

Co-Packaged Optics (CPO) Book – Scaling with Light for the Next Wave of Interconnect

January 4, 2026

Contents

1 Part 1: CPO Total Cost of Ownership (TCO) Analysis	5
1.1 Typical AI Cluster Networking Configuration and TCO	6
1.2 CPO Scale-out Power Budgets	8
1.3 CPO Scale-out Total Cost of Ownership (TCO)	15
1.4 CPO for Scale-up Networks	18
1.4.1 Scale-out vs Scale-up TAM	19
1.4.2 Copper vs Optics for Scale-Up: World Sizes, Density and Reach	20
1.4.3 Copper vs Optics for Scale-Up: Scaling Bandwidth	22
1.5 When will CPO be ready for Primetime?	23
2 Part 2: CPO Introduction and Implementation	29
2.1 What is CPO about and why is everyone so excited?	29
2.2 Co-packaged Copper	35
2.3 Past obstacles to CPO market readiness: Why only now?	40
2.4 Evolving beyond DSP-based Transceivers: From LPO to CPO	40
2.4.1 The Crusade Against DSP	45
2.5 Why CPO? The I/O challenge, BW density, and bottlenecks	46
2.5.1 Input/Output (I/O) Speedbumps and Roadblocks	49
2.5.2 Electrical SerDes Scaling Bottlenecks	51
2.5.3 SerDes Scaling Plateau as a Roadblock for Scaling NVLink	52
2.6 Wide I/O vs SerDes	56
2.7 Link Resiliency	59
3 Part 3: Bringing CPO to Market and deployment challenges	61
3.1 CPO optical engine manufacturing considerations and go to market	61
3.2 Host and Optical Engine Packaging	61
3.3 TSMC COUPE is emerging as the integration option of choice	62
3.4 Packaging the OEs with the host	67
3.5 FAUs and Fiber Coupling	68
3.5.1 Edge Coupling (EC)	68
3.5.2 Grating Coupling (GC)	69
3.6 Laser Type and Wavelength Division Multiplexing (WDM)	70
3.7 Modulator Types	72
3.7.1 Mach-Zehnder Modulator (MZM)	73
3.7.2 Micro-Ring Modulators (MRMs)	74
3.7.3 Electro-Absorption Modulators (EAM)	76
3.8 OE roadmap – scaling OEs	78
3.9 Key approaches for scaling bandwidth	78
3.10 CPO adoption pace and deployment challenges	81
3.11 Proprietary solutions vs standards	81
3.12 Serviceability and reliability	83

4 Part 4: CPO Products of Today and Tomorrow	86
4.1 NVIDIA CPO	86
4.1.1 Quantum-X Photonics	86
4.1.2 Spectrum-X Photonics	90
4.2 The Broadcom CPO Switch Portfolio	94
4.3 Intel's CPO Roadmap	99
4.4 MediaTek CPO plans	100
4.5 CPO Focused Companies	100
4.5.1 Ayar Labs	101
4.5.2 Nubis	105
4.5.3 Celestial AI	107
4.5.4 Lightmatter	117
4.5.5 Xscape Photonics	121
4.5.6 Ranovus	122
4.5.7 Scintil	123
5 Part 5: NVIDIA's CPO Supply Chain	126
5.1 Optical Engines	126
5.2 External Laser Source (ELS)	126
5.3 Fiber Attach Units (FAUs)	127
5.4 Fiber Shuffle Box	129
5.5 MPO Connectors	130
5.6 MT Ferrules	130
5.7 Manufacturing, Assembling, and the Testing Processes	131

Co-Packaged Optics (CPO)는 오랫동안 datacenter 연결성을 혁신할 것으로 기대되어 왔지만, 실질적으로 배포 가능한 제품이 2025년에야 출시되면서 시장에 나오기까지 오랜 시간이 걸렸다. 그동안 pluggable transceiver는 networking 요구사항에 맞춰 발전해왔으며, 상대적인 비용 효율성, 배포의 익숙함, 그리고 표준 기반의 상호운용성 덕분에 여전히 기본 선택지로 남아있다.

하지만 **Artificial Intelligence (AI)** workload와 함께 오는 막대한 networking 수요는 이번에는 상황이 다르다는 것을 의미한다. AI networking bandwidth roadmap은 interconnect 속도, 거리, 밀도 및 신뢰성 요구사항이 곧 transceiver가 제공할 수 있는 수준을 넘어설 것이다. CPO는 일부 이점을 제공하고 scale-out networking에 더 많은 옵션을 가져다주지만, scale-up networking에서 핵심적인 역할을 할 것이다. **CPO는 이번 10년 후반과 그 이후 scale-up networking에서 bandwidth 증가의 주요 동력이 될 것이다.**

NVLink와 같은 오늘날의 copper 기반 scale-up 솔루션은 GPU당 7.2 Tbit/s의 엄청난 bandwidth를 제공하며 – Rubin 세대에서는 곧 GPU당 14.4 Tbit/s가 될 예정이지만, copper 기반 link는 거리가 최대 2 meter로 제한되어, scale-up domain world size가 기껏해야 1개 또는 2개 rack으로 제한된다는 것을 의미한다. 또한 copper를 통해 bandwidth를 확장하는 것이 점점 더 어려워지고 있다. Rubin에서 NVIDIA는 양방향 SerDes를 통해 copper lane당 bandwidth를 또 한 번 두 배로 늘릴 것이지만, 점점 더 빠른 SerDes를 개발하여 copper에서 bandwidth를 두 배로 늘리는 것은 매우 어려운 확장 방향이며 더디게 진행되는 작업이다. **CPO는 동일하거나 더 나은 bandwidth density를 제공할 수 있으며 bandwidth 확장을 위한 추가적인 방법을 제공하면서, 동시에 더 큰 scale-up domain을 가능하게 한다.**

CPO의 필요성을 이해하기 위한 출발점은 optical communication에 transceiver를 사용할 때의 많은 비효율성과 trade-off를 고려하는 것이다. Transceiver는 더 긴 link 거리를 달성하는데 사용될 수 있지만, transceiver가 꽂히는 networking switch 또는 compute tray의 front panel에 있는 cage는 일반적으로 XPU 또는 switch Application-Specific Integrated Circuit (ASIC)으로부터 15-30cm 떨어져 있다. 이는 신호가 먼저 **Long-Reach Serializer-Deserializer (LR SerDes)**를 사용하여 해당 15-30cm 거리를 전기적으로 전송되어야 하며, 전기 신호가 optical signal로 변환되기 전에 transceiver 내부의 **Digital Signal Processor (DSP)**에 의해 복구되고 조정되어야 함을 의미한다. CPO를 사용하면 optical engine이 대신 XPU 또는 Switch ASIC 옆에 배치되므로, DSP를 제거할 수 있고 XPU에서 Optical Engine으로 데이터를 이동하는 데 더 낮은 전력의 SerDes를 사용할 수 있다. 이는 DSP Transceiver와 비교하여 데이터 전송에 필요한 에너지를 50% 이상 줄일 수 있으며, 많은 이들이 bit당 에너지 요구사항을 80%까지 줄이는 것을 목표로 하고 있다.

