된 전력 게이팅(power gating) • 후면전력공급 기술(BSPDN) - Backside Buried Rail	• 후면전력공급 기술(BSPDN) - Backside Power Via
면적 및 비용	• DP-EUV 공정 • 칩렛과 이중 집적 기술 도입 • DTCO를 넘어서 STCO 강화 • 높은 규칙성을 갖는 레이아웃 설계 • Tier 간 배선을 최소화한 3D 집적 및 적층 구현	• High-NA(0.75NA) EUV 공정 도입 - 고체형 감광제 소재

5. 요약

로직 소자 및 공정 기술은 2025년 2nm급 GAA 소자에서 2040년 0.2nm급 CFET 소자로 발전할 것으로 예측된다. 이를 위해 GAA 소자구조에서 CFET 및 Monolithic 3D 집적화 기술을 기반으로 하는 3D-VLSI 기술이 개발되어야 한다. 미세화로 인한 소자 특성 한계로 공급전압은 0.5V~0.4V로 크게 개선되지 않지만, Transconductance 등 아날로그 특성은 유지될 것이다. DRAM의 최소 반선평폭은 2025년 11nm에서 2040년 6nm로 발전할 것으로 보이며, 또한 3D DRAM과 4F² 셀과 같은 새로운 아키텍처 개발도 진행되어 셀 크기가 144nm²까지 작아질 것이다. HBM은 2025년 12단의 2TB/s, 2031년 20단의 8TB/s, 2040년 30단 이상의 128TB/s 대역폭으로 발전할 것으로 예측된다. 이를 위해서 Bump-less Cu-Cu Direct Bonding 기술과 임베디드 냉각기술이 개발되어야 한다. 플래시메모리는 2025년 321단 적층에서 2040년 2,000단 적층까지 가능할 수 있다. 이를 위해서 고종횡비 식각, 공정 미세화, 다단 적층 및 COP, CBA, Wafer Bonding 기술 등이 개발되어야 한다. 전하 저장에 기반하지 않는 새로운 방식의 차세대 비휘발성 메모리(FerAM, MRAM, PCRAM, ReRAM 등)가 개발되고 있으나, 기존 소자를 대체하기 하기에는 기술적으로 해결할 이슈가 많다.

참고문헌

- [1] IRDS (International Roadmap for Devices and Systems), 2022 Update, More MOORE, available at <https://irds.ieee.org>.
- [2] G. Yeap, et al., "5nm CMOS production technology platform featuring full-fledged EUV, and high mobility channel FinFETs with densest 0.021 um² SRAM cells for mobile SoC and high-performance computing applications," IEDM, Section 36.7, December 2019.
- [3] G. Bae, et al., "3nm GAA technology featuring multi-bridge-channel FET for low power and high-performance applications", IEDM, Session 28.7, December 2018.
- [4] M. Badaroglu, et al., "PPAC scaling enablement for 5nm mobile SoC technology," ESSDERC, September 2017.
- [5] P. Batude, et al., "Advances in 3D CMOS sequential integration", IEDM, Section 14.1, p. 1-4, December 2009.
- [6] H. Mertens, et al., "Forksheet Field-Effect Transistors for Area Scaling and Gate-Drain Capacitance Reduction in Nanosheet-based CMOS Technologies", IEEE Electron Devices Technology and Manufacturing Conference, p. 1-3, March 2024.
- [7] C.-Y. Huang, et al., "3-D self-aligned stacked NMOS-on-PMOS nanoribbon transistors for continued Moore's Law scaling", IEDM, p. 20.6.1-20.6.4, December 2020.
- [8] J. Y. Kim, et al., "The breakthrough in data retention time of DRAM using recess-channel-array transistor (RCAT) for 88 nm feature size and beyond", VLSI, p.11, June 2003.
- [9] D. C. Kau, et al., "A stackable cross point phase change memory", IEDM, pp. 617-620, December 2009.
- [10] G. Molas, et al., "Crosspoint Memory Arrays: Principle, Strengths and Challenges", IEEE International Memory Workshop, May 2020.

III. 인공지능 반도체 로드맵

1. 서론

인공지능(AI) 응용 분야 중에서 특히 머신러닝(Machine Learning: ML) 도메인은 데이터에서 학습하여 분류, 예측, 추천, 추론 등 의미 있는 결정을 내릴 수 있는 애플리케이션을 가능하게 하는 기술 분류를 의미한다. 지난 수 년 동안, AI/ML 기술 중 하나인 심층 신경망(Deep Neural Network: DNN)의 발전은 컴퓨터 비전, 음성 인식 및 언어 번역과 같은 작업에서 인간을 초월한 정확도를 입증하였다.

이 패러다임의 성공에는 세 가지 중요한 요소가 있다. 첫 째는 훈련을 위한 레이블이 붙은 대규모 데이터 셋의 용이한 사용이었다. 두 번째는 DNN의 본질적이지만 비직관적인 확장성이다. 이는 1980년대 중반부터 알려진 역전파(Back Propagation)와 확률적 경사 하강법(Stochastic Gradient Descent)을 사용한 지도 학습(Supervised Learning) 기술에 기반을 두고 있다. 현재까지 우리가 아는 것은 대부분의 경우 이러한 유형의 네트워크가 더 크게 만들어지고, 더 많은 뉴런과 더 많은 가중치를 가지며, 더 많은 데이터로 학습될 때 높은 성능을 발휘한다는 것이다. 따라서, AI/ML이 성공할 수 있었던 주요 요소 중 하나는 강력한 컴퓨팅 플랫폼이었다. 기존에 사용하던 하드웨어인 CPU와 GPU는 DNN 기법이 실제로 확장 가능하다는 것을 실증적으로 증명하는데 결정적인 역할을 하였다. 하지만 이제는 이 알고리즘의 본질적인 확장성을 명시적으로 설계된 AI/ML용 하드웨어(CPU, GPU, FPGA, AISC 등)를 통해 상업적인 응용 프로그램으로 확장하는 것이 필요하다.

인공지능 반도체 로드맵은 AI/ML 영역의 애플리케이션 벤치마킹에 초점을 둔다. 이 분야의 발전을 추적하여 CPU와 GPU 등 기존 컴퓨팅 접근법의 지속적인 개선을 가능하게 하고, 이 응용 분야를 위한 진화적 및 혁명적 컴퓨팅 접근법 모두에 대한 정확한 성능 목표를 제공하고자 한다. 특히, 관심있는 성능 지수, 예를 들어, 처리량, 지연 시간, 정확도, 에너지 효율 등의 최근 추세를 정확하게 알아봄으로써, 가까운 미래까지 이러한 추세의 미래 진화를 추정하고자 한다.

아래 그림 3-1은 ML 응용을 제작하기 위해 필요한 다양한 선택지를 보여준다. 이 그림은 ML 성능 메트릭에 영향을 미칠 수 있는 무수히 많은 요소들이 있기 때문에 단순히 반도체 칩이라는 하드웨어를 통한 개선을 정량적으로 나타내기에는 매우 어려움이 있음을 시사하고 있다.

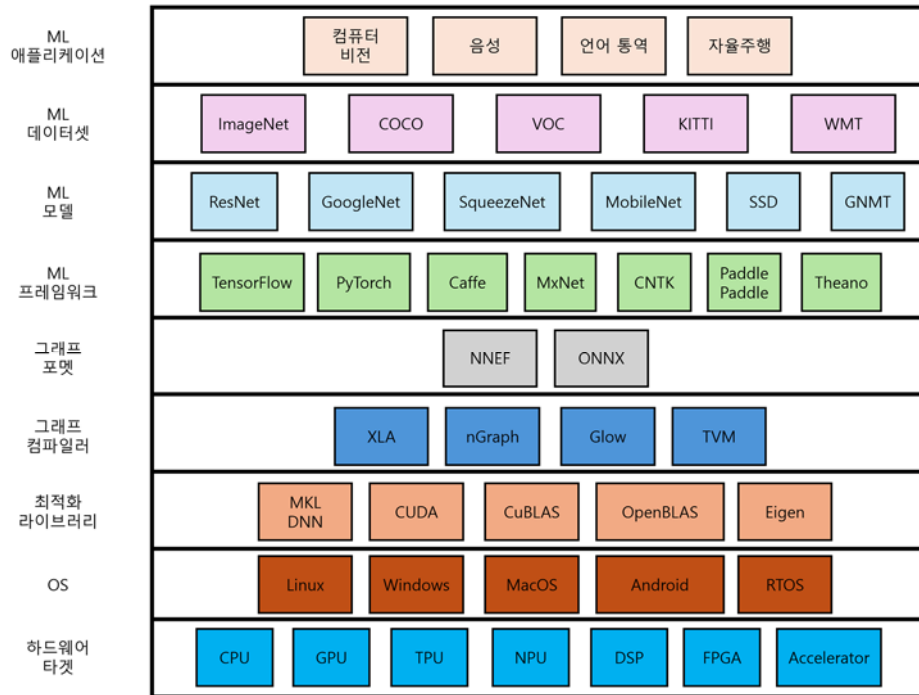


그림 3-1. AI/ML 시스템의 다양한 구성 조합

2. 응용 분야와 활용 플랫폼

AI/ML 알고리즘이 오랫동안 존재해왔지만, 특정 DNN은 지난 10년간 이미지 분류, 객체 탐지, 음성 인식, 기계 번역, 추천 시스템, 텍스트 감정 분석, 언어 모델링, 텍스트-음성 변환, 얼굴 식별, 이미지 분할, 이미지 개선^[2] 등 다양한 영역에서 상업적인 애플리케이션을 개발해왔다. 산업적으로 AI/ML은 프로세스 자동화, 인지, 사기 탐지 등의 애플리케이션에 사용된다. 스마트폰과 같은 기기와 구글, 페이스북, 유튜브, 넷플릭스, 아마존, 네이버, 카카오, KT, SKT와 같은 대형 인터넷 회사들과의 소비자 상호작용에서도 AI/ML을 활용한다.

AI/ML 애플리케이션은 응용분야별 요구에 기반하여 다양한 플랫폼에 배포된다. 이 플랫폼에서 요구되는 컴퓨팅 능력은 애플리케이션의 요구와 플랫폼의 폼팩터에 기반하여 다양하게 구성된다. 특정 AI/ML 시스템을 선택할 때 관련된 trade-off를 이해하기 위해서는 다양한 작업 부하 특성과 그 제약 사항을 이해하는 것이 중요하다.

여기에서는 DNN의 지도 학습에 초점을 맞춘다. AI/ML 기술은 훈련과 추론, 두 가지의 다른 명확한 단계를 갖는다.

- 훈련은 레이블이 지정된 데이터가 시스템에 공급되고 특정 정확도의 훈련된 모델이 구축

되는 과정이다. 모델의 훈련은 절대적인 과정이 아니며, 모델의 우수성, 데이터의 선택성, 다양한 Hyper-parameter의 세심한 조정에 따라 다양한 정확도의 모델을 출력할 수 있다. 일반적으로 더 많은 훈련 데이터와 더 큰 모델은 더 높은 성능과 정확도로 이어지지만, 이는 컴퓨팅 및 메모리 요구 사항을 증가시킨다. 모델이 극도로 커지면 훈련은 점점 더 분산되어 여러 노드가 같은 전체 작업을 연산하는 네트워킹을 포함하게 된다.

- 추론은 훈련된 모델이 배포된 후 실제 입력 데이터가 공급되어 추론이 이루어지는 과정이다.

훈련과 추론은 동일한 하드웨어에서 실행될 필요가 없으며, 이러한 계산 작업은 대상 응용 프로그램 제약 조건을 기반으로 여러 하드웨어 플랫폼에 분할될 수 있다. 대상 응용 프로그램 제약 조건은 일반적으로 지연 시간, 전력, 비용, 추론 속도, 훈련 속도, 정확도 등이 있다.

자율주행 자동차의 경우, 지연 시간은 사람이나 교통 신호에 반응하여 사고를 피하기 위해 절대적으로 중요하다. 따라서 자율주행 자동차의 추론은 자동차 안의 엣지 컴퓨터에서 실행되어야 한다. 한편, 스마트폰의 사진 태그 응용 프로그램의 경우, 지연 시간 제약이 없으므로 클라우드 서버를 추론에 사용할 수 있다. 반면에, 유해 가스를 감지하는 가스 센서라면 계산량이 많지 않을 것이기 때문에 추론을 위해 저렴한 마이크로컨트롤러(MCU)를 사용할 수 있다.

이러한 기본 분류를 사용하여, 세 가지 카테고리를 분류할 수 있다. 클라우드에서 수행되는 "고 성능" AI/ML, 엣지에서 수행되는 "실시간 추론" 및 "자원 제한 추론"이다. 자원 제한 추론은 TinyML로서 컴퓨팅 자원이 제약되는 경우이다. 이 경우 추론은 엣지에서 수행되지만, 비용, 볼륨, 면적 또는 메모리 풋프린트, 사용 가능한 배터리 전원 및 에너지량과 같은 자원이 심각하게 제한될 수 있는 MCU나 사물인터넷(IoT) 센서에서 수행된다.

훈련은 높은 계산량을 요구하므로 주로 클라우드에서 실행된다. 하지만 데이터 프라이버시의 이유로 엣지 컴퓨터에서 훈련을 실행하려는 시도가 일부 있으며, 분산 또는 연합 훈련 방법을 사용하여 가중치 업데이트 데이터를 교환하고 네트워크 규모의 이점을 누리면서 데이터 프라이버시나 다른 데이터 지역성 요구 사항을 희생하지 않는 방법도 사용되고 있다.

3. AI/ML 하드웨어

그림 3-2 는 다양한 컴퓨팅 구조 사이에서 생기는 일반적인 트레이드 오프를 시사한다. 이를 기준으로 AI/ML 하드웨어를 살펴본다. AI/ML 전용 컴퓨팅 하드웨어 시장은 빠르게 성장하고

있으며, 2025 년까지 모든 컴퓨팅 수요의 20%를 차지할 것으로 예상되는데, 이는 약 670 억 달러의 수익에 해당되는 것으로 계산된다^[3]. 앞서 언급했듯이, AI/ML 애플리케이션은 다양한 플랫폼에서 실행되며, 여기서는 이러한 플랫폼 전반에 사용되는 하드웨어를 설명한다.

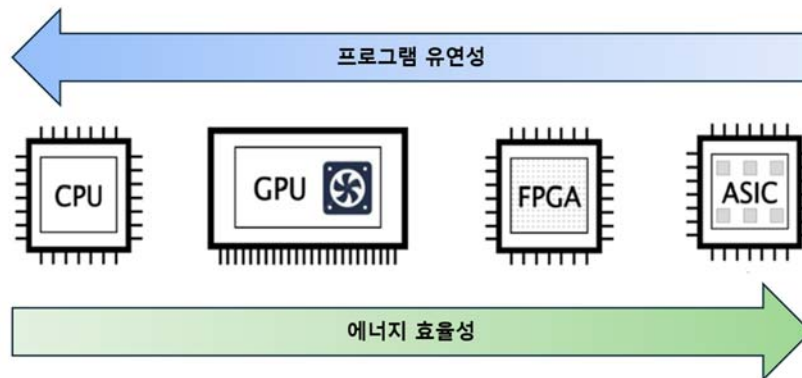


그림 3-2. AI/ML 컴퓨팅 하드웨어

(1) CPU

CPU 는 AI/ML 작업을 실행하는데 있어 기본적인 하드웨어이다. 추론은 데이터센터의 많은 코어에서 무겁게 실행되며, 스마트폰이나 IoT 기기의 간단한 CPU 에서도 실행된다. 하지만 CPU 성능 스케일링이 포화됨에 따라 더욱 높은 성능 향상을 기대하기는 어려우며, 이는 DNN 모델이 빠르게 진화함에 따라 도전적인 과제가 된다.

(2) GPU

GPU 는 AI/ML 학습 및 추론에 매우 큰 강점을 갖는 하드웨어 플랫폼이며, GPU 의 병렬 처리 능력은 DNN 의 행렬 곱셈을 가속화하는데 활용된다. 최근의 단일 GPU 는 대략 5,000 개 이상의 연산 코어를 가질 수 있으므로 고도의 병렬 작업이 가능하다. 또한 높은 내부 메모리 대역폭을 제공하므로, AI/ML의 방대한 계산량에 최적이다. NVIDIA 의 GPU 는 "Tensor Core"를 제공하여, 낮은 정밀도에서 밀집 행렬 곱셈을 효율적으로 실행할 수 있는 소형 엔진을 제공한다. 훈련은 높은 메모리 대역폭을 가진 고정밀 부동 소수점 연산을 요구하며, GPU 는 여전히 가장 인기 있는 하드웨어이다. 그러나 추론의 경우, 대부분의 GPU 가 소비전력이 매우 크기 때문에 엣지 애플리케이션에서 비효율적이다.

(3) FPGA

FPGA 는 DNN 모델의 데이터 경로를 맞춤화 할 수 있는 재구성 가능한 구조 덕분에 DNN 추론에 자주 사용된다. 또한, FPGA 는 비트 구성이 가능하므로 컴퓨팅과 메모리 효율성을 위해 낮은 정밀도에서 추론을 실행하는 데 매우 유용하다. FPGA 는 DNN 모델이 진화함에 따라 현장에서 재프로그래밍 할 수 있는 기능도 제공한다. 하지만 현재의 FPGA 는 고정밀 부동 소수점 유닛이 부족하기 때문에 훈련에는 다소 덜 적합하며, FPGA 의 전반적인 에너지 효율성과 운영 주파수는 여전히 ASIC 보다 낮다.

(4) 디지털 ASIC 가속기

AI/ML 연산에 있어서 CPU, GPU, FPGA 에서 언급된 비효율성으로 인해 AI/ML 을 위한 전문화된 디지털 ASIC 가속기가 급증하였다^[4]. 높은 수준에서 볼 때, 모든 디지털 DNN 가속기는 본질적으로 커스텀 스크래치 패드를 통해 공급되는 대규모 Multiply-Accumulate (MAC) 유닛 배열과 Pooling, Activation 등을 위한 특수 기능 유닛으로 구성된다.

다양한 프로토타입 디지털 가속기를 분류하기 위해서는 마이크로 아키텍처와 데이터 플로우라는 두가지 요소를 고려해야 한다. 우선 마이크로 아키텍처 수준에서는, MAC 유닛의 제어 패러다임에 따라 시스틀릭 배열과 SIMD/MIMD 배열의 두 범주로 분류할 수 있다. 시스틀릭 배열은 두 차원에서 파이프라인으로 연결된 MAC 유닛의 그리드로 구축된다. 데이터는 데이터 플로우 전략에 따라 수평 및 수직 방향으로 매 사이클마다 이동한다. 전체 배열은 중앙 집중식 컨트롤러를 가진 단일 행렬-행렬 곱셈 코어로 작동하며, 개별 MAC 유닛은 정지할 수 없다. 반대로 SIMD/MIMD 배열은 각각 MAC 유닛과 일부 로컬 저장소를 포함하는 처리 요소(PE)의 군집으로 구축된다. 이 배열은 전체 MIMD 유닛으로 운영되거나, 독립적으로 제어될 수 있는 여러 SIMD 유닛으로 구성될 수 있다. 이 두 접근 방식은 현재 실 제품에 사용된다.

예를 들어, 구글의 TPU^[5]는 시스틀릭 배열을 사용하는 반면, 엔비디아 딥 러닝 가속기인 NVDLA^[6]와 TESLA FSD^[7]는 SIMD/MIMD 공간 배열을 사용한다. 시스틀릭 배열은 더 간단하고 높은 계산 밀도를 제공하지만, 차원이 배열에 잘 매핑되지 않거나 데이터의 희소성이 있는 경우 활용도가 떨어질 수 있다. SIMD/MIMD 배열은 활용도가 더 좋고 매핑 유연성이 높지만 분산 제어가 필요하여 더 복잡해지는 단점이 있다^[8].

데이터 플로는 데이터의 반복 순서와 MAC 유닛 간의 분할 전략을 총칭한다. 데이터 플로우 선택은 전반적인 데이터 재사용을 결정하며, 이는 차례로 배열 내외부의 데이터 이동량에 영향을

미친다. 데이터 재사용 비율을 증가시키는 데에는 두 가지 방법이 있다. 첫째는 미니배치 내의 여러 데이터에서 가중치와 입력을 재사용함으로써, 둘째는 합성곱 커널 내의 소수의 가중치가 입력에 걸쳐 재사용되는 합성곱 레이어를 사용하여 가능하다. 가중치 재사용은 특히 중요한데, 이는 데이터를 메모리(특히 오프칩 메모리)에서 검색하는 데 드는 큰 에너지 비용을 훨씬 더 많은 수의 계산에서 데이터를 재사용함으로써 분산시키기 때문이다.

반복 순서 측면에서, 가속기는 Weight/Input/Output Stationary 등 다양한 전략이 있다. 분할 전략 측면에서는, 입력/가중치 텐서의 다른 차원들이 병렬성을 위해 MAC 유닛에 걸쳐 분할될 수 있다. 이러한 아키텍처 및 데이터플로우 개념 중 일부는 디지털 ASIC 가속기 제작 시 탐색되고 있으며, 특히 이러한 AI/ML 애플리케이션을 위해 FPGA를 프로그래밍할 때에도 활용될 수 있다.

(5) CIM

Compute-in-Memory(CIM)은 SRAM이나 RRAM과 같은 소형 NVM과 같은 디지털 메모리 장치를 사용하여 맞춤형 메모리 배열에서 다중 비트 시냅스를 인코딩하고, 이는 AI/ML 추론 및 훈련에 필요한 MAC 연산을 실제 메모리 읽기 중 수행하도록 설계된다. 이러한 접근 방식은 일반적으로 각 시냅스 가중치에 대해 여러 Bit-Line 또는 Word-Line을 포함하며, 메모리 읽기 과정 중 뉴런 자극에 의한 시냅스 가중치의 곱셈이 수행되는 방식으로 대부분 다른 비트라인과 워드라인(BL/WL)에서 2진 인수 연산을 가정한다. 따라서 이러한 CIM 매크로는 기존 메모리보다 소자 간 변동성에 더욱 의존적이지만, 개별적으로 메모리 word에 접근하는 시간을 피하거나 데이터의 MAC 연산에 필요한 회로 블록의 면적과 전력을 절약하는 장점이 있다.

(6) 아날로그 ASIC 가속기

아날로그 컴퓨팅을 기반으로 하는 새로운 가속기의 형태가 등장하고 있다. 이는 밀집된 아날로그 배열 내의 전류 합을 활용하여 대규모 벡터-행렬 계산을 매우 효율적으로 수행하므로, 높은 속도와 높은 에너지 효율성을 모두 가능하게 한다. 그러나 이러한 아날로그 ASIC 가속기는 아직은 충분히 개발되지 않은 새로운 비휘발성 메모리 기술에 의존하거나, NOR-FLASH와 같이 스케일링이 중단된 기존의 비휘발성 메모리에 의존한다. 또한, 아날로그 시스템의 본질적인 문제인 신호 대 잡음 비(Signal to Noise Ratio)가 현대 DNN 시스템에서 사용되는 매우 큰 모델을 견고하게 지원할 수 있을 정도로 높아질 수 있는지 여부는 아직 명확하지 않다. 이에 반해서, 디지털 가속기는 디지털 계산의 정밀도를 임의로 줄일 수 있어서, 원래 신경망의 정확도를 온전히 지원하면서 최고의 에너지 효율을 달성할 수 있을 것이다.

4. 최적화

(1) 데이터 양자화(Quantization)

데이터 양자화는 더 낮은 정밀도에서 추론을 실행하는 방법이다. 대부분의 현대 DNN 은 16 비트와 8 비트에서 추론을 수행하며, 실제로 일부는 4 비트까지 더 낮아진다. 낮은 정밀도는 이동해야 할 데이터의 양과 MAC 수행 시간을 줄이고, 사용 가능한 영역에서 더 많은 컴퓨팅 자원을 지원할 수 있다. 전체 비트 수와 부동 소수점 형식의 경우 지수에 할당된 비트 수는 중요한 영향을 미칠 수 있다.

(2) 프루닝(Pruning)

Pruning 은 훈련 중에 활용되는 접근 방식으로, 임계 값 혹은 블록 등 일정한 휴리스틱에 기반하여 특정 입력, 가중치 또는 둘 모두를 "0"으로 만든다. Pruning 후 추가 훈련의 수행으로 기존 네트워크와 동일한 정확도를 가질 수 있다. 이를 통해 가중치의 개수를 줄여 메모리 요구 사항을 줄일 수 있으며, 필요한 계산 수를 줄여 총 실행 시간과 지연 시간을 줄이고 처리량을 증가시킬 수 있다. 하지만 이로 인해 DNN 은 희소 행렬의 형태로 나타난다. 따라서 하드웨어가 희소성을 활용하도록 설계되지 않는 한, 이는 하드웨어 내의 컴퓨팅 유닛 활용도에 영향을 줄 수 있다.

(3) 하드웨어-소프트웨어 공동설계(co-design)

하드웨어-소프트웨어 공동설계는 하드웨어와 신경망 모델이 함께 공동으로 설계되는 접근 방식이다. 이는 현저히 높은 에너지 효율성을 달성할 수 있지만, 지원해야 할 DNN 모델의 개수가 작을 때, 그리고 맞춤형 하드웨어에 대한 긴 주기 시간을 용납할 수 있을 때 가장 잘 작동한다.

(4) 활용도(Utilization)

활용도는 CPU, GPU, 디지털 가속기 내의 디지털 처리 요소가 유용한 계산을 수행하는 시간의 비율이다. 일반적으로 데이터 플로우가 잘 못 설계되면 다음에 계산할 데이터가 아직 도착하지 않아 PE 가 유휴 상태가 될 수 있다. 디지털 시스템 내의 많은 트랜지스터에서 발생하는 대량의 누설 전력을 고려할 때, 유휴 PE 를 최소화하여 활용도를 높이는 것은 높은 에너지 효율성을 위한 핵심이다.

5. 인공지능 반도체 로드맵

그림 3-3 은 최근 CPU, GPU, FPGA 및 일부 DNN ASIC 하드웨어의 학습을 위한 TOPS/W 단위의 컴퓨팅 효율과 예측이다. 여기에서는 FP32 또는 BF16 정밀도 포맷을 사용한 Training 하드웨어에 초점을 맞춘다. 더 포괄적으로 FP16 하드웨어도 포함하지만, 일부 연구에 따르면 모든 모델을 정확하게 학습할 수 있으며, FP32 와 같은 지수 비트 수를 사용하지만 가수에서는 훨씬 적은 비트를 사용하는 BF16 포맷이 필요하다. 반면에 특정 애플리케이션 혹은 DNN 모델에서는 8 비트 정밀도에서도 학습이 가능하기도 하다. 소비 전력과 연산량은 연산하는 정밀도와 밀접한 관련이 있다. 따라서 GPU 의 경우 2040 년까지 100 TOPS/W 이상 (FP64) 또는 10 TOPS/W 이상 (FP32)의 성능을 보일 것으로 예측된다. ASIC 의 경우 FP16 에서는 100 TOPS/W 에 가까워지며, FP8 을 활용하는 경우에는 1000 TOPS/W 이상의 효율을 달성할 수 있을 것으로 전망된다.

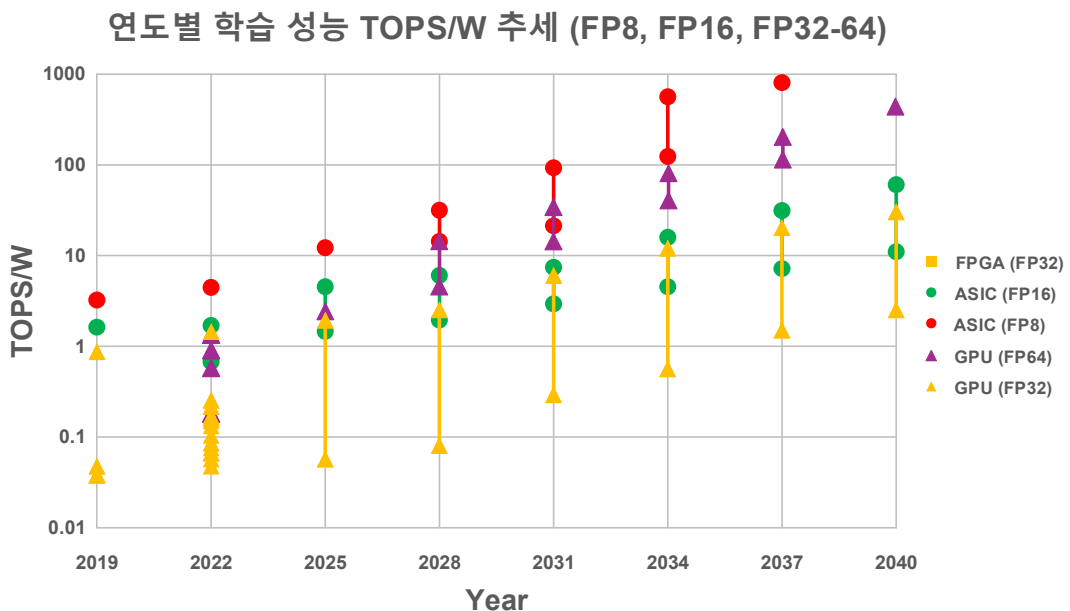


그림 3-3. 학습을 위한 하드웨어의 연산 효율 발전 동향

그림 3-4 는 그림 3-3 의 데이터를 새로 그린 것으로서, 프로세서 소비전력(W)에 대한 TOPS 와 TOPS/W 의 발전 추세를 보여준다. 타원형으로 표시된 구역은 2040 년까지 각 프로세서가 발전할 것으로 전망되는 영역을 표시한 것이다. FP8 ASIC 의 경우 적은 연산량에 비해 소비하는 전력이 줄어들어 100 TOPS/W 이상을 달성할 것이며, GPU 는 처리량(TOPS)은 높은 반면 소비전력이 높아서 그래프의 우상단에 위치하는 경향을 보일 것이다.

그림 3-5 는 유사한 플롯으로, 최근 CPU, GPU, FPGA 및 DNN ASIC 하드웨어의 추론을 위한 TOPS/W 단위 컴퓨팅 효율과 예측이다. 전반적으로 32 비트에서 16 비트, 8 비트로 줄어드는 정밀도의 변화가 분명히 드러난다. 학습에 관한 앞선 예와 마찬가지로, INT8 ASIC 은 100 TOPS/W 를 선회할 것으로 예상되며, FPGA (INT16) 또한 높은 효율을 가질 것으로 전망된다.

그림 3-6 은 시스템 소비전력(W)에 대한 TOPS 및 TOPS/W 성능의 추세를 보여준다. 마찬가지로, 2040 년까지 각 프로세서가 발전할 것으로 전망되는 영역을 타원형으로 표시하였다. INT8 ASIC 의 경우 100 TOPS/W 이상을 달성하여 FPGA 와 ASIC 은 그래프의 좌상단을 차지할 것이며, GPU 는 우상단에 위치하는 경향을 보일 것이다.

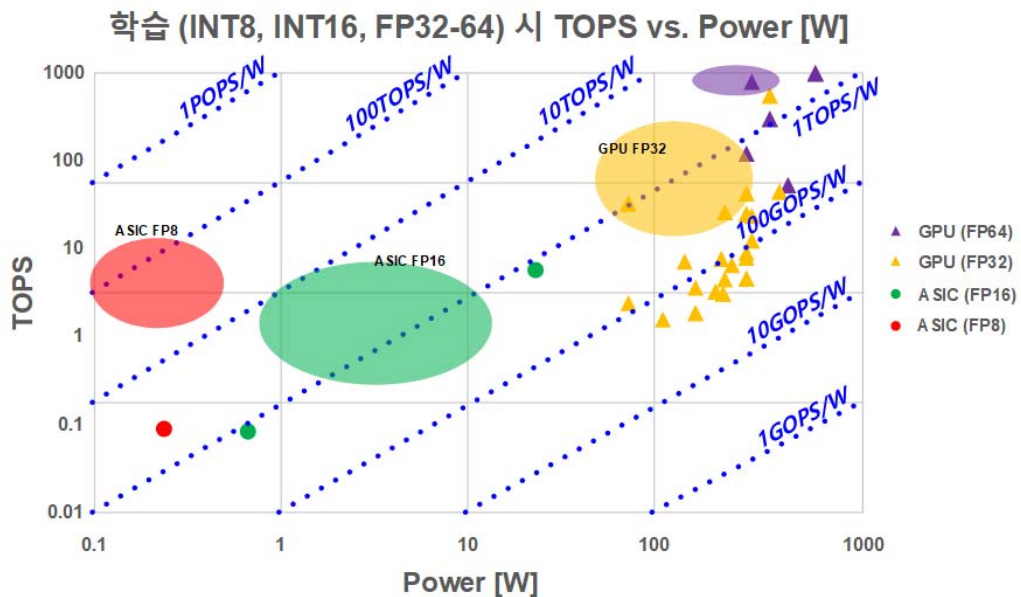


그림 3-4. 학습을 위한 하드웨어의 성능 대비 시스템 소비 전력 발전 동향

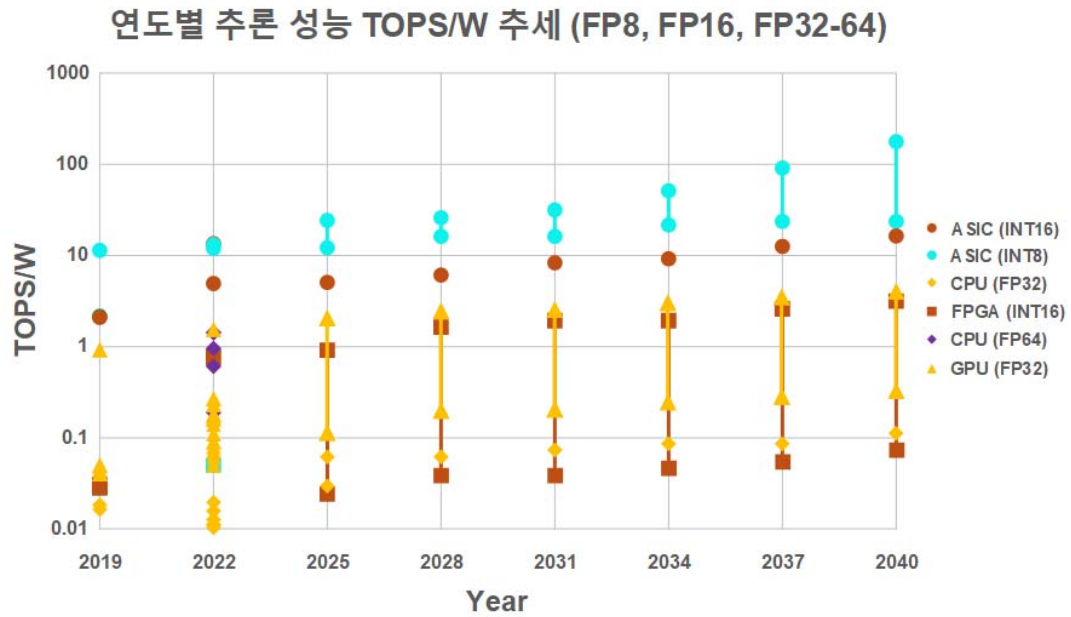


그림 3-5. 추론을 위한 하드웨어의 연산 효율 발전 동향

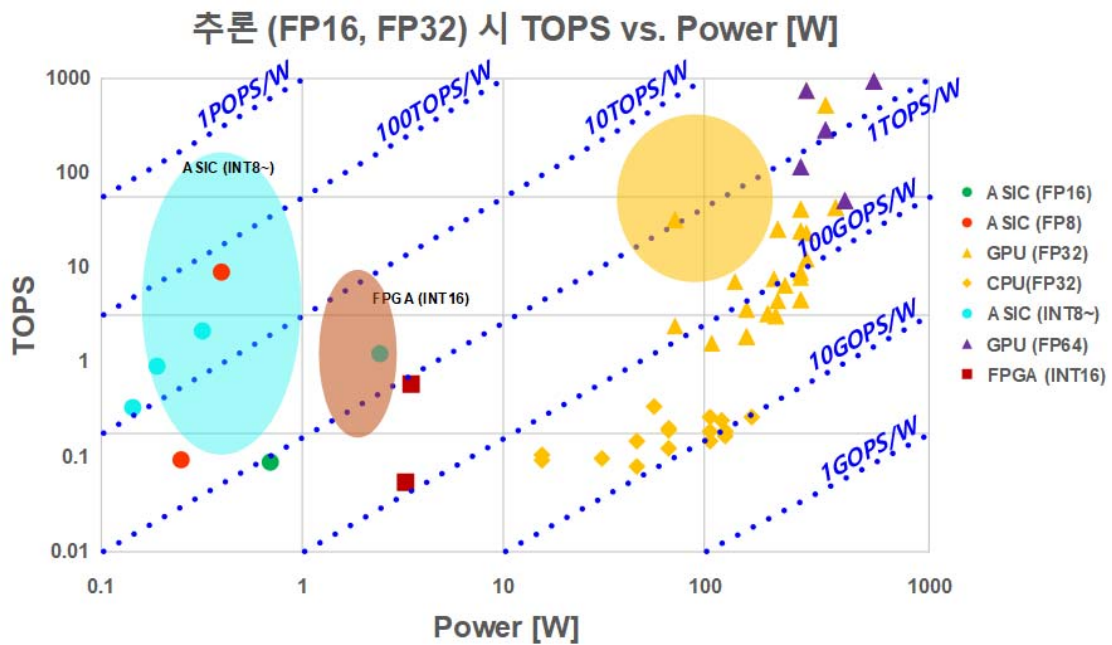


그림 3-6. 추론을 위한 하드웨어의 성능 대비 시스템 소비 전력

6. 요약

인공지능 반도체 기술은 학습과 추론으로 나눌 수 있으며 하드웨어와 정밀도에 따라 편차가 크다. 학습을 위한 컴퓨팅 능력은 2025 년 0.1~10 TOPS/W 에서 2040 년 5~1000 TOPS/W 로 발전할 전망이며, 추론을 위한 컴퓨팅 능력은 2025 년 0.1~10 TOPS/W 에서 2040 년 1~100 TOPS/W 로 증가할 전망이다. 다만, 이러한 추세는 현재의 정밀도에 기반한 것이며, 새로운 정밀도가 출현하는 미래에는 예측 값이 달라질 것으로 보인다. 결론적으로 요구되고 달성 가능한 컴퓨팅 능력은 특정 응용에 최적화되어 달라질 것으로 예상된다.

참고문헌

- [1] V. Janapa Reddi, C. Cheng, D. Kanter, P. Mattson, et al., "MLperf Inference Benchmark," <https://arxiv.org/abs/1911.02549>
- [2] <https://www.microsoft.com/en-us/research/project/project-brainwave/>
- [3] <https://www.mckinsey.com/industries/semiconductors/our-insights/artificial-intelligence-hardware-new-opportunities-for-semiconductor-companies>
- [4] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer. "Efficient processing of deep neural networks: A tutorial and survey," Proceedings of the IEEE 105(12), 2295-2329 (2017).
- [5] N. P. Jouppi et al., "In-datacenter performance analysis of a tensor processing unit" in 2017 ACM/IEEE 44th Annual International Symposium on Computer Architecture (ISCA), pp. 1-12.
- [6] <http://nvdla.org>
- [7] D. Das Sarma and G. Venkataramanan, "Compute and redundancy solution for Tesla's Full Self driving computer," IEEE Hot Chips 2019.
- [8] <https://www.sigarch.org/dnn-accelerator-architecture-simd-or-systolic/>

IV. 광연결 반도체 로드맵

1. 서론

초연결 기술에서 빅데이터 생성과 전송은 필수적 요소이다. 광연결(Optical Interconnect)은 기존 데이터 센터에서의 활용을 포함하여, 인공지능(AI)과 초고속컴퓨팅(HPC) 기능의 클라우드 네트워크 시스템에서 빅데이터를 초고속으로 전송하는 역할을 하며 더 나아가 데이터 생성과 이를 위한 연산 및 분석 프로세스를 수행한다. 광연결은 향후 IoT(Internet-of-Thing)를 위한 광센싱(Optical Sensing)과 FTTH(Fiber-To-The-Home)를 비롯하여 자동차, 항공우주, 의료, 공장 자동화, FSOI(Free Space Optical Interconnect), LiFi 등 초연결 사회를 위한 다양한 기술 분야로 그 적용을 넓혀가고 있다. 또한, 기존 반도체 소자 및 패키징 공정과의 융합과 양자컴퓨팅(Quantum computing: QC)과의 응용을 포함한 미래 기술로의 발전도 예상된다.

현재 광연결은 기존 구리 배선의 한계인 신호의 지연과 왜곡, 발열, 소비전력증가, 운영비용증가 등의 문제를 극복하는 대안으로 고려되고 있다. 근거리 통신망(LAN), 무선통신 기지국, 클라우드 데이터 센터에 연결하는 통신 네트워크, 데이터 센터 내에서의 시스템 간(Inter) 연결 및 데이터 센터 밖으로의 장거리 통신(>40km)에 적용되고 있으며, 최근 AI 와 HPC 를 포함한 초고속 데이터 전송을 위하여 데이터 센터의 연산 유닛 내(Intra) 연결과 개별 연산 유닛 간(Inter) 연결로 응용분야가 확대되고 있다.

광연결 반도체 기술 로드맵에서는 데이터 센터를 중심으로 개별 시스템 내 및 시스템 간의 전송을 위한 광 연결 기술과 이에 필요한 재료, 부품, 중간 모듈 및 시스템에 대하여 기술하고자 한다. 응용 영역에 따라서 시스템 내 광 연결에 해당하는 Inside-of-Rack 과 시스템 간 광 연결에 해당하는 Outside-of-Rack 으로 구분한다. 또한, 전송 거리에 따라 XSR(Extreme Short Reach, <1m), SR(Short Reach, <100m), DR(Data Center Reach, <500m), FR(Fiber Reach, <2km)로 구분하여 고려할 수 있다.

2. 광연결 트랜시버 발전 방향

광연결을 구성하는 기술의 분류와 본질은 전-광 및 광-전 신호를 변환하여 전송 및 수신하는 기능을 하는 CPO(Co-Packaged Optics) 개념으로 동일하나, 구현되는 제품의 용도에 따라 시스템 내(Intra-System)의 PCB 상에서 기존의 Cu 배선을 대체하는 부분을 통상적으로 Inside-of-Rack

CPO로 지칭하고, 반면에 시스템 간(Inter-System) 연결하는 송수신기를 Outside-of-Rack Pluggable 트랜시버 또는 스위치로 구분한다.

Outside-of-Rack 에 해당하는 Pluggable 트랜시버 및 스위치는 성능향상과 소형화를 목표로 기술 개발이 이루어지고 있다. 그림 4-1 은 광연결의 기술개발방향을 시스템 제공업체의 CPO(Inside of Rack) 관점에서 보여주고 있다^[1]. 1 세대(Gen-I)에서의 연산 처리된 데이터는 구리배선 기반의 PCB 를 통해서 전송되므로 신호 왜곡, 혼선 및 지연이 존재하고, 이를 극복하기 위해서는 Retimer 또는 DSP 기능이 추가적으로 필요하다.

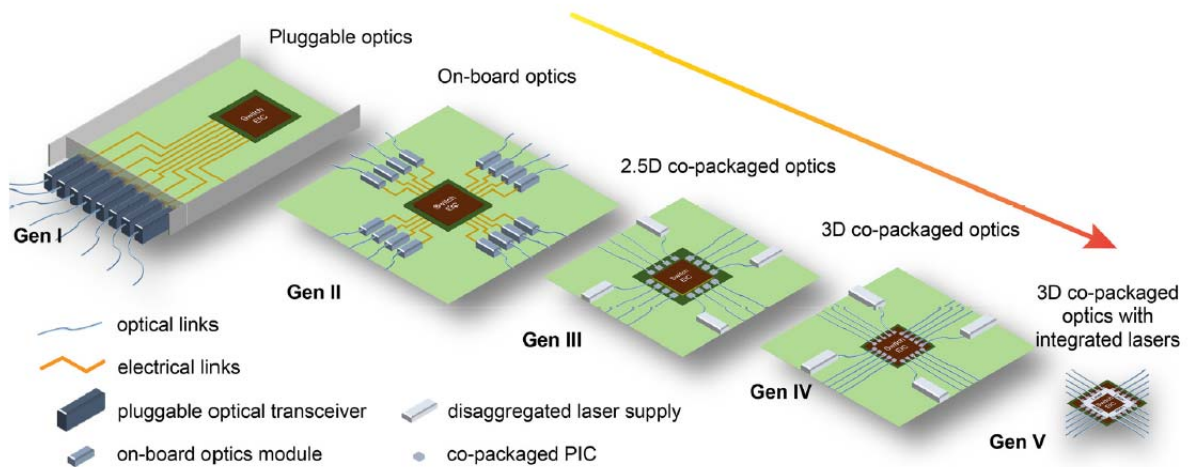


그림 4-1. 광연결 CPO (Inside-of-Rack) 발전 방향[1]

1 세대(Gen-I)에서의 고비용, 소비전력증가, 발열 등의 문제는 Cu-Link 의 길이를 줄이는 2 세대(Gen-II) 에서 OBO(On-Board Optics) 구조 적용으로 개선될 수 있다. 하지만 2 세대(Gen-II)는 현재 수요를 대처하기 위한 초고속 데이터 전송(>100Gbps/Lane) 에는 한계가 있으므로 Cu-Link 를 Optical link 로 대체하기 위한 노력이 진행중이다.

3 세대(Gen-III)의 NPO (Near-Packaged Optics) 는 PCB 의 끝단부에서 Pluggable 성격으로 광신호의 전송을 담당하는 전-광 변환기능장치(이하, 광엔진)를 연산 장치에 최근접시키고 PCB 의 Cu-link 를 Optical link 로 대체하는 방식이다. 현재 멀티모드 레이저 (VCSEL)를 기반으로 국제공동연구 (APC: Advanced Photonics Coalition, 구 COBO: Consortium for On-Board Optics) 가 추진되고 있다.

4 세대(Gen-IV)의 CPO (Co-Packaged Optics)는 연산 장치와 광엔진을 반도체 패키징을 기반으로 단일 칩렛화하여 외부의 구리배선을 완전 제거한 구조이다. 4 세대(Gen-IV)는 성능, 제조비용, 운영관점에서 다양한 연구가 이루어지고 있으며, 이종의 광부품과 반도체부품을 단일 칩렛화하는 웨이퍼수준 패키징(Wafer Level Packaging) 이나 웨이퍼수준 조립(Wafer Level Assembly) 기술이 광연결 반도체 기술의 새로운 도약 기회로 인식되고 있다.

궁극적으로는 시스템(Intra) 에서의 ELS(External Laser System)와 ILS(Integrated Laser System) 적용과 동일 반도체 기판에 Monolithic 광-전 소자 제조 기술을 결합하여, PCB 없이 광연결 송수신기를 구현하는 5 세대(Gen-V) 기술로 발전할 것이다.

3. 응용 분야와 기술 개발 과제

광연결의 목적은 빅데이터의 생성, 처리, 전송 과정에서 기존 구리 배선의 활용을 최소화 하거나 완전히 제거하여 전송능력을 획기적으로 향상하는 것이다. 여기에는 양산 적용과 시설 운용 관점에서의 경제성 및 친환경성을 확보하는 것이 필수적이며, 광연결을 데이터 센터 외의 다른 분야로 확장하기 위해서는 다양한 기술적 플랫폼을 확보해야 한다. 데이터 센터에서의 적용을 시작으로 HPC, QC(양자컴퓨팅), FSOI 및 LiFi 등의 근거리통신, 자동차 및 항공우주분야 등 다양한 응용처가 있다.

(1) 데이터센터

데이터센터는 현재의 IT 네트워킹에서 가장 빠르게 성장하는 분야 중의 하나이며, 고성능, 고밀도, 저전력, 저비용을 요구조건으로 전세계에 확산되고 있다. 이런 상황에서 연산장치 내 및 연산장치 간의 모든 영역에서 발생하는 빅데이터 전송의 병목은 큰 장애물이 되고 있으므로 반드시 해결해야 한다.

광연결의 확대 도입을 위해서는 기존에 구성되어 있는 플랫폼 인프라와의 경쟁을 피할 수 없다. 단지 성능 향상만이 아닌 새로운 플랫폼 적용이 성능, 투자, 운용 등 경제성 측면에서도 우수해야 한다. 특히, 초고속 전송을 위한 200Gbps/lane 급 요구를 위해서는 광연결 소자를 보드, 카드, 패키지, 칩렛 레벨로의 통합 여부와 데이터 네트워킹 아키텍처 결정이 관건이 될 것이다. 또한 광엔진 소형화에 의한 고밀도화로 기존 제품의 규격(Form Factor)을 최대한 활용하는 게 바람직하다.