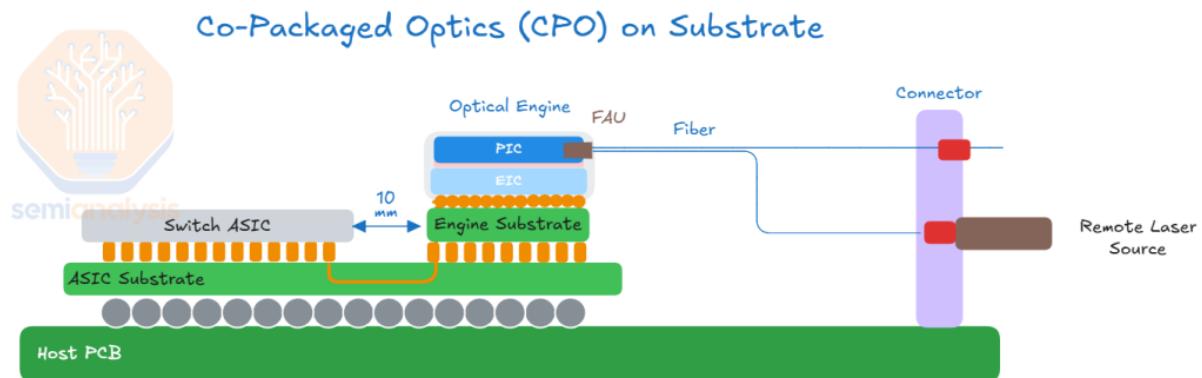


Figure 1: Source: SemiAnalysis

NVIDIA와 Broadcom의 scale-out CPO 솔루션이 더 많은 주목을 받고 있고 최종 고객들에 의해 면밀히 검토되고 있는 반면, 주요 Hyperscaler들은 이미 scale-up CPO 전략을 계획하기 시

작했으며 공급업체들과 약속을 하고 있다. 예를 들어, Celestial AI는 2028년 말까지 \$1B revenue run rate를 창출할 수 있을 것으로 추정하고 있다 - 우리는 이것이 주로 Amazon의 Trainium 4와 함께 출하되는 CPO scale-up 솔루션에 의해 주도될 것으로 믿는다. CPO에 집중하는 기업들은 이제 논문, pilot project 및 demonstration을 훨씬 넘어섰으며 대량 생산을 해결하기 위한 optical port architecture와 같은 핵심 제품 결정을 내리고 있다.

Scale-up을 위한 CPO는 이제 만약과 왜의 문제가 아니라, 언제 그리고 어떻게의 문제이다. 이러한 시스템을 대량 생산으로 가져오는 방법, 그리고 laser manufacturer와 같은 핵심 component 공급망 회사들이 충분한 생산을 늘릴 수 있는 시기의 문제이다. 이 문서는 CPO의 이점과 과제, CPO architecture가 작동하는 방식, 현재 및 미래의 CPO 제품, CPO에 집중하는 기업, CPO 관련 component 및 각각의 공급망에 대한 심층적인 논의를 제시할 것이다. 이 글은 실무자, 산업 분석가, 투자자 및 interconnect 기술에 관심이 있는 모든 사람들을 위한 가이드 역할을 하도록 의도되었다. 우리는 이 문서를 다섯 부분으로 나누었으며, 독자들은 가장 흥미롭거나 관련성이 높은 섹션에 집중할 수 있다.

Part 1: Total Cost of Ownership (TCO) Analysis에서는 CPO 채택이 scale-out 및 scale-up network의 total cost of ownership을 어떻게 변화시키는지 분석하는 것으로 시작한다. 우리는 TCO, 신뢰성, 그리고 장비 vendor bargaining power가 scale-out network에서 CPO를 채택하는 주요 고려사항이 될 것이라고 생각한다. 우리는 scale-out과 관련하여 CPO가 본격적으로 준비되었는지 살펴볼 것이며, ECOC 2025에서 발표된 Meta의 CPO scale-out switch 연구와 같은 솔루션 신뢰성에 대해 지금까지 확보한 데이터를 다룰 것이다.

Part 2: CPO Introduction and Implementation에서는 CPO가 어떻게 작동하는지 더 깊이 탐구할 것이다. 이 섹션에서는 copper에서 co-packaged copper로, 그리고 digital signal processor (DSP) optics에서 linear pluggable optics (LPO)를 거쳐 CPO로의 시장 진화를 살펴보고 CPO 채택의 동기와 논거를 다룰 것이다. SerDes scaling limit와 Wide I/O가 SerDes의 대안으로 논의될 것이다.

Part 3: Bringing CPO to Market에서는 CPO가 견인력을 얻고 시장에 출시될 수 있도록 하는 핵심 기술을 설명할 것이다. 먼저 Host와 Optical Engine packaging을 논의하고 **TSMC COUPE**를 상세히 설명하며 이것이 왜 선택되는 통합 옵션으로 부상하고 있는지 설명한다. **Fiber Attach Unit (FAU)**, **Fiber coupling** 뿐만 아니라 **Edge Coupling vs Grating Coupler**가 철저히 설명될 것이다. **Mach-Zender Modulator (MZM)**, **Micro-Ring Modulator (MRM)** 및 **Electro-Absorption Modulator (EAM)**과 같은 modulator type을 다룰 것이다. 이 섹션은 CPO가 채택되는 핵심 이유인 CPO로 bandwidth를 확장하는 다양한 방법을 설명한다.

Part 4: CPO Products of Today and Tomorrow에서는 오늘날 시장에서 사용 가능한 CPO 제품과 관련 공급망을 분석할 것이다. 주요 CPO 기업을 논의하기 전에 **NVIDIA와 Broadcom**의 솔루션으로 시작할 것이다. 우리는 Ayar Labs, Nubis, Celestial AI, Lightmatter, Xscape Photonics, Ranovus 및 Scintil을 다루며, 각 제공업체의 솔루션을 상세히 설명하고 각 기업의 접근 방식에 대한 중요한 장단점을 평가할 것이다.

Part 5: NVIDIA's CPO Supply Chain에서는 NVIDIA의 CPO ecosystem에 대한 공급망을 상세히 설명하여 이 보고서를 마무리할 것이다. Laser Source, ELS Module, FAU, FAU Alight Tool, FAU Assembly, Shuffle Box, MPO Connector, MT Ferrule, Fiber 및 E/O Testing에 대한 주요 공급업체를 명시할 것이다.

1 Part 1: CPO Total Cost of Ownership (TCO) Analysis

핵심 용어: CPO (Co-Packaged Optics)

CPO는 광학 엔진(optical engine)을 네트워크 스위치 ASIC에 직접 패키징하는 기술로, 기존의 플리그형 광 트랜시버(pluggable optical transceiver)를 제거하여 전력 소비와 비용을 크게 줄입니다. 이는 데이터센터 네트워킹의 패러다임 전환을 의미하며, 특히 AI 클러스터의 대규모 interconnect 요구사항을 충족하기 위한 핵심 기술입니다.

올해 초 NVIDIA의 GTC 2025에서 가장 큰 기대를 모은 주제 중 하나는 Jensen이 회사의 첫 CPO 지원 scale-out network switch를 발표한 것이었다. 주목할 만한 점은 scale-up의 경우 NVIDIA는 여전히 copper를 밀고 나가며 2027년과 2028년까지도 optics로 가는 것을 피하기 위해 극단적인 노력을 기울이고 있다는 것이다.

이러한 새로운 CPO 지원 switch에 대한 논의를 TCO를 검토하는 것으로 시작하여, scale-out CPO가 제공할 수 있는 비용 및 전력 절감을 분석해보자. NVIDIA의 GTC 2025 keynote는 두 가지 다른 CPO 지원 switch ASIC을 활용하는 세 가지 다른 CPO scale-out switch를 발표했다. TCO, 전력 및 배포 속도 이점이 있지만 고객들이 완전히 다른 배포 체계로 즉시 뛰어들기에는 충분히 설득력이 없으며, 우리는 첫 번째 wave의 CPO scale-out switch에 대한 채택이 제한적일 것으로 예상한다. 그 이유를 살펴보자.

Nvidia CPO Roadmap			
Switch Model	Quantum 3450 CPO	Spectrum 6810 CPO	Spectrum 6800 CPO
Launch Date	2H 2025	2H 2026	
Networking Standard	InfiniBand	Ethernet	
Switch ASIC	Quantum-3	Spectrum-6	
Throughput per Package	28.8 Tbps	102.4 Tbps	
Number of Switch Packages	4	1	4
Switch Aggregate Bandwidth	115.2 Tbps (not all-to-all)	102.4 Tbps	409.6 Tbps (not all-to-all)
SerDes speed (Gb/s uni-di)	200 Gbps	200 Gbps	
Optical Connectivity	DR Optics	DR Optics	
Physical MPO Ports	144	128	512
Bandwidth and Logical Port Configurations Available	144 Ports of 800G	512 Ports of 200G 256 Ports of 400G 128 Ports of 800G	512 Ports of 800G
Bandwidth per Optical Engine (OE)	1.6 Tbps	3.2 Tbps	
Number of OEs	72	32	128
External Light Sources (ELSS)	18	16	64
For Spectrum CPO there are 36 OEs on the package, but only 32 OEs are enabled			

Figure 2: Source: SemiAnalysis

1.1 Typical AI Cluster Networking Configuration and TCO

일반적인 AI Cluster는 세 가지 주요 networking fabric인 back-end, front-end 및 out of band management fabric을 갖는다.

기술 용어: AI Cluster의 세 가지 Networking Fabric

1. Back-end Fabric (GPU 간 통신):

- **용도:** GPU 간 scale-out communication, collective operation (All-Reduce, All-Gather 등)
- **프로토콜:** InfiniBand 또는 Ethernet (RoCE)
- **특징:** 가장 기술적으로 까다롭고 비용/전력 소비가 큼
- **대역폭:** 매우 높음 (GPU당 800Gbit/s 이상)
- **레이턴시:** 극도로 낮아야 함 (마이크로초 단위)
- **비중:** Networking 비용의 85%, 전력의 86% 차지 (3-layer GB300 NVL72 기준)
- **CPO 적용 대상:** 주요 타겟! 비용과 전력 절감 효과가 가장 큼

2. Front-end Fabric (외부 통신):

- **용도:** 클러스터와 외부 세계 간 통신 (스토리지, 인터넷, 사용자 요청)
- **프로토콜:** 주로 Ethernet
- **특징:** Back-end보다 성능 요구사항 낮음, 유연성 중시
- **대역폭:** 중간 수준 (일반적으로 100Gbit/s - 400Gbit/s)
- **비중:** Networking 비용/전력의 작은 부분 차지
- **CPO 적용:** 가능하지만 우선순위 낮음

3. Out-of-Band Management Fabric (관리 통신):

- **용도:** 시스템 모니터링, 관리, 펌웨어 업데이트, 원격 접속
- **프로토콜:** 저속 Ethernet (1Gbit/s - 10Gbit/s)
- **특징:** 메인 네트워크와 물리적으로 분리되어 장애 시에도 관리 가능
- **대역폭:** 매우 낮음, 신뢰성이 더 중요
- **비중:** 비용/전력의 극소수 차지
- **CPO 적용:** 불필요 (낮은 대역폭으로 충분)

핵심 포인트: CPO의 주요 적용 대상은 back-end fabric입니다. 이 네트워크가 전체 networking 비용과 전력의 대부분을 차지하기 때문에, CPO의 TCO 개선 효과도 여기에 집중됩니다.

가장 많이 활용되고 기술적으로 까다로운 network fabric은 back-end fabric이다. Back-end fabric은 GPU 간의 scale-out communication에 사용되어 서로 통신하고 training 및 inference를 병렬화하기 위한 collective operation에서 데이터를 교환한다. Back-end network는 일반적으로 InfiniBand 또는 Ethernet protocol을 사용한다.

까다로운 특성으로 인해 back-end network는 전체 networking 비용과 전력의 지배적인 점유율을 차지하며, NVIDIA의 X800-Q3400 back-end switch를 사용하는 InfiniBand에 배포된 3-layer GB300 NVL72 cluster의 경우 networking 비용의 85%와 networking 전력의 86%를 차지한다. CPO 기반 switch와 networking 솔루션은 back-end 및 front-end network 모두에서 사용될 수 있지만, 이 단계에서 배포의 초점은 back-end network에 있을 것으로 생각한다. 독자들은 우리의 [Optical Boogeyman Article from 2024](#) 및 [AI Networking Model](#)에서 back-end network topology, port, switch 및 transceiver 수에 대한 훨씬 더 자세한 내용을 찾을 수 있다. Total networking cost of ownership을 이해하고 싶은 분들은 우리의 [AI Neocloud Anatomy and Playbook article](#)을 읽을 수 있다.

큰 그림을 보면 networking 비용은 AI server 자체 다음으로 전체 AI cluster 비용의 두 번째로 큰 구성요소이다. 3-Layer InfiniBand network를 갖춘 GB300 NVL72 Cluster의 경우, 이는 전체 cluster 비용의 15%에 해당하며, 4-Layer network의 경우 전체 cluster 비용의 18%에 도달한다. optical transceiver는 이 비용의 상당 부분을 차지하며, 상대적으로 더 비싼 NVIDIA LinkX Transceiver를 사용할 때 3-Layer network에서 networking 비용의 60%를 차지한다. 또한 3-Layer network에서 전체 networking 전력의 45%를 소비한다.

핵심 인사이트: 네트워킹 비용 구조

AI 클러스터에서 네트워킹은 전체 비용의 15-18%를 차지하며, 그중 광 트랜시버가 네트워킹 비용의 60%, 전력의 45%를 소비합니다. 이는 CPO가 트랜시버를 제거함으로써 상당한 비용 및 전력 절감 가능성을 가진 이유를 설명합니다.

GB300 NVL72 Cluster Cost and Power Budget Breakdown, per Rack-Scale 72-GPU Server						
Item	2-Layer Network (up to 10,368 GPUs)		3-Layer Network (up to 746,496 GPUs)		4-Layer Network (up to ~53.7M GPUs)	
	Cost	Power (W)	Cost	Power (W)	Cost	Power (W)
Server	\$4,357,943	142,000	\$4,357,943	142,000	\$4,357,943	142,000
Optical Transceivers	\$363,125	4,600	\$499,925	6,184	\$568,325	6,976
Switches	\$208,681	5,214	\$312,681	8,014	\$416,681	11,606
Fiber, Cables, Software, Others	\$17,723	0	\$24,338	0	\$30,953	0
Networking	\$589,529	9,814	\$836,944	14,198	\$1,015,959	18,582
All Others	\$269,152	281	\$269,152	281	\$269,152	281
Total Power	\$5,216,624	152,095	\$5,464,039	156,479	\$5,643,054	160,863

Power for a Neocloud Giant InfiniBand cluster using X800-Q3400 Back-end switches with 144 ports of 800G
Source: SemiAnalysis Networking Model

Figure 3: Source: [SemiAnalysis AI Networking Model](#)

AI cluster의 GPU 수가 많을수록 더 많은 networking layer가 필요할 가능성이 높다. 2-layer에서 3-layer network 이상으로 가는 것은 더 높은 비용과 더 큰 전력 예산을 의미한다. CPO는 layer 수를 일정하게 유지하면서 전력과 비용을 줄이는 데 도움을 줄 수 있으며, 주어진 수의 layer를 가진 network에 연결할 수 있는 GPU 수를 확장하여 전체 전력 및 비용 요구사항을 줄일 수 있다.