O Outside-of-Rack: Pluggable

본 로드맵에서 언급하는 데이터 센터 내에서의 전송 거리는 멀티모드 레이저(VCSEL)가 적용되는 SR(~100m)과 싱글모드 레이저(EML)가 적용되는 DR(~500m) 및 FR(~2km)로 구분될 수 있다. 많은 수의 연산 장치, 서버, 스토리지 등의 연결에 QSFP (Quad Small Form-Factor Pluggable) 또는 OSFP (Octal Small Form Factor Pluggable) 광 트랜시버 구조이거나 대용량의 스위치 형태로 적용된다.

각 부품들을 CoB(Chip-on-Board) 형태로 실장하고 각각의 CWB(Cu-Wire Bond)로 연결하는 구조는 200Gbps/lane 급의 초고속 데이터 전송에 불리하므로, 웨이퍼수준 패키지(Wafer Level Package, Fan-Out Wafer Level Packaging)이나 웨이퍼수준 조립(Wafer Level Assembly) 방식으로 대체될 것이다.

데이터 센터 내의 단거리 통신(SR, <100m)에는 싱글모드(EML) 보다 멀티모드(VCSEL) 레이저가 소비전력, 비용, 광커플링 측면에서 유리하므로 주류를 이루고 있다. 현재와 같이 멀티 모드와 싱글 모드의 조합으로 유지될지 싱글 모드 위주로 전환될 것인지는 앞으로 시장 선택에 의해 결정될 것이며, 개별적인 레이저를 구동 IC 로 직접 변조 구동시키는 방식과 현재의 실리콘 포토닉스(Silicon Photonics: SiPh) 기술처럼 CWL(Continuous Wave Laser)를 간접 변조 구동 시키는 방식도 일정기간 함께 발전할 것이다. 광엔진 및 소자를 3D 칩렛 형태로 제작하는 패키징 기술로의 전환이 대세를 이루며 특성 향상 및 소형화가 이루어질 것이다.

장기적으로는 광라우팅이 전-광 및 광-전 변환을 최소화 내지 제거할 것으로 예측되며, 광데이터 스트림을 라우팅하기 위해 MEMS^[2]와 DMD 를 적용한 Hybrid E/O routing 이 시연되었다. 또한, 1ns 수준의 SOA(Semiconductor Optical Amplifier)나 MZ (Mach-Zehnder) 모듈레이터도 입증된 바 있으며, MRM(Micro Ring Modulator)의 적용 시에도 동일한 결과를 확보하였다^[3]. 궁극적으로는 광로직을 적용한 광라우팅으로 데이터 센터 내에서 중간단계에서 존재하는 전-광-전 변환없이 고속 전송하는 새로운 통신 시스템 개발을 목표로 한다.

O Inside-of-Rack: CPO

광연결 로드맵에서 NPO 적용 중간단계인 Gen-III 및 CPO 적용 최종단계인 Gen-IV 는 PCB 에서의 구리배선길이를 최소화 내지 완전 제거하는 것으로, Outside of Rack 의 구성과 역할을 PCB 상에서 Fly-over 방식보다는 매립(Embedded) 시키는 것이다. Gen-III NPO 시스템에서는 적절한 구조를 갖는 소형화된 광엔진을 연산 장치에 최근접 배치하고, Gen-IV CPO 에서는 연산 기능과

광엔진기능의 단일 칩렛화가 필요하다. 또한, PCB 에서 도파관과 레이저를 광커플링해야 하며, 레이저 모드에 따라서 광경로를 변경해주는 별도의 광학 부품 도입이 필요하다.

Outside-of-Rack 에 적용되는 Pluggable 트랜시버에서 기존 실리콘 포토닉스 확장의 주요 장애 요인 중의 하나로 싱글모드에서의 측면 광커플링 공정이 있으며, 이는 도파관을 매립(Embedded)하는 형태의 Inside-of-Rack 송수신기에서는 더욱 가중될 것이다. 이를 해결하기 위해서는 멀티모드 방식이 우선 개발되고 있다. 멀티모드나 싱글모드는 광연결 장치의 구성과 운용에서 잠재적인 경쟁 기술이지만, 단거리나 초단거리 전송(수 cm ~ 100m 이하)에서는 850 nm 의 VCSEL 을 적용하는 멀티 모드가 소비전력 및 공정 용이성 면에서 유리한 것으로 알려져 있다.

반도체의 웨이퍼수준 패키징이나 조립 공정을 도입해서 광엔진이나 실리콘 포토닉스 소자의 칩렛화를 추구하는 것은 Outside-of-Rack 의 경우와 동일하며, 각각의 레이저 소자에서 나노 구조 및 공정 기술을 기반으로 출력 에너지 밀도를 증가시키는 연구개발이 진행중이다. 장기적으로는 Outside-of-Rack 과 동일하게 광로직이 적용된 광라우팅과 관련된 신규 아키텍처의 도입으로 전-광-전 변환을 최소화 또는 제거하여 무손실 고속 전송하는 방식을 추구하고 있다.

(2) HPC(High Performance Computing)

기존 HPC 는 초고속 연산과 많은 데이터를 처리하는 이산 이벤트 (Discrete Even) 시뮬레이션, 물리적 시스템 시뮬레이션, 다항 변수가 포함된 최적화 등을 목적으로 구리 배선 저항을 최소화할 수 있는 극저온에서 주로 수행되었다. 최근에는 ML(Machine Learning)을 위한 방대한 양의 데이터를 처리, 분석하거나 거대언어모델(Large Language Model: LLM)을 운용하는 초고속 연산분야로도 역할을 확장하고 있다.

(3) 양자컴퓨팅(Quantum Computing)

Qubits (Quantum bits)는 1 과 0 만의 기존 비트와는 달리, 동시에 다수의 조합을 가지는 중첩 상태(Superposition State)를 가질 수 있으며, 양자컴퓨팅 프로세서는 이러한 Qubits 들에 대한 연산을 수행한다. 양자컴퓨팅은 현재 초기단계로서 극복해야 할 과제가 많이 있으나, 총 Qubits 수에 따라 연산능력이 기하급수적으로 증가하는 장점을 가지고 있으므로, 광연결 기반의 양자 네트워크를 구성하여 양자컴퓨터를 병렬 연결하면 현재로선 상상할 수 없을 수준의 연산 속도와 연산량을 달성할 수 있는 기술로 여겨진다.

양자컴퓨팅 분야에서 광연결 반도체는 많은 기회와 과제를 동시에 가지고 있다^[4]. 연산을 수행하는 논리 큐비트와 이를 구성하는 물리적 큐비트가 있으며, 큐비트의 생성과 제어를 위한 기술에 따라 광학, 마이크로파, 전압 또는 전류에 의한 상호 연결이 필요하다.

양자통신 시스템은 단일 광자의 손실과 교란에 취약하며, 유효 데이터 속도와 전송 거리 증가를 위해서는 큐비트의 생성과 전송 및 감지 방법의 개선이 필요하다. 또한, 양자 네트워크에서 각각의 목적에 맞는 중첩 상태를 제공하기 위하여 다른 큐비트 기술이 필요한 경우에는 다른 유형의 큐비트를 사용하여 양자 프로세서간의 통신을 수행해야 한다^[5].

양자 컴퓨팅으로의 광연결 확대를 위한 다양한 연구가 진행 중으로, 극저온 시스템의 큐비트 상태를 기존 전자장비에 전송하기 위한 레이저 소스와 극저온 광검출기 간의 통신, 초전도 큐비트의 광학 입출력 등은 실험실 차원에서 입증되어 향후 초전도 큐비트로 광학의 입력과 출력을 확장할 수 있는 가능성을 제공하고 있다^[6].

(4) 자동차와 항공우주

자동차와 항공우주 분야에서 광연결 기술은 저전력의 고속 데이터 전송을 위한 통신 기술로서 사용되며, 우수한 내구성, 내습성, 내열성, 내한성, 내충격성을 장점으로 기존 제품군을 대체하고 있다. 또한, 구리배선을 대체하는 광연결은 저전력화 및 경량화에도 기여한다.

자율주행 차량에서는 주로 무선통신을 사용하여 네트워크와 통신하지만, 운행 중에 발생하는 환경변화와 사물을 감지하고 이후 거동 분석 예측을 위해서는 다수의 자체 센서가 차량 내외부에서 네트워크를 구성하고 있다. 예를 들어, 광연결 시스템과 동일한 원리로 동작하는 LIDAR (Light Detection and Ranging)는 수 cm 부터 수백 m 거리에 해당하는 물체의 3D 형상을 실시간으로 확인할 수 있다.

항공우주분야에서의 광연결은 자동차에서의 활용뿐만 아니라, 전자기파 간섭에 대한 내성 관점에서 유리하다. 광섬유 설치가 필요 없는 FSOI (Free Space Optical Interconnect) 와 LiFi 는 기내에 적용되어 통신, 인터넷 접속 및 인포테인먼트에 응용될 수 있고 EMI 에 영향 받지 않고 저전력을 장점으로 한다^[7].

4. 광 집적회로 기술

광연결 송수신기에서 데이터 전송속도를 높이고 전력소모를 낮추기 위해서는 반도체 집적회로 기반의 광집적회로 및 소자 기술이 필요하다. 또한, 이를 광연결 및 전송에 사용하기 위해서는 전자소자와 광소자가 혼합된 광집적회로(Photonic-Integrated Circuit: PIC)의 형태로의 개발이 요구된다.

광송신 집적회로(Transmitter: TX)는 레이저 다이오드와 광신호의 변환과 전송을 효율적으로 수행하는 모듈레이터, 도파관, 멀티플렉서(MUX) 등으로 구성되고, 광수신 집적회로(Receiver: RX)는 도파관, 디멀티플렉서(DEMUX), 광검출기, 증폭기 등으로 구성된다. 초고속 데이터 전송에서의 신호 왜곡을 방지하게 위해서 Retimer 또는 DSP 가 포함되기도 하나, 소비전력 및 비용 증가의 단점이 있다. 전기적인 신호 처리와 전송은 CMOS 소자, BIPOLAR 소자, FET 소자와 같은 다양한 반도체 기술이 적용 가능하나, 본 로드맵에서는 광집적회로(PIC) 와 관련된 부문만 다루기로 한다.

○ PIC(Photonic Integrated Circuit)

광융합 패키징을 제공하는 기존의 일부 반도체 파운드리에서는 Monolithic 기반의 PIC 를 공급 중이다. 광송신 집적회로 에서는 대부분 Silicon-on-Insulator(SOI) 기반의 도파관과 변조기를 채택하고 있으나, MZ 변조기는 크기와 소비전력 측면에서, 마이크로링 변조기는 온도 안정성 문제로 기술적 극복과제가 존재한다. 이를 해결하기 위해 기존 SOI 외에 박막형태의 리튬나이오베이트(Thin Film Lithium Niobate: TFLN)로의 재료 변경도 고려되고 있는 상황이다.

PIC 제품 적용 확대를 위해서는 광연결을 위한 광커플링 기술의 획기적 개선이 필수적이다. 예를 들어, 싱글모드 광연결에서 Passive Coupling 이 가능한 플러그형 고밀도 광섬유 커넥터 (ex, Fiber Array Unit) 개발과 PIC 와의 연결 및 통합이 관건이 될 것이다.

향후에는 편광 변조, PAM(Pulse Amplitude Modulation), QAM(Quadrature Amplitude Modulation) 등의 다양한 변조 기술을 적용하여 데이터 전송속도를 향상시키려는 기술개발이 필요하며, 이를 위한 반도체 집적회로, 패키징, 모듈 기술의 도약이 요구된다.

○ 변조기(Modulator)

EOM(Electro-Optic Modulator)은 입사되는 빛의 진폭, 위상, 주파수, 편광을 변조하는 부품으로 광학효율증가, 저전력, 소형화, 원가 경쟁력을 갖춰야 한다. 크기와 소비전력에서의 문제였던 초기의 Mach-Zehnder 간섭계는 현재 소형화가 많이 진전되었고, 레이저와 한 기판에 통합하는

수준에 이르게 되었다. 100 μ m 길이의 p-i-n 방식의 MZM (Mach-Zehnder Modulator)는 5pJ/bit 수준의 위상 변조가 가능하다^[8].

직경 20 μ m 이하의 실리콘 p-i-n 링 다이오드의 파장 변조 기능은 입증되었으며, 파장 분할 다중화(WDM, Wavelength Division Multiplexing)에 적용될 수 있다. 또한, 실리콘 p-i-n 다이오드 구조에 광자 결정 (Photonic Crystal)을 통합하면 파장 변조 시의 소비전력을 줄일 수 있다.

전기 흡수 변조기(Electro Absorption Modulators)에서는 높은 On/Off 비율, 소형화, 높은 작동 주파수, 낮은 광학 및 에너지 손실 등이 주요 기술적 이슈이다. 박막 상태의 III-V 반도체나 Ge 에 전기장이 가해지면, 반도체 밸런스 및 컨덕션 밴드에 경도(gradient)를 유발하고 터널링 및 파동함수의 중첩을 증가시켜 밴드갭 근처에서 광학 흡수를 증가시키는 Bulk Semiconductor Franz-Keldysh Effect 가 발생한다. SOI 기반의 SiGe 변조기는 50 μ m 길이의 활성영역에서 3.0V 바이어스에서 5.9dB 의 On/Off 비율로 28Gb/s 에서 작동하는 것으로 보고되었다^[9].

10 μ m 의 실리콘 도파관을 가지는 플라즈모닉(Plasmonic) MZM 가 25fJ/bit 의 효율을 가지고 70GHz 까지 동작되는 것이 확인되었고, Plasmon 을 활용하는 링모듈레이터는 2.5dB loss, 12 fJ/bit 및 낮은 열 드리프트의 상태에서 100GHz 이상에서 동작한다. Plasmonic MZM 로 제작된 변조기는 0.07 fJ/bit 의 에너지 효율로 50Gb/s 속도로 동작하고, 2 fJ/bit 에너지 효율로 400Gb/s 속도에서 동작하는 것으로 확인되었다^[10].

광변조기(Molulator)의 물리적 원리로 주로 두가지가 고려된다. 첫번째는 결합된 전자-정공 쌍(Excitons)이 양자 우물에 갇힐 때 발생하는 반도체 양자 우물의 Stark 효과이다. 이는 밴드갭 근처에서 광학 흡수를 증가시키거나, 전기장이 가해지면 전자와 정공의 파동 함수의 중첩이 감소하여 밴드갭 근처에서의 광투과율을 증가시키는 효과를 발생한다. 이는 Ge 양자 우물에서 확인되었으며, 90 μ m 크기의 Ge-SiGe 양자 우물 변조기는 3V 와 4V 사이의 스윙전압으로 9dB 의 On/Off 비율과 108 fJ/bit 의 에너지 손실로 23GHz 에서 동작할 수 있다^[11]. 두번째는 초격자(Superlattice)와 같이 결합된 양자 우물을 포함하는 미니 밴드가 있는 반도체에서 발생하는 Wannier-Stark Localization 이다. 이는 적용된 전기장이 양자 우물을 분리할 만큼 클 때 광흡수가 급격히 감소하는 현상이며^[12], InGaAs/InAlAs 초격자 구조에서 1.55 μ m 파장 범위에 대해 100 μ m 길이의 도파관에 0.7V 전압을 적용하여 11dB 의 Extinction loss 가 보고된 바 있다.

O Multiplex(MUX) 및 De-Multiplex(DEMUX)

MUX 는 송신부에서 여러 파장을 결합하여 한 개의 광섬유나 도파관을 매개로 신호를 전송하기 위한 소자이고, DEMUX 는 수신부에서 각각의 파장을 분리하는 소자이다. 이를 위해서 도파관은 전체 내부 반사를 유지하면서 신호 손실을 최소화하기 위해 점진적인 곡률반경을 가지고 형성되어야 한다. 많은 파장이 단일 광섬유나 도파관에 병합됨에 따라 크기가 커지게 되므로 소형의 저손실 병합 및 분리 기능을 수행하는 MUX/DEMUX 소자 기술 개발이 필요하다.

비주기적 나노광자(Aperiodic Nanophotonic) 구조와 플라즈모닉(Plasmonic) 구조는 파장 병합 기능을 μm 수준으로 축소하여 MUX/DEMUX 소자의 소형화에 기여할 수 있으며, 기존 SOI 재료의 단점인 열감도가 낮은 광학 지수를 가진 새로운 재료 개발이 필요하다.

플라즈몬 구조: 플라즈몬 구조는 빛의 전파 방향을 조밀하게 바꾸고 작은 형상에 빛을 집중할 수 있는 특성을 가지고 있으며^[13], 재료 선정에 따라 광자와 상호작용하고 100nm 이하의 작은 도파관을 가능하게 하는 표면전자공명을 가지고 있다. 또한, 플라즈몬 구조는 광검출기의 광학 흡수를 향상시키고 도파관에서 하위 파장대의 빛을 제한하며 $1\mu\text{m}$ 이내에서 빛의 진행방향을 변경하는 것으로 알려졌다. 하지만, 플라즈몬이 공명할 때 빛으로부터 에너지를 흡수하기 때문에 상호작용하는 길이가 짧아야 한다는 제약이 있으며, 하이브리드 플라즈몬 도파관의 손실을 줄이기 위한 전기 펌핑 방식의 도입도 제안되었다^[14].

향후의 논리 연산 대응을 위해 플라즈몬 흡수기 및 증폭기는 WDM 응용에서 특정 파장을 증폭하면서 다른 파장을 흡수하고, 위상 순수 기준 소스에 따라 비위상 순수 소스의 위상을 정제하고, 증폭을 통해 입력게이트 광신호에 의해 구동되는 광신호를 변조하거나 방향성 있는 광학 절연 기능을 제공할 수 있도록 설계되었다. 플라즈모닉스는 조밀한 광자 라우팅 및 변조를 가능하게 할 뿐만 아니라 잠재적으로 광로직 기능을 가능하게 할 수 있다.

비주기적인 나노구조: 비주기적 및 동적인 광자 구조의 적용은 향후 필요한 복잡한 시스템에 통합될 소형 기능 요소/부품을 위해 필요하다. 비주기적 구조의 적용은 단일 파장 스케일의 도파관에서 고밀도 소형 라우팅을 가능하게 하며, 동적 구조는 동적 비상호성과 같이 정적시스템에서 사용할 수 없는 기능 외에 재구성 기능도 제공할 수 있다.

비주기적 광자 구조의 상호작용을 이해하는데 많은 진전이 있었으며, 나노 광자 구조의 시뮬레이션과 최적화를 가능하게 하는 초고속 수치 알고리즘의 개발을 들 수 있다^[15]. 이를 사용하면 나노포토닉 구조의 매우 큰 조합을 스캔하여 구조적 통계성 변동을 이해하고 모드 및 파장 분할 다중화 시스템을 위한 소형의 집적된 구성 요소를 설계할 수 있다.

○ 도파관(Waveguide)

실리콘 포토닉스 칩의 고집적화 및 소형화를 달성하려면 신호 혼선과 광학 손실이 적은 도파관이 필수적이다. 실리콘 기반의 도파관 주위에 유전체 메타물질의 클래딩 형성 시, Evanescent wave 손실을 1/30 로 줄이고, 도파관 곡률에 의한 굽힘 손실도 1/3 로 줄일 수 있다고 입증되었다^[16]. 또한, 반파장 피치의 초격자(Superlattice)에 의해 Crosstalk 을 30dB 이하로 줄일 수 있다. 플라즈몬 금속 도파관이 빛을 제한하고 유도하는 능력은 입증되었으며, 최근에는 TFLN 과 같이 도파관 재료 자체의 변경을 통한 성능/안정성 개선을 추구한다.

○ 광기반 스위칭(Switching) 과 라우팅(Routing)

현재 광신호는 진행하는 단계의 전기 라우터 영역에서 광-전 변환과 전-광 변환을 거쳐서 다음 단계로 전송되며, 별도로 형성된 장치나 시스템을 거치면서 많은 지연 및 대기 시간이 필요하다. 이를 극복하기 위해서는 반복적인 광-전 변환과 전-광 변환이 필요 없는 완전 광기반의 스위칭 및 라우팅(Optically-based Switching & Routing) 기술의 개발이 필요하다.

하이브리드 전-광 라우팅 기능은 MZ 간섭계와 링변조기를 포함한 MEMS 및 E/O 스위치를 사용하여 입증되었고, 광스위칭을 지원하려면 광제어 라우팅 기능을 동작 시키기 위한 새로운 재료와 장치 구조가 필요하다. 광학 기반 스위칭은 III-V 화합물 반도체 및 실리콘 기반 구조, 그리고, 마이크로 링 공진기가 있는 비정질 실리콘 기반 구조에서 입증되었으며, 이는 향후 광연결 기술의 중심이 될 것으로 기대된다^[17].

광네트워크를 운용하려면 신호 트림 라우팅 충돌을 식별하고 올바른 라우팅 대안을 결정하기 위한 광학 기반의 로직이 필요하다. 광학에서의 국부적 비선형성을 필요로 하는 많은 광학 논리 장치 및 해당 기능과^[18] 광학 편광 스위치를 기반으로 광로직도 제안되었지만^[19], 전송 속도와 변환 효율성 향상을 수반한다. 광네트워크에서 필요한 광로직에 대한 중요한 요구사항은 광신호에 노출될 때, 100GHz 이상의 속도에서 광학 특성을 변경시킬 수 있는 재료이다.

5. 광연결 반도체 로드맵

광연결 로드맵과 현재 광연결 산업의 현황을 고려하여 향후 2040 년까지의 중장기 기술개발 로드맵을 예측하면 아래 표 4-1 와 같으며, 단일 채널 (Lane, 1 optical fiber)에서 구현할 수 있는 데이터 전송속도를 기준으로 정리하였다.

중기적으로 2025년부터 PAM4 방식 기준의 200Gbps/lane 급의 도입을 시작으로 400Gbps/lane 까지 증가하며, 시스템 내에서 Gen-III NPO 구현까지 시도될 것이다. 기술 및 산업 기반으로서의 실리콘 포토닉스 산업의 표준 제정이 예상된다. 실리콘 포토닉스는 SOI 기반의 도파관이나 변조기 위주의 광집적소자 관점에서 시작되었으나, 현재는 실리콘을 포함한 Ge, GaAs, InP 등의 다양한 반도체 재료와 공정기반의 광소자를 포괄하는 더 넓은 산업적인 개념으로 진화하고 있다. 또한, FOWLP 같은 웨이퍼 레벨 패키징 기술과 결합하여 반도체소자와 광학 부품의 이종 패키징을 포함한 광융합반도체 산업으로의 성장도 예상된다.

표 4-1. 광연결 반도체 기술 로드맵

Item		Link	2025	2028	2031	2034	2037	2040
Inside of Rack	On-chip Data Rate/Lane (Gbps)	Cu	100	100	200	200	200	200
	Package-level Data Rate/Lane (Gbps)	Cu	56	56	100	100	100	100
	Package-level Energy Consumption (pJ/bit)	-	1	1	0.05	0.05	0.05	0.05
	Date Rate/Lane (Gbps, PAM4**)	Optical	-	-	400	400	800	800
Outside of Rack	Date Rate/Lane (Gbps, PAM4**)	Optical	200	200	400	400	800	800
	Date Rate/8 Lanes module (Gbps)	Optical	1600	1600	> 3200	> 3200	> 6400	> 6400
	AOC Energy Consumption (pJ/bit)	-	2	2	< 1.5	< 1.5	< 1.0	< 1.0

Optical Interconnect for system (APL 118, 220501 (2021))	Gen 2	Gen 3	Gen 4	Gen 5
	OBO(Cu) by Assembly	NPO by Min. Cu Link	CPO by Chiplet	CPO (Laser -Integ.)
Paradigm (Keyword)	*SiPh for Optical Interconnect (AI / HPC / Cloud / Datacenter)		Optical Logic & Routing (QC with Optical Link)	

* SiPh (extended) : Silicon (semiconductor)-based Optical Device, by OPTICA

** PAM4 in 2024. High order modulations are expected

일반적인 광연결 제품의 사양에서 표기되는 데이터 전송 속도는 그림 4-2 와 같이, 단위 채널(Lane) 에서의 레이저 동작 속도, 모듈레이션 방식, 제품을 구성하는 채널 개수의 세가지 변수 조합으로 결정된다. 모듈레이션 방식은 NRZ(None Return to Zero)에서 PAM4(4-Level Pulse

Amplitude Modulation)까지 제품 적용되었으나, 개선된 EIC (Electrical IC) 와 연산 로직에 기반하여 PAM16 이나 QAM (Quadrature Amplitude Modulation) 같은 High Order Modulation 방식을 적용할 경우에는 데이터 전송속도를 비약적으로 증가시킬 수 있을 것이다.

장기적으로는 고차 변조(High-Order Modulation) 방식을 포함한 800Gbps/lane 이상의 성능 확보, Gen-IV 의 CPO 적용 및 초고속 데이터전송을 위하여 전/광 변환을 최소화한 Hybrid E/O 변환 및 광로직(Optical logic)이 도입될 것으로 예상된다. 또한, 광로직을 위한 관련 기술·재료·소자·시스템과 양자컴퓨팅과의 결합에 의한 상승효과로 초고속 연산 및 데이터 전송에서 획기적인 성과가 창출될 것으로 기대된다.

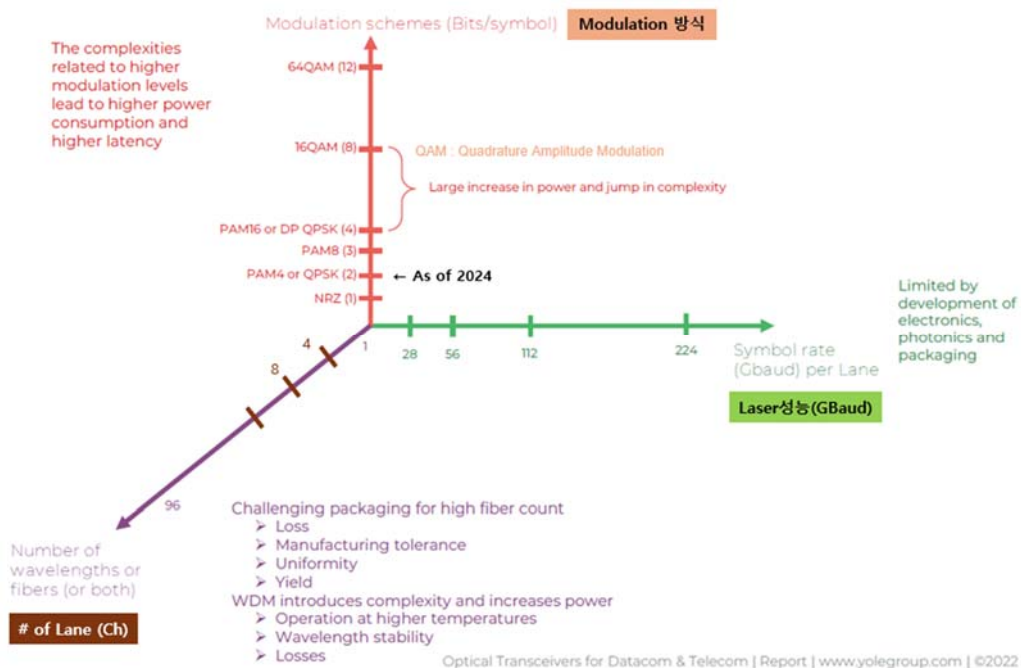


그림 4-2. 광연결 기술 및 성능 지표

이러한 광연결 기술 발전을 위한 중장기 기술개발 과제는 다음과 같다. 우선 향후 5 년 정도 기간의 중기 과제로서, 광연결이 적용되는 다양한 분야에서 데이터 전송속도 향상, 신호 왜곡 방지, 신호 대기 시간 감소 (수 μ sec $\rightarrow$ 수 nsec), 구동 전력 및 발열 감소에 의한 총소비 전력 감소, 소형 폼팩터에 의한 고밀도 제품 구현 등이 극복되어야 한다.

다른 산업분야로의 확대 전개는 광엔진과 같은 개념으로 동작하는 광센서로의 응용을 기반으로 ToF 또는 LIDAR 형태로 스마트폰이나 차량용의 3D 거리 측정에 도입되며, 항공우주, 의료, 산업현장, 가정으로의 전개를 위한 경박단소의 새로운 시스템 플랫폼 도입의 촉진을 포함한다.

현재 광연결의 주 적용 분야인 데이터 센터에서의 빅데이터 전송은 AI 와 LLM 기반의 ML(Machine Learning), 기존 슈퍼 컴퓨팅을 포함한 HPC(High Performance Computing), 다양한 형태로 구성되는 클라우드 시스템 등에 응용될 것이다. 이를 통해 효율적인 초고속 연산의 장애요인인 데이터 병목현상 해소, 소비전력 감소, 시설운영 비용절감 및 친환경 부문 등에서 획기적인 해결책을 제공한다.

표 4-2 는 중기 기술개발 과제를 정리한 것이다. 중기적 관점에서는 광엔진을 적용한 실리콘 포토닉스 분야의 기술 개발, 제품 표준의 확립과 시너지 창출이 향후 광연결산업의 장래를 좌우할 것으로 예상된다.

표 4-2. 향후 5 년간의 중기 기술 개발 과제

개발 항목	주요 아이템 과 이슈
광엔진 (전/광 변환)	기존: PIC 와 EIC 를 Chip on Board 방식으로 배치후, Cu wiring bond 수행 향후: Wafer Level Packaging 적용에 의한 단일 Chiplet 화 (2D 평면구조) Chiplet 상부에 Flipchip 형태로 추가배치하는 Wafer Level Assembly (3D 구조)
Near Packaged Optics	Gen 3 의 NPO (Multi Mode Waveguide Integration 포함) 를 위한 Optical-PCB 및 관련 광부품 개발
Silicon Photonics	반도체 재료/공정 기반의 Waveguide & Modulator 를 포함하는 진화된 광엔진 3D 패키징(Interposer) 적용으로 신호왜곡방지 및 제품 소형화/고밀도화 추구 이슈: 산업체간 협업 및 표준부재로 시너지감소 및 산업성장 장애
광커플링	저가/저손실 광학커넥터 부품 / 설비 및 공정 개발 고밀도 Multi Fiber Core 개발 및 커플링 공정 확보 이슈: 측면 광커플링 공정 확보 (Single Mode Core 9um & Waveguide < 1um) 공정 난이도 차이에 의한 Multi-mode 와 Single Mode 간 광커플링 경쟁

향후 15 년 정도의 기간 동안 극복해야 하는 장기 기술 개발 과제는 다음과 같다. 광연결의 적용 확대는 현재 빅데이터 전송에서의 많은 문제점을 개선할 수는 있으나, 데이터 센터의 구조와 데이터 전송 방식의 근본적인 문제인 반복되는 전/광 변환과 광/전 변환에 의한 신호 지연과 소비전력증가는 해결할 수 없다. 이에 대한 대책은 광학 라우팅(Optical Routing)의 도입이 있으며, MEMS 를 적용한 Hybrid E/O Routing 에서 가능성을 보여주었다. 이는 시스템간의

데이터전송은 물론 시스템내에서의 연산 장치간 데이터 전송이나 연산장치/메모리 간의 데이터 전송에도 기여할 수 있다.

이를 적용하기 위해서는 광학로직(Optical Logic)의 도입을 전제로 한 광학 정보처리 방식과 이를 위한 새로운 시스템 구조를 도입해야 하며, 다양한 광연결 기술과 전-광-전 변환을 최소화하기 위한 표준 인터페이스가 필요할 수 있다. 광학 로직이 라우팅 명령을 디코딩하고 광학 경로를 변경할 수 있다면, 신호 변환 및 대기시간과 소비에너지를 비약적으로 줄일 수 있으며, 디코딩 명령, 사용 가능한 경로 식별 및 광학 데이터 스트림을 새로운 경로로 전환하는 광학 논리를 수행하는 재료, 구조 및 장치에 대한 연구가 필요하다

광학 로직과 양자컴퓨팅의 결합은 데이터의 연산과 전송 능력을 최대화할 수 있는 조합이 될 것이며, 광연결에서 추구하는 데이터 전송속도 향상, 신호 왜곡 방지와 신호대기 시간 감소, 구동 전력 및 발열 감소에 의한 총소비 전력 감소, 소형화에 의한 고밀도 제품 구현을 가능하게 하는 게임 체인저가 될 것이다.

표 4-3. 향후 15 년간의 장기 기술 과제

개발 항목	주요 아이템 과 이슈
Silicon Photonics	3D Chiplet 제작을 위한 Wafer Level Package & Assembly 고효율 레이저 및 고밀도 저전력 모듈레이터 개발
Co-Packaged Optics	Gen 4 CPO (Multi Mode & Single Mode) 를 위한 패키징 공정 및 재료 개발 Gen 5 CPO (Laser-Integrated)를 위한 Monolithic 및 패키징 공정 개발 PCB-less 시스템 구성과 제품 구현
광학로직 및 라우팅	데이터 전송시, 전/광 및 광/전 변환 최소화 가능 재료개발 및 시스템 구성 비선형 광학재료를 적용한 광학영역에서만 디지털 신호 재생성 기술 광학영역에서 광학로직에 기반한 연산수행
표준 인터페이스	다른 광모드를 가진 개별시스템간의 데이터 전송을 위한 표준 인터페이스 확보 광자(Photon)의 파장대 변경 및 다른 시스템으로 입사하는 광편광 제어기술 기존 광정보를 다른 변조방식으로 전환하는 모듈레이터
신기술	통신용 구조 광학 (Structured light) 도입과 제품 응용 플라즈모닉과 유전체를 활용한 Waveguide 에서의 편광된 광의 회전기술 반도체 기판위의 복합 포토닉 결정에 의한 레이저 방향변경 형광 폴리머에 내장된 Au 플라즈몬 도파관의 광학이득 (Optical gain) 현상 활용

마지막으로, 광연결 반도체 분야의 미래 기술로서 구조 광학(Structured Light for Communication)기술 개발이 요구된다. 데이터 전송에서 '구조화된 빛' (Structured Light)의 일종인 OAM(Orbital Angular Momentum)의 적용은 광학 연구 및 개발영역에서 큰 발전을 이루었고^[20], OAM 광선은 강도(Intensity) 프로파일 형태에 기인하여 소용돌이 광선이라고도 불리운다. 주요 특징으로는 렌즈 및 광섬유와 같이 통신네트워크에 사용되는 많은 기존 광학 요소에 적합한 원형 대칭을 가지며, 완전한 방위각 정보를 사용할 수 있는 경우에는 빔의 방사형 강도에서의 일부 복구가 있더라도 수신기에서 OAM 모드 식별은 수행될 수 있다. OAM 기반의 모드 분할 다중화는 WDM(Wavelength Division Multiplexing)과 PDM(Polarization Division Multiplexing)과 결합하여 데이터 전송 용량을 증가시킬 수 있는 가능성이 있다.

이외에도 향후 광연결 반도체 분야의 향후 유망기술로서, 광자 신호를 광학적으로 증폭 또는 변조(파장, 편광 혹은 방향) 하거나 광자 메모리를 활성화할 수 있는 기술 개발이 필요하다. 이에 해당하는 기술은 플라즈모닉과 유전체가 도파관에서 광자의 편광을 회전시키는 현상^[21], 단일 광자가 양자점 공동시스템에서 다음 파장의 50ps 스위칭을 개시하는 현상^[22], 반도체 레이저의 복합 광자 결정이 방출된 레이저 빔의 방향을 바꾸는 현상^[23], 형광 폴리머에 내장된 Au 플라즈몬 도파관에 의한 광학 이득 (Optical Gain)^[24], 빛의 패킷 (Packets of Light) 이 비선형 공진기속에서 다양한 기능을 하는 현상 등이 있다^[25].

6. 요약

광연결 반도체는 현 반도체산업의 Si, Ge, GaAs, InP 등의 다양한 재료, 웨이퍼 가공을 위한 전공정, 패키징을 위한 후공정, 개별소자, 통합된 시스템 및 운영을 위한 모든 인프라를 활용하여, 반도체의 전기적인 기능과 광학적인 기능을 융합하는 제 4 차 산업혁명시대의 초연결을 위한 최적의 하드웨어와 시스템을 제공한다.

효율적인 빅데이터 전송과 생성을 위해 실리콘 포토닉스에 의한 고성능, 무결점, 저전력 및 친환경을 지향하며, 이종 재료/소자 간의 3D Co-Packaged Optics 적용을 시작으로 향후 전/광/전 신호 변환이 필요 없는 광로직에 의한 광라우팅 시스템의 도입과 양자컴퓨팅과의 결합에 의한 광네트워크구성은 새로운 가능성을 열어줄 것이다.

참고문헌

- [1] N. Margalit, C. Xiang, S. M. Bowers, A. Bjorlin, R. Blum and J. E. Bowers, "Perspective on the future of silicon photonics and electronics", Appl. Phys. Lett. 118, 220501 (2021); doi: 10.1063/5.0050117
- [2] N. Ledentsov, Ł. Chorchos, V. A. Shchukin, V. P. Kalosha, J. P. Turkiewicz, and N. N. Ledentsov, "Development of VCSELs and VCSEL-based Links for Data Communication beyond 50Gb/s," Optical Fiber Communication Conference (OFC) 2020, OSA Technical Digest (Optical Society of America, 2020), paper M2A.3.
- [3] Q. Cheng, et.al. "Scalable, Low-Energy Hybrid Photonic Space Switch." J. of Lightwave Technology 31, pp. 3077–3084 (2013).
- [4] [C. Gidney, & M.Ekera. How to factor 2048 bit RSA integers in 8 hours using 20 million noisy qubits. Quantum 5, pp.433 (2021).
- [5] D. Awschalom, K. K. Berggren, H. Bernien, S. Bhave, et al. "Development of Quantum Interconnects (QulCs) for Next-Generation Information Technologies." PRX Quantum 2, pp. 017002 (2021).
- [6] A. Youssefi, I. Shomroni, Y.J. Joshi, N.R. Bernier, et.al. "A cryogenic electro-optic interconnect for superconducting devices." Nature Electronics 4, pages326–332 (2021).
- [7] E. Chan, D. Koshinz, W. Krug, and H. Hager. "Wireless optical links for avionics applications". Proc. SPIE 8026, Photonic Applications for Aerospace, Transportation, and Harsh Environment II, 80260M (2011). DOI: 10.1117/12.886858.
- [8] W. M. Green, M. J. Rooks, L. Sekaric, and Y. A. Vlasov. "Ultra-compact, low RF power, 10 Gb/s silicon Mach-Zehnder modulator". Optics Express 15, pp. 17106–17113, (2007). DOI: 10.1364/OE.15.017106.
- [9] D. Feng, W. Qian, H. Liang, C.-C. Kung, Z. Zhou, Z. Li, J. S. Levy, R. Shafiiha, J. Fong, B. J. Luff, and M. Asghari. "High-Speed GeSi Electroabsorption Modulator on the SOI Waveguide Platform". IEEE J. of Selected Topics in Quantum Electronics 19, pp. 64-73, (2013). DOI: 10.1364/OE.20.022224.
- [10] W. Heni, Y. Fedoryshyn, B. Baeuele, A. Jostyn, C.B. Hoesbbacker, et al. "Plasmonic IQ modulators with attojoule per bit energy consumption." Nature Comm. 10, 1694 1-8 (2019). DOI: 10.1038/s41467-019-09724-7.

-
- [11] P. Chaisakul, D. Marris-Morini, M.-S. Rouifed, G. Isella, D. Chrastina, J. Frigerio, X. Le Roux, S. Edmond, J.-R. Coudevylle, and L. Vivien. "23 GHz Ge/SiGe multiple quantum well electroabsorption modulator". *Optics Express* 20, p. 3219, (2012). DOI: 10.1364/OE.20.003219.
 - [12] E. E. Mendez, F. Agulló-Rueda, and J. M. Hong. "Stark Localization in GaAs-GaAlAs Superlattices under an Electric Field." *Phys. Rev. Lett.* 60, 2426 (1988).
 - [13] S. A. Maier, M. L. Brongersma, P. G. Kik, S. Meltzer, A. A. G. Requicha, and H. A. Atwater. "Plasmonics—A Route to Nanoscale Optical Devices". *Adv. Mater.*, 13, pp. 1501–1505, (2001). DOI: 10.1002/adma.200390134.
 - [14] D. A. Svintsov, A. V. Arsenin, and D. Yu. Fedyanin. "Full loss compensation in hybrid plasmonic waveguides under electrical pumping." *Optics Express* 23, p. 19359, (2015).
 - [15] V. Liu, D. A. B. Miller and S. Fan. "Application of computational electromagnetics for nanophotonics design and discovery". *Proceedings of the IEEE*, vol. 101, no. 2, pp. 484–493 (2013). DOI: 10.1109/JPROC.2012.2207649.
 - [16] S. Jahani, S. Kim, J. Atkinson, J.C. Wirth, F. Kalhor, et al. "Controlling evanescent waves using silicon photonic all-dielectric metamaterials for dense integration." *Nature Comm.* 9, 1893 1-9 (2018). DOI: 10.1038/s41467-018-04276-8.
 - [17] V. R. Almeida, C. A. Barrios, R. R. Panepucci, and M. Lipson. "All-optical control of light on a silicon chip". *Nature* 431, pp. 1081–1084, (2004). DOI: <http://dx.doi.org/10.1038/nature02921>.
 - [18] Y. A. Zaghloul, A. R. M. Zaghloul, and A. Adibi. "Passive all-optical polarization switch, binary logic gates, and digital processor". *Optics Express* 19, p. 20335, 2011. DOI: 10.1364/OE.19.020332.
 - [19] J. H. Kim, Y. I. Kim, Y. T. Byun, Y. M. Jhon, S. Lee, S. H. Kim and D. H. Woo. "Optical-Amplifier-Based Devices and Their Applications". *J. of the Korean Physical Society* 45, pp. 1158–1161, 2004. PDF: <http://optiwave.com/wp-content/uploads/2014/10/xor-based-soa.pdf>.
 - [20] A. E. Willner, K. Pang, H. Song, K. Zou, and H. Zhou, "Orbital angular momentum of light for communications", *Applied Physics Reviews* 8, 041312 (2021).
 - [21] Z. Li, et al. "Controlling propagation and Coupling of Waveguide Modes Using Phase-Gradient Metasurfaces." *Nature Nanotechnology* 12, pp. 675-683 (2017).
 - [22] T. Holtz, et al. "Ultrafast All-Optical Switching by Single Photons." *Nature Photonics* 6, pp. 605–609 (2012).
-

- [23] Y. Kurosaka, et al. "On Chip Beam-Steering Photonic-Crystal Lasers." *Nature Photonics* 4, pp. 447–450 (2010).
- [24] M. Gather, K. Meerholz, N. Danz, and K. Leosson. "Net Optical Gain in a Plasmonic Waveguide Embedded in a Fluorescent Polymer." *Nature Photonics* 4, pp. 457–461 (2010).
- [25] F. Leo, et al. "Temporal Cavity Solitons in One-dimensional Kerr Media as Bits in An All-optical Buffer." *Nature Photonics* 4, pp. 471–476 (2010)

V. 무선연결 반도체 로드맵

1. 서론

현재 만물인터넷(IoE)은 더 많은 데이터 처리와 더 빠른 통신을 요구하는 다양한 애플리케이션들로 지속적으로 확장되고 있다. 대부분의 애플리케이션은 RF/마이크로파 무선통신과 고성능 백홀을 사용하여 인터넷과 클라우드 데이터 센터에 연결된다. 미래에는 RF/마이크로파 뿐만 아니라 밀리미터파(mmWave) 및 Sub-THz를 통해 무선통신이 이루어질 것이며, 하나의 안테나가 아닌 다중안테나를 통해 MIMO(massive multiple input multiple output) 통신이 이루어질 것이다. 또한 다양한 사물의 지속적 연결을 위한 IoT 센서 디바이스에 여러가지 저전력 통신 네트워크(예를 들어, LoRaWAN(Long Range Wide-Area Network), NB-IoT(Narrow Band Internet of Things), LPWAN(Low-Power Wide-Area Network) SIGFOX, LTE Cat 0들이 활용되고 확장될 것이다. 현재 위성은 모바일 기기로 방송을 전송하거나 긴급 신호 수신에 가능한 수준이며, 앞으로는 위성을 통한 음성 통신, 궁극적으로는 전체 인터넷 접속에 대한 관심이 증가하고 있다. 자율 주행 차량 및 교통 제어와 같이 빠른 통신 및 의사 결정이 필요한 응용 분야에서는 짧은 대기시간의 초저지연 무선 통신을 필요로 할 것이다.

현재 상용 5G 네트워크에서는 단말 기준 수 Gbps급 데이터 다운로드 속도가 시연·서비스되고 있으며, IMT-2020 표준에서 정의한 5G 피크 전송률은 20 Gbps이다. IoT 장치는 셀룰러 5G 네트워크를 통해 여러 LPWAN, Bluetooth, Wi-Fi 및 더 높은 데이터 속도 장치를 포함한 광범위한 네트워크를 통해 통신한다. Bluetooth 및 LPWAN 네트워크와 통신하는 IoT 장치는 일반적으로 LoRaWAN에서 2Mbps 미만 또는 100kbps 미만의 데이터 속도로 통신한다. 많은 IoT 센서와 변환기는 배터리 전원으로 작동하므로 버스트로 데이터를 전송하고 데이터를 전송하지 않을 때 대기 모드로 전환된다. 한편 마이크로파 백홀은 기존에는 수백 Mbps급이 주류였으나, 최근에는 다중 대역·고차 변조 기술 적용으로 수 Gbps~10 Gbps급 링크도 상용화되고 있다.

모바일 디바이스 최대 데이터 속도는 현재 5Gbps이며 GPS, Bluetooth, Wi-Fi, 3G 및 4G LTE 및 기타 요구 사항을 지원하기 위해 안테나 수는 13개 정도이다. 모바일 데이터 속도는 무선 기술의 발전으로 계속 증가해야 하며, 안테나 수는 MIMO 기술과 함께 계속 늘어나게 된다. 데이터 속도 증가를 지원하기 위해 전이중(Full Duplex) 통신 방식이 채택되어야 하며, 다수의 안테나를 위한 빔포밍 기술도 발전해야 한다. 전력 증폭기 및 A/D 컨버터를 포함하여 무선통신 회로는 높은 데이터 속도에서 휴대폰의 배터리 수명을 개선하기 위해 mmWave 주파수 및 광대역에서도

향상된 에너지 효율성을 제공해야 한다. 하지만, 저전력에서도 회로의 선형성을 유지해야 한다. 레이더와 같이 전파를 활용하는 센싱과 무선통신은 상호 협력하여 운영되어 센싱 및 통신 통합서비스(Integrated Sensing and Communication: ISAC)로 발전할 것으로 기대된다. 향후 우주 인터넷 망을 위한 위성통신용 통신 반도체 기술도 연구되어야 한다. 무선통신용 마이크로파 회로 기술은 향후 양자컴퓨터의 양자 상태 검출(Readout) 회로에도 적용될 것이다.

무선연결 반도체 기술의 주요 기술 키워드는 다음과 같다.

- 무선 주파수: Sub-6GHz; upper midband; mm-wave; sub-THz
- 저전력: Wake-up receiver; 고효율 PA (digital PA, Doherty PA)
- 고선형: 저전력 구간 선형성 유지; 다중 간섭환경
- 전이중 통신: 자기간섭(self-interference); On-chip circulator
- 센싱및통신 통합서비스: ISAC, Radar based communication; Communication based radar

2. 무선연결 반도체 발전 방향

ISSCC 국제학술대회에서 발표하는 무선통신 반도체 기술 트렌드는 무선통신 반도체 기술 관련 연구방향 설정에 도움을 준다. 아래는 2024년 RF 및 무선통신 기술 관련 트렌드는 다음과 같다.

그림 5-1은 RF 주파수합성기의 기술발전 동향이다. 반도체 소자와 회로 기술을 통해 적은 전력을 소비하면서도 잡음 성분인 지터를 최소화하는 방향으로 기술이 발전하고 있다^[2].