1.2 CPO Scale-out Power Budgets

올해 초 GTC 2025에서 NVIDIA의 CEO Jensen Huang은 transceiver만으로 인한 막대한 전력 소비를 CPO의 핵심 동기로 강조했다. 위 표의 rack당 전력 예산 중 일부를 사용하면, 3-layer network 의 200,000 GB300 NVL72 (rack당 72 GPU package 및 144 compute chiplet) GPU cluster는 435 MW의 Critical IT Power를 소비할 것이며 그 중 17 MW가 optical transceiver만으로 소비될 것이다. 분명히 대부분의 transceiver content를 제거함으로써 절약할 수 있는 엄청난 양의 전력이 있다.

이는 단일 800G DSP transceiver에서 사용되는 전력을 CPO 시스템 내의 optical engine 및 laser source (800G bandwidth당)가 소비하는 전력과 비교하면 쉽게 알 수 있다. 800G DR4 optical transceiver가 약 16-17W를 소비하는 반면, NVIDIA의 Q3450 CPO switch에 사용되는 optical engine과 external laser source를 합치면 800G bandwidth당 약 4-5W를 소비하여 전력이 73% 감소하는 것으로 추정된다.

기술 용어: DSP, LPO 그리고 CPO의 전력 비교

DSP (Digital Signal Processing) Transceiver: 신호 처리 기능이 내장된 전통적인 플러그형 트랜시버로, 800G당 약 16-17W 소비.

전력 소비 원인:

- **Digital signal processing chip:** 신호 품질 향상을 위한 복잡한 알고리즘 실행으로 5-6W 소비
- **Electro-optic/Opto-electric conversion (E/O, O/E):** 전기 신호와 광 신호 간 변환에 4-5W 소비
- **Integrated laser source:** 각 transceiver마다 독립적인 laser 작동에 3-4W 소비
- **Cooling 및 power management circuit:** 열 관리와 전력 변환에 2-3W 소비
- **긴 electrical signal path:** Pluggable form factor로 인한 신호 감쇠와 재생에 추가 전력 필요

LPO (Linear-drive Pluggable Optics): DSP 기능을 제거한 저전력 트랜시버로, DSP 대비 36% 전력 절감 (약 10-11W).

전력 절감 메커니즘:

- **DSP chip 제거:** Digital signal processing을 switch ASIC의 SerDes로 이동하여 5-6W 절감
- **Analog direct drive:** 복잡한 digital modulation 없이 linear drive 방식으로 laser 구동
- **Energy efficiency:** 비트당 에너지 효율이 16.3pJ/bit (DSP)에서 9.4pJ/bit로 개선
- **단점:** 짧은 전송 거리 (일반적으로 100m 이하), 높은 signal integrity 요구

CPO (Co-Packaged Optics): 광학 엔진과 외부 레이저를 스위치에 통합하여 800G당 4-5W만 소비하며, DSP 대비 73% 전력 절감을 달성합니다.

획기적인 전력 절감 원리:

- **극단적 path 단축:** Switch ASIC과 optical engine 간 electrical signal path가 수십 cm에서 수 mm로 단축되어 signal attenuation 최소화 (약 2-3W 절감)
- **Pluggable interface 제거:** Cable connector와 pluggable interface의 전력 손실 제거 (약 1-2W 절감)
- **Centralized laser:** External Light Source (ELS)가 여러 optical engine에 laser 공급하여 laser당 전력 효율 향상 (약 2-3W 절감)
- **Unified cooling:** Switch와 optical component가 단일 cooling system 공유로 overhead 감소 (약 1W 절감)
- **최적화된 power distribution:** 단일 package 내 power management로 conversion loss 최소화 (약 0.5-1W 절감)
- **최고 energy efficiency:** 비트당 에너지 효율 6.3pJ/bit 달성, DSP 대비 61% 향상

이러한 수치는 ECOC 2025에서 발표되고 제출된 Meta의 논문에서 제시된 수치와 매우 유사하다. 이 보고서에서 Meta는 800G 2xFR4 pluggable transceiver가 약 15W를 소비하는 반면 Broadcom Bailly 51.2T CPO switch 내의 optical engine과 laser source는 전달된 800G bandwidth 당 약 5.4W를 소비하여 65%의 전력 절감을 나타낸 방법을 보여주었다.

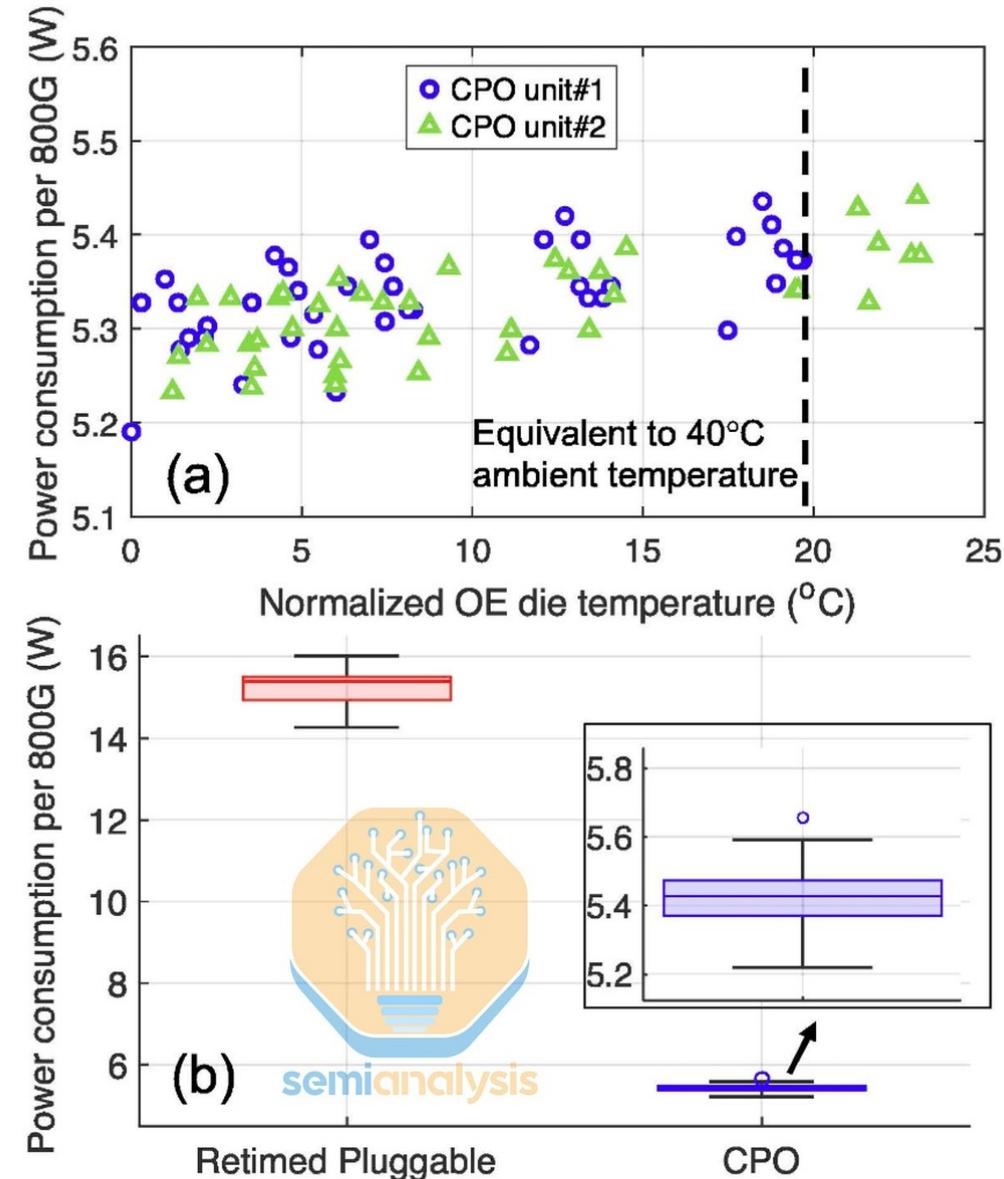


Fig. 1: a) Optics power consumption at different OE die temperatures. b) Power consumption comparison of 2x400Gbps FR4 pluggable and CPO. Inset: magnified CPO power.

Figure 4: Source: Meta

이 분석을 cluster 수준으로 확장해보자. three-Layer network에 구축된 GB300 NVL72 cluster로 전환하면, back-end network에서 DSP transceiver에서 LPO transceiver를 사용하는 것으로 이동하면 전체 transceiver 전력을 36%, 전체 network 전력을 16% 줄일 수 있음을 알 수 있다. CPO로의 완전한 전환은 DSP optics 대비 훨씬 더 큰 절감을 제공한다 (transceiver 전력을 84% 절감).

기술 용어: Optical Engine (OE)과 External Light Source (ELS)

Optical Engine (OE): 전기 신호를 광 신호로 변환하는 광학 컴포넌트로, CPO 시스템에서 스위치 ASIC에 직접 패키징됩니다. 기존 트랜시버의 복잡한 DSP 기능을 제거하여 전력 효율을 높입니다.

External Light Source (ELS): CPO 시스템에서 여러 optical engine에 레이저 광을 공급하는 중앙집중식 레이저 소스입니다. 각 트랜시버마다 별도의 레이저를 사용하는 것보다 효율적이며, 유지보수가 용이합니다.

전력 Tradeoff 상세 계산:

Networking 전력 구성 (DSP 기준선):

- Optical transceiver: 전체 networking 전력의 **45%** 차지
- Switch ASIC 및 기타 구성요소: 전체 networking 전력의 **55%** 차지

CPO 전환 시 전력 변화:

- **Transceiver 전력:** $45\% \rightarrow 45\% \times (1 - 0.84) = 7.2\%$
 - 84% 감소하지만, 원래 전체의 45%만 차지했으므로
 - 실제 절감량: $45\% - 7.2\% = 37.8\%p$ 절감
- **Switch 전력:** $55\% \rightarrow 55\% \times 1.23 = 67.7\%$
 - Optical Engine (OE)과 External Light Source (ELS) 추가로 23% 증가
 - 실제 증가량: $67.7\% - 55\% = 12.7\%p$ 증가
- **순 효과:** $7.2\% + 67.7\% = 74.9\%$
 - 전체 networking 전력 절감: $100\% - 74.9\% = 25.1\%$ 감소
 - 논문에서 제시된 23% 감소와 일치

핵심 통찰: Transceiver의 대폭적인 전력 감소(37.8%p)가 switch의 증가분(12.7%p)을 크게 상회하여, 전체적으로 약 25% 순 절감을 달성합니다. 이는 transceiver를 제거하는 것이 아니라 "더 효율적인 형태로 switch에 통합"하는 전략입니다.

핵심: CPO는 transceiver를 제거하여 큰 전력 절감을 달성하지만, 그 기능을 스위치 내부로 이동시키므로 스위치 자체의 전력 소비는 증가합니다. 이는 "전력을 없애는" 것이 아니라 "더 효율적인 방식으로 재배치"하는 것입니다.

이 전력 절감의 일부는 switch에 **Optical Engine (OE)**과 **External Light Source (ELS)**를 추가함으로써 상쇄되며, 이는 이제 전체적으로 23% 더 많은 전력을 소비한다. 아래 예에서 CPO 시나리오의 optical transceiver 전력은 server당 1,000W에 머물러 있는데, 이는 front-end networking이 여전히 DSP transceiver를 사용할 것으로 가정하기 때문이다.

GB300 NVL72 Cluster Power Budget: Traditional vs CPO Network, per Rack-Scale 72-GPU Server										
Item	DSP Transceivers (3-Layer Network)		LPO Transceivers (3-Layer Network)		CPO (3-Layer Network)		CPO (2-Layer Network)		Power (W)	Δ Power (%)
	Power (W)	Δ Power (%)	Power (W)	Δ Power (%)	Power (W)	Δ Power (%)	Power (W)	Δ Power (%)		
Server	142,000	0%	142,000	0%	142,000	0%	142,000	0%	142,000	0%
Optical Transceivers	6,184	0%	3,935	-36%	1,000	-84%	1,000	-84%		
Switches	8,014	0%	8,014	0%	9,884	23%	6,336	-21%		
Fiber, Cables, Software, Others	0	0%	0	0%	0	0%	0	0%		
Networking	14,198	0%	11,949	-16%	10,884	-23%	7,336	-48%		
All Others	281	0%	281	0%	281	0%	281	0%		
Total Power	156,479	0%	154,230	-1%	153,165	-2%	149,617	-4%		

Neocloud Giant InfiniBand cluster using Nvidia x800-Q3400 Back-end switches with 144 ports of 800G and Nvidia LinkX Transceivers. CPO used only in Back-end network.
Source: SemiAnalysis Networking Model

Figure 5: Source: [SemiAnalysis AI Networking Model](#)

NVIDIA의 CPO scale-out switch 사용은 암묵적으로 high radix network가 기본적으로 사용됨을 의미하지만, 이는 최종 사용자에게 추상화되는데, shuffle이 high radix non-CPO switch를 사용할 때 patch panel이나 octopus cable을 통해 switch box 외부에서 발생하는 것과 달리 switch box 내부에서 발생하기 때문이다.

기술 용어: High Radix Network와 Internal Shuffle

1. 핵심 개념 정의

Radix: Switch가 제공하는 port 수. High radix는 많은 수의 port를 의미합니다 (예: 128 포트 이상).

Shuffle: 여러 switch의 port를 상호 연결하여 하나의 논리적 대형 switch처럼 작동하도록 만드는 재배치 작업.