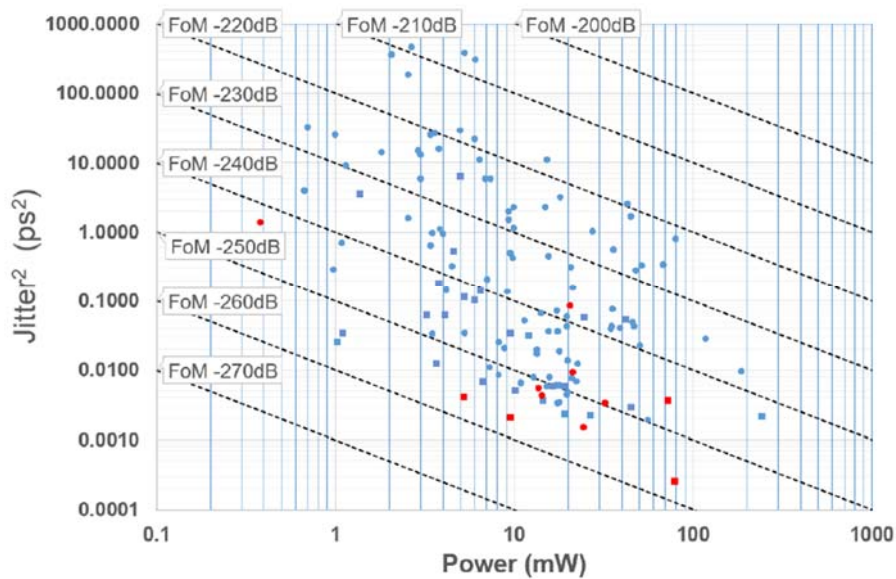


그림 5-1. RF 주파수합성기 기술 발전 동향

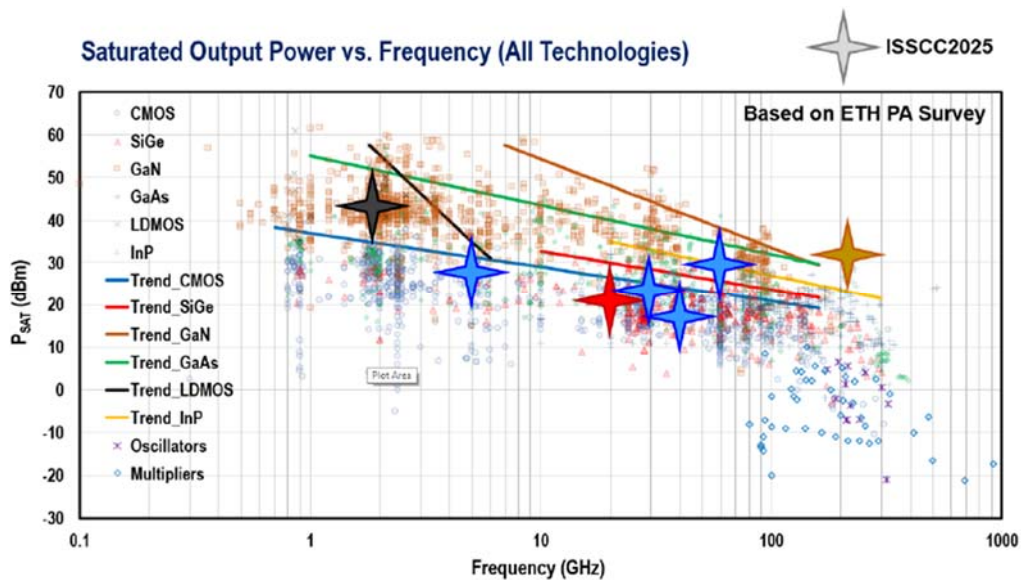


그림 5-2. 전력증폭기 기술 발전 동향

그림 5-2은 전력증폭기의 기술발전 동향을 보여주고 있다. 동작 주파수가 높아질수록 최대 출력 전력은 낮아지는 경향을 보이고 있고, GaAs 및 GaN 기반 전력증폭기가 CMOS 전력증폭기에 비해 더 높은 출력 전력을 가진다.

그림 5-3은 저전력회로의 선형성의 기술발전 동향이다. IoT 무선통신 회로들이 저전력을 소비하도록 개발됨에 따라 선형성이 낮아지는 문제가 있는데, 이런 문제를 해결하기 위해 다양한 회로 기술이 개발되면서 신호 대 간섭비 성능이 개선되고 있다.

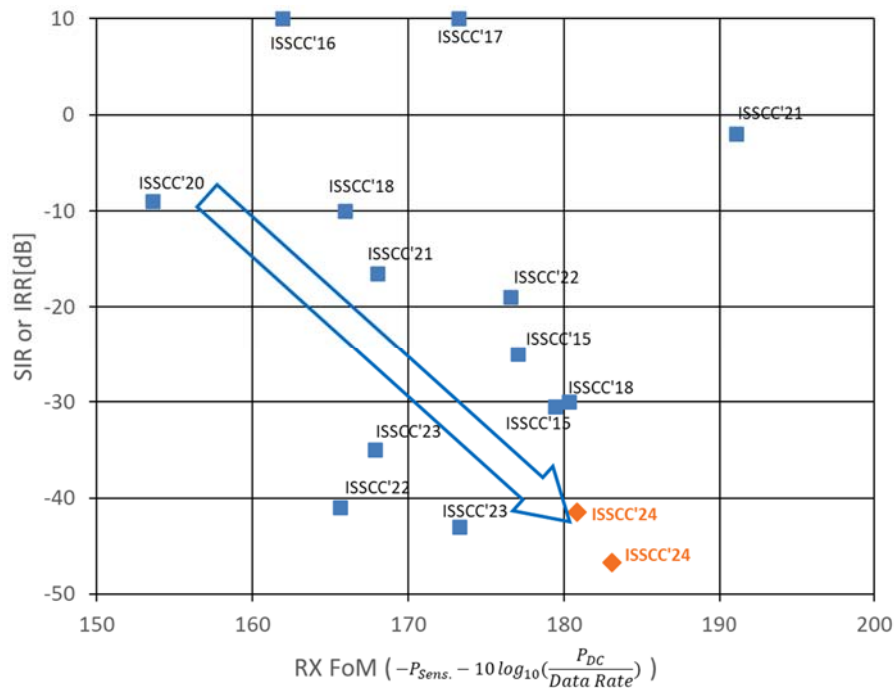


그림 5-3. 저전력 회로의 선형성 기술 발전 동향

그림 5-4는 광대역 수신기의 대역폭 증가에 따라 대역내 IIP3 선형성 특성의 기술발전 동향이다. 광대역 수신기는 불가피하게 존재하는 간섭 및 재머 신호에 의해 수신 성능 저하를 막기위해 무선통신 수신기 회로의 선형성 및 블로킹 내성 특성이 중요하다. 특히 저전력 소비하면서 광대역 수신기 특성을 확보하기 위해 다양한 회로 기술이 개발되면서 신호 대 간섭비 성능이 개선되고 있다.

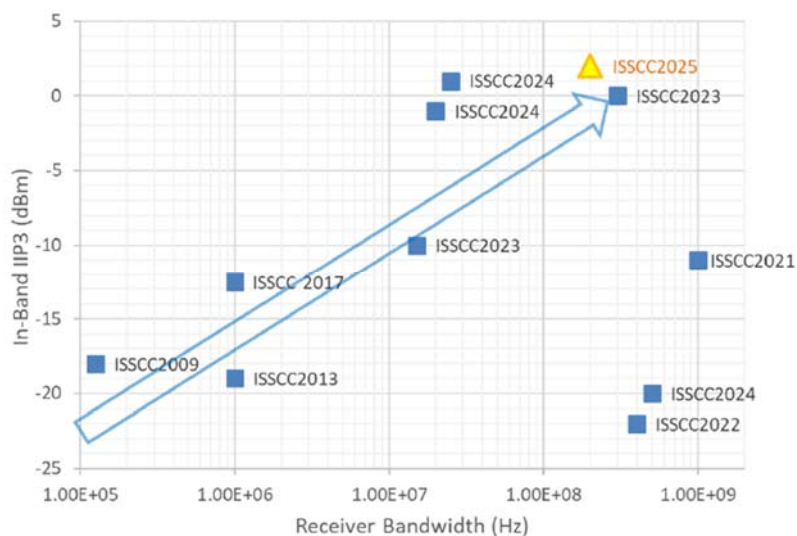


그림 5-4. 수신기 대역폭에 따른 밴드내 IIP3 선형성 기술 발전 동향

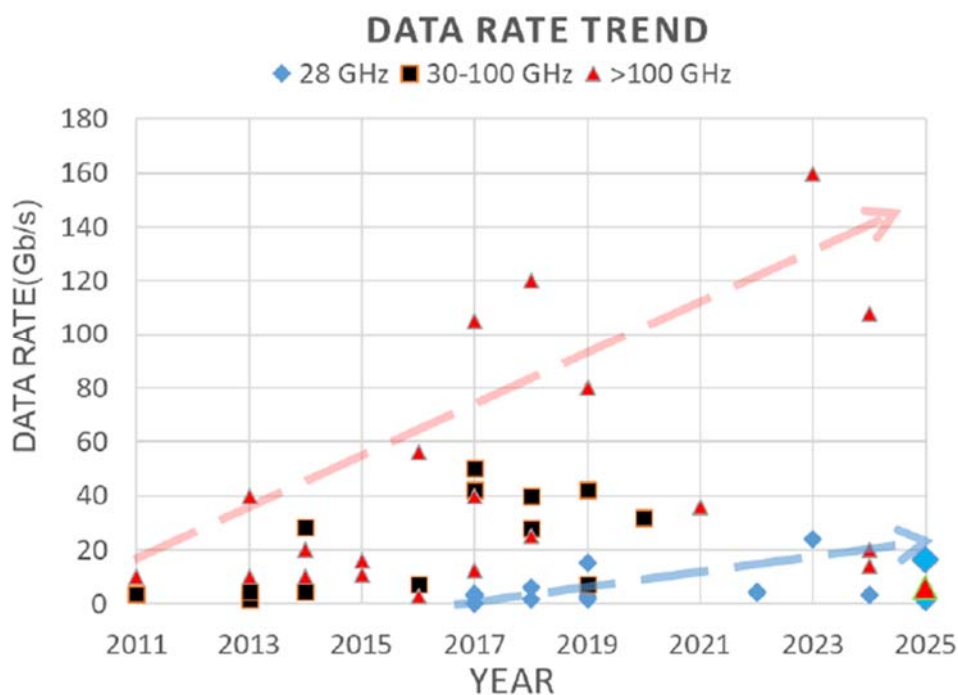


그림 5-5. 무선통신 데이터 전송속도의 발전 동향

그림 5-5는 초고속 무선송수신기에서 데이터 전송속도의 기술발전 동향이다. 무선통신 회로의 성능은 데이터 전송률(data rate) 지표로 표현할 수 있는데, 데이터 전송률은 소자 및 회로 기술 개발에 따라 꾸준히 향상되고 있다.

3. 반도체 소자 기술 발전 방향

더 높은 데이터 속도와 더 낮은 전력을 갖는 RF 통신을 위해서는 고속 저전력 회로 설계가 필요하다. QAM 등 변조 기술을 사용하는 경우, 기지국과 모바일 장치에는 저전력 디지털-아날로그 컨버터(DAC), 높은 선형성, 그리고 더 높은 샘플링 속도를 가진 고효율 전력 증폭기가 필요하다.

또한 모든 방향으로 전송하는 대신 RF 신호를 사용자에게 집중시키기 위한 고지향성 위상 배열 안테나 기술이 필요하다. 더 나아가 초광대역 수준의 변조를 지원하기 위해서는 THz 주파수에서 무선 통신 회로가 동작해야 한다. 이런 모든 무선 통신 기술을 위해서는 반도체 및 소자 기술의 발전이 전제되어야 한다.

(1) CMOS 소자

CMOS 소자에 대한 RF 성능은 Peak f_T 및 Peak G_m 값으로 확인되며, 선단공정으로 개발되는 이중게이트 FinFET과 단일게이트 FDSOI MOSFET의 평균 측정 데이터는 그림 5-6과 같다. 게이트 및 접촉 저항의 감소는 f_{MAX} 를 향상시키지만, f_T 에는 영향을 미치지 않는다. 새로운 재료의 도입과 새로운 트랜지스터 및 게이트 구조는 10nm 미만 CMOS 장치의 f_T 추세를 바꿀 수 있으므로 THz 등 차세대 무선통신 주파수 대응을 위해 연구되어야 한다.

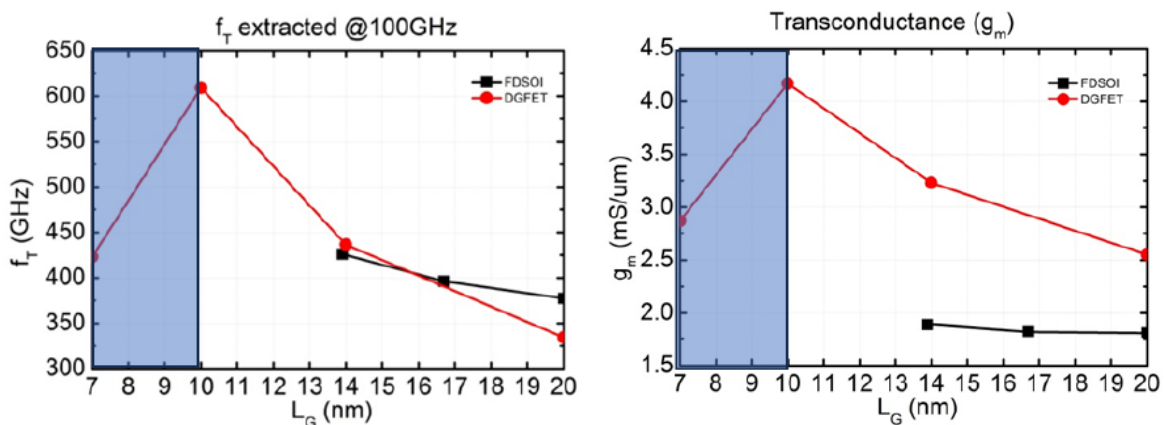
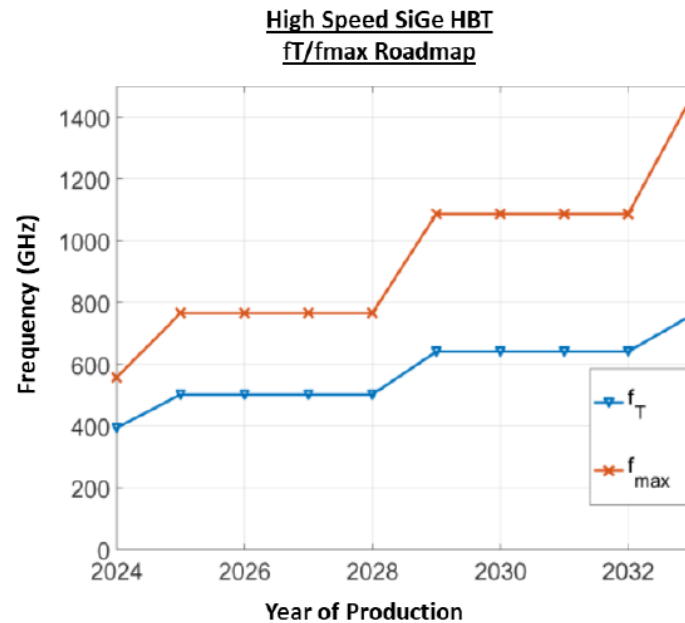


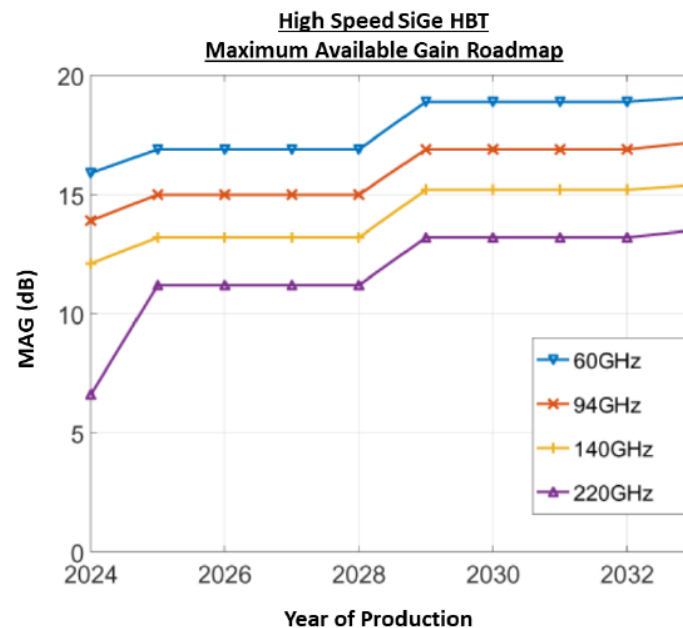
그림 5-6. CMOS 소자의 성능지표 발전 동향

(2) Bipolar 소자

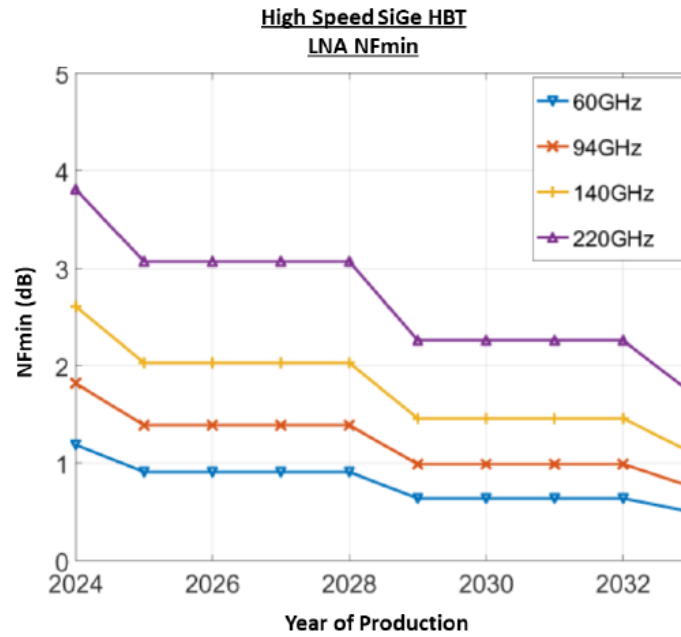
Bipolar 소자는 CMOS 소자에 비해 좀 더 고속 및 저잡음으로 동작하며 f_T/f_{MAX} 로드맵은 그림 5-7과 같다. 이러한 Bipolar 소자는 LNA, PA, VCO 및 ring oscillator의 벤치마크 회로를 통해 그 성능을 확인할 수 있다.



(a)



(b)



(c)

그림 5-7. Bipolar 소자의 성능지표 발전 동향:

(a) f_T/f_{max} 주파수, (b) 최대 가용 이득, (c) 저잡음 증폭기의 최소 잡음 지수

(3) III-V 화합물 반도체 소자

무선통신을 위한 RF/마이크로파 주파수에서는 Si 소자에 비해 III-V족 반도체 기반 FET 또는 Bipolar 소자가 우수한 성능을 발휘한다. III-V 로드맵에 따라 GaAs PHEMT, GaAs MHEMT, InP HEMT, InP HBT 및 GaN HEMT 와 같은 소자의 물리적 스케일링으로 특성(이득, 최대 전류 증폭 주파수, 최소 잡음 지수 등) 개선이 계속될 것으로 예상된다. 그림 5-8은 각 화합물 반도체 소자에 따른 밀리미터파 대역에서의 이득 동향과 전망을, 그림 5-9은 최대 전류 증폭 주파수 동향 및 전망을, 그림 5-10는 최소 잡음 지수 동향 및 전망을, 그림 5-11은 최대 증폭 주파수 동향 및 전망을 각각 보여주고 있다.

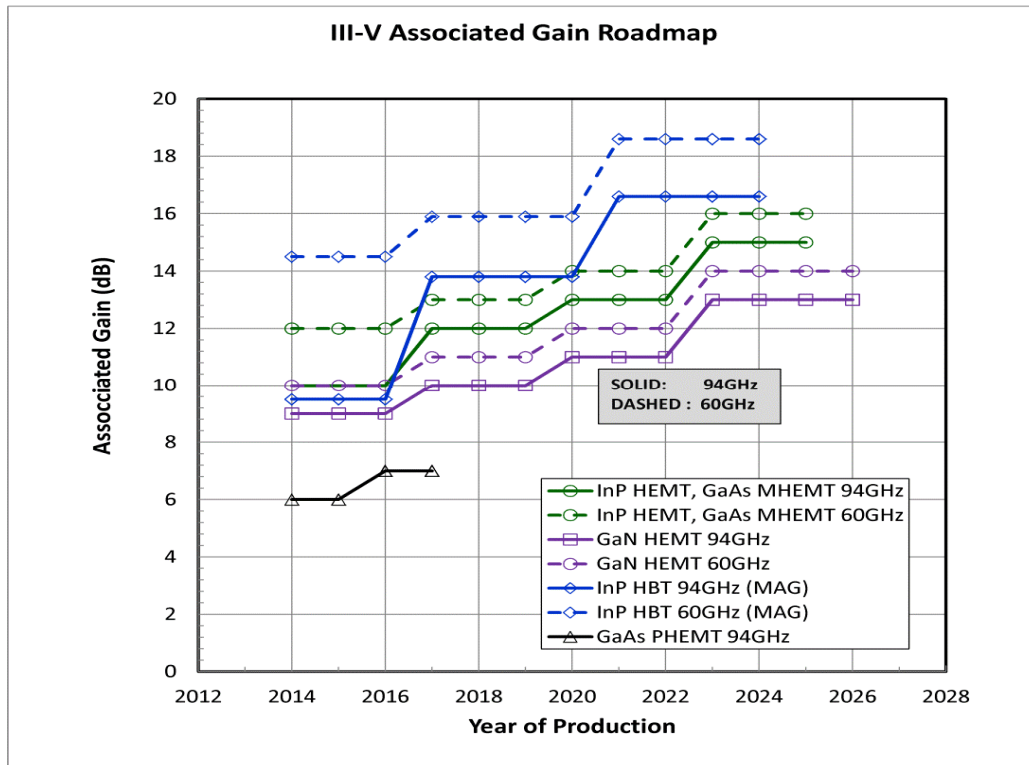


그림 5-8. III-V 화합물 반도체 소자의 증폭률 발전 동향

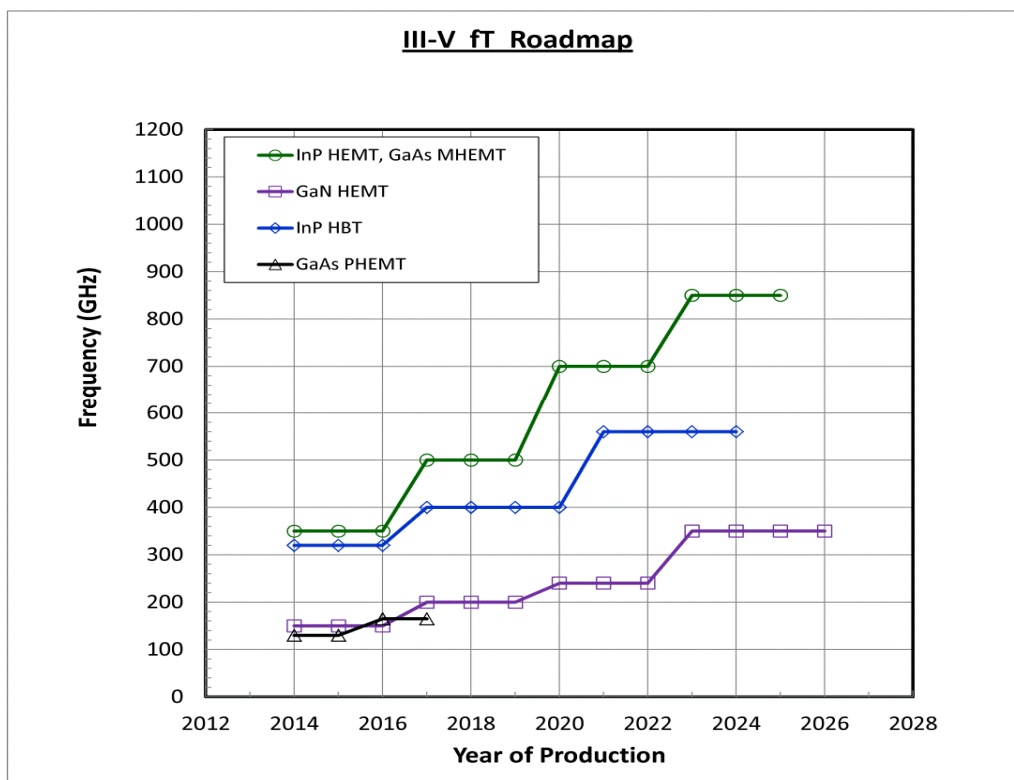


그림 5-9. III-V 화합물 반도체 소자의 f_T 발전 동향

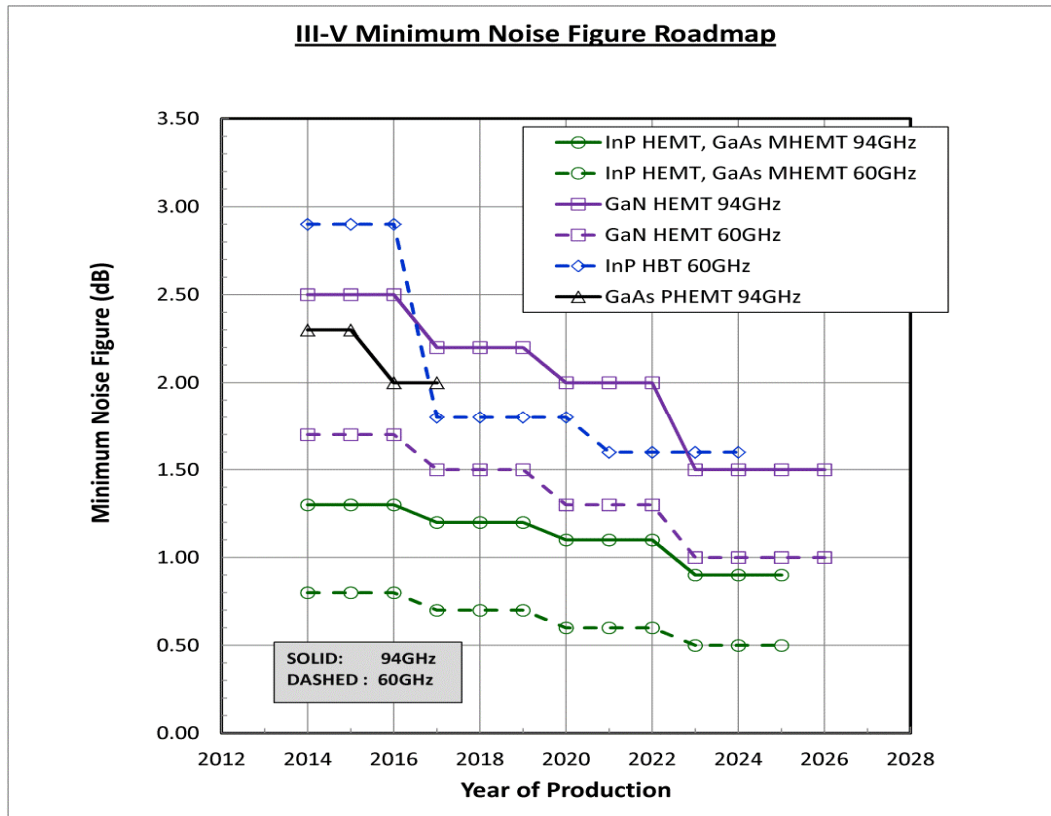


그림 5-10. III-V 화합물 반도체 소자의 잡음 특성 발전 동향

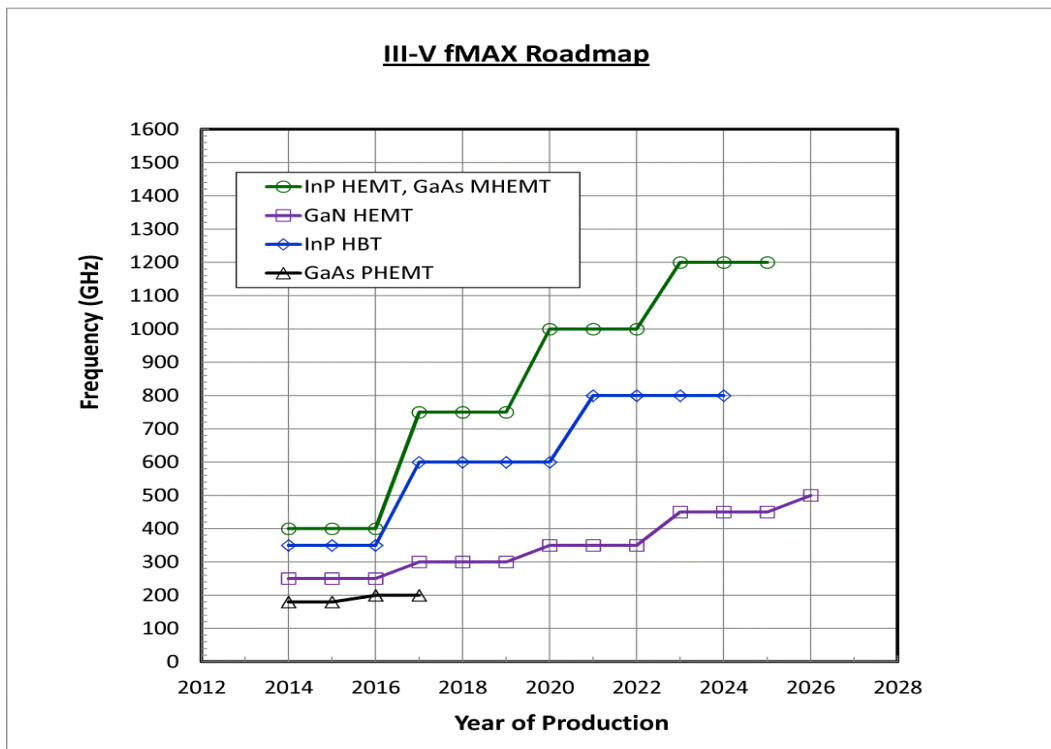


그림 5-11. III-V 화합물 반도체 소자의 최대 증폭 주파수 발전 동향

한편, 이러한 III-V 반도체 시장은 대량 생산이 가능한 소비자 시장과 소량 고가 생산이 필요한 특수 시장이 존재한다. 대량 소비자 중심 시장은 셀룰러 통신용 GaAs HBT 전력 증폭기가 대표적이고, 현재 InP HEMT, InP HBT 및 GaN HEMT가 속해 있는 소량 특수 시장은 시장 규모가 작아서 어려움을 겪고 있다.

표 5-1. 무선연결 기술 로드맵

연도	2025	2028	2030	2033	2037	2040
블루투스 EDR 전송율 (Mbps)	2/3	2/3	2/3	2/3	2/3	2/3
BLE 전송율 (Mbps)	1/2	1/2	1/2	1/2	1/2	1/2
WiFi 최대전송율(Gbps) @ 45m	2.4	3	5	10	10	10
WiGig-60GHz 전송율(Gbps)	4.6-7	4.6-7	4.6-7	8-10	15	20
3G/4G 무선주파수 (GHz)	< 4	< 4	< 4	< 4	< 4	< 4
5G/6G 무선주파수 (GHz)	< 6, 28, 39	< 6, 28, 39, > 72	< 6, 6-24, 28, 39, > 72	< 6, 6-24, 28, 39, >72	< 6, 6-24 28, 39, > 72, 120-150	< 6, 6-24 28, 39, > 72, 120-150
3G/4G(Sub-6) 전송율 (Gbps,)	7	10	20	50	100	100
5G/6G (mmWave/sub-THz) 전송율 (Gbps, peak)	10 (5G peak)	20 (5.5G)	50 (5G/6G 전환 구간)	100 (6G 초반)	300 (6G 중반)	1000 (post 6G)

4. 무선연결 반도체 로드맵

무선통신 소자 및 회로 기술에 따른 동향을 통해 무선연결 기술 발전 로드맵을 표 5-1과 같이 예상할 수 있다. 3G/4G 이동통신(주로 Sub-6GHz 대역)의 피크 데이터 전송률은 현재 수 Gbps 수준에서 향후 기지국·단말 하드웨어와 변조기술의 개선에 따라 2040년경에는 수십~100 Gbps

수준까지 향상될 것으로 예상된다. 한편 5G/6G 시스템은 mmWave 및 Sub-THz 대역을 적극 활용하여 6G 세대에서 피크 데이터 전송률 0.1~1 Tbps(100~1000 Gbps) 달성을 목표로 하고 있으며, 2040년에는 Tbps급 링크가 일부 응용 분야에서 실현될 것으로 전망된다.

5. 요약

LPWAN, Bluetooth, Wi-Fi 및 5G/6G 등 다양한 무선통신표준이 IoT 장치에 무선통신 서비스를 제공하기 위해 경쟁하고 있다. 일부 IoT 장치는 매우 낮은 전력 작동이 필요하므로 무선 통신 소자와 회로는 더 전력 효율적이어야 한다. 5G/6G 통신을 가능하게 하는 능동 위상 배열 안테나를 개발하는 데 상당한 진전이 있었으며, 이러한 고지향성 안테나는 저전력으로 장거리 통신을 가능하게 하고 간섭을 줄이며 보안을 향상시키는 방법을 가능하게도 할 수 있다. 여러 반도체 재료 기반 다양한 소자를 하나의 무선통신 시스템으로 묶기 위한 hybrid 회로 설계 및 패키징 기술에 대한 관심도 커지고 있다.

향후 5G 고도화 및 6G 시대의 무선연결 반도체 로드맵은 단순한 데이터 전송률 향상을 넘어 지연시간, 에너지 효율, 신뢰도 등 종합적인 품질 지표 개선을 목표로 한다. 6G 비전에서는 종단 간 지연을 수 ms 수준에서 수백 μ s 이하로 줄이고, 비트당 에너지 소모를 수십 pJ/bit 이하로 낮추는 것을 하나의 방향으로 제시하고 있다. 이를 달성하기 위해서는 고선형·고효율 전력증폭기, 저위상잡음 주파수합성기, 대규모 위상 배열 안테나 및 빔포밍을 지원할 수 있는 RF-SoC 플랫폼 등 무선연결 반도체 기술의 지속적인 발전이 요구된다.

향후 6G 시대에는 레이더 센싱과 통신 기능을 하나의 RF 전(front-end)과 베이스밴드 플랫폼에서 통합하는 센싱및통신 통합서비스(ISAC) 기술이 무선연결 반도체의 주요 응용 분야로 부상할 것으로 예상된다. 이를 위해서는 기존 통신용 PA/LNA, 주파수합성기 외에 고분해능 레이더를 지원하는 펄스 생성 회로, 고속·고분해능 ADC, 그리고 공통 하드웨어 자원을 동적으로 재구성할 수 있는 RF-SoC 플랫폼이 요구된다. 또한, 저궤도 위성(LEO)을 이용한 위성-지상 통합 네트워크(Non-Terrestrial Network, NTN)의 확산으로 위성 단말용 RF 프런트엔드 및 빔포밍 칩셋 수요도 크게 증가할 것이며, 이러한 영역에서도 GaN HEMT, InP HEMT 등 III-V 소자와 CMOS/SiGe BiCMOS 기반 집적 회로의 융합 설계 및 패키징 기술이 핵심 역할을 할 것이다.

참고문헌

- [1] IEEE IRDS 2024, Outside System Connectivity(OSC) Chapter," IEEE, 2024
- [2] International Solid-State Circuit Conference, 2025 PRESS KIT,
<https://submissions.mirasmart.com/ISSCC2025/PDF/ISSCC2025PressKit>.
- [3] ITU-R, Recommendation ITU-R M.2150-2: "Detailed specifications of the terrestrial radio interfaces of IMT-2020," ITU, 2023
- [4] ITU-R, "IMT towards 2030 and beyond (IMT-2030)," ITU-R WP 5D, 2023.
- [5] DigitalRegulation.org, "Overview of 6G (IMT-2030)," 2025.
- [6] M. Banafaa et al., "6G Mobile Communication Technology: Requirements, Targets, and Key Techniques," Alex. Eng. J. 2023, 64, 245–274.
- [7] IEEE 802.11 Working Group, "IEEE Std 802.11ax-2021: Telecommunications and information exchange between systems—High Efficiency WLAN," IEEE, 2021.
- [8] IEEE 802.11 Working Group, "IEEE P802.11be (Wi-Fi 7): Draft Standard for Extremely High Throughput WLAN," draft
- [9] 3GPP, "Non-Terrestrial Networks (NTN) – Overview," 3GPP Technologies, Release 17, 2024.

VI. 센서 기술

1. 서론

인공지능이 발전함에 따라 산업 전반에서 사람의 개입을 최소화하는 것이 성능, 안전성, 신뢰성 등 모든 면에서 효과적이라는 패러다임의 전환이 이루어지고 있는 지금 자동화의 필수 요소인 센서도 다양한 분야에서 발전이 이루어지고 있다. 반도체 공정의 발전과 더불어 신규 소재 합성을 통해 보다 더 정확한 결과를 얻을 수 있을 뿐 아니라 과거에는 검출하지 못했던 정보들도 확인이 가능하게 됐다.

센서는 전달하는 정보에 따라 이미지 센서, 거리 센서, 조도 센서, 비전 센서, 가스/분자 센서 등 다양한 제품들이 해당된다. 본 로드맵에서는 센서들이 정보를 제공하는 방식에 따라 이미징 센서와 검출기 센서로 나누어 현재까지 개발되어 온 센서들의 기본적인 동작 방법에 대해 간략히 소개하고 향후 발전 방향과 인공지능과의 융합까지 예측해 본다.

2. 이미지 센서

2.1 가시광 대역 이미지 센서

1) 픽셀 미세화

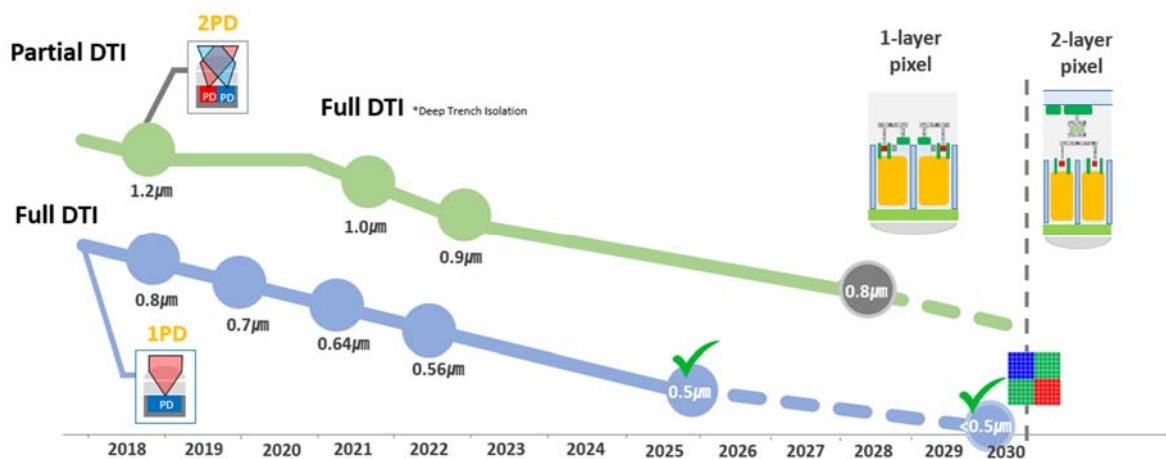


그림 6-1. 픽셀 미세화 트렌드 & 주요 기술

본 절에서는 모바일 및 자동차용 이미지 센서에 적용되는 기술의 변화 트렌드를 다룬다. 높은 해상도 및 작은 form factor를 위해 픽셀 미세화는 지속해서 요구되어 왔고 지금까지 소비자용 이

이미지센서의 발전은 픽셀 미세화 과정으로 볼 수 있다. 20년에 걸쳐 픽셀은 $5.6\mu\text{m}$ 에서 $0.5\mu\text{m}$ 까지 100배이상 미세화 되었지만 그럼에도 불구하고 화질은 열화 되지 않고 오히려 발전하였다. 이것이 가능하게 된 것은 몇 단계 기술적 점프가 있었기 때문이다.

픽셀 미세화는 모바일 시장의 요구에 의해 이끌어져 왔다. 그림 6-1은 모바일 제품을 타겟으로 한 센서의 픽셀 미세화 역사와 적용된 주요 기술을 보여준다. 1997년부터 CCD에 사용되던 pinned photo diode (PPD) 기술이 CMOS Image Sensor (CIS)에 적용된 이후 CIS의 발전은 본격화 되었다. PPD가 적용됨에 따라 photon-generated 전자의 noise-free 전송이 가능하게 되었으며 dark current 또한 현저히 감소하였다^[1]. 2009년 Back Side Illumination (BSI) 이 적용되면서 100%에 가까운 fill factor로 sensitivity [e-/Lux.sec] 가 비약적으로 상승하였다. 소자 분리 영역을 깊게 형성하는 Deep Trench Isolation (DTI) 이 적용되면서 픽셀 간 전기적, 광학적 cross talk이 감소하였다. 또한 2014년 실리콘 전면에서부터 trench를 성장시키는 FDTI가 적용되어 photodiode 간 완전한 분리가 가능하게 되었다. 하드웨어뿐만 아니라 소프트웨어적 발전으로도 픽셀의 성능은 향상되었다. 2017년 제시된 Tetra 픽셀 구조는 1x1 픽셀 단위로 구성되는 color filter array (CFA) 를 2x2 픽셀 단위로 전환하였다. 저조도에서는 하나의 color를 표현하는데 4개의 픽셀을 함께 읽어 sensitivity가 4배 상승하는 효과를 가져왔으며 고조도에서는 re-mosaic 알고리즘을 이용해 full resolution을 가능하게 하였다.

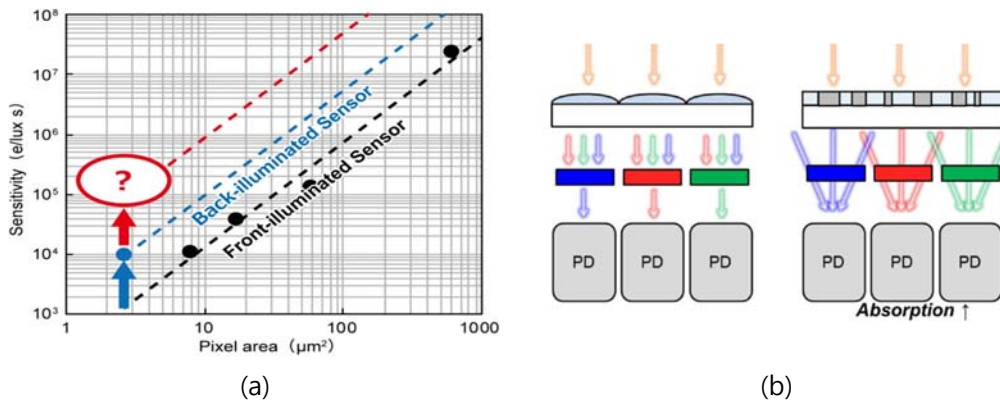


그림 6-2. 픽셀 미세화 트렌드 & 주요 기술

기술 발전으로 픽셀 미세화는 꾸준히 진행되어 왔지만 픽셀이 sub micro 수준으로 미세화 되면 sensitivity 하락, crosstalk 증가뿐만 아니라 빛의 회절 특성 또한 해결해야할 문제로 떠오름에 따라 향후 픽셀 미세화의 속도는 이전에 비해 느려질 것으로 예상된다.

그림 6-2(a)와 같이 픽셀 미세화에 따라 감소하는 sensitivity를 jump up 하기 위한 새로운 기술이 요구되는 가운데 meta optics 기술이 부상하고 있다^[2]. 대표적인 meta optic 기술 중 하나가 color sorting meta structure이다^[3]. 해당 구조물은 그림 6-2(b)와 같이 센서에 입사된 빛의 경로를 파장 별로 달리하여 CFA에 의한 광자 손실을 최소화함으로써 sensitivity를 비약적으로 향상시킬 수 있는 가능성을 보여주고 있다.

2) High Dynamic Range (HDR)

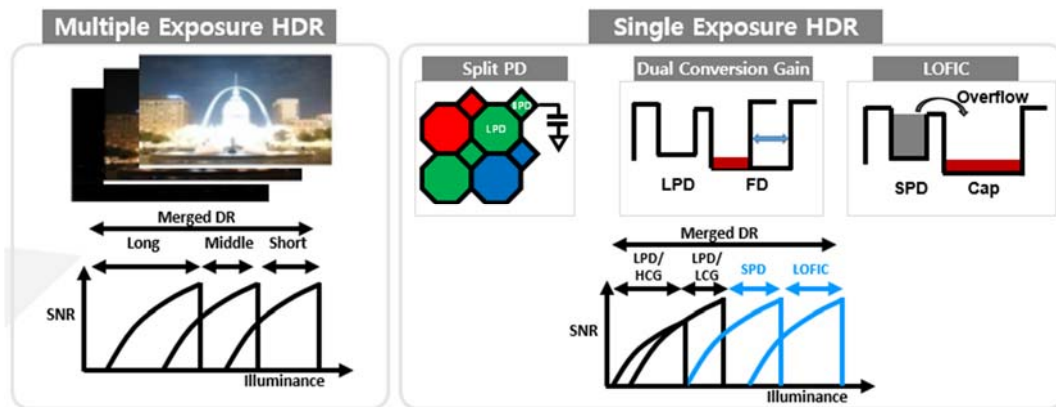


그림 6-3. HDR 주요 기술

명부와 암부를 동시에 표현하기 위해 HDR 기술은 현대 이미지센서에서 필수이다. 현재 적용되고 있는 HDR 기술은 노출을 달리하여 여러 이미지를 취득 후 합성하는 multiple exposure 방식과 노출은 한번만 하고 시그널 chain의 gain을 달리하여 다수 이미지를 생성 후 합성하는 single exposure 방식으로 나눌 수 있다 (그림 6-3). Multiple exposure 방식은 픽셀 구조가 간단하고 DR 확장 폭이 큰 장점이 있는 반면 움직이는 사물을 촬영 시 ghost와 같은 motion artifact가 발생하는 단점이 있다. Single exposure 방식은 반대로 motion artifact가 작지만 픽셀 설계가 복잡해지고 DR 확장 범위가 구조에 의해 제한되는 단점이 있다.

현재 Mobile용 센서는 Multiple exposure과 Single exposure이 상황에 맞게 사용되고 있다. 향후에는 동영상 화질에 대한 요구가 높아짐에 따라 센서에 다수의 single exposure기술이 적용될 것으로 예상되며 추가적으로 다양한 조명이 혼합된 환경에서 동영상 품질을 저하시키는 플리커를 제거하기 위한 LED Flicker Mitigation (LFM) 기술이 중요해질 것으로 예상된다. HDR과 노이즈 성능을 열화 시키지 않으면서 플리커를 얼마나 잘 억제하는지가 LFM 기술의 경쟁력 지표가 될 것이다. 차량용 센서는 요구되는 DR이 높고 motion artifact에 엄격하기 때문에 복수의 single

exposure 기술을 혼합해서 사용하고 있으며 120dB 이상의 DR을 single exposure로 구현하고 있다^{[4][5]}. 또한 LFM 기술의 적용은 필수로 여겨지고 있다.

3) Noise

설계 기술의 발전으로 회로 노이즈는 지속적으로 감소해왔다. 저조도 화질에 대한 요구는 계속되고 있으므로 Random Noise(RN)을 줄이는 노력은 계속될 것이며 향후에도 공정 및 회로 기술의 발전으로 RN은 1e- 이하 수준까지 떨어질 것으로 예상된다.

4) 전력

센서의 성능이 향상됨에 따라 증가하는 전력소모를 미세 공정 적용과 구조적 개선을 통해 억제해왔고 근래의 센서는 500mW 수준의 전력소모량을 보이고 있다. 저전력에 대한 요구는 미래에도 지속될 것이므로 우하향 기조를 유지할 것으로 예측된다.

5) Stack 공정

센서 성능 향상 요구에 따라 기능을 담기 위해 웨이퍼를 적층하는 기술이 발전하고 있다. 현재 2-stack이 시장에서 main volume을 차지하고 있고 3-stack 센서는 고성능 어플리케이션에 제한적으로 사용되고 있으나 고성능 기조에 맞춰 일반 스마트폰, 카메라도 3층 이상의 stack 구조 채택이 늘어날 것으로 예상되며, stack 구조 적용으로 확보된 공간을 활용하는 차세대 센서의 출현이 증가할 것으로 전망된다.

현재까지 개발된 가시광 센서 기술을 토대로 향후 개발 방향을 표 6-1에 예측하였다. 픽셀 미세화와 Dynamic range 확대는 지속될 것으로 예상되고 센서 성능 향상 요구에 맞춰 새로운 광학 구조와 웨이퍼 적층 기술도 함께 발전할 것으로 예상된다.