2. 구체적 예시: 512포트 달성하기

Traditional 방식 (Non-CPO High Radix):

- **하드웨어 구성:** 4개의 128포트 switch를 별도로 설치
- **External shuffle 필요:**
 - Switch A의 port 1-32 → Patch panel → Switch B의 특정 port들로 연결
 - Switch B의 port 1-32 → Octopus cable → Switch C의 특정 port들로 연결
 - Switch C와 D도 유사하게 복잡하게 교차 연결
 - 수백 개의 cable이 switch box 외부에서 얹힘
- **문제점:**
 - 물리적 복잡도: 수백 개의 cable을 수동으로 연결하고 관리
 - 설치 시간: 며칠이 소요되는 cable routing
 - 디버깅 악몽: Cable 하나가 잘못 연결되면 전체 network 문제 발생
 - 공간 낭비: Patch panel, cable tray, shuffle box가 차지하는 rack 공간
 - 전력 낭비: 각 pluggable transceiver와 긴 cable path에서 신호 손실
 - 확장성 한계: Port 추가 시 전체 cabling을 재구성해야 함
- **사용자 관점:** "내가 4개의 switch를 복잡하게 연결하고 있다"는 것을 항상 인식해야 함

CPO 방식 (Internal Shuffle):

- **하드웨어 구성:** 단일 switch box (예: Spectrum 6800)
- **Internal shuffle 메커니즘:**
 - Switch ASIC 내부에 여러 개의 작은 switching unit 통합
 - 이들을 광학적으로 switch box 내부에서 상호 연결
 - Optical engine과 optical interconnect가 shuffle 기능 수행
 - 모든 복잡한 연결이 package 내부에서 완료
- **결과:** 512포트를 제공하는 "단일 switch"처럼 보임
- **이점:**
 - **Zero external cable:** External shuffle cable 완전 제거
 - **Plug-and-play:** Switch box만 rack에 장착하면 즉시 사용 가능
 - **추상화 완성:** 사용자는 "512포트 단일 switch"로만 인식
 - **공간 절약:** Patch panel과 cable tray 불필요
 - **전력 효율:** 짧은 optical path로 신호 손실 최소화
 - **안정성:** Factory에서 검증된 내부 연결로 연결 오류 제로
- **사용자 관점:** "그냥 512포트 switch 하나"로 완전히 추상화됨

3. 실제 제품 예시

- **NVIDIA Quantum 3450 (CPO):** 800G × 144 port를 단일 switch box로 제공
- **NVIDIA Spectrum 6800 (CPO):** 800G × 512 port를 단일 switch box로 제공
- **비교:** Traditional 방식으로 512포트를 달성하려면 최소 4-8개의 switch + 복잡한 external cabling 필요

4. CPO가 Internal Shuffle을 가능하게 하는 이유

Traditional pluggable transceiver는 물리적으로 switch 전면 panel에 위치해야 하므로, 여러 switch를 통합하려면 반드시 external cable이 필요합니다. 반면 CPO는 optical engine을 switch ASIC package 내부에 통합하므로, package 내부의 optical waveguide나 optical fiber를 통해 switch unit 간 연결이 가능합니다. 이것이 "internal shuffle"의 핵심 기술적 기반입니다.

대신, 이러한 NVIDIA CPO switch는 매우 높은 port 수를 갖는 것으로 나타낸다. 예를 들어 Quantum 3450은 800G의 144 port를 제공하고 Spectrum 6800은 800G의 512 port를 제공한다. 우리가 기본적으로라는 단어를 사용하는 이유는 NVIDIA의 non-CPO InfiniBand Quantum Q3400 switch도 800G의 144 port를 제공하지만, QM9700과 같은 다른 InfiniBand switch는 800G의 32 port만 제공하기 때문이며, 전자만이 많은 수의 효과적인 port를 제공하기 위해 이 **high radix in a box**를 제공한다. 이러한 높은 port 수는 잠재적으로 고객이 three-Layer에서 2-layer network

로 network를 평탄화할 수 있게 하며, shuffle box와 patch panel 또는 다루기 어려운 octopus cable을 배포하는 번거로움을 고객에게 덜어주어 핵심 판매 포인트가 될 수 있다. 2-layer의 경우, transceiver 전력은 84% 감소하고, switch 전력은 21% 감소하며, 전체 networking 전력은 전통적인 DSP transceiver 대비 48% 감소할 수 있다.

Spectrum 6800 switch는 사용 가능한 두 가지 logical configuration 모두에서 많은 수의 port를 갖추고 있으며(800G의 512 port), 800G의 128 port, 400G의 256 port 또는 200G의 512 port를 제공하는 Spectrum 6810과 비교할 때 특히 이를 가능하게 한다. Spectrum 6810을 사용하는 800G의 128 port 옵션의 경우, network는 2-layer network에서 최대 8,192 GPU를 연결할 수 있는 반면, 800G의 512 port를 갖춘 Spectrum 6800은 131,072 GPU를 연결할 수 있다.

핵심 인사이트: Network Scaling 수학

L-layer network에서 k port의 switch로 지원 가능한 최대 host 수: $\text{Max Hosts} = k^L$

Port 수의 지수적 효과:

- 2-layer network에서 128 port → 최대 $128^2 = 16,384$ hosts
- 2-layer network에서 512 port → 최대 $512^2 = 262,144$ hosts
- Port를 2배로 늘리면 → 지원 가능한 host가 2^L 배 증가
- 2-layer에서 port 2배 증가 → host 4배 증가!

Port 증가 방법:

- Internal shuffle (CPO 스위치)
- Breakout cable (800G → 2x400G)
- Twin-port transceiver

이것이 CPO의 high radix가 network layer를 줄이고 연결 가능한 GPU 수를 극적으로 늘리는 이유입니다.

간단히 언급하자면, L-layer network에서 k port의 switch를 사용하여 지원할 수 있는 최대 host 수는 다음과 같다: 마법은 port 수 k가 layer 수로 거듭제곱된다는 사실에서 나온다. 따라서 2-layer network의 경우, internal shuffle (Spectrum 6800의 경우처럼), breakout cable 또는 twin-port transceiver를 사용하여 port당 bandwidth의 절반을 할당하여 (즉, 800G port를 두 개의 400G port로 분할) logical port 수를 두 배로 늘리면 지원되는 host가 4배가 된다!

지금까지 이 섹션에서 논의된 전력 절감, 3-layer CPO network의 경우 23%, 2-layer CPO network로 내려가는 경우 48%는 환상적으로 들리지만, 문제는 networking이 3-layer network의 경우 전체 cluster 전력의 단지 9%에 불과하다는 것이다. 따라서 결국 CPO로 전환하는 영향은 적어도 scale-out network의 경우 상당히 희석된다. 3-layer network에서 CPO를 사용하도록 전환하면 networking 전력이 23% 감소하지만 전체 cluster 전력 절감은 단지 2%에 불과하다. 2-layer network로 이동하면 networking 비용이 48% 낮아지지만, 전체 cluster 전력 절감은 단지 4%에 불과하다.

핵심 인사이트: 전력 절감의 희석 효과

CPO는 네트워킹 전력을 23-48% 절감하지만, 네트워킹이 전체 클러스터 전력의 9%에 불과 하므로 전체 클러스터 관점에서는 2-4%의 전력 절감만 달성합니다. 이는 scale-out 네트워크에서 CPO 채택이 더딘 주요 이유 중 하나입니다. GPU 자체가 전력의 대부분을 소비하기 때문에, 네트워킹 개선의 효과가 희석되는 것입니다.