표 6-1. 가시광 센서 기술 로드맵

		2025	2028	2031	2034
Pixel shrink		0.5um	0.5um	0.5um	<0.5um
Optics		Micro Lens (ML)	ML+Meta Structure	ML+Meta Structure	Meta Structure only
HDR (w/ LFM)	Mobile	100dB	110dB	110dB	>110dB
	Auto	135dB	140dB	160dB	>160
Wafer stack		2	3	4	4

6) 차세대 구조

(1) Global Shutter, Hybrid-Global Shutter

기존의 Rolling shutter 방식은 각 row의 노출시간에 차이가 있어 움직이는 사물을 촬영하는 경우 jello, wobbling과 같은 rolling artifact가 발생한다. 반면 Global shutter(GS) 방식은 전 픽셀의 노출 시간이 동일하므로 rolling shutter 대비 빠르게 움직이는 사물 촬영에 확실한 우위가 있다. 하지만 픽셀 크기 증가, rolling shutter 대비 열세인 noise 등의 단점을 가지고 있어 스마트폰, 카메라 등 소비자 시장보다는 산업 영역에 제한적으로 사용되었다.

근래에는 3-stack, 고성능 cap 적용하여 RS 대비 열세인 성능을 개선하면서 동시에 상황에 맞게 셔터 방식을 바꾸어 rolling shutter와 GS의 장점을 극대화하기 위한 hybrid global shutter 구조가 보고되었다^[6]. 향후 rolling shutter 대비 열세인 해상도, 노이즈 성능을 개선시키는 방향으로 기술이 발전하고, 모바일 영역까지 GS 구조의 채택이 늘어날 것으로 예상된다.

(2) Digital Pixel Sensor (DPS)

Digital Pixel Sensor(DPS)는 픽셀 내부에 ADC를 집적한 구조이다. 때문에 global shutter가 기본적으로 가능하며 새로운 방식의 HDR 기술 적용, high frame rate가 가능하다. 구조의 복잡성으로 인한 픽셀 미세화 열세, 전력소모 등의 문제로 모바일 센서에 DPS를 적용하기에 제약이 있었지만 최근에는 3-stack이 적용되어 모바일, AR 등 소비자 기기를 타겟으로 한 DPS 센서가 보고되고 있다^[7]. 향후에는 DPS 성능을 모바일에 적용 가능한 수준으로 향상시키기 위한 기술발전이 계속될 것으로 보인다.

(3) Photon Counting Sensor (PCS)

Photon counting sensor(PCS)는 광자를 감지할 때마다 pulse로 변환하는 픽셀을 기반으로 한 이미지 센서이다. 단일 광자 검출이 가능하므로 매우 어두운 환경에서도 이미지 획득이 가능하고^[8] 동작에 결함이 없을 경우 dark random noise를 0.0e- 수준으로 낮출 수 있다. 즉 저조도 성능이 매우 우수하다. 또한 광자의 총량만을 정보로 가질 수 있는 기존 적분 방식과 달리 광자의 도달 시간 또한 정보로서 이용할 수 있는 잠재력이 있다. PCS가 이미지센서로써 범용적으로 사용되기 위해서는 여러 기술적 과제를 해결해야 한다. 빛이 많은 상황, 즉 광자가 많은 상황에서는 개별 광자를 세는 것은 saturation 문제가 생길 수 있다. 따라서 photon counting 방식에 적합한 데이터 처리 알고리즘이 발전해야 한다. Photon을 감지하는 픽셀은 구조가 복잡하여 크기가 크며 소비전력이 크다. 픽셀의 성능을 유지하면서 픽셀 크기와 전력소모를 줄일 방안이 마련되어야 한다. 그리고 상용화를 위해서는 비용, 균일한 성능, 수율 등의 문제도 해결해야한다. 단기간에 소비자

시장에서 상용화 되기에는 해결해야할 문제의 양과 난이도가 높다. 하지만 저조도 성능 향상에 있어 명확한 잠재력을 가지고 있어 관련 기술 개발은 계속될 것으로 보인다.

2.2 비가시광 대역 이미지 센서

비가시광 대역 이미지 센서는 400nm~700nm 가시광 대역의 파장대를 제외한 단파장의 UV, 장파장의 적외선 대역의 넓은 범위를 말한다. 가시광 대역의 이미지 센서가 발전함에 따라 다양한 산업에 적용될 수 있는 비가시광 대역의 이미지 센서도 주목을 받기 시작했다. 군사 목적으로 적외선 대역의 열감지 센서의 개발을 시작으로 항공 우주, 산업, 의료, 안전 및 보안, 자율주행 등 여러 분야에서 널리 적용되고 있다. 그림 6-4는 비가시광 대역을 파장에 따라 UV, NIR, SWIR, LWIR로 나누고 각 대역을 검출하기 위해 사용되는 주 소재들을 보여준다.

가시광 대역 이미지 센서의 빠른 발전 속도는 실리콘 기반의 반도체 공정 기술에 의해 가능했다. 가시광 대역의 빛의 흡수율이 높은 실리콘 덕분에 이미 개발된 반도체 공정 기술을 통해 빠르게 센서 제작이 이루어졌다. 더 나아가 픽셀단과 Readout 회로단이 전부 실리콘 기반의 반도체로 이루어져 높은 Quantum Efficiency (QE)의 센서 제작이 가능했다. 하지만 적외선 (NIR, SWIR, LWIR) 파장은 실리콘 소재에서의 흡수율이 낮고 UV 파장은 흡수율은 높지만 단파장의 특성 때문에 침투 깊이가 짧아 검출율이 낮아 센서의 제작이 쉽지 않다. 본 절에서는 비가시광 대역 이미지 센서의 현재까지의 개발 과정을 소개하고 현재의 문제점, 개선 방향과 향후 발전 방향을 예측해본다.

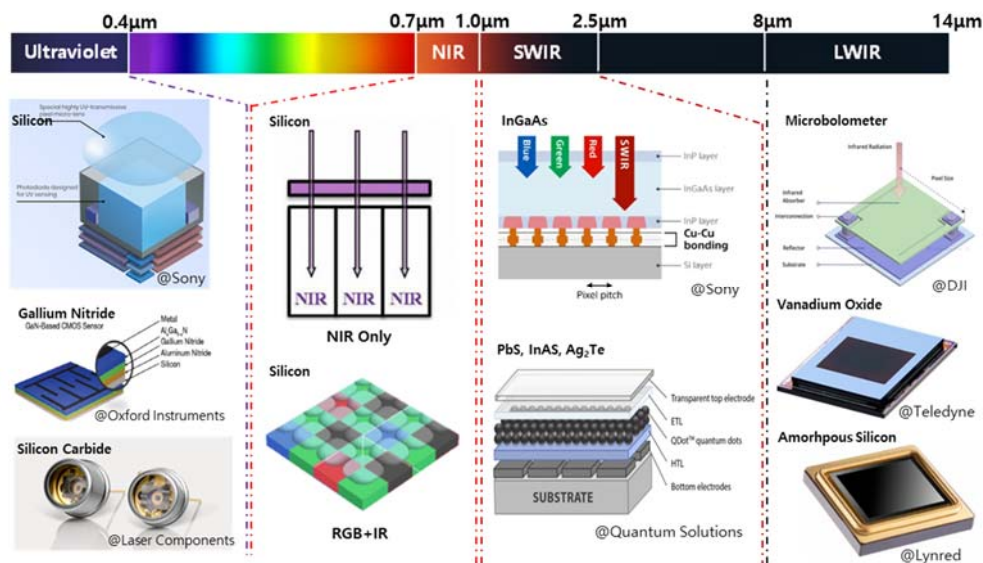


그림 6-4. 비가시광 대역 이미지 센서들의 흡광 소재

1) Ultraviolet (UV)

UV는 200~400nm 대역을 말하며 실리콘 소재에서 높은 흡수율을 보이기 때문에 실리콘 기반의 이미지 센서가 주를 이루고 있다. UV 파장의 특성은 실리콘 기반의 센서 제작에 도움이 되기도 걸림돌이 되기도 한다. 실리콘의 UV 파장의 흡수 계수 (absorption coefficient) 는 10^6 으로 매우 높아 실리콘 기반의 센서로 제작하기 적합해 보이지만 너무 높은 흡수율은 센서 안쪽까지 깊숙이 침투하는데 방해된다. 그 때문에 광자가 전자로 전환되기 전에 소재에 흡수가 이루어져 UV 센서들이 낮은 QE 를 보인다. 더 나아가 UV 센서에 필수인 모듈 렌즈, 마이크로 렌즈, 센서 커버 글라스의 재질 선택에도 UV 파장의 고려가 필요하다. 유리, 플라스틱, PMMA (Polymethyl Methacrylate) 같은 소재들은 UV 파장의 흡수율이 높아 센서까지 도달하는 것을 방해하기 때문에 UV 흡수율이 낮은 소재들로 대체하는 것이 필요하다. 최신 UV 이미지 센서의 스펙은 2/3", 8.13MP, 2.74 μ m pixel pitch, EQE 47.5% @300nm, Global Shutter 동작, 최대 12bit dynamic range로 [9] 향후 실리콘 기반의 UV 이미지 센서의 발전 방향은 고화소, 더 높은 QE, 그리고 소형화에 맞춰질 것으로 예상된다.

UV 이미지 센서의 발전은 흡광 소재 개발에 더 집중될 것으로 보인다. UV 대역의 낮은 검출율을 보완하고자 Perovskite Quantum Dot (PQD), Silicon Carbide (SiC), Zinc Oxide (ZnO), Gallium Nitride (GaN) 등 소재의 개발이 진행되고 있다. PQD는 ABX_3 의 구조를 가지도록 소재를 합성하는 방식으로 A, B, X의 물질에 따라 PQD 소재의 발광 파장이 정해진다. PQD를 사용한 UV 센서는 PQD를 photodiode에 도포하는 방식으로 제작이 되며 UV에 노출된 PQD가 광에너지 하향 변환을 통해 600nm 대역의 파장을 방출하도록 소재의 개발이 이루어진다. SiC, ZnO, GaN 물질들도 넓은 에너지 밴드갭을 가져 UV 센서로 많이 연구되고 있다. 다만 앞서 언급한 소재들은 UV 검출은 가능하지만 빛의 세기를 세밀하게 측정하는 것은 어려워 detector 형태로 사용된다. 이 소재들을 이미지 센서로 활용하기 위해선 dynamic range, 안정성, 신뢰성 등 여러 방면의 연구가 필요할 것으로 보인다.

2) Near Infrared (NIR)

NIR은 700~1000nm 대역을 말하며 실리콘에서 가시광 대비 낮은 흡수율을 보인다. 실리콘의 NIR 파장 대역의 흡수 계수는 $10^2 \sim 10^3$ 으로 가시광선 대비 10% 정도이다. 물론, Gallium Arsenide (GaAs), Germanium (Ge), Indium Phosphate (InP) 등과 같이 NIR 파장의 흡수율이 높은 소재들도 존재한다. 다만 Readout 회로가 실리콘으로 제작되어 있어 다른 소재 간 접합 시 높은 효율을 내

기 어려워 낮은 흡수율을 보이더라도 NIR 대역에서는 실리콘 소재의 사용이 지속되고 있다.

NIR 대역을 사용하는 응용처 중 최근 들어 가장 각광받고 있는 것은 Light Detection and Range (LiDAR) 이다. NIR이 LiDAR에 사용될 수 있는 이유는 Single Photon Avalanche Photodiode (SPAD) 소자의 개발 덕분이다. SPAD는 통상적인 photodiode의 reverse bias 동작 범위를 사용하지 않고 breakdown voltage 이하의 전압을 주어 동작시키는 소자로 광자의 양을 판별하기보단 광자의 유무를 판별하는 소자로 사용된다. SPAD 소자의 동작을 위해선 소자뿐만 아니라 NIR 대역의 active 광원과 검출된 광자의 거리를 정확히 측정할 수 있는 알고리즘도 시스템의 중요한 역할을 수행한다.

NIR 이미지 센서는 실리콘 기반으로 발전이 예상된다. Readout 회로와 픽셀 부가 이미 실리콘 기반으로 발전이 많이 되어 있는 만큼 신규 소재의 개발은 가능성이 낮을 것으로 판단된다. 보안, 검사, 검출 등 시장에서 요구하는 사항에 맞춰 QE를 극한으로 끌어올리는 시도와 함께 픽셀 미세화를 통한 고화질 센서 개발이 이루어질 것으로 예상된다.

NIR 대역이 주는 이미지 정보가 극저조도와 같은 특수한 상황에서 중요하다고 판단됨에 따라 RGB+IR 센서의 개발도 이루어지고 있다. 최근까지도 RGB와 IR 정보가 필요한 상황에서 두개의 다른 센서를 사용해왔다. 하지만 다른 센서를 사용하는 경우 두 센서 간 광축의 차이로 발생하는 상이 때문에 효율적인 정보 활용이 어려웠다. RGB+IR 이미지 센서는 통상적으로 GRGB 컬러 필터를 사용하는 RGB 센서의 B 또는 R pixel 하나를 IR필터로 교체하여 제작한다. 하나의 센서로 동일한 광축으로 측정된 4개의 파장 대역의 이미지를 획득할 수 있다. 현재 5MP, 2.1 μ m pixel pitch, QE >35% @940nm 성능의[10] 센서 출시를 앞두고 있으며 미래에는 고화소, 픽셀 미세화, QE 향상시키는 방향으로 발전할 것이 예상된다.

3) Short-wave Infrared (SWIR)

SWIR은 1.0~2.5 μ m 대역을 말하며 NIR 대역과 마찬가지로 눈으로 식별이 불가능한 영역이다. NIR 대역과 다른 점은 수분은 통한 흡수가 대부분 이루어져서 대기를 통과하며 대부분 감쇄되어 지표면에서 소량만 검출이 된다는 장점이 있다. 그 때문에 SWIR은 active 광원을 사용해야 하며 주변 광에 대한 간섭을 최소화할 수 있기 때문에 좀 더 간단한 시스템 구축이 가능하다. 수분에 의한 흡수가 이루어진다는 장점 때문에 NIR 보다 각막과 망막에는 좀 더 안전한 것으로 알려져 있다.

실리콘의 SWIR 흡수 계수는 $10^{-5} \sim 10^{-4}$ 로 이 파장 대역은 거의 흡수되지 않고 투과된다. 때문

에 실리콘 기반의 photodiode로는 검출이 불가능하여 Indium Gallium Arsenide (InGaAs) 흡수층을 통한 센서 개발이 주를 이루었다. 다만 앞서 언급했듯이 Readout 회로는 실리콘 소재로 제작되어 있어 InGaAs로 제작된 흡수층과 실리콘으로 제작된 Readout 회로단의 접합 공정이 필수이며 높은 공정 난이도 때문에 센서 제작의 수율이 높지 않다. 공정 난이도, 낮은 수율과 더불어 4인치 InGaAs 웨이퍼와 8~12인치 Readout 회로단의 공정 호환성이 좋지 않아 InGaAs 센서는 높은 가격을 형성하고 있다. InGaAs를 대체할 수 있는 물질에 대한 연구는 지속적으로 진행되고 있지만 현재로서는 InGaAs 보다 높은 QE를 달성하기는 어려워 보인다.

InGaAs를 대체할 수 있는 흡수층으로는 Germanium (Ge)와 Quantum Dot (PbS, InAs, Ag₂Te) 가 가장 유력하다. Ge는 SWIR 대역에서 흡수 계수가 10⁴로 InGaAs와 비슷해 InGaAs를 대체할 소재로 각광받아왔지만 실리콘과 Ge의 격자 상수 (lattice constant) 차이에서 생기는 접합면에서의 격자 불일치로 높은 성능을 기대하기 어렵다. Ge가 가지는 단점들을 극복하고자 실리콘 공정과 가장 호환성이 높은 QD 흡수층에 대한 연구도 활발히 진행되고 있다. QD는 구성 소재 자체에서 결정되는 흡수 계수, 에너지 준위와 더불어 QD의 입자 크기에 따라 생기는 양자구속효과로 흡수하는 파장 대역을 조절할 수 있다. SWIR 영역에서 활성화 파장 대역을 가지는 Lead Sulfide (PbS) QD 가 가장 먼저 제품화되었지만 InGaAs의 성능에는 미치지 못한다.

SWIR 이미지 센서의 향후 발전은 QE, 해상도의 증가와 더불어 Restriction of Hazardous Substances (RoHS) 요건 충족하는 QD 소재 개발에 있을 것으로 보인다. 현재 InGaAs를 사용하는 SWIR 센서는 독보적인 성능을 보이지만 비싼 가격으로 접근성이 떨어진다. 비교적 저렴한 QD 소재의 성능 향상이 이루어진다면 다양한 시장에서의 활용처가 확대될 것으로 보인다. 물론 RoHS 조건을 충족하지 못하는 PbS QD의 대체 소재 개발이 우선시되어야 하며 현재 InAs, Ag₂Te, 등 몇 가지 소재들이 가능성을 보이고 있다.

4) Long-wave Infrared (LWIR)

LWIR은 8~14μm 대역을 말하며 다른 대역들과는 다르게 열에 의해 발생하는 파장이다. LWIR 이미지 센서가 차용하는 microbolometer 방식은 주변 온도에도 민감하게 반응하기 때문에 오차 범위가 넓어 정확한 온도를 측정하는 응용처에는 사용할 수 없다는 단점이 있다. 때문에 대략적인 온도 표현, 온도 감지 정도로 사용처가 제한된다.

Microbolometer는 LWIR 파장에 민감하게 반응하도록 만들어야 해서 여러 공정 과정을 거친다. LWIR 대역을 흡수할 수 있는 소재인 Vanadium Oxide (VOx) 또는 Amorphous Silicon (a-Si) 재질로 이루어진 흡수막, 이 흡수막이 온도에 민감하게 반응할 수 있도록 공중에 띄우는 pillar 구조, 효

율을 높이기 위한 반사판 그리고 주변 온도에 영향을 최소화하기 위한 진공 밀봉 구조까지 필요하다. 이 때문에 공정 과정보다 까다롭고 복잡한 구조 때문에 Pixel Pitch를 줄이는데 어려움이 있다. VOx는 감도, 노이즈 성능이 좋지만 고온 안정성이 확보되지 않아 10년 주기로 센서의 교체가 필요하다. a-Si는 안정성 면에서 VOx에 앞서 고온 측정에 주로 사용되어 왔지만 감도와 노이즈가 커서 일반적인 사용처에 사용되기는 어렵다는 단점이 있다.

현재 LWIR 이미지 센서 시장의 과반 이상을 차지하는 소재는 VOx이다. 다만 a-Si의 가격 경쟁력 덕분에 VOx의 성능을 따라잡기 위한 a-Si의 연구가 계속되고 있다. Ge이 넓은 적외선 대역에서의 흡수 특성을 보이고 있어 a-Si 과의 합성 구조가 대체 물질로 연구가 진행되고 있다. 소재뿐만 아니라 픽셀 미세화를 달성하기 위해 pillar 구조를 더 간단히 할 수 있는 공정 방법들도 시도되고 있다^[11].

표 6-2에서는 UV, NIR, SWIR, LWIR 대역에서 사용되는 흡광 소재들과 그 소재를 사용한 이미지 센서의 발전 방향 및 이미지 센서로의 발전 가능성을 예측하였다. 이미 이미지 센서로 개발이 된 소재의 경우 시장의 요구에 따라 더 높은 해상도를 달성하는 방향으로 발전할 것으로 예측되지만 단일 소자 (unit detector) 로만 개발되어 있는 소재들은 array-type으로 발전하기 위한 연구가 지속될 것으로 예상된다.

표 6-2. 비가시광 대역 이미지 센서의 향후 발전 예측

	Sensing Material	2025	2028	2031	2034	2037	2040
Ultraviolet	Silicon	8MP	8MP	8MP	8MP	12MP	12MP
	GaN	Unit Detector	Unit Detector	Array-type	Array-type	Array-type	Array-type
	SiC	Unit Detector	Unit Detector	Unit Detector	Array-type	Array-type	Array-type
	PQD	Unit Detector	Unit Detector	Unit Detector	Unit Detector	Unit Detector	Array-type
Near-Infrared	Silicon (NIR)	12MP	12MP	12MP	12MP	12MP	12MP
	Silicon (RGB+IR)	5MP	5MP	5MP	8MP	8MP	12MP
Shortwave-Infrared	InGaAs	5MP	5MP	5MP	8MP	8MP	8MP
	Quantum Dot	PbS	6MP	6MP	8MP	8MP	8MP
		InAs	-	0.3MP	2MP	4MP	4MP
		Ag ₂ Te	-	-	0.3MP	1MP	2MP
Longwave-Infrared	Vanadium Oxide	1.3MP	1.3MP	1.3MP	1.3MP	3MP	3MP
	Amorphous Silicon	0.8MP	0.8MP	0.8MP	0.8MP	1.3MP	1.3MP

2.3. 비전 센서 (이벤트 기반)

이벤트 기반 비전 센서(Event-based Vision Sensor) 또는 다이내믹 비전 센서(Dynamic Vision Sensor)는 기존의 프레임 기반 카메라와 근본적으로 다른 방식으로 작동한다. 전체 픽셀 배열을 고정된 간격으로 샘플링하는 대신, EVS는 비동기적(asynchronous)으로 동작한다. 각 픽셀은 빛의

밝기 변화에 독립적으로 반응하며, 로그 스케일상에서 설정한 임계값을 초과하는 빛의 변화가 발생할 때만 이벤트를 생성하고 이벤트가 발생한 픽셀의 위치와 정보(밝아져서 발생한 이벤트인지 혹은 어두워져서 발생한 이벤트인지)만 전송하는 센서가 EVS이다. 빛의 변화가 있는 픽셀의 위치와 정보만 전달되기에 빠른 frame rate 과 낮은 전력 소비가 특징이다. 그림 6-5는 일반적인 CIS 카메라의 이미지와 EVS 카메라의 이미지 차이를 비교한 그림이다. CIS 와 달리 EVS는 움직이는 사물 촬영 시, 빛의 변화가 많이 발생하는 사물의 Edge 부분에서 주로 많은 이벤트 데이터가 발생하고 빛 변화가 없는 배경 부분은 이벤트 정보가 거의 없다.



그림 6-5. EVS와 CIS의 차이와 EVS의 기본 동작 방식

앞서 언급한 바와 같이 빠른 frame rate 이 장점인 EVS는 고속으로 움직이는 물체를 감지하거나 트래킹 하는데 활용될 수 있어 로봇/전장/IoT 등 다양한 응용처에 적용될 수 있으며, 최근에는 모바일 CIS 센서와 EVS 를 함께 활용하여 CIS에서 고속 물체 촬영 시 발생하는 이미지 블러 현상을 개선하거나 이미지/동영상 화질 개선용으로도 응용 연구가 많이 이뤄지고 있다(그림 6-6, 6-7 참조).

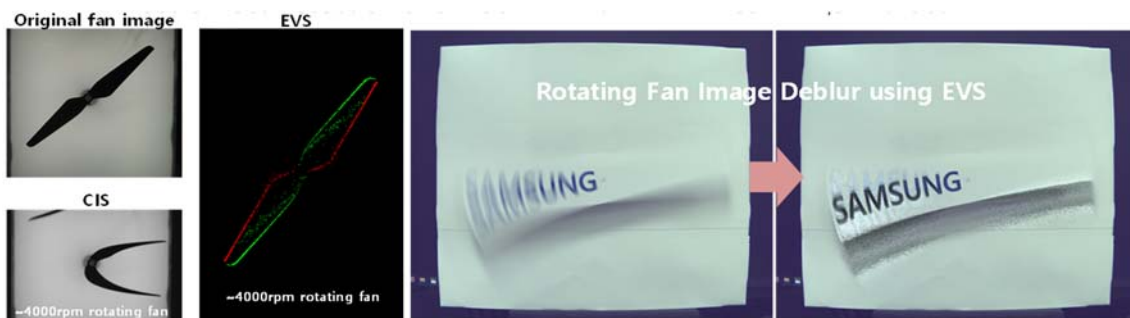


그림 6-6. 높은 frame rate 동작으로 Blur 최소화

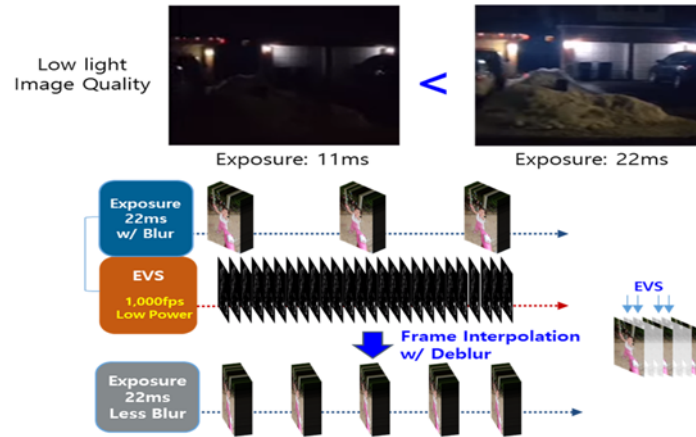


그림 6-7. EVS를 사용한 저조도 화질 개선 원리

HW 측면에서는 최근 CIS 이미지 센서 업체들이 CIS 와 EVS 를 하나의 센서에서 Hybrid 형태로 구현하는 연구를 진행하고 있고(그림 6-8 참조), 개별 센서들로 구현했을 때 극복하기 어려웠던 센서 간 disparity 및 small form factor 등의 이슈를 극복하여 응용 성능을 높이려는 노력을 꾸준히 진행해오고 있다.

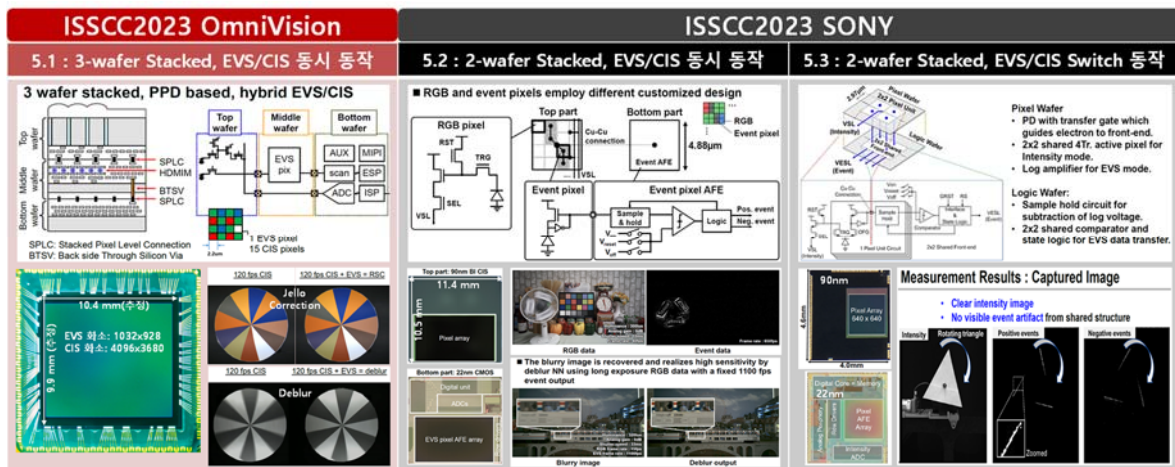


그림 6-8. EVS 관련 지속적인 연구 성과

EVS 센서의 향후 발전 방향에 대한 예측을 표6-3에 정리하였다. 이벤트 기반 센서는 개별 픽셀 별로 빛 변화를 감지하는 회로를 가지기에 픽셀 사이즈가 CIS 대비상대적으로 크다. 따라서, 응용 다변화를 위해 픽셀 shrink 통한 고화소 센서 구현이 필요하며, 저조도 성능 및 HDR 향상을 위한 소자 및 회로 개선 연구가 지속적으로 필요하다. 기능 측면에서는 비동기적으로 발생하는 이벤트 정보를 효과적으로 처리할 수 있는 전처리 IP 개발(Event Signal Processing) 및 I/F protocol 개발

과 On-sensor AI 기능이 꾸준히 요구될 것으로 보인다.

표 6-3. EVS 향후 발전 예측

	2025	2028	2031	2034	2037	2040
Pixel Size	4.86 μ m	< 4 μ m	< 3 μ m		< 2 μ m	
Resolution	1280x720	2560x1440	4K (3840x2160)		8K (7680x4320)	
Dynamic Range	86dB	100dB	110dB		120dB	
ESP 및 I/F On-sensor AI	Anti-flicker, Event Rate Control, MIPI, I2C, etc.	← + I3C, SPI, etc.	← + Simple AI functions (Motion Detection, Face Presence Detection, etc.)		← + Mature AI functions (Face Identification, Scene Parsing, etc.)	

3. 기타 검출기 센서

이미지 및 비전 센서와 같은 광학 정보 검출 센서 외에 온습도, 압력, 가속도 등 다양한 물리량을 측정하는 검출 센서들이 산업 전반에 걸쳐 줄곧 활용되어 왔다. 최근에는 반도체 공정 기술의 고도화와 신소재 개발에 힘입어, 단순 물리량 측정을 넘어 이온 및 전해질 조성, 유기화합물 농도 등 정밀 화학 분석까지 가능한 반도체 소자가 등장했다 (그림 6-9 참조). 이러한 소자를 기반으로 하는 센서들은 광학 센서를 보조하여 복합적인 환경 모니터링을 가능케 할 뿐 아니라, 정밀 농업, 바이오센싱 등 센서 산업의 영역을 확장하고 있다.

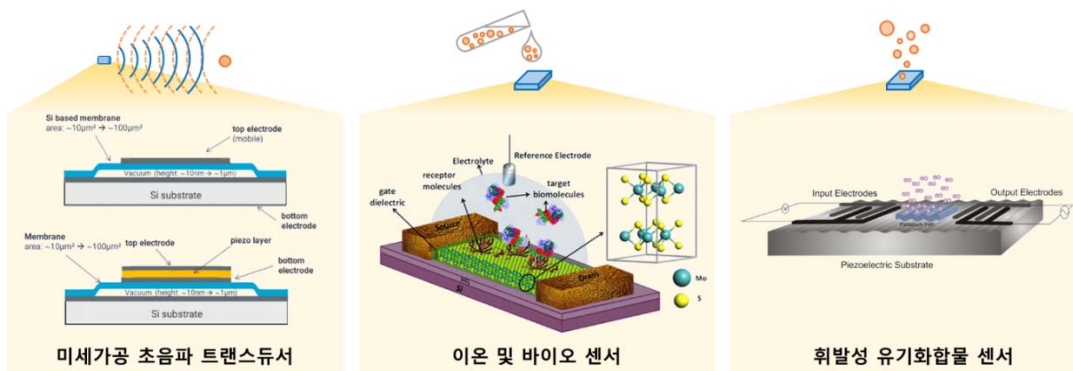


그림 6-9. 반도체 공정 기술의 고도화에 따라 등장한 검출기 센서[12]-[14]

본 절에서는 이미 상용화되어 널리 사용되고 있는 온도, 가속도 센서와 같은 전통적인 비광학 물리량 센서들을 다루는 대신, 최근 반도체 공정, 소재, 패키징 기술의 고도화에 따라 새롭게 등

장하거나 응용분야가 다각화되고 있는 센싱 기술들을 중점적으로 살펴본다. 특히, MEMS (micro-electro-mechanical systems) 공정의 발전으로 상용화가 진행중인 미세가공 초음파 트랜스듀서 (micromachined ultrasound transducer, MUT), 나노 물질을 활용하여 특정 이온 및 생체분자를 특이적으로 검출하는 이온 및 바이오 센서, AI 기술 적용이 점차 활발해지고 있는 휘발성 유기화합물 센서의 작동 원리와 주요 응용 분야, 향후 기술 발전 방향에 대해 자세히 살펴본다. 마지막으로 표 6-4에 검출기 센서들의 성능 향상을 위해 기술적으로 극복해야할 과제들을 단기/장기로 나누어 정리하였다.

1) 미세가공 초음파 트랜스듀서

초음파 센서는 인간의 가청 범위를 넘어서는 20kHz 이상의 음파를 송신하고, 대상 물체에서 반사되어 돌아오는 초음파 신호를 분석함으로써 거리, 속도, 위치, 움직임 등을 감지하는 기술이다. 전통적인 초음파 트랜스듀서의 경우, 압전 소자에 전압을 인가하여 초음파를 발생시키거나 외부 초음파에 의한 압전 소자의 물리적 진동에 의해 생성되는 전하를 분석하는 등, 압전 소자 자체가 진동자 역할을 한다. 반면, 최근 MEMS 공정의 발전으로 등장한 MUT의 경우 별도의 미세진동막을 제작하고 이를 진동시키거나 진동에 의해 발생하는 전기적 신호를 분석하여 초음파를 송수신할 수 있다. 유전체 미세진동막을 설계하고 초음파 송수신에 정전력 및 정전용량을 활용하는 구조를 정전용량형 미세가공 초음파 트랜스듀서 (capacitive micromachined ultrasonic transducer, CMUT), 박막 압전 소자 기반의 미세진동막을 설계하고 초음파 송수신에 압전 효과를 활용하는 구조를 압전형 미세가공 초음파 트랜스듀서 (piezoelectric micromachined ultrasonic transducer, PMUT)라고 한다. CMUT은 미세진동막 설계에 자유도가 높아 넓은 대역폭을 지원할 수 있는 대신, PMUT은 입력 전압이 낮고 에너지 효율이 높아 저전력 센서 개발에 유리하다. 현재 8인치 파운드리 공정을 기반으로 한 단일 MUT element와 2D MUT array 제조 기술은 현실화 단계에 있다^[14].

초음파는 빛이 닿지 않는 밀폐 공간이나 먼지가 많은 환경에서도 안정적으로 작동하므로 극한 환경에서 광학 센서를 보완하거나 대체하는 역할을 할 수 있으며, 지금까지 다양한 산업 환경과 자동차, 가전제품에 널리 활용되어 왔다. 한편, 수 MHz 이상의 고주파 초음파를 활용하면 인체 내부의 장기나 태아의 모습을 실시간으로 영상화할 수 있어, 의료 이미징 및 진단 장비의 핵심 기술로도 활용되고 있다.

초음파 센서는 전통적인 거리 감지 기술에서 출발했으나 반도체 공정과 집적회로 기술의 융합을 통해 고정밀 초소형화 센서로 진화하고 있으며, 그 응용분야는 향후 웨어러블, 로봇틱스, 스마트 모빌리티, 헬스케어 분야 전반으로의 확대가 기대된다.

2) 이온 및 바이오 센서

기존의 MOSFET (metal-oxide-semiconductor field-effect transistor) 구조에서 금속 게이트를 전해질과 reference electrode로 대체하면 전해질의 전위가 게이트 전압으로 작용할 수 있다. 이때 전해질의 전위는 이온의 농도에 따라 변화하며, 그에 따라 드레인 전류의 세기 또한 변화한다. 즉, 드레인 전류의 세기 변화를 분석함으로써 전해질 내 이온 농도를 측정할 수 있는데, 이를 Ion-Selective FET 혹은 ISFET이라고 한다. ISFET은 기존 MOSFET 공정과의 호환성이 높아 대량 생산과 집적화 측면에서 매우 유리하며, 현재 상용화되어 있다^[16].

최근 연구에서는 target 이온에 대한 선택성과 민감도를 높이기 위한 나노물질 합성이 활발히 진행되고 있다. 그래핀, MoS₂ 등의 2차원 나노소재는 전기 전도성과 화학 반응성이 매우 우수하여, 이를 활용한 ISFET은 미세한 이온 농도 변화도 감지할 수 있는 고감도 센서로 주목받고 있다. 또한, 나노소재 표면에 DNA와 같이 생체 분자와 결합하는 receptor를 결합함으로써, 이온뿐 아니라 특정 생체 분자를 검출하거나 그 농도를 측정할 수 있는 바이오센서 (BioFET) 로도 확장되고 있다. 이로써 혈액, 타액, 땀 등 다양한 체액에서 생화학적 지표를 직접 측정하는 기술이 현실화되고 있으며, 스마트 헬스케어 시스템의 핵심 센서로 부상하고 있다. 또한, 의료 이외의 분야에서도 농업·수질 관리·산업 공정 제어용 소형 이온 센서로 활용되어, 환경 내 오염물질이나 영양분의 변화를 정밀하게 측정할 수 있는 등 응용 가능성이 높다.

ISFET 및 BioFET은 반도체 소자 기술에 신소재 기술을 접목함으로써 화학적 및 생리학적 신호의 전기적 신호 변환을 가능케 하였으며, 향후 개인 맞춤형 건강관리, 질환 예측, 환경 모니터링 등 다양한 영역에서도 반도체 공정 기술이 핵심 역할을 계속할 것을 시사한다.

3) 휘발성 유기화합물 센서

전자코 (E-nose)라고도 불리는 휘발성 유기화합물 (volatile organic compounds, VOC) 센서 설계에 다양한 방식을 적용할 수 있으며, 현재 활발하게 상용화된 방식은 금속 산화물 기반의 방식이다. 이는 기존 반도체 공정과의 호환성이 높아 소형화와 대량생산에 유리하기 때문이다. 그러나 VOC에 대한 선택성 부족과 낮은 전력 효율 등의 한계가 있어, 최근에는 압전 소자의 표면 음향파 (Surface Acoustic Wave, SAW) 방식이 차세대 VOC 센서 기술로 떠오르고 있다. 이는 압전 소자 양쪽에 빗살 패턴의 전극 (Interdigital Transducer, IDT)을 위치시키는데, 한 쪽의 IDT에 전압이 인가되면 발생하는 전기적 신호가 압전 효과에 의해 SAW가 발생한다. 이는 반대쪽의 IDT로 전파되어, 다시 전기적 신호로 변환될 수 있다. 이때 SAW가 전파되는 압전 소자의 센서 표면에 target VOC를 선택적으로 흡착할 수 있는 polymer를 코팅하면 주변 환경의 VOC 분자가 이 층에 흡착

되고 압전 소자의 표면 물리적 특성이 변화하게 된다. 그에 따른 SAW의 전파 속도, 위상, 진폭 변화 패턴의 변화를 분석함으로써, 흡착된 VOC의 종류와 농도를 정밀하게 감지할 수 있다.

지금껏 금속 산화물 기반 방식 대비 CMOS 공정과의 낮은 호환성을 해결하기 위한 연구가 계속되어 왔으며, 쿼츠와 같은 압전 결정 대신 실리콘 기반의 SAW 센서 양산 공정이 점차 현실화되고 있다^[17]. 최근에는 AI 기술이 눈에 띄게 발전함에 따라 이를 활용한 미세한 SAW 변화 측정 혹은 복합적인 SAW의 분석 연구도 등장하고 있다. 한편, 앞서 소개한 ISFET과 유사하게 높은 반응성, 재현성, 내구성 확보를 위한 신규 polymer 소재 개발 연구도 진행되고 있다. 이러한 연구들은 SAW 기반 VOC 센서의 대량생산은 물론, 미세한 농도 측정 및 다중 VOC 혼합 환경에서의 개별 VOC 검출 등으로 이어질 수 있다. 덕분에 현재는 환경 모니터링, 공기질 관리, 산업용 가스 검출 등 비교적 제한적인 분야에 상용화된 VOC 센서는 고정밀 다중 VOC 감지 기술을 기반으로 향후 질병 진단 및 생체 가스 분석 등 의료 분야로의 응용이 기대된다. 예를 들어, 호흡 내 아세트론, 암모니아 등 특정 VOC의 조합은 각종 대사 질환이나 호흡기 질환의 조기 징후로 활용될 수 있어, 비침습적 진단 플랫폼으로 확장될 가능성을 보여준다.

이처럼 VOC 센서는 압전 소자, 고분자 화학, 신호처리 기술이 융합된 복합 반도체 센서로, 기존 물리량 중심의 감지 영역을 화학적 분석 단계로 확장하고 있다. 나아가 AI 기반 신호 해석과 나노물질 기반의 소재 혁신이 결합되면서 환경 및 의료 분야 전반에서 고감도 센싱의 핵심 기술로 주목받고 있다.

표 6-4. 기타 검출기 센서 기술적 극복과제 (로드맵)

	단기(2026~2031년)	장기(2032~2040년)
초음파 센서	• PMUT용 압전 박막 소재 개발 및 개선 • MEMS 기반 CMUT/PMUT 공정 수율 향상	• 초음파 송수신부와 신호처리 회로의 집적화 • 고해상도 다채널 어레이 및 실시간 이미징 기술 고도화
이온 및 바이오 센서	• 나노소재 기반 표면 개질로 이온 검출 선택성 및 정밀도 향상 • CMOS 공정 응용한 소형 ISFET 공정 확보	• 생체 분자 검출용 바이오센서 상용화 • 재현성, 내구성 및 오염 저항성 개선 • 나노소재, receptor 및 AI 활용한 혼합물 내 복합 이온 농도 분석 기술 개발
휘발성 유기화합물 센서	• SAW 기반 센서와 CMOS 공정 호환 • IDT 및 센서 구조 최적화 • VOC 흡착층의 반응 특성 개선	• 다중 VOC 감지형 고해상도 SAW 어레이 개발 • 재현성, 내구성 및 오염 저항성 개선 • AI 기반 복합 VOC 분석 알고리즘 개발

4. AI 시대를 위한 센서 미래 기술 방향

(1) In-Sensor DNN

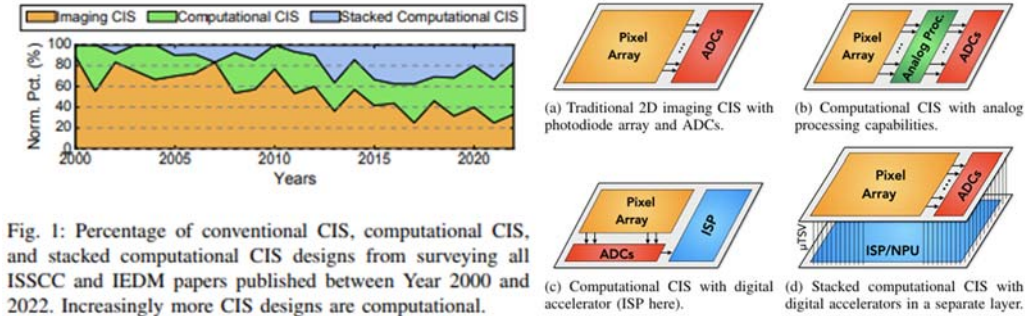


그림 6-10. CIS 이미지 센서와 AI의 융합 발전 방향^[18]

In-Sensor DNN이란, 이미지 센서 (CIS) 내부 또는 픽셀/근접 로직층에 DNN (또는 그 일부)을 통합해, Raw 프레임 대신 저차원 특성/메타데이터를 직접 출력할 수 있도록 지원하는 것을 말한다 (그림 6-10). 이를 통해, 이미지 기반 딥러닝 모델의 전부 또는 일부 레이어를 센서단에서 처리하여 저전력 및 저 지연으로 처리되도록 할 수 있다. 최근 CIS는 아날로그/디지털 연산 블록, 3D 스택킹 로직층, 전용 SRAM/DSP를 센서에 통합해 “컴퓨팅형 CIS(computational CIS)”로 진화하고 있다. 이러한 산업 동향은 ISSCC/IEDM (2000-2022) 분석을 통해서도 확인할 수 있는데, 컴퓨팅형/스택형 CIS 비중이 지속해서 증가하고 있다. 이는 CIS가 “영상 취득 장치”에서 “작업 특화 컴퓨터 노드”로 변화하고 있음을 나타내고 있다.

이를 위해, 기존 전통 CIS에서 아날로그 연산 통합, 디지털 ISP/가속기 통합, 로직층을 분리해 적층하는 등 점차 발전하는 양상을 보이고 있다. 이러한 변화는 센서 내부 TSV/하이브리드 본딩 경로의 데이터 이동 에너지는 $\sim 1\text{pJ/Byte}$ 에 불과하지만, MIPI CSI-2로 외부전송하는 전통적인 방식의 경우 $\sim 100\text{pJ/Byte}$ 수준인 것이 지속적인 관심을 끌고 있는 핵심 요소이다. 즉, 센서 안에서 집약 처리하고 메타데이터만 내보내면 에너지 이득이 $\sim 100\text{x}$ 이상 가능하다는 것이다^[19]. 이는 현재 발전중이지만 한계를 보이고 있는 MIPI 대역폭 문제에서도 In-Sensor DNN의 근거를 찾을 수 있는데, 25년 발표된 MIPI C-PHY Version 3.0의 경우 각 레인 성능을 약 30~35% 향상시켰으나, 이는 빠르게 증가하고 있는 고해상도/고 FPS 탑재 조건에, 멀티 카메라/멀티 노출 HDR/듀얼 ISO등 급격히 증가하고 있는 데이터 양에 맞춘 전송 요구량을 모두 만족시키기에는 부족한 면이 있다.

이러한 기술적 장점 및 배경에 근거하여, 다양한 In-Sensor DNN을 위한 Stack형 센서가 연구/

개발되고 있다. 최근에는 3-Wafer Stack을 통해 픽셀/로직/DNN 구조를 통합하여 제품화하는 상용화 제품 개발이 시도되고 있다^[20]. 앞으로 이러한 연구는 지속 발전하여 센서 단에서 점차 복잡한 추론까지 지원하면서 메타데이터를 전송하여 출력도 축약하여 대역폭을 절감하는 쪽으로 지속 연구될 것으로 보인다.

(2) 저전력 기술

저전력 기술은 항상 동작이 필요한 응용(예: 스마트폰 얼굴 인식, 도어락, 카메라 기반 사용자 감지 등)에서 센서가 초저전력 대기(AON, always-on)로 환경 변화를 감지하고, 필요 시에만 고전력 프로세서/고성능 모드를 깨우는 시스템 설계 일체를 뜻한다. 이에 (i) 센서 자체 상황 인지(Wake-on-Motion), (ii) 부분 판독/ROI (windowing) 로 불필요한 데이터 이동 축소, (iii) 경량 AON 모델 (양자화, 축약 구조) 운용, (iv) 센서+메모리+프로세서의 3D 적층 통합이 포함된다. 이러한 기술들은 이미 다양한 센서 제조사에서 상용화하여 제품화 중에 있다.

이는 센서의 항상 대기가 실시간 배터리 수명에 치명적이므로, 센서 단에서 먼저 가볍게 판단(예: 사람/얼굴/동작 존재 여부) 하여 상위 프로세서를 깨우는 방식이 요구됨에 근거한다. 연구 수준에서도 아날로그/혼합 신호 CNN을 센서에 통합한 AON CIS가 1.12mW, 120 fps로 AON동작이 가능함이 연구된 바 있다^[21]. 이러한 AON이외에도 연구되고 있는 대표적인 방법 중 하나는 부분 Readout/ROI 게이팅으로, 전체 장면이 아니라, 관심 영역만 선택적으로 판독하여 메모리, I/F 전력을 줄이는 방식으로 구현되고 있다.

(3) Multi-Sensor Fusion

멀티 센서 퓨전은 카메라, 라이다, 레이더, IMU, GNSS, 도로 인프라 센서 등 이기종 신호를 시간, 공간 정합하여 정확성과 강건성을 높이는 기술이다. 최근 딥러닝 기반 자율주행/로보틱스 영역에서는 data-level/feature-level/decision-level의 다층에서 다양한 신호들을 융합하는 연구가 지속, 발전하고 있다.

이는 단일 센서가 가지고 있는 원리적 한계 (occlusion, 악천후, 거리, 해상도 등)에 근거한다. 다수의 연구들은 멀티 센서 융합이 단일 센서 대비 객체 탐지의 재현율/정확도를 유의하게 높이고 있음을 보이고 있다. 현재 멀티 카메라 이미지를 BEV (Bird Eye View) 로 변환하여 Transformer 모델을 사용하여 결합하는 것에서 더 나아가서, 추후 불확실도 지도/센서 건강도 등 품질 추정 정보까지 활용하여 Fusion이 발전할 것으로 예상된다. 이 후, V2X 기반 Cooperative Perception (CP) 등의 방법이 발전하여 차량-차량/차량-인프라 간 시야 공유로 원거리, 비가시 Occlusion을 보완하

는 등 다양한 센서 및 장치에서의 정보를 융합하는 방향으로 Multi-Sensor Fusion이 발전할 것으로 보인다^[22].

표 6-5에서는 이미지 센서의 AI 응용 및 발전 방향을 예측해보았다. 이미지 센서 내부 또는 픽셀/근접 로직층에 DNN을 탑재하여 더 높은 FPS를 달성하고 AON 대기전력 최소화, 출력 축약률 최대화를 달성하는 방향으로 발전할 것으로 예상된다. 더 나아가 이미지 센서와 더불어 다양한 센서들과의 Fusion을 통해 응용처에서 요구하는 정확도 및 재현성을 향상시키는 방향으로 발전을 예상한다.

표 6-5. 센서의 AI 응용 및 발전 방향 예측

	2025	2028	2031	2034	2037	2040
In-Sensor DNN AI FPS	~30 fps	~45 fps	~60 fps	~ 90 fps	~120 fps	~120-180 fps
출력 축약률 (대역폭 절감)	10^3 - 10^5 x	10^3 - 10^5 x (확대)	10^4 - 10^6 x	10^5 - 10^6 x	10^5 - 10^7 x	10^6 - 10^7 x
AON 대기전력 (센서단)	~7mW	~5mW	~2-3mW	~1-2mW	~1mW	<1mW
Sensor Fusion	카메라-라이다/레이더의 Feature-level fusion		센서/장치 내 품질 추정 정보까지 Fusion에 활용		Cooperation Perception(CP) 표준화	

5. 요약

센서는 지속적인 발전을 통해 사람이 감지할 수 있는 정보를 보다 더 세밀하고 정확하게 측정, 기록할 수 있게 되었다. 이미지 센서는 인간의 눈을 따라잡기 위해 지속적으로 발전할 것이며 더 나아가 센서는 인간의 눈, 코, 혀, 피부, 등으로 감지하지 못하는 정보들까지 기록할 수 있는 새로운 센서의 개발까지도 확인할 수 있을 것으로 보인다.

현재 센서의 개발은 파장 또는 입자 중 한 가지의 정보를 더 높은 효율로, 더 세밀하게 기록할 수 있도록 센서의 성능을 높이는 방식에 집중되어 있다. 이러한 개발 방향이 점차 성능의 최대치를 향해 가면서 각기 다른 센서가 검출하는 정보를 모아 활용도를 높일 수 있는 DL 기술의 개발도 지속적으로 이루어질 것으로 보인다.

참고문헌

- [1] A. Theuwissen, "Major Consumer Image Sensor Innovations Presented at IEDM," 2024 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 2024, pp. 1-4.
- [2] M. Miyata, "Dispersion-engineered metasurfaces for high-sensitivity color image sensors," *Opt Rev* 31, pp. 290-298, 2024.
- [3] C. Choi et al., "Optical design of dispersive metasurface nano-prism structure for high sensitivity CMOS image sensor," 2023 International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 2023, pp. 1-4.
- [4] <https://semiconductor.samsung.com/kr/image-sensor/automotive-image-sensor/isocell-auto-1h1/>
- [5] <https://www.ovt.com/products/ox03j10/>
- [6] H. Shim et al., "6.1 A 3-Stacked Hybrid-Shutter CMOS Image Sensor with Switchable 1.2 μ m-Pitch 50Mpixel Rolling Shutter and 2.4 μ m-Pitch 12.5Mpixel Global Shutter Modes for Mobile Applications," 2025 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 2025, pp. 1-3.
- [7] T. -H. Tsai et al., "6.4 A 400 $\times$ 400 3.24 μ m 117dB-Dynamic-Range 3-Layer Stacked Digital Pixel Sensor," 2025 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 2025, pp. 120-122.
- [8] <https://global.canon/en/technology/spad-sensor-2023.html>
- [9] <https://www.sony-semicon.com/en/products/is/industry/uv.html>
- [10] <https://www.sony-semicon.com/en/news/2025/2025100201.html>
- [11] Yadav, PV Karthik, et al. "Advancements of uncooled infrared microbolometer materials: A review." *Sensors and Actuators A: Physical* 342 (2022): 113611.
- [12] <https://vermon.com/cmut-pmut-technologies/>
- [13] D. Sarkar et al., "MoS₂ field-effect transistor for next-generation label-free biosensors," *ACS nano*, vol. 8, no. 4, pp. 3992-4003, 2014.
- [14] Y. Wang et al., "Sensitive Materials Used in Surface Acoustic Wave Gas Sensors for Detecting Sulfur-Containing Compounds," *Polymers* 16(4),457 (2024).
- [15] <https://vermon.com/cmut-pmut-technologies/>

-
- [16] <https://www.lionix-international.com/mems/mems-solutions/mbari-isfet-for-deep-sea-ph-sensor/>
 - [17] <https://www.isit.fraunhofer.de/en/mems/Acoustic-systems-and-microdrives/saw.html#cop16718282>
 - [18] <https://horizon-lab.org/pubs/iccad23.pd>
 - [19] Y. Feng, T. Ma, A. Bolor, Y. Zhu and X. Zhang, "Invited Paper: Learned In-Sensor Visual Computing: From Compression to Eventification," 2023 IEEE/ACM International Conference on Computer Aided Design (ICCAD), San Francisco, CA, USA, 2023, pp. 1-9.
 - [20] R. Nakamura et al., "A Novel 1/1.3-inch 50 Megapixel Three-wafer-stacked CMOS Image Sensor with DNN Circuit for Edge Processing," 2024 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 2024, pp. 1-4.
 - [21] J. Choi, S. Lee, Y. Son, and S. Y. Kim. "Design of an Always-On Image Sensor Using an Analog Lightweight Convolutional Neural Network," Sensors. MDPI AG, May 30, 2020.
 - [22] Y. Ji, Z. Zhou, Z. Yang, Y. Huang, Y. Zhang, W. Zhang, L. Xiong, Z. Yu, "Toward autonomous vehicles: A survey on cooperative vehicle-infrastructure system," iScience, vol. 27, no. 5, 2024.

VII. 유선연결 반도체 기술

1. 서론

반도체 시스템 산업의 목표는 많은 양의 데이터를 빠른 시간 내에 처리하여 반도체 시스템의 성능을 향상시키는 것이다. 단위 시간당 보낸 데이터의 양을 데이터 처리량 (throughput) 이라 정의하는데, 높은 데이터 처리량을 달성하려면 단위 시간당 송수신하는 데이터 비트가 많아야 한다. 이를 데이터 전송속도 (bit rate)라고 하며, 기존 반도체 시스템에서 데이터 전송 속도가 높아짐에 따라서 전통적 성능 향상 방법으로 한계를 겪고 있다. 칩의 성능 향상 대신, 칩을 물리적으로 보호하던 패키징 기술이 반도체 시스템 성능 향상의 새로운 해결책으로써 관심을 받고 있다.

아래의 그림은 SerDes의 여러 표준에서 규정한 데이터 전송 속도를 2000년부터 2024년까지 표현한 것이다. 각 표준에서는 꾸준히 데이터 전송 속도를 높이도록 규정하고 있으며 매년 데이터 속도의 증가량은 선형적으로 증가하는 것이 아니라 지수적으로 증가함을 보이고 있다. 이러한 기술적 경향은 유선 연결에 요구되는 주파수 대역 (Bandwidth) 또한 지수적으로 증가함을 의미한다. 전송선 이론에 따르면, 모든 금속 인터커넥트는 기생 저항 및 커패시턴스 성분으로 인해 저역통과필터 (low pass filter)처럼 동작한다. 다시 말하면, 고속 직렬 링크에 요구되어지는 신호 전송 속도를 달성하기 위해서는 대책이 필요하다.

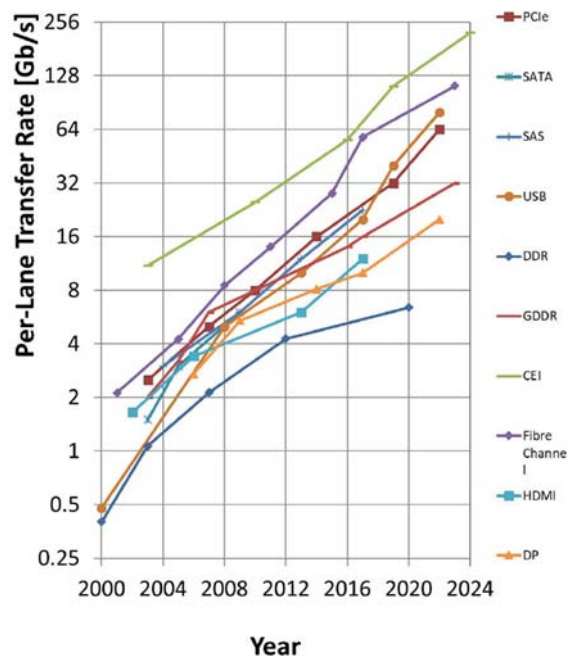


그림 7-1. SerDes 규격 데이터 전송속도 경향

2. 유선 연결 기술 동향

유선 연결은 반도체 시스템에서 반도체 칩 간의 통신을 금속 배선을 이용하여 연결하는 것으로 규정할 수 있으며, 수준에 따라 다음과 같이 정의할 수 있다.

(1) 이종 집적 (Heterogeneous Integration)²을 통한 시스템 설계이다. 이종 집적은 반도체 패키징 수준에서 집적이 이뤄지며, 대표적으로 인터포저 기반의 시스템 집적과 칩렛 기술 기반의 시스템 집적이 있다. 인터포저 기반의 집적은 기존 패키지 기판을 대신하여 고집적 배선 기술이 가능한 물질을 이용하여 칩 간의 연결을 달성한다.

아래의 표는 이종 집적과 단일 집적의 차이를 비교하고 있으며 이종 집적이 차세대 반도체 시스템에 왜 적합한지를 보여준다. 이종 집적은 반도체 집적의 한계를 극복할 수 있게 함으로써 반도체 시스템의 성능 개선에 자유도를 안겨주었다.

표 7-1. 집적 방식 비교

Material	Heterogeneous Integration	Homogeneous Integration
Component origin	Different technologies & process	Typically different
Integration level	Package-level (advanced packaging)	Substrate-level (PCB)
Coupling tightness	High	Moderate to low
Performance optimization	High	Moderate
Interconnect density	High	Lower (bond wires, traces)
Application domains	AI Chips, Smartphones, 3D ICs	RF modules, hybrid ICs

위에서 소개했듯이, 이종 집적의 대표적인 핵심 추진 요소는 인터포저³이다. 반도체 시스템을 계층으로 표현할 때, 계층 사이에 존재하며 그 정의는 다음과 같다.

² 이종 집적은 다양한 반도체 기술, 재료, 혹은 기능 블록(예: 로직, 메모리, 센서, RF 등)을 단일 패키지나 시스템 수준에서 집적하는 기술을 의미한다. 이를 통해 각 소자의 장점을 극대화하고, 고성능·저전력·소형화된 시스템을 구현할 수 있으므로 차세대 반도체 시스템에서 각광받고 있는 집적 방식이다.

³ 기판 기반의 신호/전원 배선을 중간 매개체를 활용하여 배선의 집적도를 높일 수 있는데, 이때 중간 매개체를 인터포저라고 정의한다. 어떤 물질을 사용하냐에 따라 집적도 및 전기 성능이 달라지며 가격 또한 다양하므로 시스템의 요구에 맞게끔 선택된다.

아래 표는 인터포저의 물질에 따른 장점 및 한계점을 나타내고 있다. 인터포저의 발전 방향을 시스템의 집적도를 증가시키는 방향으로 맞춰어져 있지만, 인터포저의 물질에 따른 장/단점이 명확하게 규정되기 때문에 아래와 같이 시스템에 요구되어지는 특성에 따라 정해진다.

표 7-2. 물질에 따른 인터포저 비교

Material	Key Benefits	Challenges
Silicon	- High density - Mature	- High cost - Lossy at high frequency
Glass	- Low loss - Low CTE	- Brittle - TGV manufacturing
Organic (ABF, BT)	Low cost	Lower precision
Polymer (LCP, PI)	Excellent RF performance	Processing complexity

표 7-3. 패키지 주요 인터커넥트 비교

Feature	Wire bonding	C4 bump	C2 bump	Hybrid bonding
Material	Au/Cu wire	SnPb/SnAg solder	Cu pillar + solder	Cu-Cu direct
Signal speed	< 1 GHz	10-20 GHz	30-50 GHz	> 100 GHz
I/O density ⁴	Low	Medium	High	Ultra high
Elec. parasitics	High L, R	Moderate	Low	Very low
Reliability	Excellent	Good (with underfill)	Moderate (solder fatigue)	High (no solder fatigue)

고속 시스템용 패키지의 유선 연결을 위한 인터커넥트는 크게 4종류가 있다. 개발 순서대로 (1) Wire bonding (WB), (2) Controlled collapse chip connection (C4) bump, (3) Chip connection (C2)

⁴ 입/출력 신호의 집적도인 I/O density는 단위 면적당 신호 배선의 갯수를 의미한다. 단순히 공간 효율성으로 볼 수도 있지만, 집적도의 향상으로 인해 신호/전원 무결성 성능이 개선되므로 반도체 시스템 설계 과정에서 중요한 지표이다.

bump, (4) Hybrid bonding이다. 아래 표에 정리되어 있듯이 wire bonding은 신뢰도가 높은 반면에 전기적 기생 성분 값이 크므로 전송할 수 있는 신호 대역이 1 GHz 아래의 대역이다. C4 bump는 주석-납 합금을 이용해 WB 보다 더 짧으면서 더 작은 기생 성분 값을 가지면서, 전송 신호의 대역 폭은 대개 10-20 GHz 대역이다. C4 범프의 집적도를 높이기 위해서 구리 기둥 (Cu pillar)을 이용하여 더 높은 집적도를 위해 C2 범프가 소개되었다. 마지막으로, 유전체와 구리를 동시에 접합시키는 hybrid bonding이 제안되어 가장 높은 집적도의 인터커넥트를 달성하였다.

인터포저에서 주요 연결 요소는 실리콘 관통 전극 (Through Silicon Via, TSV)⁵이며, 기존 인터커넥트인 wire bonding (WB)에 비해 길이가 굉장히 짧다. 길이가 짧아짐에 따라 기생 인덕턴스와 저항 값을 매우 낮은 수준으로 저감할 수 있으므로 신호 전달 특성을 개선할 수 있다. 이를 통해 TSV를 사용하면 반도체 시스템의 집적도와 성능을 동시에 개선할 수 있다. 실리콘 인터포저에서 사용되는 TSV는, 유리 기판에서는 TGV (Through Glass Via)에 해당된다. TSV와 마찬가지로 TGV 또한 수직 인터커넥트로서 사용된다. TSV와 TGV의 차이를 아래 표를 통해 비교했으며 대표적인 차이는 물질의 특성 차이에 기인한다. 이는 실리콘과 유리의 유전 상수값이 달라서 발생하는 현상으로 유전 상수는 고주파 신호에 얼마나 물질이 잘 반응하느냐는 물질 특성을 의미한다. 이러한 차이로 인해 실리콘과 유리가 사용되는 영역이 결정되며, 추가적으로 유리의 경우 패널 레벨의 공정이 가능해서 가격적인 이점 또한 존재한다.

표 7-4. TSV와 TGV 비교

Item	TSV (Through Silicon Via)	TGV (Through Glass Via)
Electrical loss	Higher	Lower
Crosstalk isolation	Moderate	Excellent
Thermal path	Good	Poor
Mechanical stress	High (Cu-Si mismatch)	Moderate
Frequency suitability	Digital/High-performance logic	RF/mmWave/optical
Manufacturing cost	Expensive (Wafer-level)	Lower (Panel-level possible)

⁵ Chip간 신호선 연결을 위한 수직 인터커넥트 기술로써 실리콘에 구멍을 생성하여 금속 물질을 메워서 수직 인터커넥트 연결을 완성한다. TSV의 크기 및 물질은 TSV의 형성 단계에 따라 결정되며, 이로 인해서 이들의 전기적 성능 또한 결정된다.

인터포저 기반의 집적 기술은 신호 뿐만 아니라 전원 전달 성능도 개선시킨다. 반도체 시스템에서 동시에 많은 트랜지스터가 동작하면 순간적으로 많은 전류를 사용하는데, 이 때 기생 저항에 의해서 전압이 순간적으로 변하는 잡음이 발생하게 된다. 이러한 잡음을 감쇄하기 위해서 반도체 칩 가까이에 큰 값의 기생 커패시턴스를 위치시킨다. 기존 패키지 기판에서 만들던 전원/접지 평면을 인터포저에 배치시킴으로써, 거리가 가까워지게 되어 기생 인덕턴스가 감소하여 전원 전달 성능이 대폭 개선된다. 위에서 TGV의 손실이 작다고 소개했지만, 전원 무결성 관점에서는 저감된 감쇄 성능으로 인해 공진이 눈에 띄게 관측된다고 알려져있다. 다시 말하자면, 실리콘 기판에서는 큰 전기적 손실로 인해 저감된 공진이 유리 기판에서는 공진이 감쇄되지 않아 신호에 다시 역으로 영향을 주는 경우가 발생한다.

(2) 실리콘 영역의 일부 개선을 통한 시스템 설계이며, 칩렛⁶이 대표적인 기술이다. 칩렛 기술 기반의 시스템 집적은 기존에 단일 공정으로 제작하던 실리콘 칩 (Monolithic Si chip)을 여러 공정으로 각각의 칩 조각 (chiplet)을 생산하여 이들을 패키지 단계에서 집적하는 것이다. 인터포저 기술은 패키지 기판 수준에서, 칩렛 기술은 실리콘 칩 수준에서 집적도를 높이는 것으로 설명할 수 있다.

표 7-5에서 볼 수 있듯이, 칩렛간 고속 인터커넥트 통신 규격은 (ㄱ) Bunch of Wire (Bow), (ㄴ) Advanced Interface Bus (AIB), (ㄷ) Universal Chiplet Interface express (UCIe) 이다. UCIe는 직렬 링크에 대한 통신을 규정하고 있으며 최대 속도를 32 Gbps로 정의하는 반면에, BoW와 AIB는 여러개의 채널을 이용하여 equalization⁷ 없이 single-ended 신호⁸ 전송방식을 기반으로 데이터를 송/수신한다.

⁶ 단일 공정 대신, 여러 공정을 사용하여 칩의 생산 비용을 낮추는 기술로써 대표적으로 Core와 입/출력 회로를 분리하여 목적에 맞게끔 입/출력 회로를 구성한다. 대표적인 예시는 인텔의 Stratix 10이 있으며 통신 규격에 따라 입/출력 회로를 이더넷, PCIe 등으로 구성한다.

⁷ 고속 채널을 통해 데이터 전송 시에 채널의 기생 저항(R), 인덕턴스 (L), 커패시턴스 (C)에 의해서 데이터 전송 과정에서 손실을 겪게 된다. 이러한 손실이 없는 채널은 없으므로, 이를 보상하기 위해서 송/수신단에서 추가적인 회로를 이용하여 손실을 보상하는 기술을 equalization (등화)이라 한다.

⁸ 단일 금속을 이용하여 데이터를 전송하며, 통상적으로 기준 (reference) 신호는 공통 평판 (plane) 형태로 이용한다. 집적도를 높일 수 있다는 장점이 있지만, 차동 (differential) 시그널링 방식에 비해 외부 잡음에 취약하다.

신호 무결성 (Signal Integrity) 관점에서 가장 적합한 표준은 UCle이며, 그 이유는 차동 (differential) 신호 전송 방식을 이용하여 외부 잡음에 대한 내성이 우수하며 equalization과 encoding⁹을 지원하기 때문이다. UCle의 또다른 타 인터커넥트 표준과의 차별점은 CDR (Clock Data Recovery)¹⁰를 채용함으로써 클록 신호에 대한 공간/시간적 여지를 줄였다.

결론적으로, 칩간 고속 인터커넥트 통신 규격 중 UCle가 신호 전송 능력에 대한 여러 기능들을 지원하도록 정의하므로 유한한 주파수 대역에 대한 신호 전송에서 타 표준에 비해 강점들을 가지고 있다. 이뿐만 아니라, 거리에 대한 규격도 가장 멀리 (up to 10 mm) 지원하므로 고성능 칩렛 아키텍처에서 적절하다는 평을 받고 있다.

표 7-5. 칩간 고속 인터커넥트 표준 비교

Feature	BoW	AIB 2.0	UCle1.1
Type	Parallel	Parallel	Serialized
Lane speed	0.5-16 Gbps	1-2 Gbps	Up to 32 Gbps
Latency	1-2 ms (low)	< 1ms (very low)	10-20 ns
Signaling	Single-ended or differential	Singled-ended (Parallel I/O)	Differential (High-speed serial)
Clocking	Forwarded clock or source-synchronized	Forwarded clock	Embedded clock
Equalization	Not required	Not required	Equalization & Encoding
Clock Data Recovery (CDR)	Not required	Not required	Required
Protocol support	Protocol agnostic	Memory-mapped	PCIe / CXL / Memory-mapped
Distance support	Up to mm	< 2mm	Up to 10 mm

⁹ 인코딩은 데이터 비트를 목적에 맞게끔 심볼(symbol)로 변환하는 과정이며, 목적은 신호 무결성 향상부터 오류 정정 (error correction)까지 다양하며, 신호 전송 성능 향상을 위해 사용되는 인코딩은 대표적으로 128b/130b, 64b/66b, TMDS 등이 있다.

¹⁰ 클럭-데이터 복원 회로를 의미하며, 클럭 신호를 데이터 신호에 임베딩하여 수신 회로에서 클럭 신호를 복원한다. 클럭 신호를 따로 수신받지 않아도 되는 장점이 있다. 혹은 Source synchronous처럼 클럭 신호에 맞춰서 데이터 신호의 타이밍을 조절하지 않아도 된다.

표 7-6. 칩간 고속 인터커넥트 표준 비교

Feature	PCIe	Ethernet	USB
Type	칩 간/보드내 통신	네트워크 통신용	주변기기용 단거리 통신
Lane speed	PCIe 6.0: 64 GT/s	up to 224 Gbps	USB3.2: 20 Gbps
Latency	~ 100 ns	~ μ s	10~100 μ s
Signaling	PCIe 6.0: PAM-4	NRZ, PAM-4	NRZ (일부 PAM-3)
Clocking	Embedded CLK	Embedded CLK	Source-synchronous
Equalization	CTLE + DFE + Tx FIR	CTLE + FFE + DFE + Adaptive EQ	Simple CTLE or Pre-emphasis ¹¹
Distance support	Up to 1 m	cm ~ km	~ 2 m (passive) ~ 5 m (active)

(C) 실리콘 영역의 개선을 통한 시스템 설계이며, SerDes¹² 회로 개선이 대표적인 사례이다. SerDes 기술은 고속 데이터 전송을 위해서 필수적이며 많은 양의 디지털의 신호를 고속 아날로그 신호로 변환하는 역할을 한다.

표 7-6은 SerDes 표준 중 대표적인 고속 표준들을 비교하였다. PCIe는 칩-to-칩/보드 내의 통신을 규정하며 이더넷과 USB는 조금 더 큰 범위의 통신 방식을 규정한다. 표준 별 데이터 전송 속도는 아래와 같으며 지속적으로 증가하고 있다. 앞서서 간단히 소개했듯이 데이터의 전송 속도 증가는 요구하는 주파수 대역 폭의 증가와 직접적으로 연관되므로 데이터의 전달 능력을 강화하기 위해 주파수 대역을 적게 사용하는 PAM 신호 전송 방식을 사용하는 추세이다. 클럭 신호 또한 데이터 채널에 embedding하는 방식을 사용하는데, 이는 신호선 감소, 시스템 단순화, 그리고 신호-클럭 신호 간의 위상 불일치 문제를 해결할 수 있기 때문이다. Equalization도

¹¹ 송신단에서 고/저주파 신호들을 증폭시켜 신호 왜곡에 감쇄시키는 회로

¹² Serializer/Deserializer는 고속 입/출력 인터페이스 회로를 말하며 칩/보드 간 데이터를 빠르고 효율적으로 주고받기 위한 회로 요소이며 핀 수와 배선 수를 줄이면서도 고속 데이터 전송을 실현하기 위함이 목적이다.

포함하는 추세이며 CTLE¹³는 모두 채용하는 상황이며 DFE/FFE¹⁴는 선택적으로 사용하는 상황이다.

3. Chiplet 로드맵

표 7-7. 칩렛 로드맵

연도	2025	2029	2033
산업전망	Chiplet 상용화 전환점	Chiplet 확장기	Polyolithic SoC
인터커넥트	UCle기반 short reach (SR) PHY (~2 mm)	2D (UCle SR) +3D TSV Hybrid	3D TSV +Hybrid bonding +Optical I/O
패키징	2.5D 구조 중심	CoWoS, Foveros, SoIC	Wafer on Wafer
설계 자동화	Chiplet-aware 설계	Chiplet기반 설계 자동화	AI-assisted Co-design
전력 관리	Chiplet간 전력분산구조	칩렛별 DVFS	DVFS 고도화

연도	2037	2040
산업전망	범용 설계/생산	OS-like chip management
인터커넥트	Optical I/O TSV +Si Bridge +MCM 구조	전기+광학+양자 인터페이스 통합
패키징	Flexible substrate	Interposer-less fully 3D vertical integration
설계 자동화	Chiplet통합 최적화 자동화	다중 칩렛 동적 매핑 시스템
전력 관리	Chiplet별 전력 도메인 분리	패키지레벨 전력 공유

(1) 산업 전망

¹³ 고속 채널에서의 신호 왜곡을 주파수 영역에서 보상하는 회로

¹⁴ 기생 저항 및 커패시턴스에 의해 발생하는 왜곡을 시간 영역에서 보상하는 회로

표 7-7에서 언급한 바와 같이, 반도체 유선 연결 산업은 향후 Chiplet 기술의 상용화와 생태계 확장기를 거치며 급격한 변화를 맞이할 것으로 예상된다. 이러한 변화를 기반으로, 다수의 이기종 칩을 통합한 Polyolithic SoC 구조로 진화하고, 표준화된 인터페이스를 중심으로 한 범용 설계·생산 체계가 확립될 전망이다. 궁극적으로는 운영체제(OS) 수준에서 칩 간 자원과 기능을 관리·조율하는 Chip Management 체계로 전환되어, 반도체 시스템의 유연성과 확장성이 한층 강화될 것으로 보인다.

(2) 인터커넥트

표 7-7에서 알 수 있듯이, UCle 기반의 단거리 고속 전송 구조를 중심으로, 2D 및 3D TSV 하이브리드 방식으로 확장되고 있다. 이후에는 하이브리드 본딩과 광학 I/O, 그리고 실리콘 브릿지·MCM 구조의 결합을 통해 전기적·광학적 통합이 가속화된다. Si Bridge¹⁵-MCM 구조는 하이브리드 패키징 구조로써, 실리콘 브릿지를 이용해 칩간 고속 구간을 직접 연결하고 그 외 주변 회로나 저속 신호들은 MCM(Multi-Chip Module)¹⁶ 기판을 통해 연결하는 방식이다. 궁극적으로는 전기·광학·양자 신호가 하나의 통합 인터페이스로 융합되어, 칩 간 경계가 사실상 사라진 완전 통합형 인터커넥트 구조로 진화할 것으로 예상된다.

(3) 패키징

패키징 기술은 초기에는 2.5D 구조 (CoWoS¹⁷, Foveros¹⁸, SolC¹⁹ 등) 중심으로 발전하며, 칩 간 연결 밀도와 신호 무결성을 동시에 확보하는 방향으로 진화하고 있다. 이후에는 Wafer-on-Wafer 및 Flexible Substrate 기술이 적용되어, 다이 적층의 자유도와 패키지 구조의 유연성이 크게 향상된다. 궁극적으로는 인터포저가 필요 없는 완전한 3D 수직 통합 구조로 발전하여, 칩 간 물리적 경계가 사라지고 시스템 수준의 집적이 가능해질 것으로 예상된다.

¹⁵ 실리콘 브릿지 기술은 실리콘 인터포저의 대표적인 단점인 비싼 가격 문제를 해결하기 위해 소개되었으며, 칩 인터커넥트의 필요한 부분에 대해서 실리콘 조각 (bridge)를 매립하여 신호를 전달한다. 대표적으로 Intel의 EMIB (Embedded Multi-die Interconnect Bridge)가 있다.

¹⁶ 여러 개의 칩을 하나의 패키지 기판 위에 배치하여 단일 시스템처럼 동작하도록 하는 패키징 기술이며, 이를 통해 칩간 배선 길이를 줄이고 신호 지연과 전력 손실을 최소화할 수 있다.

¹⁷ CoWoS (Chip on Wafer on Substrate): TSMC의 실리콘 인터포저 기반 2.5D 패키징 기술

¹⁸ Foveros: Intel의 TSV기반 3D 패키징 기술

¹⁹ SolC (System on Integrated Chips): TSMC의 Hybrid bonding 기반 3D 패키징 기술

(4) 설계 자동화

초기에는 Chiplet 특성을 인식하는 설계 (Chiplet-aware design) 중심으로 발전하여, 이기종 칩 간 인터페이스와 전력 구조를 고려한 시스템 수준의 설계를 가능하게 한다. 이후에는 Chiplet 기반 설계 자동화와 AI 보조 설계(Co-design)가 결합되어, 설계 효율과 최적화 수준이 비약적으로 향상된다. 궁극적으로는 다중 Chiplet을 동적으로 매핑하고 시스템 전체를 최적화하는 자율적 설계 관리 체계로 진화하여, 복잡한 SoC 수준의 설계도 자동화가 가능해질 것으로 예상된다.

(5) 전력 관리

초기에는 Chiplet 간 전력 분산 구조를 중심으로 발전하여, 각 칩 간의 전력 공급 경로와 효율적 배분을 고려한 구조적 설계가 이루어진다. 이후에는 칩렛별 DVFS²⁰ (Dynamic Voltage and Frequency Scaling)이 적용되어, 칩별 부하와 동작 조건에 따라 전압·주파수를 독립적으로 제어할 수 있게 된다. 궁극적으로는 패키지 전체 수준에서 전력을 공유·조율하는 통합 전력 관리 체계로 진화하여, 시스템 단위의 효율과 에너지 활용도가 극대화될 것으로 예상된다.

표 7-8. 칩렛 예상 수치 로드맵

연도	데이터 전송속도	I/O 집적도	Latency
2025	32 Gbps	1.35 TB/s per mm ²	1.5 ns
2029	64 Gbps	5 TB/s per mm ²	1.2 ns
2033	96 Gbps	10 TB/s per mm ²	0.8 ns
2037	128 Gbps	100 TB/s per mm ²	0.6 ns
2040	256 Gbps	500 TB/s per mm ²	0.5 ns

(6) 데이터 전송속도

표 7-8에서 정리한 바와 같이 2025년, 2029년도는 각각 UCle 2.0 및 3.0표준 문서를 기준으로 이미 명시된 지원가능한 데이터 전송속도이며, 2033년에는 패키징 기술 및 신호/전원 무결성 성능 향상으로 달성가능할 것이라 예상된다. 2037년도의 128 Gbps는 고대역폭 패키징 기술 및 높은 집적도를 통해서 100 Gbps 이상의 데이터 전송속도를 달성할 것이라 기대하며 2040년도에는

²⁰ 시스템의 부하나 동작 상태에 따라 전압(Voltage)과 클럭 주파수(Frequency)를 동적으로 조절하여 전력 소모를 최적화하는 기술이다. 이를 통해 성능 요구가 낮을 때는 전력을 절감하고, 부하가 높을 때는 성능을 향상시켜 효율적인 전력-성능 균형을 달성한다.

패키징 기술의 성숙도에 따라 200 Gbps 이상의 데이터 전송 속도를 달성할 수 있을 것이라 예상된다.

(7) I/O 집적도

UCle 1.0 표준 문서 기준으로 1.35 TB/s per mm²로 I/O 집적도가 명시되어 있으므로, 2025년의 I/O 집적도 값은 해당 값으로 표기하였으며, 2029년도는 패키징 기술 발전으로 인해 더 작은 신호 배선 간 거리 달성을 통해 집적도가 약 4배 향상될 것이라 예상된다. 2033년은 광 I/O 도입으로 인해 약간의 집적도 향상이 예상되며, 2037년에는 광/전자 혼합 기술의 성숙도에 따라 달라지겠지만 급격한 집적도 향상이 예상된다. 2040년도에는 패키징, I/O 집적 기술 등의 성숙도에 따라 집적도의 성능이 결정될 것이라 예상된다.

(8) Latency

UCle 1.0 표준 문서 기준 2 ns을 목표로 한다고 명시되어 있으므로, 2025년도의 값은 short reach 기준 equalization까지 고려하여 1.5 ns으로 설정하였다. 다음 연차인 2029년도에는 PHY 경로 최적화 및 SR보다 더 짧은 trace가 고려될 것으로 예상되어 1.2 ns으로 설정하였고 2033년도에는 SerDes-to-Photonics 전환을 통해 latency가 감소하여 0.8 ns으로 명시하였다. 그 이후는 패키징 기술의 성숙도에 따라 결정될 것으로 예상되며 광/전기 I/O간의 변환 시간으로 인해 극단적으로 latency가 줄어들지는 않을 것이라 예상된다.

4. SerDes 로드맵

표 7-9. SerDes 로드맵

연도	2025	2029	2033
Equalization	AI기반 파라미터 튜닝	DSP기반 Equalization AI-assisted 튜닝	Adaptive equalization with ML
신호전송	112 Gbps PAM4	224 Gbps PAM4/6	448 Gbps PAM6/8+FEC
공정 노드	7/5 nm	3/2 nm FinFET→GAA	2nm GAA
I/O 아키텍처	Chiplet과 Die-to-die interface로 최적화 IP제공	DSP기반 Equalizer	고/저속 다중 인터페이스

연도	2037	2040
Equalization	AI-native self-learning 구조 자동 튜닝	AI-native equalization
신호전송	512 Gbps PAM-16, QAM	1 Tbps Coherent modulation (QPSK, 16-QAM)
공정 노드	1.4 nm - sub-1nm GAA	Sub-1nm GAA +Monolithic 3D IC
I/O 아키텍처	DSP +AI equalizer +Optical driver	Photonic-first Electric channel (저속)

(1) Equalization

표 7-9에서 볼 수 있듯이, Equalization 기술은 고속 신호 전송에서 채널 손실과 왜곡을 보상하기 위해 발전해왔으며, 초기에는 AI 기반 파라미터 튜닝과 DSP 기반 보정 알고리즘이 중심이 된다. 이후에는 기계학습(ML)을 활용한 적응형 (Equalization with ML) 구조로 발전하여, 채널 특성에 따라 실시간으로 보정 값을 자동 조정할 수 있을 것이다. 궁극적으로는 AI 네이티브(self-learning) 이퀄라이저가 적용되어, 신호 환경 변화에 따라 스스로 학습하고 최적화되는 자율적 신호 보정 시스템으로 진화할 것으로 예상된다.

(2) 신호전송

신호 전송 (Signaling) 기술은 데이터 전송 속도를 높이기 위해 PAM-4 기반의 다중 진폭 변조에서 시작하여, PAM-6, PAM-8 등으로 확장되며 FEC (Forward Error Correction)와 함께 고속 신호의 신호 무결성을 향상시킨다. 이후에는 PAM²¹-16 및 QAM (Quadrature Amplitude Modulation) 과 같은 고차 변조 방식이 적용되어, 동일 대역폭에서 더 많은 비트를 전송할 수 있을 것이다. 궁극적으로는 1 Tbps급 전송을 지원하는 Coherent Modulation (QPSK, 16-QAM 등) 이 도입되어, 전기적 신호와 광 신호가 융합된 초고속·초대역폭 전송 구조로 진화할 것으로 전망된다.

(3) 공정 노드

반도체 공정 노드 (Process Node)는 미세화에 따라 7/5 nm FinFET 구조에서 시작하여, 3 nm ~ 2 nm 세대의 GAA (Gate-All-Around) 트랜지스터로 전환되며 성능과 전력 효율을 동시에 개선할 것이다. 이후에는 1.4 nm 이하 초미세 GAA 공정과 3D 적층 (Monolithic 3D IC) 기술이 결합되어, 수평적 미세화 한계를 넘어 수직적 집적을 통해 밀도를 높일것으로 예상된다. 궁극적으로는 서브 1 nm급 GAA 기반의 완전 3D 구조로 진화하여, 기존 평면적 스케일링 대신 공간적 통합 (Spatial Integration) 중심의 성능 향상이 이루어질 것으로 예상된다.

(4) I/O 아키텍처

I/O 아키텍처는 초기에는 Chiplet 간 인터페이스 최적화와 Die-to-Die 연결용 IP 제공을 중심으로 발전하며, 고속·저지연 통신을 위한 물리 계층 (PHY) 효율 향상에 초점이 맞춰질 것이다. 이후에는 DSP 기반 Equalizer 및 Optical Driver 기술이 결합되어, 전기적 신호 품질을 개선하고 광 인터페이스와의 연동을 지원하는 하이브리드 구조로 확장될 것으로 예상된다. 궁극적으로는 Photonic-first 아키텍처로 진화하여, 전기·광 통합 채널을 기반으로 초고속, 저전력, 장거리 전송을 구현하는 차세대 I/O 시스템이 구축될 것으로 예상된다.

(5) 데이터 전송속도

2025년도 값은 이미 발표된 PCIe 6.0 표준 문서 기준 값을 사용하였으며, 2029년도 값은 PCIe 7.0 표준 문서에서 확정된 값이다. 마찬가지로, PCIe 8.0 표준에서는 256 GT/s를 목표치로 제시되어서 2033년도의 값으로 표기하였다. PCI-SIG 그룹에서 매년 표준 발표시에 2배로 올리는 경향이 유지된다는 가정하에 2037, 2040 년도 값을 명시하였으며 이는 인터커넥트 기술의 성숙도에 따라 조정될 여지가 있을 것이라 예상된다.

²¹ PAM (Pulse amplitude modulation): 진폭을 이용하여 다수의 비트를 송/수신하는 변조 방식

Coherent modulation: 진폭/위상/편광 정보를 이용하여 다수의 비트를 송/수신하는 광학적 변조 방식

표 7-10. SerDes 예상 수치 로드맵

연도	데이터 전송속도	Energy per bit	Latency
2025	64 GT/s	0.80 pJ/bit	1.8 ns
2029	128 GT/s	0.50 pJ/bit	1.2 ns
2033	256 GT/s	0.35 pJ/bit	0.9 ns
2037	512 GT/s	0.25 pJ/bit	0.7 ns
2040	1024 GT/s	0.20 pJ/bit	0.5 ns

(6) Energy per bit

표 7-10에서 정리한 바와 같이 1 비트당 소모 에너지는 2025년부터 2040년까지 공정, 패키징, 회로, 신호처리 각 영역의 기술적 발전을 통해 지속적으로 개선될 것으로 예상된다. 2025년의 0.80 pJ/bit 수치는 UCIe 및 PCIe 표준 문서에서 제시한 0.5 ~ 1.0 pJ/bit 범위를 기반으로 설정하였다. 2029년의 0.50 pJ/bit 값은 [14]-[15]에서 언급된 바와 같이, 고집적 패키징 기술을 적용하여 데이터 전송 거리를 단축함으로써 달성 가능한 수준으로 보고된다. 여기서의 패키징 조건은 더 짧은 전송 경로와 높은 I/O 밀도를 갖는 advanced packaging 구조를 의미한다. 2033년의 0.35 pJ/bit는 패키징 기술의 성숙과 함께 회로 효율이 향상되더라도, equalization 복잡도의 증가로 인해 전력 절감 폭이 제한될 것으로 예상된다. 2037년에는 광 인터페이스와의 결합을 통해 equalization 회로가 제거되며 소모 전력이 추가로 감소하여 약 0.25 pJ/bit 수준이 가능할 것으로 전망된다. 2040년에는 전자-광 통합 인터페이스의 고도화에 따라 0.20 pJ/bit까지 개선될 것으로 예측된다.

그러나 이 이상의 수준에서는 전자-광 변환 과정이 에너지 소모의 병목으로 작용할 가능성이 높으며, 향후 소자 수준의 혁신이 전체 전력 효율 향상의 핵심 요인이 될 것으로 전망된다.

(7) Latency

UCIe1.0-2.0/PCIe 6.0의 공식 목표가 1.5-2.0 ns이라고 명시되어 있으므로 2025년도는 약 1.8 ns의 latency로 값을 명시하였으며, 2029년도는 패키징의 고집적도를 통해 물리적 길이 감소로 인해 손실 또한 감소하여 equalization 회로가 간소화된다고 가정하면 1.2 ns의 latency가 예상된다. 2033년도의 latency 값은 패키징 기술의 지속적인 발전을 통해 실리콘 칩간 연결 거리가 감소하여 0.9 ns 수준으로 감소할 것으로 전망된다. 2037년도에는 광 인터페이스의 도입으로 인해 equalization 및 DSP 회로가 제거됨에 따라 추가적인 latency 개선이 있을 것으로 예상되어 0.7 ns의 신호 latency가 예상되며, 2040년도에는 패키징/신호처리/회로 전반적인 개선을 통해 신호 전송의 상한 선인 latency가 propagation delay로 수렴할 것이라 예상된다.

5. 결론

반도체 유선 연결은 고속 채널의 기생 성분으로 인해 한정되는 주파수 대역을 이용하여 신뢰성 있는 데이터 송/수신을 달성하기 위해서 여러 기술들이 소개되고 있다. (1) 이중 집적 기술을 통해 칩간의 거리를 줄이거나 신호/전원 배선의 단위면적당 집적도를 향상시켜 고속 시스템의 채널의 주파수 대역을 향상시킨다. (2) 칩렛 기술을 이용하여 반도체 시스템의 소요 비용을 개선시키면서 자유도를 증가시켜 반도체 시스템의 성능 향상에 기여를 하며 (3) SerDes 영역에서는 신호의 요소들을 잘 활용하여 신호를 신뢰도있게 전송하는 방법을 통해 반도체 시스템의 성능 향상에 기여하고 있다. 위와 같은 접근 방식은 계속해서 적용될 것이라 예상된다.

참고문헌

- [1] ISSCC 2016 Press Kit, Figure 1: Per-pin data rate vs Year for a variety of common I/O standards, International Solid-State Circuits Conference (ISSCC) Press Kit, 2016.
- [2] Z. Chen, Y. Zhang, J. Liu, and M. Li, "Challenges and prospects for advanced packaging," in *Materials Today Physics*, vol. 38, p. 101288, 2024.
- [3] K. Salah, A. Henini, N. Tabet, and A. Bouzouraa, "TGV versus TSV: A comparative analysis," in 2016 International Conference on Electrical and Information Technologies (ICEIT), Marrakech, Morocco, May 2016, pp. 613–618.
- [4] M. Yousuf and M. E. Hossain, "Performance analysis of through silicon via (TSV) and through glass via (TGV) for different materials," in *Proceedings of the 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, Cox's Bazar, Bangladesh, Feb. 2019, pp. 1–5.
- [5] OCP ODSA Workgroup, "Bunch of Wires (BoW) PHY Specification, Revision 1.1," Open Compute Project Foundation, Sept. 2022.
- [6] CHIPS Alliance, "Advanced Interface Bus (AIB) Specification, Revision 2.0," CHIPS Alliance, Sept. 2021.
- [7] Universal Chiplet Interconnect Express (UCIe) Consortium, "UCIe™ Specification, Revision 1.1," UCIe Consortium, July 2023.
- [8] PCI Special Interest Group (PCI-SIG), "PCI Express® Base Specification, Revision 6.0," PCI-SIG, Jan. 2022.

-
- [9] IEEE Standard for Ethernet, "IEEE Standard 802.3-2022 (Ethernet)," IEEE Computer Society, Piscataway, NJ, USA, Nov. 2022.
 - [10] S. R. Vangal, A. Keshavarzi, J. Howard, et al., "An Interconnect Fabric for Multi-Die Packages Using Embedded Silicon Bridge," in IEEE Journal of Solid-State Circuits, vol. 53, no. 1, pp. 275–284, Jan. 2018.
 - [11] C.-C. Liu, C.-C. Hsieh, T.-C. Chen, and TSMC Research Team, "TSMC 3D IC and Advanced Packaging Technologies for High-Performance and Low-Power Applications," IEEE International Electron Devices Meeting (IEDM), Washington, DC, USA, Dec. 2011, pp. 511–514.
 - [12] R. Mahajan, R. Sankman, K. Chang, M. Radojcic, and J. Knickerbocker, "Embedded Multi-die Interconnect Bridge (EMIB) and Foveros Technologies for High-Density Integration," in IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA, Dec. 2019, pp. 1961–1964.
 - [13] T.-Y. Chiang, C.-C. Chen, C.-C. Hsieh, et al., "TSMC 3D SolC™ Technology for High-Performance and Energy-Efficient Computing," in IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA, Dec. 2019, pp. 1971–1974.
 - [14] UCle Consortium, "How UCle Will Enable Broad Chiplet Adoption," UCle Express Official Blog, May 2023. [Online]. Available: <https://www.uciexpress.org/post/how-ucie-will-enable-broad-chiplet-adoption>. [Accessed: Oct. 10, 2025].
 - [15] P. Mota, "UCle: An Open Standard Interface for Chiplet-Based Multi-Die Integration," Proceedings of the 2023 Chiplet Summit, San Jose, CA, USA, Jan. 2023. [Online]. Available: https://chipletsummit.com/proceeding_files/a0q5f000001WuE0/20230126_A-201_Mota.PDF. [Accessed: Oct. 10, 2025].

VIII. PIM(Processing-In-Memory) 기술

1. 서론

인공지능의 급격한 발전은 학문의 영역에서의 신기함을 넘어서, 실생활에서의 편리함 그리고 업무 효율성 향상을 가져왔다. 그리고 2025년 현재는, 인공지능의 활용도가 높아지고 다양해짐에 따라 인공지능의 시대가 도래하였다. 이러한 인공지능의 수요증가에 따라 하드웨어 역시 변화하게 되었는데, 특별히 Large language model (LLM)이라고 불리는 거대 인공지능의 등장은 Generative pre-trained transformer (GPT), Copilot, Gemini 와 같은 서비스되는 인공지능을 등장시켰고 이를 위해서 Data center에 들어가는 Server 향 Graphic processing unit (GPU) 나, Mobile 기기에 들어가는 Neural processing unit (NPU) 와 같은 형태의 hardware의 수요를 폭발적으로 증가시켰다. 이러한 neural network 와 hardware의 관계는 서로가 서로를 발전시키며 일상에 빠르게 접목되고 있으며 인공지능의 연산을 위한 패러다임 전환의 hardware 구조로 Processing-in-memory (PIM)의 구조가 제안되어 왔다. 본 로드맵에서는 PIM의 발전 방향성에 대해 다루도록 한다.

2. 인공지능 하드웨어 관련 최신 기술 동향

1) Workload: Inference mode for Transformer

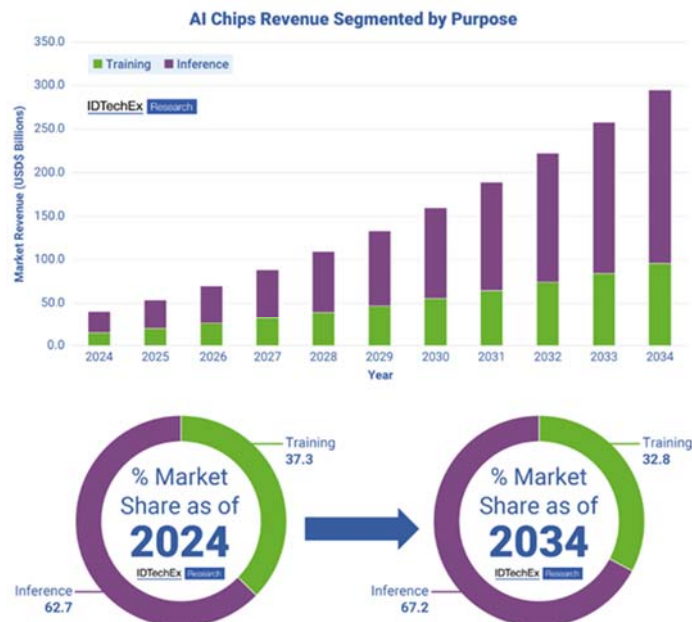


그림 8-1. Inference hardware 수요 증가^[1]

LLM의 발전 및 이를 토대로 하는 Service의 증가는 Inference hardware의 수요를 폭발적으로 증가시켰다(그림 8-1). 특별히 이러한 Service 중심의 LLM 같은 경우 Transformer 로 이루어져 있으며 답변을 생성하기 위한 생성형 인공지능이기 때문에 Decoder 와 같은 형태로 이루어져 있다. 이러한 Decoder의 특징은 현재 생성되는 output이 다시 input 으로 전달되는 특징을 가지고 있는데, 결국 문장에서 단어가 하나씩 추가되는 형태이다(Next token prediction). 따라서 이러한 연산은 매번 모든 문장을 input으로 새롭게 Transformer 가 연산하는 것이 아니라, 새롭게 추가된 단어에 대해서만 Transformer 연산을 하는 형태의 연산을 통해 Hardware의 연산량을 줄일 수 있다.