GB300 NVL72 Cluster Power Budget Breakdown, per Rack-Scale 72-GPU Server						
Item	2-Layer Network (up to 10,368 GPUs)		3-Layer Network (up to 746,496 GPUs)		4-Layer Network (up to ~53.7M GPUs)	
	Power (W)	%	Power (W)	%	Power (W)	%
Server	142,000	93%	142,000	91%	142,000	88%
Optical Transceivers	4,600	3%	6,184	4%	6,976	4%
Switches	5,214	3%	8,014	5%	11,606	7%
Fiber, Cables, Software, Others	0	0%	0	0%	0	0%
Networking	9,814	6%	14,198	9%	18,582	12%
All Others	281	0%	281	0%	281	0%
Total Power	152,095	100%	156,479	100%	160,863	100%

Power for a Neocloud Giant InfiniBand cluster using X800-Q3400 Back-end switches with 144 ports of 800G
Source: SemiAnalysis Networking Model

Figure 6: Source: SemiAnalysis AI Networking Model

GB300 NVL72 Cluster Power Budget: Traditional vs CPO Network, per Rack-Scale 72-GPU Server								
Item	DSP Transceivers (3-Layer Network)		LPO Transceivers (3-Layer Network)		CPO (3-Layer Network)		CPO (2-Layer Network)	
	Power (W)	Δ Power (%)	Power (W)	Δ Power (%)	Power (W)	Δ Power (%)	Power (W)	Δ Power (%)
Server	142,000	0%	142,000	0%	142,000	0%	142,000	0%
Optical Transceivers	6,184	0%	3,935	-36%	1,000	-84%	1,000	-84%
Switches	8,014	0%	8,014	0%	9,884	23%	6,336	-21%
Fiber, Cables, Software, Others	0	0%	0	0%	0	0%	0	0%
Networking	14,198	0%	11,949	-16%	10,884	-23%	7,336	-48%
All Others	281	0%	281	0%	281	0%	281	0%
Total Power	156,479	0%	154,230	-1%	153,165	-2%	149,617	-4%

Neocloud Giant InfiniBand cluster using Nvidia x800-Q3400 Back-end switches with 144 ports of 800G and Nvidia LinkX Transceivers. CPO used only in Back-end network.
Source: SemiAnalysis Networking Model

Figure 7: Source: SemiAnalysis AI Networking Model

전체 cluster 자본 비용을 살펴볼 때도 비슷한 상황이다.

1.3 CPO Scale-out Total Cost of Ownership (TCO)

Transceiver를 CPO 솔루션과 비교할 때 비용 세부사항을 간략히 살펴보자. 첫 번째 NVIDIA CPO Switch인 Quantum X800-Q3450 CPO는 각각 1.6Tbit/s로 작동하는 72개의 optical engine을 사용할 것이다. 이후 버전의 Quantum CPO switch는 각각 3.2Tbit/s에서 작동하는 36개의 optical engine으로 전환할 가능성이 있으며, 단위당 (FAU 포함) ~\$1,000의 비용으로 시스템당 총 OE 비용은 \$36K가 된다.

이를 관점에서 보기 위해, 전통적인 optical transceiver module을 대신 사용한다면 총 비용이 얼마나 될지 고려해보자. non-CPO X800-Q3400은 72개의 OSFP cage를 갖추고 있으며, 800G에서 144 port를 제공하기 위해 1.6T twin-port transceiver가 사용된다.

기술 용어: Twin-Port Transceiver와 OSFP Cage

1. OSFP (Octal Small Form Factor Pluggable) Cage

정의: Switch 전면 패널에 있는 물리적 slot으로, pluggable optical transceiver를 삽입하는

socket입니다.

특징:

- 물리적 크기: QSFP보다 큰 form factor, 고전력/고대역폭 transceiver 지원
- 전력 처리: 최대 15-20W transceiver 지원 가능
- Hot-swappable: 시스템 가동 중에도 transceiver 교체 가능
- 1 cage = 1 transceiver: 물리적으로 하나의 cage는 하나의 transceiver를 수용

2. Twin-Port Transceiver의 개념

정의: 하나의 물리적 transceiver 모듈 내부에 **2개의 독립적인 광 채널**을 통합한 transceiver입니다.

구체적 예시: **1.6T Twin-Port = 2 × 800G**

Traditional Single-Port 방식:

- 1개의 800G transceiver = 1개의 OSFP cage 점유 = 1개의 800G port 제공
- 144개의 800G port를 만들려면 → 144개의 OSFP cage 필요
- 문제: Switch 전면 패널에 144개의 cage를 물리적으로 배치하기 어려움

Twin-Port 방식:

- 1개의 1.6T twin-port transceiver = 1개의 OSFP cage 점유 = **2개의 800G port 제공**
- 144개의 800G port를 만들려면 → 72개의 OSFP cage만 필요 ($144 \div 2 = 72$)
- 이점: 동일한 bandwidth를 절반의 물리적 공간에 구현

3. X800-Q3400의 Port 구성 계산

- 물리적 구성: 72개의 OSFP cage
- 사용 transceiver: 72개의 1.6T ($2 \times 800G$) twin-port transceiver
- 논리적 결과: $72 \times 2 = 144$ 개의 800G port
- 총 bandwidth: $144 \times 800G = 115.2$ Tbps

4. Twin-Port의 내부 구조

하나의 1.6T twin-port transceiver 내부:

- 2개의 독립적인 laser: 각각 800G 신호 생성
- 2개의 독립적인 modulator: 각각 데이터를 광 신호로 변환
- 2개의 fiber output: 물리적으로 2개의 광섬유 출력
- 공유 DSP chip: 하나의 DSP가 두 채널 모두 처리 (비용 절감)
- 공유 cooling: 단일 heat sink로 두 채널의 열 관리

5. 비용 계산 예시 (X800-Q3400)

- 1개의 1.6T DR8 twin-port transceiver 가격: ~\$1,000
- 필요한 transceiver 수: 72개 (cage 수와 동일)
- 총 transceiver 비용: $72 \times \$1,000 = \$72,000$
- 제공되는 논리적 port: 144개의 800G port
- Port당 비용: $\$72,000 \div 144 = \sim \$500/\text{port}$

6. Twin-Port vs CPO 비교

Traditional Twin-Port:

- 72개의 1.6T transceiver = \$72,000
- 각 transceiver는 독립적으로 교체 가능
- Vendor 선택의 자유, 가격 협상 가능

CPO:

- 72개의 1.6T optical engine + ELS = ~\$36,000 (원가)
- 하지만 switch vendor margin (60%) 적용 시 = \$80-90,000
- Switch에 통합되어 교체 불가, vendor lock-in

핵심 통찰: Twin-port transceiver는 물리적 cage 수를 절반으로 줄여 **높은 port 밀도**를 달성하는 중간 단계 기술입니다. CPO는 이를 한 단계 더 발전시켜 transceiver 자체를 제거하고 switch 내부에 통합하지만, vendor margin과 lock-in 이슈가 발생합니다.

일반적인 1.6T DR8 transceiver의 비용을 \$1,000로 가정하면, 이 switch를 채우기 위한 총 transceiver 비용은 \$72,000가 되며, 이는 CPO switch에 동일한 양의 bandwidth를 제공하는 데 필요한 optical engine 및 ELS의 추정 비용 \$35-40k의 두 배이다.

핵심 인사이트: CPO 비용의 합정 - Vendor Margin

Component 비용 vs 최종 구매 가격:

- 72개 1.6T OE + ELS 원가: ~\$36k
- 60% gross margin 적용 시 최종 가격: \$80-90k
- 비교: 72개 1.6T DR8 transceiver: \$72k (margin 포함)

핵심 문제:

- CPO component가 switch에 통합되면서 모든 component가 switch vendor의 margin 적용 대상
- Transceiver는 별도 구매 시 경쟁 시장에서 가격 협상 가능
- CPO는 vendor lock-in 심화 → 협상력 상실
- Fiber shuffle 등 추가 component도 margin stack 적용

결론: CPO의 기술적 비용 절감이 vendor margin과 협상력 상실로 희석됩니다. 실제 TCO 절감은 기대보다 작을 수 있으며, 이는 scale-out에서 빠른 채택이 어려운 중요한 경제적 이유입니다.