구체적으로는 Transformer에서 이전 Token 들에 대한 Key와 Value 값을 hardware의 Cache에 저장해 둔 다음 (KV Cache), 새로운 Token에 대해서 연산이 수행될 때 KV Cache 를 불러와 연산하게 되면 추가된 Token에 대한 연산만을 수행하게 되기 때문에 연산량을 크게 줄일 수 있는 원리이다. 이때 추가된 Token 은 Vector 이므로, 추가된 Token에 대한 연산은 Vector와 Matrix 곱으로 나타낼 수 있으며, 일반적으로 GEMV (General matrix vector multiplication) 연산에 해당한다. 특별히 GEMV 와 같은 경우 일반 선형대수 연산 (Basic linear algebra subprograms, BLAS)의 GEMM (General matrix-matrix multiplication) 연산에 비해 연산량이 적게 된다. 따라서 GEMV의 연산에 있어 연산 속도의 제한을 거는 것은 연산기의 동작속도이기 보다는 연산에 필요한 Data를 연산기로 읽어오거나 연산한 Data를 메모리에 저장하는 Memory와의 통신 속도가 된다. 즉, GEMV 같은 경우에는 Memory bandwidth 가 일반적으로 연산 속도에 제한이 되며, 이를 극복하기 위한 방안이 필요하다.

Challenges: The End of Moore's Law and Scaling

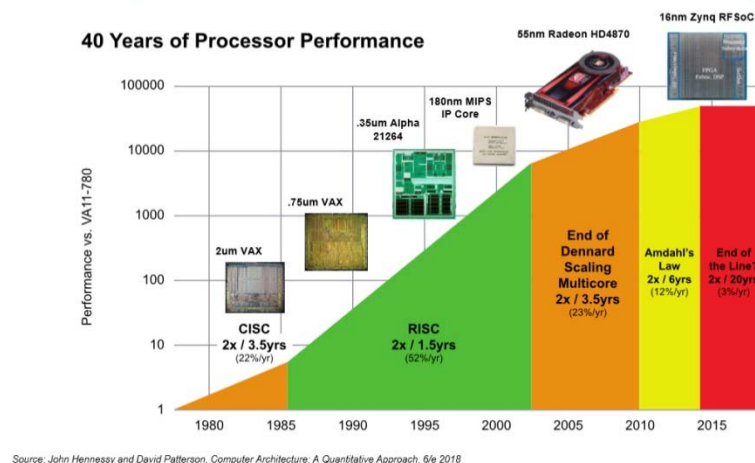


그림 8-2. Moore's 법칙 둔화^[2]

2) Computer Architecture: From Von-Neumann to Non-von-Neumann

이러한 Workload의 변화는 GPU와 같은 하드웨어의 발전을 통해 이뤄진 면이 있지만, 발전된 Workload의 변화는 GPU와 같은 형태의 하드웨어의 수요를 크게 증가시키고, 하드웨어의 더 높은 성능을 요구하였다. 즉, 하드웨어의 발전을 통해 발전한 소프트웨어가 특정 하드웨어의 수요를 증가시켜 하드웨어 구조 변화와 성능향상을 이끌게 되었다.

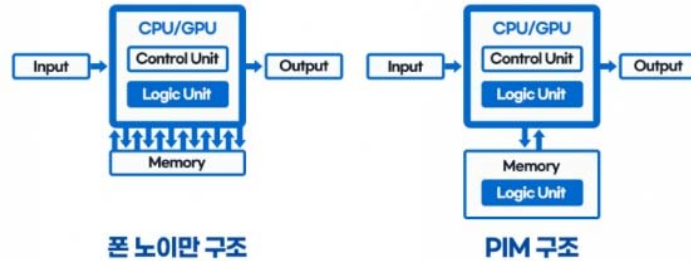
하지만 하드웨어의 발전은 소프트웨어의 폭발적인 발전과 다른 양상을 띄고 있었는데, 이는 Transistor의 scaling 이 한계에 봉착하였기 때문이다. 무어의 법칙(Moore's law)라고 불리는 Transistor 가 작아지는 경향과 데나드 스케일링 (Dennard's scaling) 이라고 불리는 컴퓨터 성능의 향상 경향은 이미 둔화되거나 더 이상 작동하지 못하는 상황에 이르렀다(Figure 2). 그러다 보니 소프트웨어의 발전을 따라가기 어렵게 되었고 하드웨어의 새로운 혁신이 필요하게 되었다.

Transistor 자체의 발전이 둔화된 상황에서 시도되는 혁신은 컴퓨터 구조의 변화였다. 인간의 모든 논리의 연산이 가능한 Turing machine 으로부터 Turing machine의 가장 효율적인 구조라고 여겨지는 Von-Neumann 구조가 지금까지의 컴퓨터 구조의 철학이었는데, 새로 등장한 중요한 Workload로 요구되는 인공지능은 그 철학과 다른 부분이 많았기 때문이다. 인간의 뇌로부터 구조가 시작된 Neural network 가 인공지능의 중요한 block 이 되고, neural network의 거대함이 성능에 중요한 지표가 된 Workload의 특성은 복잡한 논리로 이루어진 세계에서 단순한 논리의 수많은 반복으로 workload의 큰 변화를 가져왔다. 따라서 아주 복잡한 논리를 연산할 수 있는 구조에서 단순한 논리의 빠른 처리를 요구하는 하드웨어의 설계방향이 바뀌게 되었다.

이러한 상황에서 Von-Neumann 구조에서 Non-von-Neumann 구조로의 변화는 자연스러운 시도이 되었고, Data-centric 한 특성과 맞물리어 Von-Neumann 구조로부터 벗어나려는 시도는 계속 되어왔다. 특별히 Data-centric 한 인공지능의 특성은 기존의 Von-Neumann 구조에서 연산기와 memory 사이의 통신 속도인 memory bandwidth 가 전체 시스템 성능의 병목현상을 가져왔다. 즉, 전체 시스템의 연산 속도가 연산기 자체의 속도가 아닌 저장장치의 읽고 쓰는 속도에 한계가 걸리게 되었다. 게다가, 연산기와 저장장치의 분리는 통신속도의 제한 뿐만 아니라 큰 전력소모를 야기하게 되었다. 즉 Von-Neumann의 처리 방식의 많은 데이터 이동은 전체 연산 속도의 제한과 많은 전력소모를 야기하게 되었다.

[PIM 기술 개념도]

PIM(Processing-in-Memory)은 메모리 내부에 연산 작업에 필요한 프로세서 기능을 더한 차세대 신개념 융합기술



* Control Unit : 제어부 * Logic Unit : 연산부

그림 8-3. PIM 기술 개념도^[3]

이를 극복하려는 시도로 PIM의 구조가 등장하게 되었는데(그림 8-3), PIM의 연산 가능한 Workload를 제한하지만 특정 연산에 높은 성능을 보이는 형태의 Non-von-Neumann 구조를 기초로 한다. 특별히 AI의 구조상 앞서서 말한 GEMV 연산이 현재로는 그 목표가 되는 것이 타당한데, 이는 현대 하드웨어 변화가 Workload의 변화에 뿌리를 두고 있기 때문이다. 따라서 앞선 내용의 PIM의 등장 배경을 요약하면 아래와 같다.

(1) 단순 연산 반복에 적합한 Workload의 변화에 따른, 단순 연산 특화 하드웨어 요구

(2) Data 이동 최소화를 통한 전력 효율, 연산 속도 향상 가능성

이러한 배경을 기초로 한 PIM 기술은 현재 연산기와 저장장치를 물리적으로 구분하였던 Von-Neumann 구조에서 벗어나서 연산기와 저장장치의 통합을 말한다. 즉 메모리 안에서의 연산 (Processing-in-memory) 라는 이름 그대로, 저장장치 내부에서 연산을 하는 것을 말하며, 이를 통해 저장장치와 연산기의 통합을 만들어 Data 이동의 최소화 및 연산 속도의 향상을 가져올 수 있다. 또한 저장장치 내부에 연산기를 집어넣기 때문에 복잡한 연산을 지원하기 어렵지만, 단순 연산에 특화될 수 있다. 이러한 PIM 연산에 대해 다음 장에서 자세히 다루도록 한다.

3. PIM 기술 로드맵

PIM 기술은 단순히 보기에는 메모리의 연산기를 합치는 것 같아 보이지만, 실제로는 고려할 것이 매우 많다. 먼저는 저장장치에 해당하는 memory 가 계층구조로 되어있으며, 그 계층구조가 생기게 된 합당한 이유가 있기 때문이다. 따라서 계층에 따른 memory의 특성이 다르며, 각 memory에 연산기를 넣는 것이 합당한지 다시 살펴보아야 한다. 또한 이러한 memory 자체의 특성을 고려할 때 다양한 형태의 연산기 접합 형태가 있을 수 있게 된다. 이러한 형태를 명확하게 설명하기 위해 분류가 필요하다.

또한, PIM의 등장 배경에서 근본적으로 중요한 점은, Workload의 변화가 하드웨어의 변화를 이끌었다는 점이다. 문제는 Workload 가 소프트웨어를 의미한다는 것인데, 구현 자체가 쉬운 소프트웨어의 특성 상 workload의 변화는 매우 쉽고 빠르게 일어나며 Neural network 분야의 발전만 보더라도 그렇다. 따라서 본 장에서는 먼저 1) memory의 종류와 특성을 먼저 살펴보고, 2) PIM 기술을 정의 및 분류한 다음, 3) Workload인 AI의 발전동향을 살펴볼 것이다. 그 이후에 살펴본 지식을 바탕으로 4)로드맵을 설명하려 한다.

1) Memory의 종류와 특성

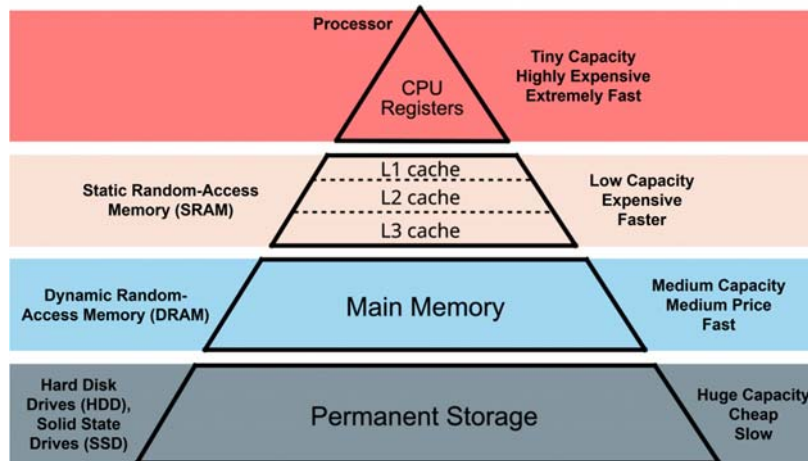


그림 8-4. Memory hierarchy ^[4]

그림 8-4는 일반적인 컴퓨터의 메모리 계층도를 나타낸다. 잘 알려져 있듯, Von-Neumann 구조를 기반으로 연산기에 해당하는 Central processing unit (CPU)가 가장 상위에 위치해 있고, CPU에 가까운 순으로 Register, Static random access memory (SRAM), dynamic RAM (DRAM), Storage 역할을 하는 Hard disk drive (HDD), Solid-state drive (SSD) 가 위치해 있다. 이러한 구조도는

Processor의 memory의 효율적인 관리를 위해 생겨났는데, 중요한 것은 이러한 관리 구조가 각 memory의 물리적 구현에까지 영향을 미친다는 것이다. 즉, Register와 SRAM 같은 경우는 CPU와 같은 die 로 만들어지지만 DRAM과 SSD는 일반적으로 각각의 다른 die로 만들어져 Board 상에서 연결된다. 또한 물리적인 위치 차이와 함께, 각각의 생산비용 감소를 위해 SRAM, DRAM, SSD 가 다른 공정을 통해 만들어지고 각 공정의 특성이 모두 다른 특성을 가지고 있다.

이러한 환경 속에 PIM 기술은 어려움을 겪게 되는데, 단순한 저장부와 연산부의 통합이라는 명제는 어떤 저장부와의 통합할 것인지를 어렵게 한다. 앞서 설명한 대로, **데이터 이동의 최소화**라는 목적이라면 큰 데이터를 저장하는 저장부에 연산기를 위치하는 것이 알맞을 것이지만 위 구조도에서 보듯, 큰 데이터를 저장하는 DRAM과 SSD 같은 경우 CPU와 다른 chip으로 구현되기 때문에 연산에 대한 고려가 없는 공정으로 구현되어 있으며, 따라서 일반적으로 Transistor의 성능이 CPU에 비해 좋지 못하고 이로 인한 연산 성능이 떨어지게 된다. 따라서 저장부에 위치시킨 연산기의 성능이 좋지 못하게 되고, 이 때문에 memory bandwidth를 해결하려 했던 목적에서 벗어나, 연산기 자체 효율의 문제에 부딪히게 되는 어려움을 겪게 된다. 따라서 이러한 상황을 명확히 하기 위해, 본 로드맵에서는 다음 장에서처럼 PIM 기술에 대한 용어를 정의하도록 한다.

2) PIM 기술의 정의 및 분류

위와 같은 논의에 따라, 기술적 차이를 명확히 하기 위해 PIM 기술에 대한 세부적인 용어를 아래와 같이 정의한다.

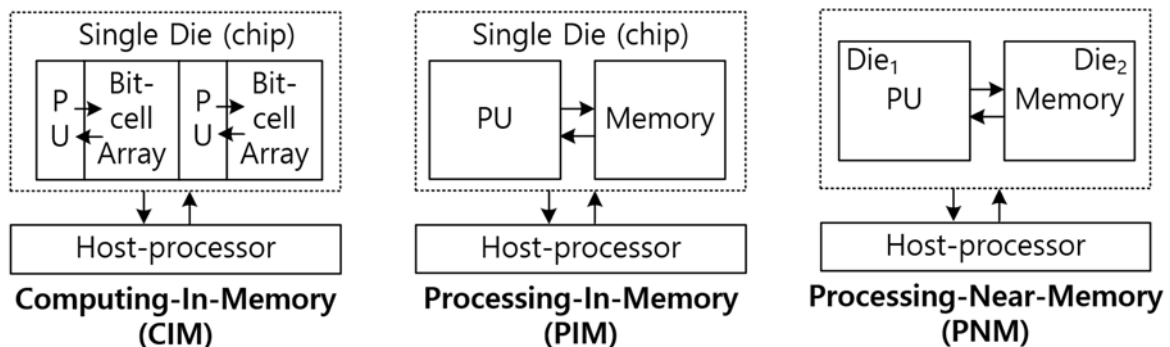


그림 8-5. PIM 기술의 정의 및 분류

(1) Compute-in-memory (CIM)

Memory와 연산기가 같은 Die에 위치하면서도, **Memory array 자체**를 연산기로 활용하는 경우.

대표적으로는 (a) Analog computation을 기반으로 하는 기술: SRAM의 구조에서 Capacitor를 주

가하는 방식^{[5]-[6]}이나, 저항성 메모리 (Resistive-RAM (RRAM), Magneto-resistive RAM (MRAM) 등)의 bit-cell의 저항을 이용해 연산하는 방식^{[7]-[9]}. 또는 (b) Digital computation을 기반으로 하는 기술: SRAM의 구조에서 Bit-cell array들 사이에 Logic을 첨가하여 연산하는 방식^{[10]-[11]}이 있다.

(2) Process-in-memory (PIM)

Memory와 연산기가 같은 Die에 위치하면서도, Memory array의 변경없이 가까운 곳에 연산기를 위치시켜 연산하는 경우.

대표적으로는 DRAM 기반으로 DRAM의 Bit-cell array에 Processing unit을 위치시켜 연산하는 경우가 있으며, High-bandwidth memory (HBM)을 활용한 HBM-PIM^[12], AiM^[13], UPMEM-PIM^[14] 등이 있다.

(3) Process-near-memory (PNM)

연산기가 Host-processor와 Memory와 다른 Die에 위치하지만 Memory die와 같은 Package에 있거나, memory의 controller에 위치하는 경우.

대표적으로는 Compute express link(CXL)를 활용한 CXL-DRAM 시스템에서 CXL controller 안에 연산 logic을 넣거나, Multi-chip packaging에서 memory-die와 processing die를 넣어서 packaging 된 형태가 있다.

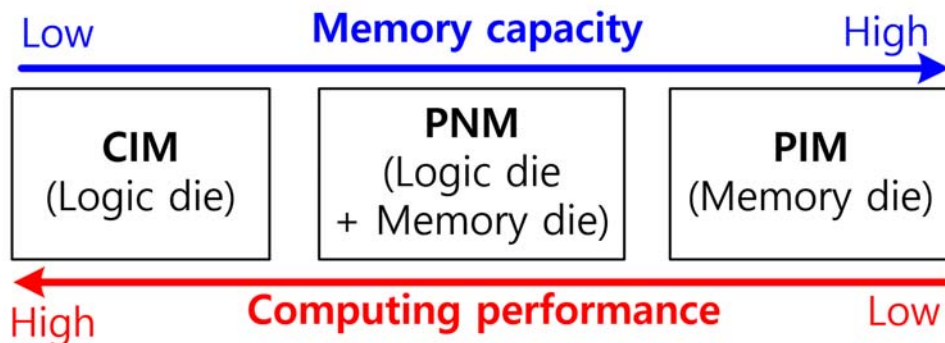


그림 8-6. PIM 기술 분류에 따른 각 기술의 특성

이러한 PIM 기술의 분류는 그림 8-4에서 나타낸 Memory 구조도와 관련이 깊다. 그 이유는 memory의 종류에 따라 공정이 달라지고, 구현될 수 있는 연산기의 특성 또한 달라지기 때문이다. 예를 들어 CIM 같은 경우 Memory를 자체를 연산기로 활용하기 때문에 SRAM과 같은 Logic 연

산에서 연산기가 구현될 수 있다. 이 경우 Logic 공정의 특성이 연산기에 적용될 수 있기 때문에 빠른 연산을 지원할 수 있게 된다. 일반적으로 Logic 공정의 경우 Transistor의 성능이 좋고, 사용할 수 있는 Metal layer의 개수가 많아 구현의 자유도가 높다. 따라서 CIM 같은 경우 연산기 구현에 알맞은 기술이다. 하지만 메모리의 입장에서 보기에 큰 비용을 치러야 하는데, 우선 SRAM 자체가 용량이 작고 Logic 공정이 비싼 편이기 때문에 생산 비용이 많이 들게 된다. 따라서, **CIM 같은 경우 연산면에서는 훌륭하지만 데이터의 양 면에서는 좋지 못하다.**

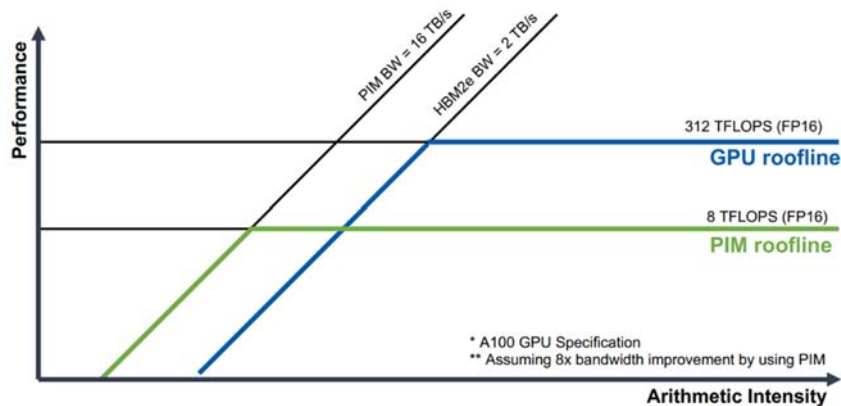


그림 8-7. Roofline of GPU and PIM^[15]

PIM 기술의 경우는 CIM과는 반대의 경향을 보인다. DRAM의 경우 생산 단가를 낮추기 위해 일반적으로 Logic 공정에 비해 낮은 성능의 Transistor와 적은 수의 Metal layer를 가진다. 이러한 특성은 SRAM에 비해 큰 메모리 용량을 제공하지만, 연산기를 구현하는 면에서는 CIM에 비해 불리한 면을 가지고 있다. 따라서 매우 단순한 연산은 PIM 기술을 사용할 수 있지만, 연산이 복잡해지거나, 빠른 연산이 필요한 경우에는 PIM 기술은 연산기의 성능 제한으로 인해 제약될 수 있다. 따라서 **PIM 같은 경우 데이터의 이동을 줄인다는 면에서는 큰 장점을 갖지만 연산면에서 좋지 못하다.** 이러한 특성은 그림 8-7과 같은 성능 전체 시스템 성능 곡선을 야기한다. 즉, PIM이 가지는 Memory band-width의 장점이 부각되는 Arithmetic intensity 가 낮은 경우에는 DRAM 내에서의 연산이 많아져 PIM의 roofline (초록색)이 더 좋은 성능을 보이지만, Arithmetic intensity 가 높은 경우에는 Logic die의 활용성이 부각되는 GPU roofline 이 더 높은 형태를 보이게 된다. 이러한 경향은 2장의 1)에서 설명한 GEMV 연산에 PIM이 장점을 나타낼 수 있는 경향으로 나타난다.

PNM 같은 경우는 CIM과 PIM의 중간에 위치한다. 연산기에 해당하는 부분과 Memory에 해당하는 부분을 다른 Die에 위치시킴으로 각각에 유리한 특성을 모두 가질 수 있다. 하지만, 이러한 구현의 경우 다른 Die끼리 어떻게 빠르게 데이터를 통신시킬 수 있는가의 문제를 가지게 된다. 즉, Host-processor와 Memory도 다른 Die이고, 연산기와 Memory도 다른 Die 일 때, Host-processor에서 Memory를 접근하는 것보다 연산기에서 Memory를 접근시키는 것이 얼마나 큰 장점을 가지는가가 PNM의 장점 요소가 된다. 따라서, PNM 기술 같은 경우 연산기와 Memory를 얼마나 물리적으로 가깝게 (빠르고 저전력으로) 위치시키느냐가 기술의 핵심이 될 수 있으며, packaging의 형태나 CXL과 같은 기술을 응용할 때 그 장점이 극대화될 수 있다.

3) AI Algorithm과 가속기의 Memory 기술 동향

(1) Multi-modal model과 Multi-head latent attention^[16]

Vision transformer의 등장은 자연어 처리의 대표적 model이었던 transformer를 image classification의 영역으로 확장시킬 수 있게 하였다. Transformer의 이러한 다양한 분야로의 확장은 실제로 Multi-modal model로의 발전으로 확장되었고, 이는 서비스되는 LLM이 할 수 있는 역할의 범주를 넓히게 되었다. 그러나 이러한 LLM의 다양한 쓰임새와는 별도로 LLM의 크기를 감당하기 위한 hardware의 부담은 줄지 않았는데, 이를 극복하기 위해 제안된 model 구조의 변화가 중국의 Deep-seek로 알려진 Multi-head latent attention (MLA) 이다^[16]. MLA 같은 경우 기존의 Transformer에서 Key와 Value 값을 작은 fully connected layer를 사용해 작은 차원의 행렬로 압축시켜 사용하는 기술인데, 이를 통해 hardware 연산에 필요한 행렬 크기를 줄임으로 하드웨어의 부담을 줄였다. 즉, 작은 차원에서의 행렬 곱셈을 추가하여 연산량이 줄어드는 경우이다. 이러한 큰 행렬 연산을 줄이기 위한 노력은 Software에서도 이뤄지며, 이에 따라 Workload는 시간이 지나면서 계속 변화할 가능성이 크다. 따라서, 다양한 크기의 행렬 곱에 대한 지원은 하드웨어에 필수적인 요소이다.

(2) 4세대 HBM



그림 8-8. HBM4^[18]

최근 Samsung과 SK Hynix와 같은 메모리 업체가 수요가 급증한 HBM의 새로운 버전을 발표하였다(그림 8-9 참조). 주목해야 할 것은 기존의 Through silicon via (TSV) 기법을 활용한 적층 memory 구조와 함께 가장 아래에 놓여있는 Base die라고 불리는 chip을 Logic 공정을 이용해 생산한다는 것이었는데^[17], 이를 통해 Memory와 Host-processor의 통신 속도를 높이거나 Supply를 낮춰 전력소모를 줄이는 형태의 기술을 선보인다는 것이었다. 이러한 Memory 내부에 Logic die의 추가는 일반 Logic 공정을 사용하는 Host-processor를 제작하는 Fabless 업체 (Nvidia) 에게도 큰 기회가 될 수 있었으며, 실제로 뉴스를 통해 Nvidia와 메모리 업체들과 협력을 검토하는 기사가 발표되었다^[19].

PIM 기술에 분류에서 언급하였듯, DRAM과 Logic은 공정면에서 추구하는 바가 다르게 된다. 이러한 차이를 인정하고, 하나의 공정에서 모든 문제를 해결하기 보다 Packaging 기술을 통해 문제를 해결하려는 시도는 PIM 기술에 있어서도 큰 의미를 가진다. 이러한 기술은 PNM 과 같은 형태로 PIM 기술에 적용될 수 있으며 PIM의 발전방향에 큰 의미가 있을 것으로 생각된다.

(3) SRAM Scaling의 둔화와 Design technology co-optimization (DTCO)

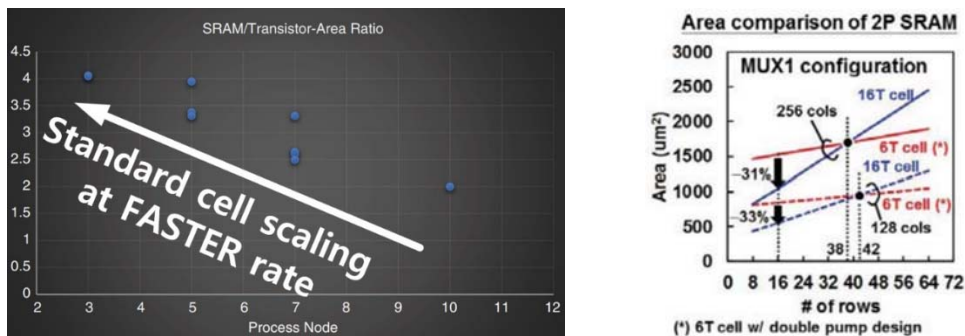


그림 8-9. SRAM Scaling의 둔화^{[20], [21]}

Logic 공정이 미세해짐에 따라 Transistor의 크기는 계속 줄어들었는데, 이러한 scaling 은 Logic 공정 내에서 SRAM 역시도 작게 만들었다. 하지만 집적도를 위해 Logic 공정 내에서도 다른 Design rule check (DRC)를 적용받는 SRAM의 경우 일반 Logic에 비해 scaling 이 둔화되었는데 특별히 다른 DRC를 적용받기 때문에 일반 Logic 과 어느 정도의 거리를 띄워줘야 하는 문제는 SRAM의 집적도를 낮게 만들었다^[21]. 그림 8-9에 나와있듯, 이러한 문제에 의해 SRAM의 전체 용량이 작을 때 Logic과 같은 DRC를 적용 받는 18Transistor (일반적으로 12 transistor 형태)가 Logic과 다른 DRC를 적용 받는 일반 6T SRAM 보다 더 면적밀도가 좋은 구간이 나타났다. 이는 Transistor의 크기 매우 작아 짐에 따라 DRC를 위해 띄운 거리 내에 많은 Transistor 가 집적될 수

있기 때문이며, 이를 통해 고속으로 동작하는 12T SRAM 등이 미세공정으로 갈수록 효율성이 좋아지게 되었다.

한편 DTCO 기술을 통해 특정 IP의 성능을 끌어올리는 시도도 연구되고 있는데^{[22], [23]}, DTCO 같은 경우 memory와 같이 사용처가 매우 많고 구조가 고정된 형태의 IP에서 큰 효과를 발휘한다. 실제로 CIM-macro의 SRAM을 Foundry optimization을 통해 Area를 30% 줄일 수 있었으며, Memory를 활용하는 PIM 기술의 경우 DTCO의 활용에 적합한 구조로 생각된다^[22].

(4) Training processor로의 발전 가능성: Zero-infinity 기술의 의의

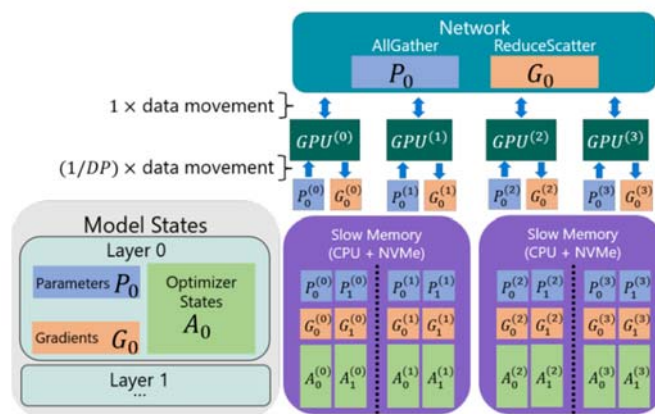


그림 8-10. Zero infinity 기술^[25]

최근 Microsoft로부터 ZeRo^[24], ZeRo-infinity^[25]와 같은 학습 기법이 발표되었다. 학습에는 Optimizer라는 특정한 function이 필요로 하는데, 이 Optimizer가 Neural network를 학습할 때 State라고 하는 큰 Data의 저장이 필요로 한다. 이 때문에 Neural network의 학습을 위해 GPU에 매우 큰 memory가 필요 된다. 이를 극복하기 위해 제안된 ZeRo-infinity 기술은 DRAM뿐만 아니라 SSD역시 memory로 사용하는 기술을 제안하였다. 이는 엄청나게 큰 LLM을 여러 개의 GPU에 어떤 식으로 나눌 것인지에 대한 내용을 넘어, 각 GPU가 가지는 Memory 크기에 대한 한계를 어떻게 뛰어 넘을 것인지에 대한 구체적인 방법을 제시하였다.

특히 DRAM을 넘어 SSD까지 활용하는 형태의 학습방법은 GPU가 가지는 메모리(DRAM)에 학습 모델의 크기가 제한되지 않고, SSD를 활용하여 적은 수의 GPU를 가지고도 큰 LLM을 학습시키는 것이 가능 해졌음을 의미한다. 우리가 이를 통해 알 수 있는 것은, 학습에 있어서도 인공지능이 연산부가 아닌 Memory의 중요성이 높다는 것이다. 특히 inference가 아닌 Training이

갖는 의미는 Foundation model 구축이 있는데, 이는 Model을 통한 독자적인 생태계를 구성할 수 있음을 의미한다. 즉, 이미 만들어진 model을 연산만 하는 inference와는 달리, 새로운 model을 만들 수 있다는 것은 기술력 뿐 아니라 정보의 보안과 시장의 독점성과 같은 경제적으로 중요한 가치를 가진다. 따라서 PIM의 로드맵에서는 Model training 역시 target application으로 잡는 것이 기술적으로 보나 경제적으로 보나 바람직하다.

4) 로드맵

앞선 논의에 따라 본 장에서는 PIM 기술의 로드맵을 제시한다. PIM 기술 같은 경우 Workload의 변화에 그 뿌리를 둔 hardware의 변화이므로 현재는 Hardware의 변화가 Workload의 변화에 종속되는 경향을 갖는다. 하지만 이것은 반도체의 생산을 경제적인 가치로 여기는 분야에 적합하지 못하다. 따라서 **본 로드맵의 최종 목표는 Workload를 Hardware에 종속되게 만드는 것에** 있으며, neural network의 특성상 근본적이면서도 데이터 양에 직접적으로 연관되어 있는 행렬 곱셈 (GEMV, GEMM) 을 목표로 잡는 것이 바람직하다. 즉, **PIM 로드맵의 최종 목표는 hardware 적인 기술을 통해 빠르게 곱셈하려는 행렬의 크기를 쉽게 증가시킬 수 있게 하는 것에** 있으며 이러한 hardware 기술을 통해 앞으로 새롭게 생성되는 Neural network (또는 Neural network의 방향성) 가 아주 큰 행렬의 곱셈에 종속적이게 만드는 것에 있다. 이러한 최종 목표를 가지고 앞으로 약 9년 동안(~2034년)에는 현재 생성되어 있는 GPU 생태계에서의 훌륭한 부품의 역할을 그리고 그 이후 6년(~2040년)에서는 GPU를 뛰어 넘는 핵심적인 hardware의 역할을 하는 것이 로드맵의 내용이다.



그림 8-11. PNM 구조와 Roof-line

먼저 로드맵으로 제안하는 구조의 기술을 Figure 11의 왼쪽에 나타냈다. 왼쪽과 같이 제안된 이유는 PIM 기술의 극대화를 위해서는 data 이동이 적어야 하며 이를 위해 메모리 계층 구조에서 상대적으로 data가 큰 DRAM에 연산기가 위치하는 것이 알맞다고 판단하였다. 또한 우리나라의

기술이 DRAM에 강점을 보이는 만큼 DRAM에 연산기가 붙는 것은 합리적일 것이라 판단하였다. 하지만 제조 원가 절감과 같은 공정의 고려 사항이 있기 때문에 일반 Logic 공정과 같은 높은 수준의 연산기를 DRAM에 배치시키는 것은 어렵다. 따라서 현재 개발되고 있는 4세대 HBM 기술과 발전 방향을 같이 하여, 연산기를 DRAM die가 아닌 Base die에 배치하는 것이 합리적일 것으로 판단하였다. (그러나 TSV의 bandwidth를 고려하여 아주 작은 연산을 DRAM die에 넣는 PIM 기술이 고려될 수 있다). 따라서 현재 연산기와 저장부가 다른 Die에 위치하기 때문에 **PNM 구조**로 보는 것이 합리적인 발전 방향으로 생각된다.

그러나 앞서 3-3)-(1)에서 논의한 것처럼 Software 역시도 행렬 곱 연산을 줄이기 위한 방향으로 연구되고 있으므로 다양한 행렬 크기에 대해 연산 효율을 낼 수 있는 Scalable 한 구조가 필요하다. 이를 위해 **Base-die를 통한 PNM Cube 간의 연결을 제안**하고, 이를 Base-die 를 통해 이뤄지는 구조를 제안한다. 이러한 **Scalability와 Bank/row/column Parallelism의 극대화를 위해 Base-die에 SRAM을 두고 CIM 구조를 활용**하는 것이 타당할 것으로 생각된다. 특별히 CIM 같은 경우 DTCO를 통해 면적/전력 효율이 극대화될 수 있으며 PNM을 위하지 않더라도 일반 Host-processor에서 Foundry 제공의 Macro 형태로 사업이 가능할 수 있다. 마지막으로 이러한 기술을 통해 그림 8-11의 오른쪽 그림처럼 기존 GPU보다 memory bandwidth도 높고 연산속도도 빠른 Roofline을 달성하는 것이 로드맵의 최종 목표이다. 이러한 구조적인 논의와 함께 작성한 2040년까지의 로드맵 표는 아래와 같다

표 8-1의 가장 왼쪽은 현재 Silicon으로 proof된 Samsung의 HBM-PIM^[12]을 2025년의 기술로 참고하였다. 본 PIM 기술의 최종 목표는 PIM을 이용한 독자적인 인공지능 생태계 조성이기 때문에, 최종적으로는 Inference 와 training까지 application을 목표로 하였으며, GPU의 대체를 목표로 하기 위해 Memory의 표준인 JEDEC 이 아닌 Host-processor의 PCI를 host-connection으로 설정하였다. 특별히 현재로부터 Base-die 가 일반 Logic 공정으로 바뀔 수 있기 때문에 연산기의 Clock을 크게 높였으며 (0.3GHz → 8GHz), 이에 따른 Throughput과 bandwidth 도 높여서 제시하였다. 마지막으로 발전되는 Packaging 기술을 고려하여 Memory density를 높여서 설정하였다.

표 8-1. PIM 기술 로드맵

	현재 ^[12]	~2028	~2031	~2034	~2037	~2040
Application	Inference	Inference	Inference	Inference / Training	Inference / Training	Inference / Training
Precision	FP16	FP16	FP32	FP64	FP64	FP64
Target 연산	GEMV	GEMV	GEMV	GEMM	GEMM + Activation	GEMM + Activation
Host-connection	JEDEC 표준 command	JEDEC 표준 Command	JEDEC 표준 command	JEDEC 표준 Command	PCI	PCI
Memory Density	6GB/cube	10GB/cube	12GB/cube	15GB/cube	20GB/cube	25GB/cube
Scalability	No	No	Yes	Yes	Yes	Yes
Processing Operation speed	0.3GHz	1GHz	2GHz	4GHz	6GHz	8GHz
Peak Throughput [TFLOPS/cube]	1.2	5	10	20	35	50
Bandwidth [TB/s/cube]	0.3	1	2	4	6	8

4. 해결해야 할 기술

(1) CIM-PIM TSV connection 기술

먼저 CIM-PIM을 TSV로 연결하는 기술의 발전이 필요하다. 물론, 현재 HBM4 를 통해서 알 수 있듯 상용화가 될 만큼 충분히 성숙된 기술로 보이지만, Logic die와 DRAM die가 서로 다르기 때문에 이 두개의 다른 종류의 칩을 연결할 때 얼마나 작은 Impedance로 연결될 수 있는지가 중요할 것이다. 특별히 제안하는 기술도 HBM처럼 빠른 속도의 통신을 요구하고 Data 이동량이 많아 전력소모가 클 것으로 예상되기 때문에 TSV 기술의 발전은 PIM 기술의 발전에 매우 중요할 것으로 생각된다.

(2) CIM-PIM 구조의 역할 분담과 발열 문제

현재 메모리의 구조를 제시하고 로드맵을 제안하였지만, Base-die (CIM)과 DRAM-die (PIM)의 구체적인 역할 분담이 명확하지 않다. DRAM-die에 연산기를 넣는 것이 효율적인지, 아니면 연산을 CIM에 모두 맡기고 큰 DRAM memory로 사용하는 것인지도 연구가 필요하며, Base-die의 CIM 역

시 Buffer 역할을 하는 SRAM의 크기와 연산속도에 대한 연구가 필요하다. 각 회로기술과 더불어 각 Die 가 어떤 형태의 역할을 하는 것이 미래 지향적이면서도 효과적인지 연구가 필요하다.

또한 발열 문제를 빼놓을 수 없는데, Scalability 를 위한 옆 Cube와의 통신과 Host-processor와의 통신을 위해 Base-die를 가장 아래에 두었기 때문에 연산기를 Base-die로써 가장 아래에 두었을 때 생기는 발열은 큰 문제가 될 수 있다. 따라서 이를 위해 일반적인 HBM 위에 CIM-processor 를 가장 위층으로 올리는 형태와 같은 발열에 대한 고려도 해결해야할 기술이다.

(3) Host-processor의 software stack 고려한 운영 방식

Hardware 자체의 효율 대한 연구와는 별도로, 실제 메모리를 사용하는 Host-processor 가 운용되는 Software 를 고려하는 것을 필수적이다. 중요한 것은 메모리에 특정 연산을 집어넣어 Host-processor와는 독립적으로 내부적인 함수를 정의하였다면, 이를 사용하는 Host-processor의 Software 와의 결합에 대한 연구/고려는 매우 중요하다. 현재 CPU/GPU 와 같은 processor의 software는 견고하게 구축이 되어있기 때문에 이를 변화시키는 것은 쉬운 일이 아니며 그렇기에 이미 구축된 Software의 변화 필요성에 대한 분명한 Hardware의 장점이 확보되어야 한다. 게다가, 단순한 DRAM 이 아닌 Scalable한 memory 를 Software에서 사용가능한지에 대한 논의는 반드시 필요하다. 따라서 Hardware 자체내의 발전 뿐 아니라, Software 역시 수용가능한 변화를 가정해야 하며 이에 대한 연구/고려가 매우 필요하다.

(4) Base-die끼리의 저전력 초고속 통신 기술

더 높은 memory bandwidth의 요구가 PIM 기술을 연구하게 만들었듯, 현재 상용화 된 Host-processor 와 DRAM의 통신 속도는 계속 연구되고 발전하고 있다. 제안된 로드맵 구조에서는 Host-processor 와 DRAM 사이의 통신 이외에도 Scalability 확보를 위한 PIM-Cube 사이에도 통신을 가정한다. 따라서 통신을 위한 회로가 더 많이 필요하며 이에 따라 저비용의 통신회로가 필요할 것으로 보인다.

5. 요약

PIM 기술은 구체적으로 CIM, PIM, PNM 으로 분류할 수 있다. 이러한 분류에 따르면 CIM은 저장부보다는 연산부에 특화되어 있고, PIM은 연산부보다는 저장부에 특화되어 있다. TSV와 같은 차세대 Die 연결 기술은 PNM 구조를 통해 CIM과 PIM의 각각의 장점을 극대화할 수 있을 것으로 보이며, 본 로드맵에서는 이러한 PNM 기술을 미래형 PIM 연산기로 제안한다. 또한 PNM끼리의

연산을 통해 Scalability을 확보하며, 이를 통해 GPU에 쓰이는 중요한 부품에서, 독립적인 host-processor로 발전하는 것을 목표로 한다. 그래서 최종적으로는 Training까지 담당함으로써 PIM 기술만을 통해 neural network가 감당 가능하게끔 인공지능 생태계를 조성하여 Data-centric한 시대에 가장 중요한 Hardware가 되는 것을 최종 목적으로 한다.

참고문헌

- [1] <https://www.signalintegrityjournal.com/articles/3428-the-age-of-artificial-intelligence-ai-chips-to-2034>
- [2] John hennessy and David Patterson, Computer Architecture: A Quantitative Approach, 6/2 2018
- [3] 삼성전자
- [4] <https://spear-itn.eu/memory-hierarchy-how-does-computer-memory-work/>
- [5] [D. Bankman et al., "An Always-On 3.8μJ/86% CIFAR-10 Mixed-Signal Binary CNN Processor with All Memory On Chip in 28nm CMOS," ISSCC, pp. 222-223, 2018.
- [6] H. Jia, et al., "A Programmable Neural Network Inference Accelerator Based on Scalable In-Memory Computing," ISSCC, pp. 236-237, 2021.
- [7] J. -H. Yoon, et al., "A 40-nm, 64-Kb, 56.67 TOPS/W Voltage-Sensing Computing-In-Memory/Digital RRAM Macro Supporting Iterative Write with Verification and Online Read-Disturb Detection," in IEEE Journal of Solid-State Circuits, vol. 57, no. 1, pp. 68-79, Jan. 2022.
- [8] S.-C. Jung, et al., "A crossbar array of magnetoresistive memory devices for in-memory computing," Nature 601, 211–216, 2022.
- [9] S. Kwon, et al., "37.3 Monolithic in-Memory Computing Microprocessor for End-to-End DNN Inferencing in MRAM-Embedded 28nm CMOS Technology with 1.1Mb Weight Storage," ISSCC, pp. 1-3, 2025.
- [10] H. Mori, et al., "A 4nm 6163-TOPS/W/b 4790–TOPS/mm²/b SRAM Based Digital-Computing-in-Memory Macro Supporting Bit-Width Flexibility and Simultaneous MAC and Weight Update," ISSCC, pp. 132-133, 2023.
- [11] H. Fujiwara, et al., "A 3nm, 32.5TOPS/W, 55.0TOPS/mm² and 3.78Mb/mm² Fully-Digital Compute-in-Memory Macro Supporting INT12 × INT12 with a Parallel-MAC

-
- Architecture and Foundry 6T-SRAM Bit Cell," ISSCC, pp, 572-573, 2024.
- [12] Y. -C. Kwon, et al., "A 20nm 6GB Function-In-Memory DRAM, Based on HBM2 with a 1.2TFLOPS Programmable Computing Unit Using Bank-Level Parallelism, for Machine Learning Applications," ISSCC, pp, 350-351, 2021.
- [13] S. Lee, et al., "A 1ynm 1.25V 8Gb, 16Gb/s/pin GDDR6-based Accelerator-in-Memory supporting 1TFLOPS MAC Operation and Various Activation Functions for Deep-Learning Applications," ISSCC, pp, 176-177, 2022.
- [14] UPMEM, "UPMEM Website," <https://www.upmem.com>, 2020.
- [15] 2025PIM 워크샵, 노원우 교수 발표자료
- [16] F. Meng, et al., "TransMLA: Multi-Head Latent Attention Is All You Need," arXiv preprint arXiv:2502.07864.
- [17] <https://news.skhynix.co.kr/skhynix-tsmc-hbm4-mou/>
- [18] SK Hynix
- [19] <https://www.autodaily.co.kr/news/articleView.html?idxno=535316>
- [20] K. Wilcox, et al., "Design Technology Co-Optimization: Tradoffs for Building the Best Processor," in IEEE Solid-State Circuits Magazine, vol. 11, no. 4, pp. 58-66, Fall 2019.
- [21] H. Fujiwara, et al., "A 5nm 5.7GHz@1.0V and 1.3GHz@0.5V 4kb Standard-Cell-Based Two-Port Register File with a 16T Bitcell with No Half-Selection Issue," ISSCC, pp, 340-341, 2021.
- [22] G. Jedhe, et al., "A 12nm 137 TOPS/W Digital Compute-In-Memory using Foundry 8T SRAM Bitcell supporting 16 Kernel Weight Sets for AI Edge Applications," IEEE Symposium on VLSI Technology and Circuits (S. VLSI), pp. 1-2, 2023.
- [23] S. Hsu, et al., "2.4GHz, Double-Buffered, 4kb Standard-Cell-Based Register File with Low-Power Mixed-Frequency Clocking for Machine Learning Accelerators," IEEE Symposium on VLSI Technology and Circuits (S. VLSI), pp. 22-23, 2022.
- [24] S. Rajbhandari, et al., "ZeRO: Memory Optimizations Toward Training Trillion Parameter Models", arXiv: 1910.02054.
- [25] S. Rajbhandari, et al. "ZeRO-Infinity: Breaking the GPU Memory Wall for Extreme Scale Deep Learning", arXiv: 2104.07857.

IX. 반도체 패키징 기술

1. 서론 및 기술 정의

반도체 패키징 기술은 오랫동안 “칩을 보호하고 기판에 연결하는 하드웨어”로 인식되어 왔다. 그러나 미세공정 노드의 비용과 난이도가 지속적으로 증가하고, 시스템 수준의 요구 사항이 복잡해지면서, 패키지는 이제 성능과 신뢰성을 동시에 좌우하는 핵심적 플랫폼으로 그 역할이 확장되고 있다. 특히 이종집적(heterogeneous integration)의 확산은, 서로 다른 공정과 기능을 갖는 여러 종류의 다수 칩을 단일 패키지 내에서 집적함으로써, 기존 단일칩 SoC 방식만으로는 달성하기 어려운 시스템 수준의 성능 최적화를 가능하게 한다.

최근 AI, HPC (high performance computing) 등으로의 응용분야의 확대는 패키지 기술의 중요성을 더욱 부각시키고 있다. 초고대역·초저지연·고전력 밀도를 요구하는 AI 가속기 및 HPC는 단일 대형 SoC 구조만으로 성능 확장이 어려운 상황에 직면하고 있으며, 이에 따라 칩렛 기반 구조, 2.5D/3D 적층, 고밀도 RDL, 개방형 Die-to-Die 인터페이스 등이 중요한 필수 기술 요소로 자리 잡고 있다. 또한 GPU, HBM, 네트워크 칩을 하나의 패키지에서 통합하는 구조가 점차 보편화되리라 예상되면서, 패키징은 단순한 후공정을 넘어 AI·HPC 시스템 구조를 결정하는 핵심 아키텍처로 인식되고 있다.

본 로드맵에서는 이러한 기술적 흐름을 바탕으로 패키징 기술 가운데 다섯 가지 주요 기술을 정의하고자 한다. 첫째, 단일 칩을 하나의 패키지로 구성하는 Single-Chip 구조와, 다수의 칩을 하나의 모듈로 통합하는 Multi-Chip 구조를 소개한다. 둘째, 패키지 내부 배선 및 인터커넥트 관점에서 전통적인 2D 패키지와, 고밀도 인터포저나 브리지를 사용하는 2.xD, 그리고 수직 적층 구조인 3D 패키지를 구분하여 설명한다. 셋째, 웨이퍼 혹은 패넌 단위에서 여러 칩을 동시에 패키징하는 Fan-out Wafer-Level/Panel-Level 패키지(FO-WLP/PLP) 기술을 다룬다. 넷째, HPC 및 데이터센터 패키징에서는 칩렛 기반 이기종 통합, 초고대역 메모리(HBM) 결합, 미세 피치 인터커넥트 및 Die-to-Die 표준, 고열밀도 대응 패키지·냉각 구조 등 고성능 컴퓨팅 시스템을 구성하는 핵심 패키징 기술을 다룬다. 다섯째, 고전력·고밀도 환경에서 필수적인 열 관리 구조와 함께 패키지 전반을 뒷받침하는 모델링, 시뮬레이션, Co-Design 기술을 포함한다.

본 문서는 반도체공학회 주관으로 최근 반도체 패키지 기술의 주요 변화를 정리하고, 공개 자료를 바탕으로 향후 로드맵을 제시하는 것을 목표로 한다. 1장에서는 패키징 기술의 정의와 역할을 정리하고, 2장에서는 주요 패키지 기술별 현재와 향후 발전 방향을 기술한다. 3장에서는 공개된

자료를 기반으로 고성능 패키지의 정량적 로드맵을 표로 정리하며, 4장에서는 향후 업데이트를 위해 필요한 제언과 후속 작업의 필요성을 간단히 정리하고자 한다.

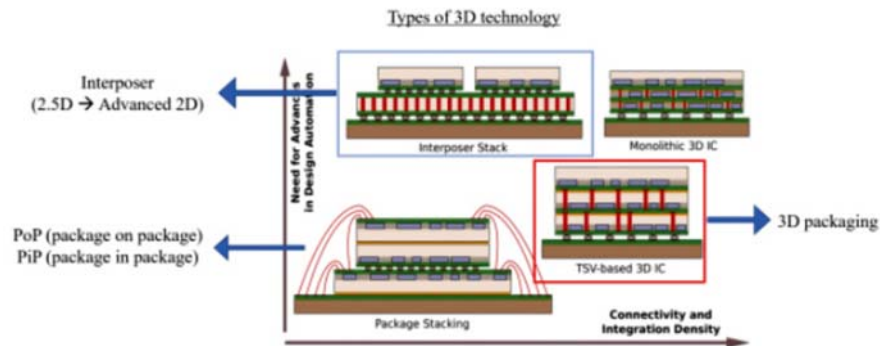


그림 9-1. 삼차원 구조 패키지 정의^[1]

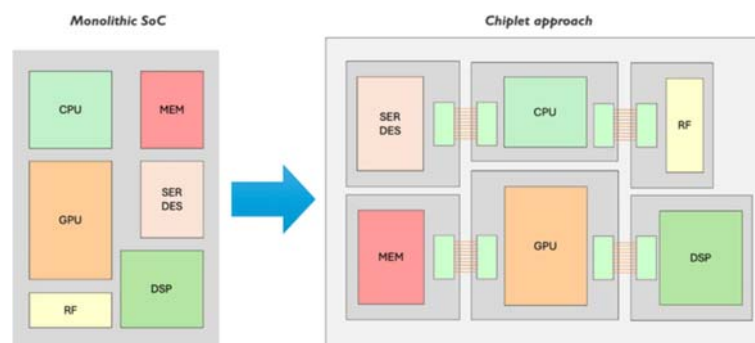


그림 9-2. 단일칩 SoC와 Chiplet 개념도^[2]

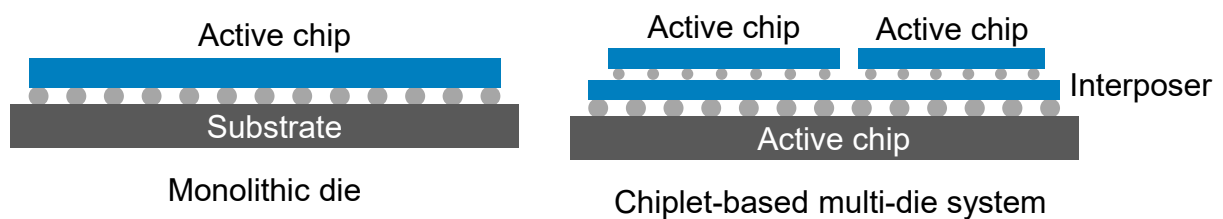


그림 9-3. 단일 die와 Chiplet 기반 multi-die 개념도

2. 반도체 패키지 핵심 기술 요약

(1) Single-Chip 및 Multi-Chip Integration

Single-Chip 패키지는 오랜 기간 반도체 산업의 기본 형태로 자리하여, 하나의 실리콘 다이에 가능한 많은 기능을 집적하는 방향으로 발전해왔다. 그러나 공정 미세화가 수 nm 급 이하로 진입하면서, 미세화에 따른 비용 증가와 설계 복잡도 상승, 대형 다이의 수율 저하 문제가 동시에 부각되고 있다. 특히 대형 SoC의 경우 단일 결함이 전체 칩의 수율에 직접적인 영향을 미치기 때문에, 비용 및 제조 리스크 측면에서 구조적인 한계가 점차 뚜렷해지고 있다.

이러한 배경에서 Multi-Chip Integratio과 칩렛(chiplet) 개념이 대안적인 집적 방식으로 주목받고 있다. Multi-Chip 패키지는 기능별·공정별로 분리된 여러 개의 칩을 하나의 패키지 내부에서 상호 연결함으로써, 단일 SoC 수준의 시스템 기능을 구현하는 구조이다. 예를 들어, 고성능 연산을 담당하는 로직 다이는 선단 공정을 적용하고, I/O, 아날로그, 전력 관리 회로는 상대적으로 성숙한 공정 노드에 구현하는 방식이 활용되고 있다. 이러한 분리는 공정 비용 부담을 완화하는 동시에, 설계 재사용성과 플랫폼화 측면에서도 유리한 구조로 평가된다.

멀티칩 통합에서 핵심 요소는 다이 간 인터커넥트와 패키지 구조이다. 2.5D 인터포저, RDL 기반 브리지, 패키지 내부 고밀도 라우팅 기술을 활용하면 칩 간을 수십 μm 수준의 피치로 연결할 수 있으며, HBM(High Bandwidth Memory)과 로직 칩을 동일한 인터포저 상에 집적하는 구조가 대표적인 사례로 활용되고 있다. 이러한 구조는 패키지 내부에서 높은 메모리 대역폭과 효율적인 전력 전달을 동시에 달성할 수 있는 기반을 제공한다. 또한 하이브리드 본딩(Cu-Cu direct bonding)과 같은 3D 인터커넥트 기술을 적용하면, 기존 범프 방식보다 훨씬 미세한 피치의 수직 연결이 가능해져 IO 밀도와 대역폭을 추가적으로 향상시킬 수 있다.

한편, Multi-Chip 구조는 열·전력·기계적 측면에서 새로운 공학적 과제를 동반한다. 각 칩의 발열 특성과 전력 프로파일이 상이하고, 열팽창계수(CTE)가 다른 재료들이 동시에 사용되기 때문에, 패키지 내 온도 분포, warpage, 인터커넥트 신뢰성이 서로 복합적으로 연관된다. 특히 미세 피치 범프 및 하이브리드 본딩 인터페이스는 기계적 응력과 계면 결함에 상대적으로 민감할 수 있어, 칩-패키지-기판을 아우르는 열-기계 연성 해석과 설계 최적화가 중요하다. 또한 웨이퍼 박막화(thinning)와 다이 절단(singulation) 공정에서도 기계적 신뢰성 이슈가 발생할 수 있으며, 이를 완화하기 위해 저 CTE 소재, 고강도·저변형 패키지 레진, 웨이퍼 레벨 적층 공정 등이 지속적으로 연구되고 있다. 최근에는 적층 제조(Additive Manufacturing) 기반의 미세 배선·전극 구조 형성 기술도 연구 단계에서 검토되고 있다.

HPC·AI·데이터센터 응용이 확대되면서, 멀티칩 통합은 고성능 시스템 구현을 위한 중요한 패키징 아키텍처 중 하나로 정착되고 있다. 증가하는 메모리 대역폭, 전력 밀도, 칩 간 통신 요구를 단일 대형 SoC 구조만으로 모두 충족하기는 점차 어려워지고 있으며, 이에 따라 산업 전반에서는 칩렛 기반 구조를 중심으로 한 제품 개발 전략이 확대되고 있다. GPU, HBM, 가속기 칩을 인터포저 또는 브리지 구조에서 통합하는 방식은 이미 다양한 고성능 서버 및 AI 시스템에서 활용되고 있으며, 향후에는 하이브리드 본딩 기반 3D 적층, 개방형 Die-to-Die 인터페이스(UCIe 등), 저유전·저손실 기판 기술 등이 패키지 성능을 좌우하는 중요한 기술 요소로 작용할 것으로 전망된다.

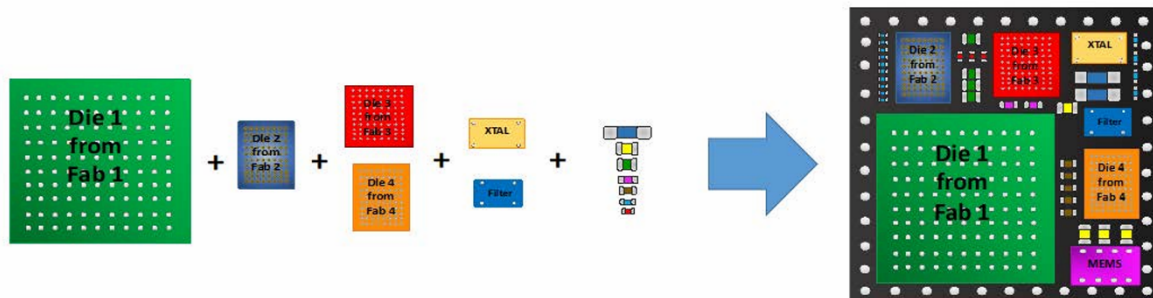


그림 9-4. 이기종 통합 및 SiP (System in Package) (출처: ASE)

(2) 2D / 2.xD / 3D 아키텍처와 인터커넥트 로드맵

2D 패키지는 하나 또는 여러 개의 active die를 패키지 기판 위에 나란히 배치하고, 기판 배선을 통해 상호 연결하는 구조이다. 전통적인 플립칩 BGA 패키지가 이에 해당하며, 다이와 패키지 기판 사이에는 수십에서 수백 μm 수준의 피치를 갖는 범프가 사용된다. 이러한 다이-패키지 범프 피치는 최근 고성능 응용을 중심으로 약 수십 μm 후반~100 μm 내외 수준에서 완만한 개선 양상을 보이고 있으며, 향후에도 급격한 축소보다는 점진적인 개선이 예상된다. 대신, 더 미세한 피치와 높은 I/O 밀도를 요구하는 영역은 인터포저, 브리지, RDL 레이어 등 패키지 내부 구조로 점차 이동하는 추세이다.

2.xD 패키지는 2D 구조의 한계를 보완하기 위한 중간 단계의 아키텍처로, 문헌에 따라 2.1D, 2.3D, 2.5D 등으로 구분되어 사용된다. 구조적으로는 유기 기판 기반의 고밀도 배선 구조(2DO, Organic 기판)와 실리콘 또는 글라스 기판 인터포저 구조(2DS)로 나눌 수 있다. 2DS 구조에서는 실리콘 BEOL과 유사한 미세 배선을 통해 수십 μm 이하의 피치로 고속 인터커넥트를 구현할 수 있으며, HBM과 로직 칩, 또는 CPU·GPU·I/O 칩 등의 복수 칩을 단일 인터포저 위에 집적하는 구조가 활용되고 있다. 2DO는 고밀도 유기 기판과 임베디드 브리지를 이용하여, 2DS 대비 다소 완

만한 미세화 수준에서 경제성과 확장성을 중시한 고밀도 패키징 솔루션을 제공한다.

3D 아키텍처는 수평 방향의 확장 대신 다이를 수직으로 적층하여 면적 효율과 집적도를 향상시키는 접근 방식이다. TSV를 이용한 3D DRAM 및 HBM 스택, 로직-메모리 적층, 그리고 Cu-Cu 하이브리드 본딩을 통한 로직-로직 적층 구조 등이 대표적인 사례이다. 최근 상용 및 준상용 기술에서는 10 μm 내외 수준의 하이브리드 본딩 피치가 적용되고 있으며, 연구 단계에서는 이보다 더 미세한 피치의 본딩 기술도 지속적으로 보고되고 있다. 이러한 초미세 인터커넥트는 단위 면적당 I/O 밀도를 크게 높이고, die-to-die 대역폭 밀도와 에너지 효율 향상에 기여하는 기반 기술로 활용되고 있다.

인터커넥트 로드맵을 종합적으로 보면, 다이-패키지 범프 피치는 비교적 완만한 속도로 개선되는 반면, 다이-다이 area 인터커넥트와 Fan-out RDL은 보다 빠른 속도로 미세화가 진행되는 이른바 “양극화된 스케일링” 경향을 보인다. 이는 패키지 구조가 점차 계층화되면서, 각 계층별로 요구되는 해상도와 기능이 구분되고 있음을 의미한다. 향후에는 패키지 내부의 로직-메모리 및 로직-로직 인터커넥트는 하이브리드 본딩 기반의 초미세 피치 방향으로 발전하고, 패키지 외부 인터페이스(보드-모듈 측)는 상대적으로 완만한 스케일링과 신호 무결성 확보에 중점을 두는 구조로 발전할 것으로 예상된다.

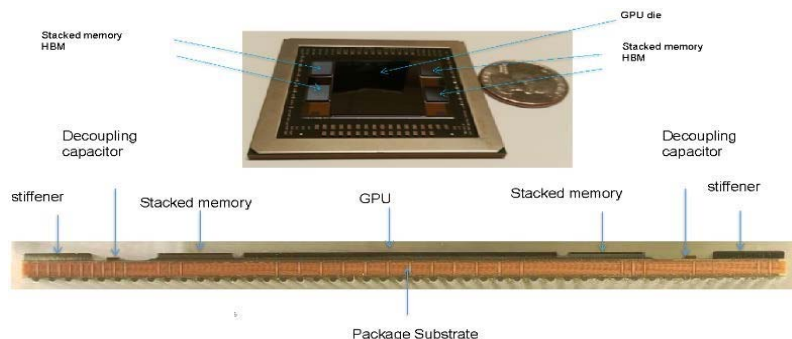


그림 9-5. AMD 피치GPU: 실리콘인터포저2.5D 패키지의HBM (출처: ASE)

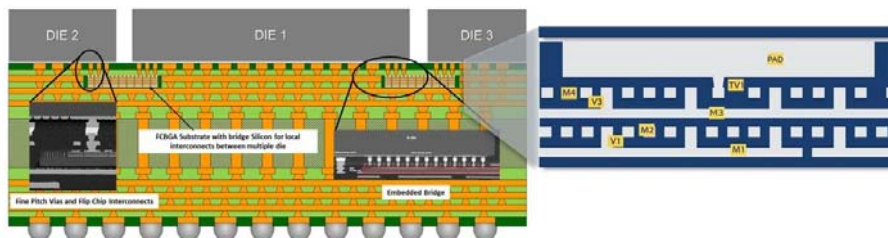


그림 9-6. Intel EMIB Technology (출처: Intel)

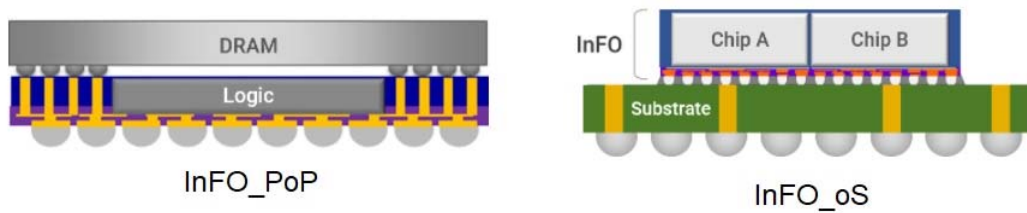


그림 9-7. InFO-PoP (package-on-package) and InFO-oS (assembly-on-substrate) (출처: TSMC)

(3) Fan-out Wafer-Level 및 Panel-Level 패키징

Fan-out Wafer-Level Packaging(FO-WLP)은 die를 몰드(compound) 내부에 임베드한 뒤, 재구성된 웨이퍼(reconstituted wafer) 상에 재배선(RDL, Redistribution Layer)을 형성하여 I/O를 다이 외곽으로 확장하는 패키징 구조이다. 이 방식은 기존 Fan-in WLCSP와 달리 다이 가장자리 내의 I/O 수에 제한을 받지 않으며, 소형 다이에서도 비교적 높은 I/O 수와 표준 패키지 피치를 동시에 구현할 수 있다는 특징을 갖는다. 또한 별도의 유기 서브스트레이트 없이 패키지가 구성되므로, 패키지 두께 감소와 함께 전기적·열적 경로를 단순화할 수 있는 장점이 있다.

Fan-out 기술의 진화는 주로 RDL의 라인/스페이스(Line/Space) 해상도와 적층 레이어 수의 고도화를 중심으로 이루어지고 있다. RDL 해상도가 향상될수록 패키지 내부에서 고속 신호 라우팅, 차동 신호 구조, 분포형 전력망 설계 등이 보다 용이해지며, 복수의 로직 칩과 메모리, RF 모듈을 단일 Fan-out 패키지 내에서 집적할 수 있는 설계 자유도도 함께 증가한다. 이러한 특성으로 인해 Fan-out 패키지는 모바일, RF, 고집적 시스템 반도체 등 다양한 응용 분야에서 활용 범위가 확대되고 있다.

Panel-Level Packaging(PLP)은 이러한 Fan-out 구조를 기존의 원형 웨이퍼 대신 대면적 사각 패널 포맷으로 확장한 방식이다. 패널 기반 공정은 동일한 공정 사이클에서 더 많은 다이를 동시에 처리할 수 있어, 생산성 향상과 비용 효율성 측면에서 장점을 가질 수 있다. 반면, 패널 면적이 커질수록 warpage, 수축·팽창 제어, 다이 위치 정렬 오차(die shift), 장비 및 재료 표준화 등 공정적인 난제가 함께 증가한다. 이에 따라 PLP는 현재 단계적으로 적용이 확대되는 추세에 있다.

향후 Fan-out 및 Panel-Level 패키징 기술은 RDL 해상도의 점진적인 미세화, 대면적 공정에서의 공정 균일성 및 warpage 제어 기술의 개선, 그리고 장비·재료·공정 표준화의 진전이라는 방향으로 발전할 것으로 예상된다. 이러한 흐름은 패키지 레벨에서의 집적도 향상과 생산성 개선, 그리고 비용 효율성 확보를 동시에 추구하는 방향으로 이어질 것으로 전망된다.

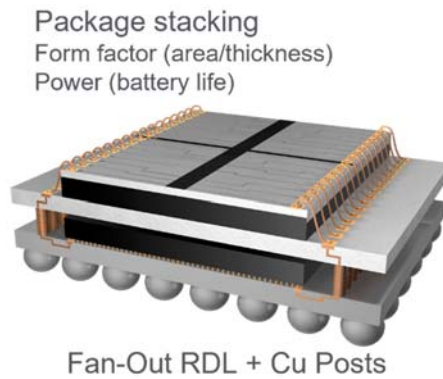


그림 9-8. Fan-out 패키지 개념도 (출처: Applied Materials)

(4) High Performance Computing 및 Data Center

HPC(High Performance Computing)와 데이터센터는 대규모 연산과 데이터 처리를 수행하는 대표적인 시스템 응용 분야로, 초고속 데이터 처리, 낮은 지연 특성, 대용량 메모리 접근 성능을 공통적으로 요구한다. 최근에는 단일 대형 칩의 성능 확장 한계로 인해, 다수의 칩을 패키지 내부에서 통합하는 칩렛 기반 구조가 주요한 시스템 구성 방식으로 활용되고 있다.

HPC 시스템에서는 높은 메모리 대역폭과 낮은 지연 특성을 확보하기 위해 프로세서와 메모리를 패키지 내부에서 근접 배치하거나 직접 연결하는 구조가 중요하다. 이에 따라 2.5D 인터포저 기반 구조에서 HBM과 로직 칩을 통합하는 패키징 방식이 GPU 및 AI 가속기를 중심으로 널리 활용되고 있다. 이러한 구조는 기존 기판 기반 연결 대비 신호 지연과 손실을 줄이고, 메모리 접근 효율을 향상시키는 데 기여한다.

한편, 칩렛 기반 구조가 확대됨에 따라 패키지 내부의 고대역 신호 전달을 위한 고밀도 인터커넥트 기술과 Die-to-Die 인터페이스의 중요성도 함께 증가하고 있다. 또한 고집적 패키지 구조에서는 전력 소모 증가와 함께 열·기계적 부담이 동시에 커지므로, 냉각 구조와 전력 전달 경로를 포함한 패키지-시스템 수준의 통합 설계가 중요한 요소로 인식되고 있다.

HPC 및 데이터센터용 패키징에서는 단순한 칩 집적을 넘어, 연산 성능, 메모리 계층, 전력 전달, 냉각 구조가 상호 연동된 시스템 최적화 관점이 중요하게 작용한다. 특히 다수의 연산 칩과 메모리, 인터페이스 칩이 동시에 집적되는 구조에서는 전기적 성능뿐 아니라 열 확산 경로, 전력망 구조, 패키지 변형 등 다양한 물리 현상이 서로 영향을 미치므로, 패키지 수준에서의 통합 설계 접근이 요구된다.

종합하면 HPC 및 데이터센터 패키징에서는 칩렛 기반 다이 통합, 고대역 메모리 결합, 고밀도 인터커넥트, 열·전력 대응 구조가 핵심 기술 요소로 작용하고 있으며, 패키징 기술은 시스템 성능과 확장성을 뒷받침하는 중요한 기반 기술로 자리잡고 있다.

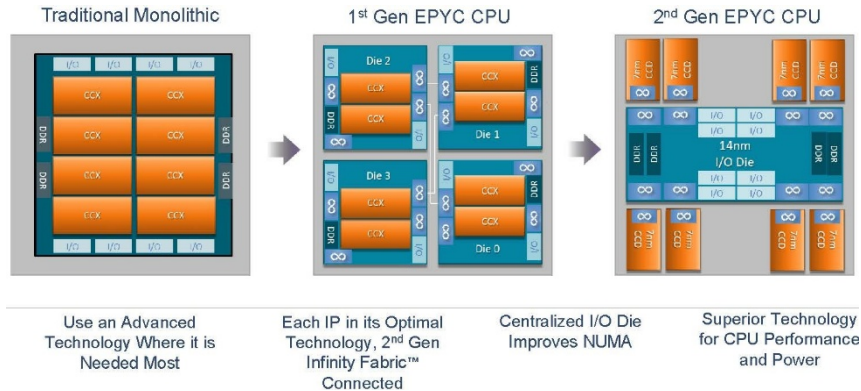
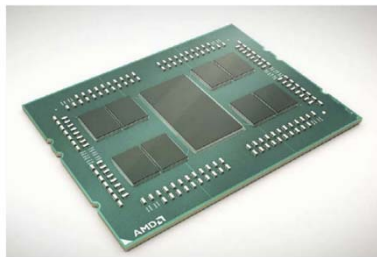
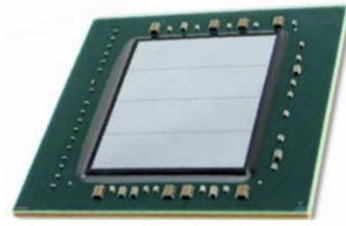


그림 9-9. Chiplet evolution (출처: AMD)



AMD EPYC Partitioned Server Processor



Xilinx Virtex Partitioned FPGA

그림 9-10. HPC 및 데이터센터용 다이 분할(Die Partitioning) 기반 고성능 프로세서 예시

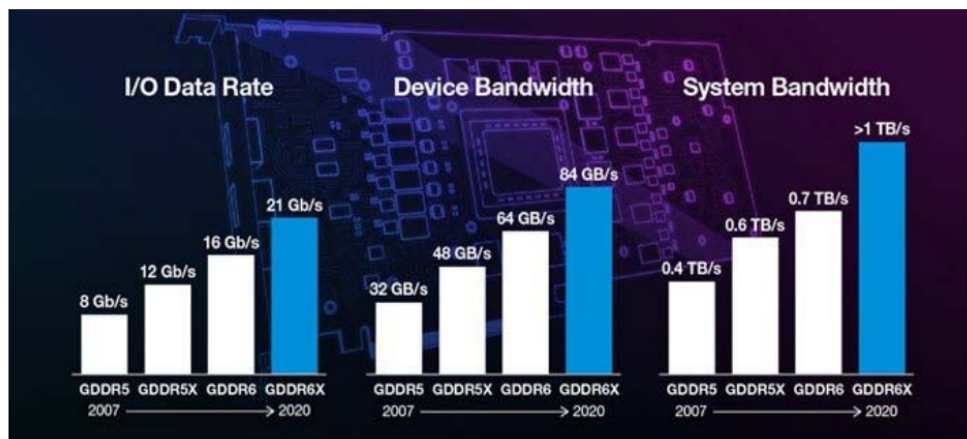


그림 9-11. HPC 및 데이터센터용 GDDR 메모리의 세대별 전송 속도 및 대역폭 (출처: Micron)

(5) 열관리 · 모델링 · Co-Design 통합 기술

고성능 패키지의 고집적화와 함께 전력 밀도는 지속적으로 증가하고 있으며, 이에 따라 열 관리는 단순한 부가 요소가 아니라 패키지 설계의 핵심 제약 조건 중 하나로 인식되고 있다. 고집적 로직, 메모리, 인터페이스 칩이 단일 패키지 내에 동시에 집적되는 구조에서는, 칩 내부 발열 특성뿐 아니라 패키지 구조, 열전도 재료, 전력 전달 경로, 냉각 구조가 서로 강하게 연동되어 열 거동을 결정하게 된다. 따라서 열 관리는 히트싱크나 팬과 같은 단일 부품의 문제가 아니라, 패키지-시스템 수준에서 통합적으로 고려되어야 할 설계 요소로 확장되고 있다.

특히 3D 적층 구조와 HBM 기반 패키지에서는 로직 다이, 메모리 스택, 인터포저, interconnect가 복합적인 열 전달 경로를 형성한다. 이러한 구조에서는 단순한 상부 냉각 방식만으로 내부 온도 분포를 균일하게 제어하기 어렵기 때문에, 열 확산 경로 설계, 열전도 재료 선택, 패키지 구조 최적화 등이 함께 검토된다. 일부 고성능 패키지 응용에서는 액체 냉각과 같은 적극적인 냉각 방식이 검토되고 있으며, 향후에는 집적 구조에 적합한 다양한 냉각 아키텍처가 단계적으로 도입될 것으로 예상된다.

이와 같은 복잡한 패키지 구조에서는 경험이나 규칙 기반 설계만으로는 성능과 신뢰성을 동시에 만족시키기 어려우며, 이에 따라 Modeling & Simulation(M&S)과 Co-Design 기반 설계 인프라의 중요성이 크게 증가하고 있다. 전기적 신호 무결성(SI/PI), 열 해석, 기계적 변형 및 warpage 해석은 상호 독립적인 문제가 아니라 서로 강하게 연동된 현상으로 나타나므로, 칩-패키지-기판을 관통하는 멀티피직스·멀티스케일 해석이 요구된다.

Co-Design은 칩 설계와 패키지 설계를 동시에 고려하는 통합 설계 방법론으로, I/O 구조, 인터커넥트 배치, 전력망 구성, 열 경로, 기판 배선 규칙 등을 하나의 설계 문제로 묶어 최적화하는 접근을 의미한다. 이를 위해서는 칩, 패키지, 보드 설계 데이터가 상호 연동되는 환경과, 서로 다른 물리 도메인 간 영향을 직관적으로 분석할 수 있는 설계 플랫폼이 필요하다. 향후 패키지 기술이 더욱 고집적·고출력 구조로 발전함에 따라, 열관리와 M&S, Co-Design은 개별 기술을 넘어 패키지 로드맵을 뒷받침하는 핵심 설계 기반으로 자리잡을 것으로 전망된다.

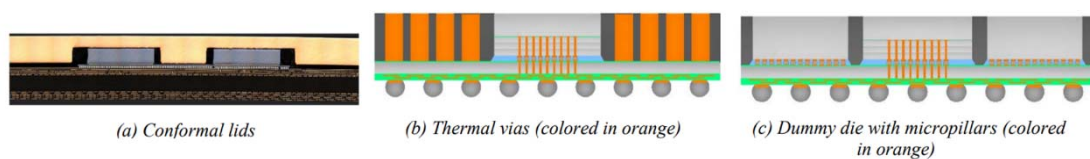


그림 9-12. 방열을 위한 패키지 구조 예시

3. 반도체 패키지 로드맵 예측

표 9-1은 공개된 산업 자료와 학술 발표를 기반으로, 반도체 패키징 기술의 발전 방향을 정성적·정량적으로 정리한 전망치이다. 각 수치는 기술적 가능성의 상한을 제시하기보다는, 현재 상용화 수준과 공정·비용·신뢰성 제약을 고려하여 비교적 보수적으로 설정하였다. 특히 인터커넥트 피치, Fan-out RDL, 패키지 I/O 및 냉각 항목은 이미 보고된 실증 사례 또는 로드맵 수준에서의 점진적 개선 추세를 반영한 값이다. 따라서 본 표의 수치는 확정적 예측이 아닌 기술 발전의 방향성과 상대적 속도를 이해하기 위한 참고 지표로 활용하는 것이 바람직하다.

표 9-1. 반도체 패키지 로드맵

항목	2020년	2030년	2040년
Die-Substrate Interconnect Pitch (Flip-chip bump)	약 80-100 μm 수준 (패키지 별 상이)	약 60-90 μm 수준 (고성능 패키지 중심)	약 40-80 μm 수준 (완만한 개선)
Die-to-Die Interconnect Pitch (μ -bump / Hybrid Bonding 등)	μ -bump : 40-50 μm	Hybrid bonding: 10 μm 급	Hybrid bonding: 5-10 μm 급 (공정/비용 제약)
Fan-out RDL L/S	5-10 μm 급	~2 μm 급 고도화 (고급 PKG 기준)	~1 μm 급 목표로 개발 지속 (전망)
Fan-out RDL Layer 수	2-4 층 수준 (응용별 상이)	4-6 층 수준 (고급 PKG 기준)	6-8 층 가능성 (전망)
Package IO / High-Bandwidth IO (HPC 기준)	32-56 Gb/s 급 전기 적 IO, EMIB·CoWoS 활용	64-112 Gb/s 급 전기적 IO, Photonic IO 확산	112-224 Gb/s 급 전기적 IO, Photonic IO 상용화 가능성
Package Cooling (HPC 기준)	공랭·히트스프레더 중심 예: 200-250 W 수준	액체냉각/Vapor chamber 등 적극 냉각 확대 예: 400-600 W 수준	3D 적층 대응 초고효율 냉각 필요 예: 800 W+수준

4. 요약 및 차후 로드맵 과제

본 보고서에서는 반도체 패키징 기술의 주요 변화를 Multi-Chip 및 칩렛 기반 집적, 2D에서 2.xD·3D로의 아키텍처 확장, Fan-out 및 Panel-Level 공정의 고도화, HPC·데이터센터 응용 확대, 그리고 열관리·모델링·Co-Design 통합 기술이라는 다섯 가지 핵심 축을 중심으로 정리하였다.

Single-Chip 기반 집적 방식은 공정 비용 증가와 대형 다이 수율 한계로 인해 구조적 제약이 점차 커지고 있으며, 이에 따라 칩렛 기반 Multi-Chip Integration이 새로운 시스템 집적 방식으로 확산되고 있다. 이와 함께 패키지 구조는 2D에서 2.xD 및 3D 구조로 확장되고 있으며, 인터포저, Fan-out RDL, 하이브리드 본딩 기반 인터커넥트 미세화는 고대역·저지연 특성 구현을 위한 핵심 요소로 자리 잡고 있다. 또한 Fan-out 및 PLP 공정은 패키지 미세화와 생산성 향상을 동시에 추구하는 기술로, 적용 범위가 점진적으로 확대되고 있다.

HPC 및 데이터센터 분야는 이러한 패키징 기술 변화를 가장 빠르게 요구하는 대표적인 응용 영역으로, 칩렛 기반 구조와 HBM 결합, 고밀도 인터커넥트, 전력·냉각 통합 설계가 시스템 성능과 확장성을 좌우하는 핵심 요소로 작용하고 있다. 아울러 고집적·고전력 구조로 갈수록 열관리, 멀티피직스 모델링, Co-Design 기반 통합 설계 환경은 패키지 성능과 신뢰성을 결정하는 필수 기반 기술로 인식되고 있다.

이와 같은 기술 흐름을 종합할 때, 반도체 패키징은 더 이상 단순한 후공정 기술이 아니라, 시스템 성능·전력·열·신뢰성·비용을 동시에 좌우하는 핵심 시스템 통합 기술로 자리매김하고 있으며, 향후 반도체 로드맵에서도 독립적인 기술 축으로 지속적인 고도화가 요구되는 영역이라 할 수 있다.

본 보고서는 공개 자료를 기반으로 정리하였으며, 본문에 포함된 모든 그림의 저작권은 각 원저자에게 있다. 향후 보다 발전된 패키지 로드맵을 구성하기 위해서는 다음과 같은 주제들이 추가적으로 중점 검토될 필요가 있다.

- 다중 스택 및 대면적 인터포저 기반 초고대역 패키지 기술
- HBM·칩렛을 연결하는 저지연·고대역 Die-to-Die 인터커넥트
- Fan-out 및 PLP 기반 2.5D·3D 패키징 공정 고도화
- 고전력·고밀도 환경 대응 패키지 냉각 및 신뢰성 구조 확보

참고문헌

- [1] 좌성훈, "첨단 반도체 패키징기술 개발 표준 트렌드", 소재·부품·장비 표준동향 리포트, 2022
- [2] E.-L. Bukshish, et al. "Design Considerations of an Analog Voltage Mode Readout Circuit for the CMOS-SOI-MEMS Gas Sensor Dubbed GMOS," Micromachines, 2025.
- [3] R. Mahajan, et al, "Embedded Multi-Die Interconnect Bridge (EMIB) – A High Density, High Bandwidth Packaging Interconnect", Electronic Components and Technology Conference (ECTC), 2016.
- [4] T. Dillinger, "Highlights of the TSMC Technology Symposium 2021 – Packaging", 2021.
- [5] M. Liu, "Unleashing the Future of Innovation," ISSCC, 2021.
- [6] J. H. Lau, "Recent Advances and Trends in Advanced Packaging," IEEE Transactions on Components, Packaging and Manufacturing Technology, vol. 12, no. 2, pp. 228-252, 2022.
- [7] J. H. Lau, "Recent Advances and Trends in Chiplet Design and Heterogeneous Integration Packaging," ASME. J. Electron. Packag. 2024.
- [8] TSMC, "CoWoS® Advanced Packaging Technology Overview,"

X. 양자컴퓨팅 반도체 기술

1. 서론

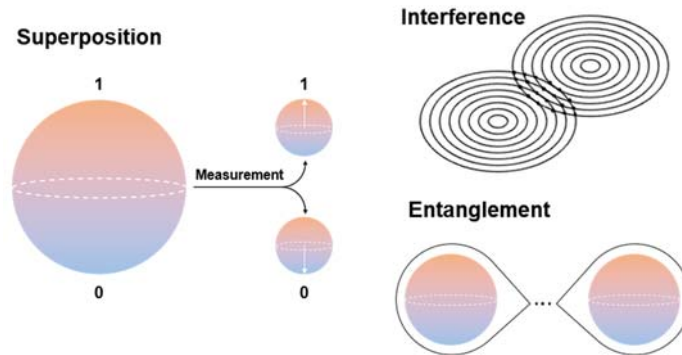


그림 10-1. 큐비트의 양자역학 매커니즘

양자컴퓨터란 큐비트를 기본 단위로 중첩, 얽힘, 간섭 등의 양자역학 매커니즘을 이용하여 확률적이고 가역적인 연산 방법을 사용하는 컴퓨팅 기술이다. 중첩이란 큐비트에서 0과 1의 상태가 공존하는 현상으로, 이는 n 개의 큐비트로 2^n 개의 상태를 동시에 표현할 수 있어 병렬 연산 처리를 가능하게 한다. 양자 얽힘은 서로 다른 큐비트가 연결되어 한쪽 큐비트의 상태가 결정될 때 다른 큐비트의 상태가 이와 관련되어 결정되는 현상이다. 얽힘 현상을 이용해 병렬성을 구현할 수 있다. 양자 상태들이 서로 더해지거나 상쇄되는 간섭 현상을 통해, 양자 특정 결과의 정답 확률을 증폭시키고 오답 확률을 약화시킨다. 이러한 양자 현상을 제어하여 논리적 상호작용으로 양자 게이트 연산을 구현하고, 복잡한 문제를 고전 컴퓨터보다 훨씬 빠르게 처리할 수 있다 (그림 10-1 참조).

큐비트의 구현 방식으로는 표 10-1과 같이 플랫폼에 따라 크게 전기적 방식, 광학 방식, 이온 트랩 방식이 있다. 전기적 방식으로 구현되는 초전도 큐비트는 조셉슨 접합을 사용한 초전도 회로를 매체로, 마이크로파 펄스를 이용하여 큐비트 상태를 조절하고 공진자를 통한 마이크로파 읽기를 수행한다. 초전도 큐비트는 결맞음 시간이 비교적 짧아 오류정정을 위한 추가 기술이 요구되고 극저온 동작을 위한 냉각 장치를 필요로 한다. 그러나 기존 CMOS 반도체 공정과의 높은 호환성, 대규모 집적 가능성으로 인한 상용화의 용이함, 빠른 제어 속도 등을 강점으로 바탕으로 산업계와 학계에서 가장 활발히 연구가 진행되고 있는 플랫폼 중 하나이다. 이와 같은 전기적 방식으로 구현되는 큐비트로는 처전도 큐비트 외에도 반도체 양자점 내 전자의 스핀 상태를 이용하는 스핀 큐비트 또한 존재한다.

광자(photon)을 통해 구현되는 광자 큐бит은 레이저, 광섬유, 비선형 광학 소자 등을 이용하여 편광, 위상, 경로, 시간 모드 등으로 양자 상태를 표현한다. 광자 큐бит은 일반적으로 상온에서 동작 가능하여 확장이 용이하며, 통신 모듈과 통합에 유리하나 광손실 문제를 해결해야 하고 큐비트 간 상호작용 구현, 얽힘 제어가 어려워 논리게이트 구현에 어려움이 있다.

이온 트랩 큐бит은 전자기적 포텐셜을 이용하여 이온을 포획하고 레이저를 통해 양자 상태를 제어하여 전자기적 방식과 광학적 방식이 혼합되어있다. 이온 트랩 큐бит은 긴 결맞음시간과 높은 정밀도로 양자 계산과정의 오류 수정이 효율적으로 이뤄질 수 있으나 속도가 느리고, 레이저 및 진공 인프라 시스템이 복잡하여 상용화를 위해서는 기술적 도전이 필요하다.

표 10-1. 양자컴퓨터 큐비트 플랫폼의 종류

Qubit Platforms			
Qubit Type	Superconducting Qubits	Trapped Ion Qubits	Photonic Qubits
Medium	Superconducting circuits	Trapped ions	Photons
Structure	Circuits composed of Josephson junctions and capacitors	Ions trapped in electromagnetic fields and controlled by lasers	Optical systems using photon states
Qubits temperature	10mK ~ 20mK (Dilution cryostat)	4K	Room temperature
Best two qubits Gate fidelities ^[1]	99.68% (IBM Heron r2)	99.84% (56q, Quantinuum)	99%
Coherence time	Short (ns)	Long (>s)	Long (>μs)
Gate time ^[1]	20ns to 300ns	0.1μs to 4μs	<1ns
Noise Sensitivity	High (Flux noise, Charge noise)	Relatively low	Low
Readout Method	Resonator-based	Fluorescence detection	Photodetector
Scalability	High (High Commercial Viability)	Low (Each ion needs to be individually controlled with lasers)	High (Generated and manipulated on chips)
Advantages	• High scalability • High fidelity • High gate speeds • Integration with existing semiconductor tech	• Extremely high fidelity • Long coherence time • High-quality entanglement	• Extremely high gate speeds • No cryogenics or vacuums required • Can leverage existing CMOS fabs
Disadvantages	• Requires cryogenic • Short coherence times • High sensitivity to noise	• Scalability challenges • Low gate speeds • Ultra-high vacuum required	• Noise from photon loss • Each program requires its own chip

표 10-2. 큐비트 관련 국내외 연구기관 및 기업

Qubit Platforms				
Qubit Type		Superconducting Qubits	Trapped Ion Qubits	Photonic Qubits
Research Institutions and Companies	Overseas ^[1]			
	Domestic	KRISS, SKKU, UNIST, KISTI, SNU	SNU, KIST	ETRI, KIST, KRISS

큐비트 종류에 따른 국내외 연구기관 및 기업은 표 10-2에서 보이는 바와 같이 초전도 큐비트를 연구하는 IBM, Google, Intel, Rigetti, D-Wave, KRISS 등과 광학 방식 큐비트를 연구하는 Xanadu, PsiQuantum, ETRI, 이온 트랩 방식을 연구하는 IonQ, Quantinuum 등이 있다. Google에서는 랜덤 양자 회로를 구현하며 양자 우월성을 입증하였고, 2020년 Intel과 QuTech에서 1.1K에서 동작하는 2개의 스핀 큐비트 게이트를 발표하였다. 국내에서는 원자 및 광자 기반의 양자 시뮬레이터 연구와 포토닉스 기반 양자컴퓨팅 기술 연구가 이루어지고 있다. KRISS, 서울대학교, KAIST 간의 컨소시엄에서는 Hubbard 모델 문제 해결을 위한 대규모 큐비트 양자 시뮬레이터 개발 연구가 진행 중이며, KIST에서는 광자 기반 솔루션을 활용하여 분자 에너지 구조 문제 해결을 위한 연구가 진행되고 있다. 또한 ETRI는 2019-2020년 기간 동안 집적형 광자 소자를 이용한 양자 프로세서 및 양자 네트워크용 확장형 양자컴퓨팅 기술을 연구하였고, 현재는 양자컴퓨팅 성능 평가 및 검증 시스템 개발을 병행하고 있다^[2].

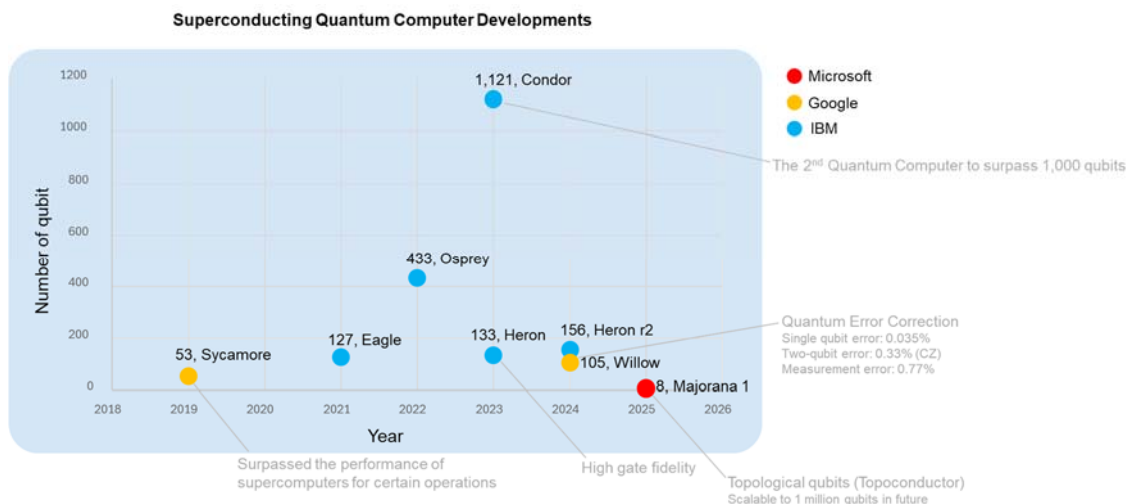


그림 10-2. 초전도 큐비트 양자컴퓨터 개발 현황

2. 초전도 큐비트 국내외 기관 기술 동향

초전도 큐비트는 반도체 공정 기술을 이용하여 구현되므로 높은 공정 성숙도와 우수한 집적 가능성, 반도체 호환성을 가져 반도체 회로 기반 양자컴퓨터의 실질적인 구현 가능 기술로 주목받고 있다. 초전도 큐비트 기반 양자컴퓨터는 개발이 지속됨에 따라 스케일(큐비트 개수)이 증가해왔으며 게이트 충실도, 오류정정 기술 등의 성능도 계속해서 발전하고 있다.

그림 10-2와 같이 2019년 Google에서 발표한 Sycamore는 53개의 큐비트를 구현하였으며 이는 특정 연산에 대해 현대 슈퍼컴퓨터를 능가하는 성능을 보였다. 2021년에는 IBM에서 127개의 큐비트를 구현한 Eagle을, 2022년에는 433개의 큐비트를 구현한 Osprey를 선보였으며 2023년에 각각 1,121개와 133개의 큐비트를 구현한 Condor와 Heron을 발표했다. Condor는 2번째로 1,000개 이상의 큐비트를 구현하여 하드웨어 기술력을 확장한 양자컴퓨터이며, Heron은 높은 게이트 연산 충실도를 보여주었다. 2024년에는 IBM에서 156개의 큐비트를 구현한 Heron r2를 발표했으며, 같은 년도 Google에서 105개 큐비트의 Willow를 발표했다. Willow는 정확한 큐비트 에러 정정률 지수가 발표되어 있다. 2025년에는 Microsoft에서 8개의 큐비트를 구현한 Majorana 1을 발표했다. 이는 향후 1백만 개의 큐비트로 확장될 수 있으며 위상적 구조로 정보를 저장하여 Topological qubits이라 불린다.

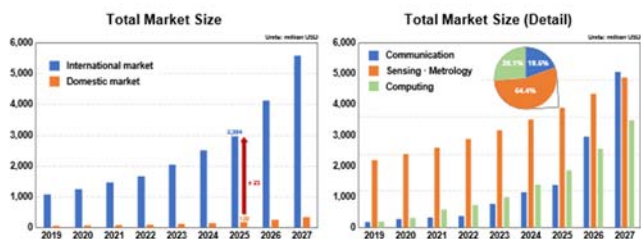
최근 전 세계적으로 양자정보통신기술 분야의 시장이 빠르게 성장하고 있으며, 특히 양자컴퓨팅(Quantum Computing) 부문이 전체 시장 성장의 핵심 동력으로 주목받고 있다. 그림 10-3은 양자컴퓨터 산업 시장의 전망 및 국내외 동향을 보여준다. IITP의 기술 로드맵에 따르면, 2027년까지 글로벌 양자컴퓨팅 시장 규모는 약 1,400억 원 수준으로 성장할 것으로 전망되며, 하드웨어, 소프트웨어, 클라우드 서비스 분야 모두에서 높은 연평균 성장률(CAGR)을 보일 것으로 예측된다^[2].

국외에서는 Google, IBM, Intel, D-Wave, Microsoft 등 주요 기업이 양자컴퓨터 하드웨어 개발과 더불어 클라우드 기반 양자컴퓨팅 서비스로 연구 범위를 확장하고 있다. Google은 Sycamore 프로세서를 통해 양자우위(Quantum Supremacy)를 달성했으며, TensorFlow Quantum과 qsim 등 양자머신러닝 및 시뮬레이션 플랫폼을 선보였다. IBM과 Microsoft는 각각 IBM Quantum System 및 Azure Quantum을 통해 클라우드 환경에서 접근 가능한 양자컴퓨팅 서비스를 운영 중이다. Intel과 QuTech은 Cryo-CMOS 기반 제어 칩과 스핀 큐비트(two-spin qubit) 연구를 병행하고 있으며, D-Wave는 양자 어닐링(quantum annealing) 클라우드 플랫폼을 상용화하였다. 이처럼 해외 주요 기업들은 하드웨어-소프트웨어 통합형 생태계를 중심으로 클라우드 접근성, 양자 시뮬레이션, Cryo 제어 기술 분야에 주력하고 있다.

Market Forecast by Segment for Quantum Information and Communication Technologies

		2019	2020	2021	2022	2023	2024	2025	2026	2027	CAGR
Quantum Communication	International	80.2	121.7	142.9	158.6	323.6	481.5	581	1,238	2,708	50.5%
	Domestic	4.2	6.5	9.3	10.5	21.9	33.2	40.8	88.8	154	37.0%
Quantum Sensing / Metrology	International	919	1,000	1,091	1,207	1,326	1,483	1,623	1,810	2,032	10.4%
	Domestic	85.6	72.8	81.1	91.5	102.5	115.4	130.5	148.5	169.9	12.6%
Quantum Computing	International	84.8	133.7	245.4	312	415.5	585.7	780.1	1,075	1,454	42.6%
	Domestic	1.02	1.74	3.49	4.68	6.85	9.96	14	19.9	27.6	31.0%
Total	International	1,084	1,255	1,483	1,676	2,065	2,531	2,944	4,123	5,394	22.8%
	Domestic	70.6	81	93.9	106.7	131.1	158.5	185.3	257.1	351.6	22.2%

Source: ICT R&D Technology Roadmap 2025, ITR.



Market Forecast for Quantum Computing

		2019	2020	2021	2022	2023	2024	2025	2026	2027	CAGR
Hardware	International	70.8	107.4	185.4	223	274.7	358.6	447.7	573.2	883.5	33.0%
	Domestic	0.93	1.40	2.73	3.35	4.40	6.22	8.50	12.96	13.18	40.8%
Software	International	13.2	23.9	45.0	72.2	109.4	163.9	238.0	342.8	494.0	36.1%
	Domestic	0.16	0.31	0.63	1.08	1.75	2.79	4.28	6.34	9.20	66.1%
Cloud Service	International	0.8	2.4	9.0	16.8	31.4	56.2	94.4	158.6	276.3	107.6%
	Domestic	0.01	0.03	0.13	0.25	0.50	0.96	1.70	2.93	5.25	119.8%
Total	International	84.8	133.7	249.4	312.0	415.5	585.7	780.1	1,075	1,454	42.6%
	Domestic	1.02	1.74	3.49	4.68	6.85	10.0	14.0	19.9	27.6	31.0%

Source: ICT R&D Technology Roadmap 2025, ITR.



International research trends

Classification	Research Trends
HUAWEI 2018	- Launched a 42-qubit quantum-computing simulation cloud service.
Google 2019	- Achieved quantum supremacy by verifying random quantum-circuit sampling on a 53-qubit processor.
Volkswagen 2019	- Tested real-world traffic-optimization problems on a quantum computer.
intel 2019	- Developed Cryo-CMOS chips for quantum-processor control. (with Microsoft & QuTech)
intel 2020	- Unveiled a two-spin-qubit gate operating at 1.1 K.
NIST 2020	- Performed quantum simulation of quantum-magnetism models with alkaline-earth-metal atoms.
MPQ 2020	- Demonstrated quantum simulation of the Hubbard model using alkali-atom platforms.
Google 2020	- Released software quantum-simulators—TensorFlow Quantum for quantum machine learning and qsim.
D-WAVE 2020	- Introduced the Leap 2 quantum-annealing cloud platform.

Domestic research trends

Classification	Research Trends
KAIST 2018	- Development of a Rydberg-atom quantum simulator and 20-qubit-scale quantum-ising model computations.
Research on quantum simulators centered on atomic and photonic platforms is underway.	
KRIS 2019	- Research on large-scale qubit quantum simulators for solving the Hubbard model.
KIST 2019	- Photonic-based researches on molecular energy-structure problem solving.
KRIS 2019	- Ytterbium-atom quantum platform development for Transmon physical-qubit CNOT dual-qubit gates and quantum simulation of the Chem-Hubbard lattice model.
KIST 2019	- Fabrication of diamond color-center physical-qubit nanostructures and spin-resonance measurements.
KAIST 2019	- Development of a sodium-atom quantum platform implementing dual-qubit operations in Si and GaAs quantum-dot physical qubits, and simulating strongly correlated and topological systems.
ETRI 2019	- Research on new quantum algorithms via machine learning and optimization of high-reliability quantum hardware.
ETRI 2019	- Research on photonic-integrated-circuit quantum processors and network-based quantum-computing scale-up technologies.
KAIST 2020	- Development of a lithium-atom quantum platform for spin-model quantum simulations.
ETRI 2020	- Development of a quantum-computing performance evaluation and verification system.

Main service offered by International quantum computing companies

Classification	Main Service
IBM USA	- Provides more than 15 quantum computer systems. - Offers services such as the IBM Q Hub, enabling case studies and application opportunities. - Major clients: Daimler, Wells Fargo, Delta Air Lines.
Google USA	- Provides various platforms to support algorithm development and diverse research opportunities. - Simultaneously advancing hardware development and preparing cloud-based services.
Microsoft USA	- Scheduled to initiate cloud-based quantum computing research services via Azure.
amazon USA	- Preparing to provide quantum computing services via AWS (Amazon Web Services). - the world's largest cloud service
intel USA	- Developing semiconductor-based quantum processors. - Preparing to offer cloud-based quantum processor services. (co-developed with QuTech) - Providing open-source development tools and services for researchers.
IONQ USA	- Offering practical quantum computing services with trapped-ion quantum processors. - Major clients: Microsoft (Azure), Amazon (AWS)
PsiQuantum USA	- Developing photonic-based quantum processor hardware and software for quantum computing implementation
Alibaba China	- Providing cloud-based quantum computing services enabling direct execution of computations via Alibaba Cloud (since 2018)

Main service offered by domestic quantum computing companies

Classification	Main Service
LG전자	- Provides cloud access to IBM Quantum processors and the Qiskit SDK for developing quantum applications in AI, IoT, and related domains.
kt cloud	- Integrates Microsoft Azure Quantum into its cloud platform, enabling enterprises to provision Azure Quantum workspaces and execute quantum workloads.
QUANOVA COMPUTING	- Offers hardware-agnostic HI-VQE hybrid algorithms for quantum-chemistry simulations, achieving chemical accuracy on NISQ devices.
sd	- Supplies quantum-grade ODM hardware (Qubit Controller, Time-Tagging, and Coincidence Counting units) and is developing superconducting, neutral-atom, and diamond-NV quantum computers for commercial cloud deployment.
KQC Korea Quantum Computing	- Korea's official IBM Quantum partner, delivering cloud access to IBM Quantum systems and planning an on-premises IBM Quantum System Two in Busan by 2028 with IBM watsonx AI integration.

그림 10-3. 양자컴퓨터 산업 시장 전망 및 국내외 동향

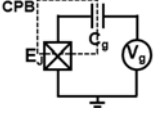
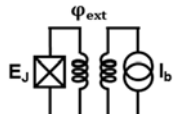
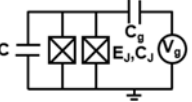
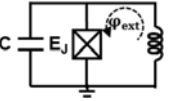
국내에서는 KAIST, KRISS, KIST, ETRI 등 주요 연구기관이 원자·광자 기반 양자 시뮬레이터와 양자 검증·평가 시스템 개발을 중심으로 연구를 수행하고 있다. ETRI는 포토닉스 집적회로(Photonic Integrated Circuit)를 기반으로 한 확장형 양자컴퓨팅 기술 및 성능 검증 시스템 연구를 진행 중이며, KRISS와 KAIST는 Hubbard 모델 해석을 위한 양자 시뮬레이터 및 양자소자 플랫폼 개발에 집중하고 있다. 국내 기업 중에서는 LG전자, KT, 퀀텀바이오(Qunova), SDT, KQC 등이 IBM Quantum 시스템 및 양자 SDK를 활용한 AI·보안·클라우드 연계 서비스 개발을 추진하고 있다. 종합적으로, 해외 연구는 상용화와 클라우드 확장 중심의 플랫폼 구축에 초점을 두고 있는 반면, 국내 연구는 광자·원자 기반의 양자 시뮬레이터 및 시스템 검증 인프라 구축 중심으로 진행되고 있다. 향후에는 Cryo-CMOS, 양자 네트워킹, 양자 보안 등 시스템 통합형 기술 개발과 산업 응용 확장이 주요 연구 방향으로 이어질 것으로 예상된다.

3. 초전도 큐비트 양자컴퓨터 기술 발전 방향

(1) 초전도 큐비트 회로

초전도 큐비트의 종류는 회로 구성에 따라 다음과 같이 전하 큐비트, Flux 큐비트, Transmon 큐비트, Fluxonium 큐비트가 있다(표 10-3 참조). 가장 기본적인 형태의 전하 큐비트 회로는 축전기와 조셉슨 접합으로 이루어진 회로이다. 고속 제어가 가능하고 다중 큐비트 시스템에 적합하지만 전하 잡음에 민감하고 큐비트 간 cross-talk 등이 더 많이 발생하여 큐비트 제어 복잡성이 크다. Flux(자속) 큐비트는 처전도 루프에 흐르는 전류가 조셉슨 접합에 의해 방해받는 구조이다. 빠른 게이트 연산이 가능하나 Transmon 큐비트는 조셉슨 접합의 에너지가 축전 에너지보다 크도록 설계하여 전하 잡음 민감도를 최소화하고 결맞음 시간을 늘린 큐비트이다.

표 10-3. 초전도 큐빗의 종류

Superconducting Qubits			
Qubit Type	Structure	Advantages	Disadvantages
Basic	Charge Qubit Josephson junction with a cooper pair box (CPB) 	• High bandwidth frequency • Suitable for multi-qubit systems	• High sensitivity to charge noise • Complexity of qubit control • Higher potential for cross-talk between qubits
	Flux Qubit Josephson junction with a loop structure 	• Long coherence times • High anharmonicity	• Difficulty in control and fabrication • High sensitivity to flux noise
Derived	Transmon Qubit Josephson junction with a large shunt capacitor 	• Long coherence times • Easy control • Scalability • Low sensitivity to noise	• Low anharmonicity
	Fluxonium Qubit Josephson junction with a capacitor and large superinductor 	• Long coherence times • High anharmonicity • Low sensitivity to flux noise	• Difficulty in control and fabrication • Low bandwidth frequency

* E_J = Josephson Energy

* V_g = External Voltage

* C = Capacitance

* C_J = Capacitance connected in parallel with the Josephson junction

* C_g = Gate Capacitance

* I_b = Bias Current

* ϕ_{ext} = External Magnetic Flux

(2) 초전도 큐빗 제어회로 및 센싱회로 구조

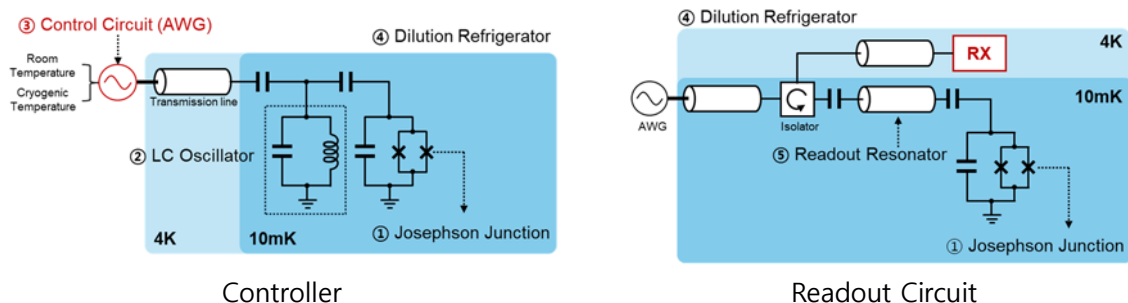


그림 10-4. 큐빗 제어회로 및 센싱회로의 구조

초전도 큐빗을 제어하기 위해서는 제어 신호를 생성하고 전달하는 제어회로(Controller)와, 큐빗

의 상태를 읽어오고 판별하는 센싱회로(Readout Circuit)가 필요하다.

제어 및 센싱회로를 이루는 구성요소는 그림 10-4에 표시되었다. 큐비트의 상태를 조절하고 측정하기 위해서는 큐비트를 이루는 조셉슨 접합과 이를 초전도 상태를 유지하기 위한 극저온 냉각 장치가 필요하며, 이 외에도 펄스 생성 및 송신을 위한 전자회로, LC Oscillator, 리드아웃 공진기 등이 필요하다. AWG(Arbitrary Waveform Generator)를 포함하는 상태 제어 전자회로는 LC oscillator를 통해 마이크로파를 전송하고, 공진 주파수를 통해 큐비트 상태를 제어할 수 있도록 한다. 또한 마이크로파 펄스를 다시 생성하고 전달하고 리드아웃 공진기를 통해 돌아온 반사파를 수신단 회로를 통해 계측하여 큐비트의 상태를 간접적으로 측정할 수 있다.

(3) 큐비트 제어 방법

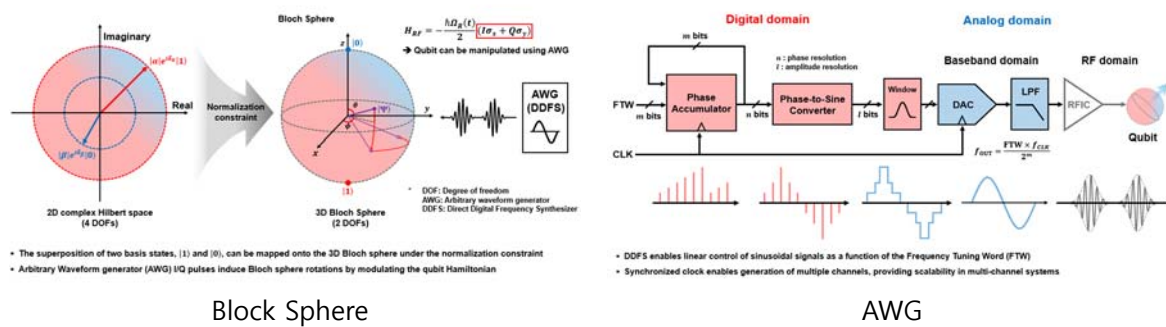


그림 10-5. 큐비트 제어의 기본 원리

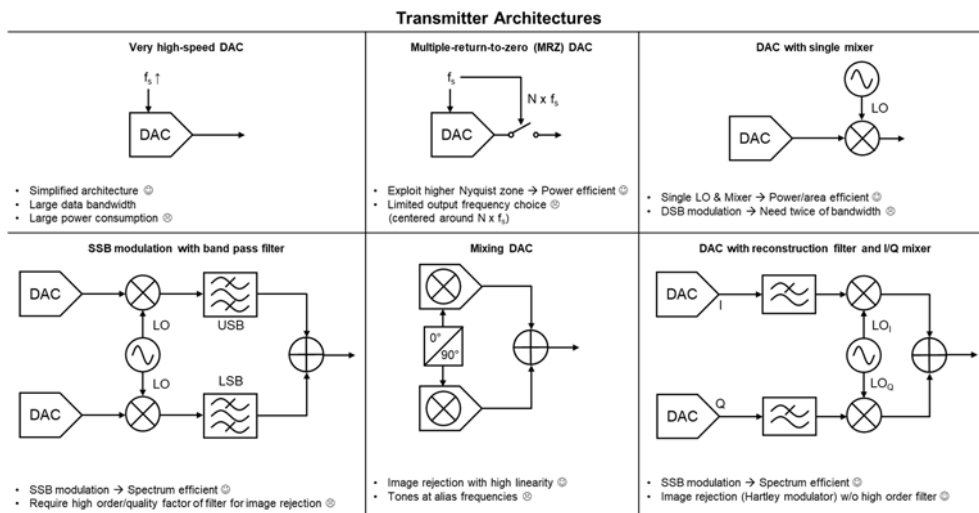


그림 10-6. 큐비트 제어기의 블록 다이어그램^[3]

큐빗은 중첩 상태로 존재하기 때문에 4개의 자유도를 가지지만, 전체 확률은 1보다 작다는 정규화 조건과 전역 위상은 큐빗 상태 측정 결과에 물리적인 영향을 주지 않음을 고려하면 큐빗의 상태는 Bloch Sphere 위의 한 점으로 시각화할 수 있다. 구면 위 벡터의 방향이 큐빗의 확률 진폭과 위상을 나타낸다. 따라서 그림 10-5와 같이 AWG로 마이크로파펄스를 발생시켜 큐빗 상태 조절이 가능하다.

제어회로에서 디지털 방식으로 마이크로파 펄스를 생성하는 경우 신호 전송을 위한 송신기 구조는 기본적으로 디지털 신호를 아날로그 신호로 변환하기 위한 DAC(Digital-to-Analog Converter)를 포함하며 경우에 따라 잡음 신호를 차단하는 필터, 신호를 원하는 주파수 대역으로 보내기 위한 믹서, 이에 필요한 주파수를 생성하는 LO(Local Oscillator) 등으로 구성된다. 송신기 구조로는 그림 10-6과 같은 구조들을 채택할 수 있다^[3].

우선 고속 DAC 구조는 넓은 대역폭을 가진 DAC를 사용하여 간단한 구조로 구현이 용이하나 전력소모가 크다는 단점이 있다. Multi-return-to-zero(MRZ) DAC를 사용하는 구조는 고차 Nyquist 대역을 활용하여 고속 DAC보다 낮은 샘플링 속도를 사용할 수 있으므로 전력 소모를 줄일 수 있다. 그러나 불필요한 주파수 성분이 발생하며, 선택할 수 있는 출력 주파수 대역이 제한된다는 한계점이 있다. 이에 비해 DAC와 함께 단일 LO 및 믹서를 사용하는 구조는 하나의 RF 채널을 통해 각각의 큐빗 주파수 채널에 대한 출력 신호를 생성할 수 있다. 또한 단일 LO 및 믹서를 사용하여 전력 및 칩 면적 효율을 기대할 수 있으나, Double-sideband(DSB) 방식을 사용하여 DAC에서 실제 필요한 대역폭보다 큰 샘플링 주파수를 필요로 하게 된다. Single-sideband(SSB) 방식을 사용하면 DSB 방식에 비해 양측 대역 중 필요한 대역만 합성하여 출력하므로 동일 신호 생성 시 필요한 샘플링 주파수를 절반으로 줄일 수 있다. 그러나 이미지 신호 제거를 위해 고차 및 고품질 필터 설계가 필요하다. Mixing DAC를 사용하는 구조는 우수한 선형성으로 이미지 신호를 제거할 수 있으나 에일리어싱 주파수로 인한 불필요한 톤이 함께 생성된다. DAC 뒤에 reconstruction 필터를 추가하고 I/Q 믹서를 사용하는 구조를 통해 고차 필터 설계 없이 SSB 방식을 구현할 수 있다.

다중 큐빗 제어로의 확장성

실제 양자 프로세서에는 수십~수천 개의 큐빗이 동시에 동작해야 하므로, 앞서 설명한 큐빗 제어 기술을 바탕으로 다중 큐빗 제어로 확장되어야 한다. 이는 여러 큐빗을 병렬로 제어하면서도 각 큐빗 간의 간섭이나 주파수 중첩을 최소화하는 기술이 포함된다.

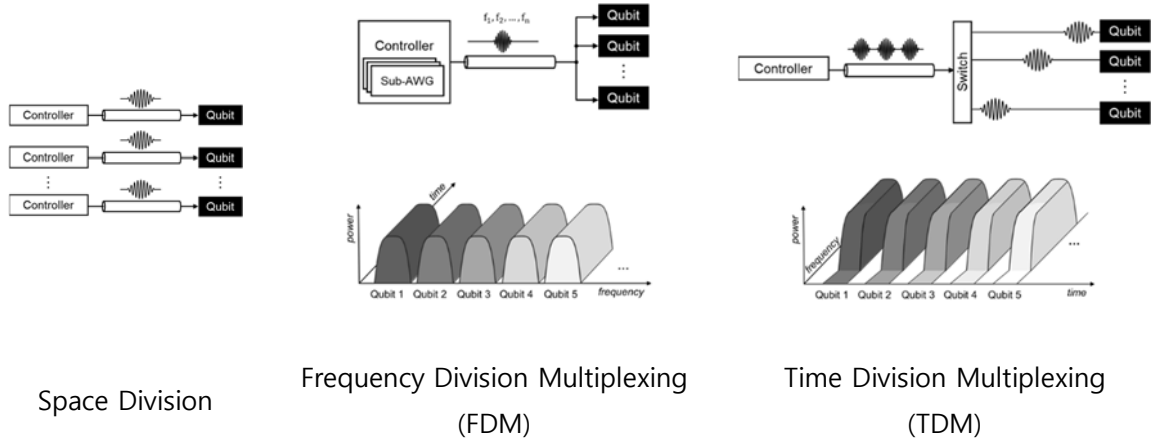


그림 10-7. 다중 큐비트 제어 방식

여러 개의 큐비트를 제어하는 방법으로는 그림 10-7에서 보이듯이 크게 3가지 방법이 있다. 첫 번째로 공간 분리(Space Division) 방법은 각 큐비트에 대한 driveline을 분리하여 신호를 전송하는 방법이다. 단일 큐비트 제어를 단순히 여러 개로 확장시킨 구조로, 설계 요구조건이 비교적 간단하나 큐비트와 controller 사이에 케이블 연결이 많아져 잡음과 cross-talk 발생이 커질 수 있고 확장성 측면에서 불리하다.

두 번째로 주파수 분할 다중화(FDM) 방법은 각 큐비트마다 제어(공진) 주파수를 다르게 하여 여러 개의 큐비트에 하나의 controller 및 driveline을 연결하는 방법이다. Controller는 각 큐비트의 공진 주파수에 맞는 다중 주파수가 합성된 제어 신호를 전송하고, 이 신호는 주파수에 따라 목표 큐비트에 전달되어 상태 제어가 가능하다. 이 방법은 동시에 다중 큐비트 제어가 가능한 공간 분리 방법의 장점을 유지하면서도, 하나의 driveline을 사용하여 케이블 연결 수를 최소화할 수 있다. 그러나 여러 주파수를 동시에 전송하므로 스펙트럼이 퍼지고 겹치면서 채널 간 간섭과 불필요한 상태 전이가 발생할 수 있다. 이를 방지하기 위해 pulse shaping 기법을 필요로 한다.

여러 큐비트에 하나의 controller 및 driveline을 연결하는 또다른 방법으로 시간 분할 다중화(TDM) 방법이 있다. 하나의 controller에서 시간 차이를 두고 여러 큐비트에 대한 신호를 하나씩 생성하여 전송한다. 하드웨어 리소스를 줄일 수 있으나, 양자 연산 실행 시 소요 시간이 늘어날 수 있다.

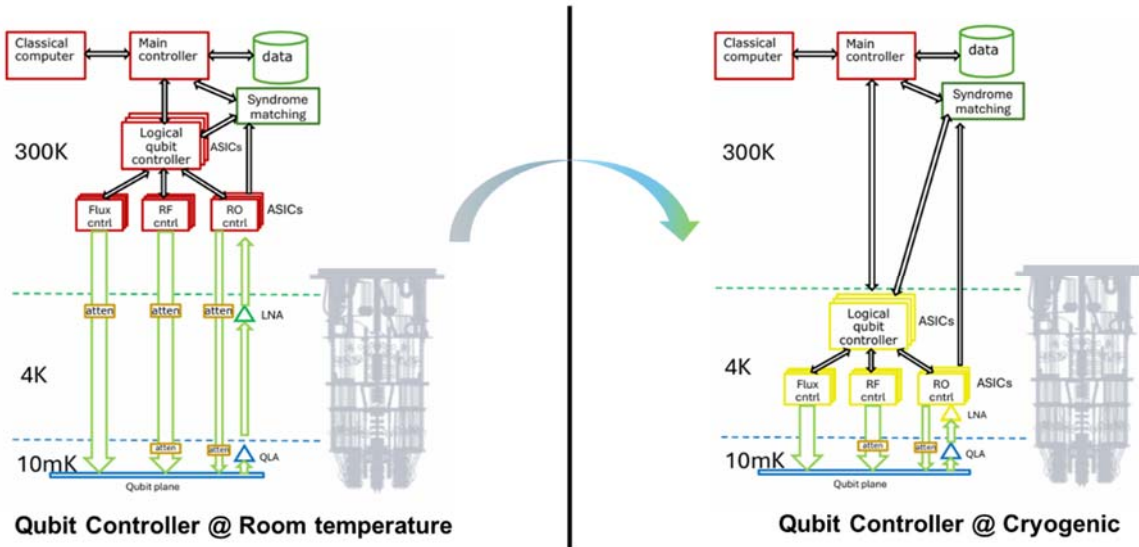


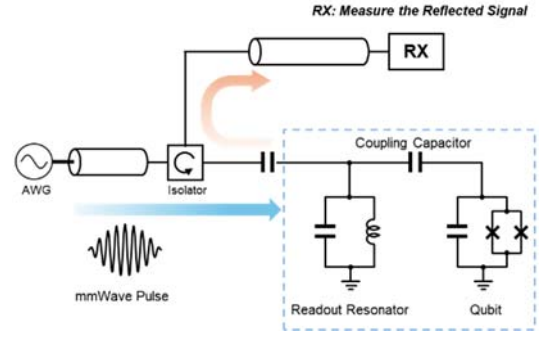
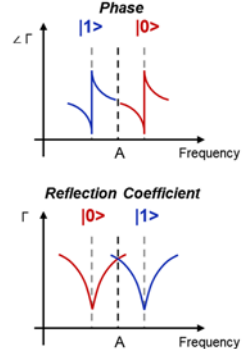
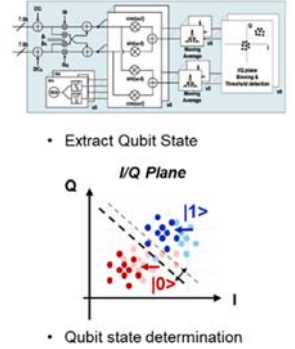
그림 10-8. 극저온 큐비트 컨트롤러^[4]

큐비트 컨트롤러의 극저온 환경 배치 필요성

초기 양자컴퓨터 시스템에서는 큐비트 컨트롤러와 전자 장비를 모두 상온에 두고 동축 케이블을 통해 극저온 환경에 위치한 큐비트를 구동하는 구조가 일반적이었다. 이러한 상온 컨트롤러는 구현이 비교적 간단하고 전력 소모 제약이 완화된다는 장점이 있으나 케이블에서 발생하는 신호 왜곡, 상온에서부터 발생하는 열 잡음에 의한 SNR 감소 등의 한계가 있다. 이를 극복하기 위해 최근에는 큐비트와 물리적으로 가까운 극저온 환경에서 동작 가능한 Cryo-CMOS 기반 컨트롤러 연구가 활발히 진행되고 있다(그림 10-8 참조). 극저온 컨트롤러는 케이블 연결 문제를 해결하여 신호 왜곡과 SNR 감소를 방지하며, 대규모 큐비트 집적화에 유리하다. 그러나 전체 시스템을 극저온에서 동작시키기 위해서는 전력 소모를 매우 낮춰야 하므로, 향후 저전력 컨트롤러에 대한 연구가 지속적으로 진행되어야 한다.

(4) 큐비트 측정 방법

표 10-4. 큐비트 리드아웃 방식: 공진기를 이용한 반사파 측정

Dispersive Readout [5],[7]		
- Qubit states are indirectly measured through the transmission and reflection measurement of mmWave pulses Readout resonator is connected indirectly through a coupling capacitor Resonance frequency of the readout resonator is shifted depending on the qubit state - After I/Q demodulation, we can measure qubit state		
Readout Mechanism	Reflected Signal	Measure Qubit State
		 • Extract Qubit State • Qubit state determination

큐비트 리드아웃 방식 중 가장 대표적인 것은 마이크로웨이브 펄스를 공진기에 보내서 반사파를 활용하는 것으로, 큐비트 상태에 따라서 변화하는 공진기의 특성을 감지한다. AWG에서 생성한 마이크로파 펄스를 큐비트와 약하게 커플링시킨 공진기에 보내면 큐비트 상태에 따라 공진기의 공진 주파수가 미세하게 달라진다. 공진기의 유효 공진 주파수가 미세하게 변하면서 반사 특성 역시 변하게 되고, 이로 인해 반사파에서는 위상, 진폭 변화가 발생한다. 수신기에서는 반사된 신호를 처리하고 I/Q demodulation으로 복원하면 위상, 진폭 정보를 추출하여 큐비트의 상태를 판별할 수 있다(표 10-4, 10-5 참조).

표 10-5. 반사파 측정 방식 리드아웃 기술의 주요 구성요소와 역할

Key Components of Qubit Readout for Superconducting Qubit	
① Parametric Amplifier	JPA (Josephson Parametric Amplifier) or TWPA (Traveling-wave Parametric Amplifier) Quantum-limited amplification with a gain, Cryogenic, near-quantum-limited amplification of reflected signal.
② Low-noise Amplifier	HEMT LNA (Low-noise Amplifier) Further amplification with minimal added noise (typically at 4 K or RT).
③ Mixer	Down-conversion of RF signal to IF or baseband using a local oscillator.
④ ADC	Digitization of the analog IF/baseband signal for digital processing.
⑤ I/Q Demodulation	Extraction of I/Q components to analyze qubit state via amplitude/phase.

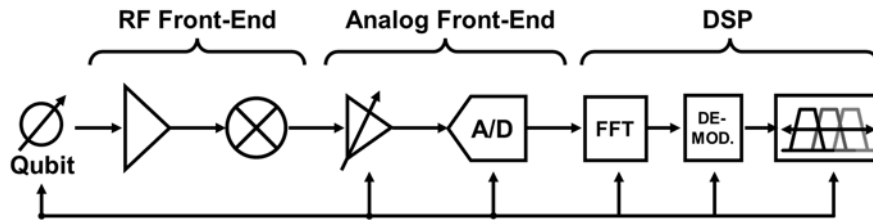


그림 10-9. 반사파 리드아웃 체인 구성

반사파 측정 방식 리드아웃 기술의 주요 구성요소와 역할은 다음과 같다. 그림 10-9에 도시된 리드아웃 체인은 큐빗에서 반사되어 돌아온 미세한 마이크로파 신호를 증폭하고, 이를 분석 가능한 형태로 변환하여 큐빗의 상태를 판별하는 기능을 수행한다.

가장 앞단에는 Parametric amplifier가 위치하며, 극저온 환경에서 반사된 신호를 양자한계(quantum limit) 수준까지 증폭한다. 이 증폭기는 양자역학적으로 허용되는 최소한의 잡음만을 추가하므로, 리드아웃 신호의 SNR을 극대화하는 데 필수적이다. 일반적으로는 Josephson Parametric Amplifier (JPA) 또는 Travelling-wave Parametric Amplifier (TWPA) 구조가 사용된다.

다음 단계에서는 저잡음 증폭기(LNA, Low-Noise Amplifier)가 배치되어 리드아웃 체인 전체의 SNR을 유지하면서 상온의 리드아웃 회로까지 신호를 안정적으로 전달하기 위한 충분한 이득을 제공한다. 이후 Mixer 단계에서는 고주파 신호를 중간주파수(IF)로 하향 변환하여 신호의 위상과 진폭 정보를 추출하기 쉽게 만든다. 이 과정을 통해 고속 ADC를 사용하지 않고도 상대적으로 낮은 샘플링 속도의 ADC로 신호를 획득할 수 있다. 변환된 신호는 ADC를 통해 디지털 도메인으로 변환된 뒤, DSP 블록에서 필터링, 평균화, 판별 등의 처리가 수행된다.

다수의 큐빗을 동시에 읽어내기 위해서는 주파수 분할 다중화(FDM, Frequency Division Multiplexing) 기술이 적용되어, 각 큐빗의 리드아웃 주파수를 분리하여 병렬적으로 판독할 수 있다. 마지막으로, I/Q 복조(I/Q Demodulation)는 반사된 신호의 위상과 진폭 성분을 분리함으로써 큐빗의 상태($|0\rangle$ 또는 $|1\rangle$)를 구분하는 역할을 한다. 이 복조 과정은 아날로그 도메인에서 수행되는 경우 ADC 이전에 이루어지며, 디지털 도메인에서는 DSP 단계에서 동일한 연산으로 구현될 수 있다.

표 10-6. 반사파 측정 방식의 최근 기술 동향

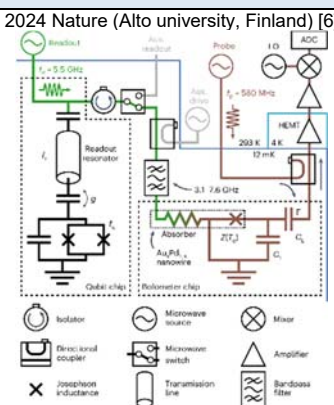
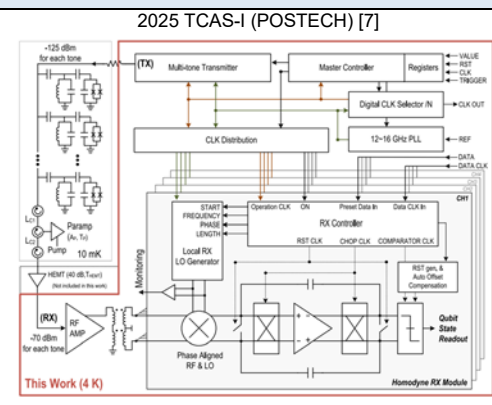
Trends of Readout Technique			
2024 Nature (Alto university, Finland) [6]  Legend: Isolator, Directional coupler, Josephson inductance, Microwave source, Microwave switch, Transmission line, Bandpass filter, Mixer, Amplifier.		2025 TCAS-I (POSTECH) [7]  This Work (4 K)	
Conventional	Parametric amplifier (4-8 GHz)	Conventional	- Heterodyne receiver - ADC + DSP (FFT Processor)
Proposed	- Bolometer (Sub-GHz Probe Tone) - Low power - Scalability ↑	Proposed	- Without ADC and DSP - SNR ↑

표 10-6은 초전도 큐비트의 반사파 측정 방식의 최근 기술 동향을 보여준다. 2024년 Nature^[6]에 발표된 기술로는 볼로미터(들어온 마이크로파 신호의 흡수 에너지를 감지하는 소자) 기반 리드아웃 방식이 있다. 기존 Parametric amplifier의 구조보다 확장성 측면에서 유리하며, Parametric amplifier는 전압 증폭이었던 반면 볼로미터는 에너지, 즉 전력을 직접 흡수해 온도로 측정하여 잡음 역시 낮출 수 있다. 측정 결과 단일 큐비트에서 fidelity는 62% 이며, fidelity 99% 달성 및 다중 큐비트로의 확장성을 위한 설계 목표 수치가 이론 분석이 제시되어 있어 향후 새로운 리드아웃 방식 개발의 기반이 될 것이다.

2025년 TCAS-I^[7]에서는 아날로그 Homodyne 구조의 Cryo-CMOS 리드아웃 기술이 발표되었다. 기존 Heterodyne receiver 구조에서는 상온의 ADC, DSP 사용으로 인해 큐비트에서 RT 장비까지의 케이블 연결에 의한 지연 시간 및 decoherence 발생 문제가 있었다. 이를 해결하기 위해 ADC와 DSP 없는 Homodyne 구조의 리드아웃 회로를 개발하였으며, 전력 소모 문제 역시 개선하였다.

4. 양자컴퓨팅 반도체 로드맵

표 10-7. 큐비트 컨트롤러/리드아웃 State-of-the-art

	Intel 2020 JSSC [8]	Google 2023 JSSC [9]	Tsinghua University 2023 JSSC [11]	POSTECH 2025 JSSC [11]	POSTECH 2025 TCAS-I [7]	Tsinghua University 2024 ISSCC [12]
Architecture	Controller (SSB)	Controller (Direct Conversion)	Controller (Polar Modulation)	Controller (SSB)	Receiver (Homodyne)	Controller ASIC + Readout ASIC
Technology	Intel 22nm FinFET	28nm Bulk	28 nm Bulk	40nm bulk CMOS	40nm CMOS	28nm Bulk CMOS
Temp.	3 K	3 K	3.5 K	3.5 K	4 K	3.5 K
Number of qubit	32 / channel	2-qubit patch	2	4 / channel (Total 5-8 qubits in practical)	1 / channel	2
Channel	4	2 XY, 2 Z, 1 g	2 XY, 2 Z	4	4 (FDM)	2 XY, 4 Z (Controller) 2 (Readout)
SFDR	> 42 dB	-	> 40 dB	> 40 dBc	-	-
Noise Temperature / Noise Figure	-	-	-	-	NT: 440 K (RT Sim)	NF: 1.6 dB
Fidelity	-	-	-	-	93 % (circuit- only fidelity)	0>: 94.8 % 1>: 92.6 %
Power	12 mW / qubit	< 4 mW / qubit	13.7 mW / qubit	6.5 mW (Sine with DRAG) 9.1 mW (Rased-cosine with DRAG) 4.2 mW (Rectangular)	(TX) 0.9 mW / qubit (RX) 2.4 mW / qubit	(TX) 6.3 mW / qubit (RX) 11 mW / 2 qubits
LO generation	External (6 GHz – 14 GHz)	External	Injection Locking (4 – 6 GHz)	PLL (4.5 – 8 GHz)	DDS + PLL (4 – 6 GHz)	External (TX: 4 – 6 GHz) (RX: 5 – 7 GHz)
Pulse shaping	- Rectangular - Cosine - Raised Cosine - Triangular - Gaussian	Raised cosine, DRAG	-	- Sine with DRAG - Raised- cosine with DRAG - Rectangular	-	DRAG

	2016-2020	2021-2023	2024-2025	2026-2028	2029-2031	2032-2035	2036-2040
Hardware Innovation	초전도·스핀 큐비트 검증 및 안정화	Cryo CMOS 제어기 고도화 Error mitigation 하드웨어	1k급 칩 Condor/Willow 공개 Cryo Readout 통합	수천 큐비트 모듈형 QPU (Nighthawk) / Multiplexing 기술 확립	200 Logical Qubit FTQC 프로토타입 완성	10 ⁴ 논리 큐비트 시스템 안정화	수십만 큐비트급 Quantum-Centric Supercomputer 통합
Error Correction & Control Systems	게이트 충실도 검증 결맞음 시간 확보	Error Mitigation 및 Dynamic Circuit 기술	Surface / LDPC 코드 적용 시작	QEC 자동화 Resilient Level 진입	완전 패시형 오류보정 루프 / FTQC 완성	코드 최적화 및 표준화	QEC 운용 기반 상용 플랫폼 정착
Software Innovation & Quantum Software Stack	Qiskit-Cirq 등 SDK 출시	Qiskit Runtime / Quantum Serverless 환경 구축	QEC Library 및 Logical Simulator 도입	자동화 컴파일 / 리소스 매핑 SW 발전	Fault-tolerant ISA 확립	Quantum Algorithm 최적화 / Compiler 표준화	AI-Quantum 통합 소프트웨어 스택 완성
Workload Orchestration & Applications	클라우드 원격 실험 (IBM QE 초기 서비스)	하이브리드 연산 (QPU + HPC) 프레임워크 정착	응용별 벤치마크 / QPU-HPC 통합 실험	다중 QPU 연산 및 자동 스케줄링	클라우드 리소스 스케줄러 및 하이브리드 실행 완성	화학·재료·AI 응용 양자우위 확립	Quantum-Centric HPC 서비스 상용화
요약 단계	기초 검증 및 SDK 형성	오류완화 및 하이브리드 연산 정착	1k급 칩·초기 QEC 실험	수천 큐비트 확장 자동 보정 진입	FTQC 1세대 완성	양자우위 응용 정착	Quantum-Centric 시대 도래

그림 10-10. 양자컴퓨팅 기술 발전 양상 예상

표 10-7은 초전도 큐비트 컨트롤러/리드아웃 회로의 최신 연구 성과들을 비교한 것이다. 컨트롤러 성능 측면에서는 일부 연구에서 SFDR 40 dB 이상의 성능이 보고되었고, 리드아웃 회로의 경우 회로 단독 93 %의 정확도 등이 측정되었다. 전력 소모는 구현 방식 등에 따라 큐비트당 수 mW에서 10 mW으로 보고되고 있다. 전반적으로 집적도 향상 및 저전력, 고신뢰도 동작을 동시에 달성하는 것이 차세대 큐비트 시스템을 위한 핵심 과제로 정리된다.

그림 10-10은 현재까지의 양자컴퓨팅 기술 발전 현황과 앞으로의 발전 양상 예상도를 보여준다. 2016~2023년에는 초전도·스핀 큐비트의 검증 및 안정화가 이루어지고, IBM의 Falcon, Google의 Sycamore가 개발되었다. Qiskit, Cirq, Q# 등 주요 SDK가 출시되며 초기 개발 환경이 형성되었고, IBM은 Qiskit Runtime과 Quantum Serverless 환경을 구축해 양자-고전 하이브리드 연산 체계를 확립하였다. 2024~2025년에는 양자 프로세서의 집적도와 신뢰성이 향상되며, IBM은 Heron(133Q), Condor(1,121Q) 칩을 공개하여 Cryo-Readout 및 제어 통합을 구현하였다. Google은 Surface/LDPC 코드 기반의 논리 큐비트 실험을 진행하고, Microsoft와 Quantinuum은 QEC Library 및 논리 시뮬레이터를 도입하였다. QPU-HPC 통합과 응용별 벤치마크 실험이 본격화되어 오류 관리가 가능한 중규모 양자 연산이 실현되었다.

2024년 이후에는 초전도 큐비트의 확장성과 집적도를 대폭 향상시키는 데 초점을 맞추고 있으며, 2024년에는 133큐비트급 Heron 프로세서가, 2025년에는 1,121큐비트의 Condor 프로세서가 공개되었다. 이 시기를 기점으로 양자 프로세서의 신뢰성과 제어 효율이 크게 개선되었으며, Cryo-CMOS 기반 제어 회로와 Cryogenic 리드아웃 기술이 통합되어 안정적인 양자-고전 연산이 가능해졌다.

2026~2028년에는 수천 큐비트급 모듈형 아키텍처인 Nighthawk 시리즈가 등장하여, 고충실도 커플러를 통해 여러 칩을 연결하는 다중칩 구조를 구현할 것으로 예상된다. 이 단계에서 자동화된 오류보정(QEC)과 Resilient Level 운영 기술이 확립되며, 양자 시스템의 규모 확장과 연산 안정성이 동시에 확보될 전망이다. 2029~2031년에는 약 200개의 논리 큐비트를 안정적으로 운용하는 내결함(Fault-Tolerant) 양자컴퓨터가 실현될 것으로 보이며, 양자 명령어 집합(ISA)과 클라우드 기반 리소스 스케줄러가 표준화되어 상시 운용이 가능한 양자 시스템이 구축될 것으로 예상된다. 2032~2035년에는 1,000 논리 큐비트 이상의 대규모 시스템이 안정화되어, 화학, 재료, 인공지능 최적화 등에서 고전 슈퍼컴퓨터를 능가하는 양자우위가 입증될 전망이다. 이와 함께 AI-Quantum 통합 소프트웨어 스택이 완성되어 양자-고전 연산이 자율적으로 최적화되는 지능형 하이브리드 구조가 확립을 기대할 수 있다. 2036~2040년에는 수십만 큐비트 이상의 초전도 기반 Quantum-Centric 슈퍼컴퓨터가 완성되며, HPC 및 AI 인프라와의 통합을 통해 범용 양자 플랫폼으로 발전할 것으로 예상된다. Cryogenic 칩렛 기술이 상용화되면서 양자 SoC 생산이 가능해지고, 글로벌 데이터센터에는 양자 코프로세서(QPU)가 탑재되어 클라우드 기반 Quantum-Centric Computing 시대가 본격적으로 도래할 것으로 전망된다.

5. 요약

양자컴퓨팅은 큐비트의 양자역학적 현상을 제어해 확률적, 가역적 연산으로 고전 컴퓨터 대비 훨씬 우수한 성능과 연산 속도를 낼 수 있는 차세대 컴퓨팅 기술이다. 큐비트 종류 중 초전도 큐비트는 반도체 공정 친화성과 집적 용이성, 빠른 게이트 연산 속도로 산업계와 학계에서 매우 활발히 연구되고 있다. 세계 시장의 성장과 함께 국내 초전도 큐비트 양자컴퓨터 또한 연구와 개발이 지속적으로 이루어지고 있으며, 이의 상용화를 위해서는 초전도 큐비트 자체의 개발 뿐만 아니라 극저온 환경 큐비트 제어 및 리드아웃 회로의 개발 또한 필수적이다.

IBM 등 세계 주요 기업은 2030년대 초반까지 수백만 물리 큐비트와 및 수천 논리 큐비트 규모의 완전 오류정정형 양자컴퓨터 완성을 목표로 하고 있다. 향후 초전도 큐비트 기반 집적화, Cryo-CMOS 제어 및 리드아웃 회로, 양자 네트워킹 등 시스템 통합 연구가 양자컴퓨팅의 산업화 핵심 요소가 될 것이다.

참고문헌

- [1] Olivier Ezratty, Understanding Quantum Technologies, Seventh Edition, 2024.
- [2] 정보통신기획평가원(IITP), ICT R&D 기술로드맵 2025.
- [3] J. P. G. van Dijk et al., "Designing a DDS-Based SoC for High-Fidelity Multi-Qubit Control," IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 67, no. 12, pp. 5380-5393, Dec. 2020.
- [4] Sadhu, Bodhisatwa, et al. "Cryogenic CMOS Circuits for Future Scaled Quantum Computing Systems: Challenges and Solutions." 2025 IEEE Custom Integrated Circuits Conference (CICC). IEEE, 2025.
- [5] J. Park et al., "A Fully Integrated Cryo-CMOS SoC for State Manipulation, Readout, and High-Speed Gate Pulsing of Spin Qubits," in IEEE Journal of Solid-State Circuits, vol. 56, no. 11, pp. 3289-3306, 2021.
- [6] Gunyhó, A.M., Kundu, S., Ma, J. et al. "Single-shot readout of a superconducting qubit using a thermal detector", Nature Electronics, vol. 7, pp. 288–298, 2024.
- [7] D. Minn et al., "A 40 nm Cryo-CMOS Homodyne-Demodulation Readout SoC for Superconducting Qubits," in IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 72, no. 3, pp. 1003-1016, 2025.
- [8] J. P. G. Van Dijk et al., "A Scalable Cryo-CMOS Controller for the Wideband Frequency-Multiplexed Control of Spin Qubits and Transmons," IEEE J. Solid-State Circuits, vol. 55, no. 11, pp. 2930-2946, Nov. 2020.
- [9] J. Yoo et al., "Design and Characterization of a <4-mW/Qubit 28-nm Cryo-CMOS Integrated Circuit for Full Control of a Superconducting Quantum Processor Unit Cell," IEEE J. Solid-State Circuits, vol. 58, no. 11, pp. 3044-3059, Nov. 2023.
- [10] Y. Guo et al., "A Polar-Modulation-Based Cryogenic Transmon Qubit State Controller in 28 nm Bulk CMOS for Superconducting Quantum Computing," IEEE J. Solid-State Circuits, vol. 58, no. 11, pp. 3060-3073, Nov. 2023.
- [11] K. Kang, S. Bae, D. Minn, J. Lee and J. -Y. Sim, "A Quantum Controller IC With DRAG Generation in 40-nm Cryo-CMOS for Scalable Superconducting Quantum Computing," in IEEE Journal of Solid-State Circuits, vol. 60, no. 8, pp. 2832-2841, Aug. 2025.
- [12] Y. Guo et al., "29.4 A Cryo-CMOS Quantum Computing Unit Interface Chipset in 28nm Bulk CMOS with Phase-Detection Based Readout and Phase-Shifter Based Pulse Generation," 2024 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 2024,

pp. 476-478.

XI. 결론 및 향후 계획

반도체 기술은 단순한 산업적 중요성을 넘어서, 국가 경제, 사회, 국방 등 다양한 분야에 직간접적으로 큰 영향을 미치는 핵심 기술입니다. 따라서, 반도체 기술의 발전 전망을 세우고 이에 대한 로드맵을 수립하는 것은 우리나라 반도체 산업의 경쟁력을 확보하고, 기술적 우위와 주도권을 유지하는데 매우 중요한 과제이다.

이에 반도체공학회는 향후 15년을 전망하는 반도체 기술 발전 로드맵을 수립하기 위한 연구를 진행하였다. 반도체 기술의 범위는 매우 넓으며, 공정, 소자, 회로, 시스템, 소프트웨어, 응용분야 등 다양한 도메인의 기술이 상호 연계되어 있어, 이를 종합적으로 로드맵으로 제시하는 작업은 상당히 방대한 작업이 될 수 밖에 없다. 이러한 점을 고려하여, 반도체공학회는 「반도체기술로드맵 2025」에서 다룬 4가지 주요 반도체 기술 분야인 소자 및 공정 기술(More Moore), 인공지능 반도체 기술(AI Semiconductor), 광연결(Optical Connectivity) 반도체 기술, 무선연결(Wireless Connectivity) 반도체 기술을 중심으로 센서 기술, 유선연결 반도체 기술, Processing-In-Memory(PIM) 기술, 패키징 기술, 양자컴퓨팅 기술 등 총 9개 분야로 로드맵 수립 범위를 확장하였다. 이를 통해 전반적인 반도체 기술 발전의 방향성을 다루고자 하였다. 본 로드맵에서는 각 기술 분야에 대한 핵심 기술 개발 과제, 발전 동향 및 향후 발전 방향을 예측하고 제시하였다.

반도체공학회는 이번 두 번째 로드맵 발표를 통해, 우리나라 반도체 기술의 중장기적인 발전 방향을 제시하는 로드맵 수립을 지속적으로 추진할 예정이다. 이를 통해 반도체공학회는 우리나라 반도체 기술 및 산업 발전은 물론, 반도체 인력양성 전략 수립에도 기여하고자 한다.

반도체기술로드맵 2026 저자목록

○ 반도체공학회 회장:

신현철(광운대학교)

○ 반도체공학회 비전전략위원회 위원장:

남일구(부산대학교), 양준모(나노종합기술원)

○ 반도체공학회 비전전략위원회 위원:

이규호(연세대학교), 송진호(라이팩㈜), 문현원(대구대학교), 고주현(삼성전자), 김준석(삼성전자), 신승철(삼성전자), 박준용(경희대학교), 창동진(충남대학교), 홍슬기(서울과학기술대학교), 윤상원(서울대학교), 권지민(UNIST), 김주성(이화여자대학교), 남재원(서울과학기술대학교)