그러나 이것은 switch vendor의 margin을 고려하지 않은 것이다. 60% gross margin을 적용한다면, 최종 구매자에게 optical engine 비용은 \$80k-\$90k USD가 되어 transceiver에 상응하는 것보다 더 높은 비용이 된다. Fiber shuffle과 같은 다른 component들도 이러한 margin stack의 대상이 될 것이다. 이는 transceiver에 지불하는 비용과 switch vendor가 취하는 margin에 따라 CPO로 전환할 때 비용 절감이 극적이지 않을 수 있는 이유를 설명한다.

아래 표에서 볼 수 있듯이, 3-layer network에서 transceiver에서 CPO로 전환할 때, CPO component에 추가로 취해지는 margin이 switch 비용을 81% 증가시켜 transceiver를 구매하지 않음으로써 얻는 86% 절감에서 차감된다. 총 networking 비용은 여전히 DSP transceiver를 사용하는 것보다 CPO에서 31% 낮지만, 전력의 경우와 마찬가지로 server rack이 cluster TCO에서 차지하는 지배적인 점유율은 전체 cluster 비용이 단지 3%만 감소함을 의미한다.

3-layer 대신 2-layer로 network를 평탄화하면 더 많은 비용 절감을 제공할 수 있다. 최대 7%의 전체 cluster 비용 감소, transceiver 비용은 86% 감소, 그리고 전체 networking 비용은 46% 감소한다.

GB300 NVL72 Cluster Capital Costs: Traditional vs CPO Network, per Rack-Scale 72-GPU Server								
Item	DSP Transceivers (3-Layer Network)		LPO Transceivers (3-Layer Network)		CPO (3-Layer Network)		CPO (2-Layer Network)	
	Cost	Δ Cost (%)	Cost	Δ Cost (%)	Cost	Δ Cost (%)	Cost	Δ Cost (%)
Server Cost	\$ 4,357,943	0%	\$ 4,357,943	0%	\$ 4,357,943	0%	\$ 4,357,943	0%
Optical Transceivers	\$ 499,925	0%	\$ 434,146	-13%	\$ 71,525	-86%	\$ 71,525	-86%
Switches	\$ 312,681	0%	\$ 312,681	0%	\$ 564,821	81%	\$ 359,965	15%
Fiber, Cables, Software, Others	\$ 24,338	0%	\$ 24,338	0%	\$ 24,338	0%	\$ 17,723	-27%
Networking Cost	\$ 836,944	0%	\$ 771,164	-8%	\$ 660,684	-21%	\$ 449,213	-46%
All Others	\$ 269,152	0%	\$ 269,152	0%	\$ 269,152	0%	\$ 269,152	0%
Total Cost	\$ 5,464,039	0%	\$ 5,398,259	-1%	\$ 5,287,779	-3%	\$ 5,076,308	-7%

Neocloud Giant InfiniBand cluster using Nvidia x800-Q3400 Back-end switches with 144 ports of 800G and Nvidia LinkX Transceivers. CPO used only in Back-end network.
Source: SemiAnalysis Networking Model

Figure 8: Source: [SemiAnalysis AI Networking Model](#)

그렇다면 CPO가 한편으로는 최대 7%의 비용 절감과 최대 4%의 전력 절감만을 제공하지만, 다른 한편으로는 어려운 현장 서비스, 신뢰성과 blast radius에 대한 우려 (정당하든 아니든), 그리고 여러 transceiver vendor와 가진 협상력의 상실에 대한 우려를 제기한다면 왜 GPU cloud에 의해 채택되고 있는가? 간단한 답은 아직 널리 채택되지 않고 있다는 것이다. 우리는 가까운 장래에 hyperscaler 내에서 scale-out CPO 시스템에 대한 빠른 채택 곡선을 기대하지 않는다.

1.4 CPO for Scale-up Networks

핵심 인사이트: Scale-up이 CPO의 Killer Application인 이유

Scale-out vs Scale-up 비교:

- Scale-out:** 서버 간 통신 (GPU당 800Gbit/s), 전체 클러스터 전력의 9%만 차지 → CPO 효과 희석
- Scale-up:** 동일 시스템 내 GPU 간 통신 (GPU당 7,200Gbit/s - 9배 더 높음), 전력과 비용에서 훨씬 큰 비중 → CPO 효과 극대화

Scale-up 네트워크는 더 높은 bandwidth, 더 낮은 latency를 요구하며, copper의 물리적 한계(2m 도달거리)를 극복해야 합니다. CPO는 이러한 문제를 해결하면서 world size(단일 scale-up domain의 GPU 수)를 획기적으로 늘릴 수 있는 유일한 솔루션입니다.

반대로, 우리는 scale-up을 위한 CPO를 killer application으로 본다. 앞서 언급했듯이, 주요 hyperscaler는 이미 10년 말까지 CPO 기반 scale-up 솔루션 배포를 위해 공급업체와 약속을 하고 있다. 현재 기존의 copper 기반 scale-up paradigm은 copper cable의 제한된 도달 거리 (lane당 200Gbit/s로 실행될 때 기껏해야 2 meter) 및 lane당 bandwidth를 두 배로 늘리는 것의 증가하는 어려움으로 인해 한계에 도달하고 있다. CPO는 bandwidth density 요구사항을 충족하고, 미래에도 bandwidth를 확장하기 위한 여러 방법을 제공하며, 훨씬 더 큰 scale-up world size를 가능하게

함으로써 이러한 문제를 해결할 수 있다.

CPO가 scale-up networking에 배포되면, scale-up domain은 더 이상 interconnect 도달 거리에 의해 제한되지 않을 것이다. 원칙적으로 고객은 scale-up domain을 임의로 큰 크기로 성장시킬 수 있을 것이다. 물론 scale-up domain은 all-to-all connection을 허용하는 single-tier fan-out network 내에 유지하고 싶다면, scale-up domain 크기는 switch radix에 의해 제한될 것이다.

1.4.1 Scale-out vs Scale-up TAM

Scale-up fabric의 networking 요구사항은 back-end scale-out network의 요구사항보다 훨씬 더 까다롭다. GPU 간 또는 switch link는 GPU가 상호 연결되어 memory와 같은 resource를 coherently 하게 공유할 수 있도록 훨씬 더 높은 bandwidth와 더 낮은 latency를 필요로 한다.

예를 들어, NVIDIA Blackwell의 5세대 NVLink는 GPU당 900GB/s의 uni-directional bandwidth를 제공한다.

기술 용어: NVLink, NIC, SerDes

NVLink: NVIDIA의 독점 GPU 간 interconnect 기술

- GPU 간 고속 직접 통신 제공
- Blackwell (5세대): GPU당 900GB/s
- Scale-up 네트워크에 사용되며, memory coherency 지원
- Scale-out (100GB/s)보다 9배 높은 bandwidth

NIC (Network Interface Card):

- GPU를 scale-out network에 연결하는 어댑터
- CX-8: GB300 NVL72에 사용, GPU당 100GB/s 제공
- RDMA를 통한 낮은 latency 통신 지원

SerDes (Serializer/Deserializer):

- 병렬 데이터를 직렬로 변환하여 고속 전송
- Line speed: SerDes가 동작하는 속도 (lane당 200Gbit/s, 224Gbit/s 등)
- Higher line speed → 더 높은 bandwidth density, 하지만 signal integrity 도전

이는 back-end scale-out network (GB300 NVL72용 CX-8 Network Interface Card (NIC) 사용)에서 GPU당 100GB/s보다 GPU당 9배 더 많은 bandwidth이다. 이는 또한 host로부터 훨씬 더 높은 shoreline bandwidth density에 대한 필요성을 만들어내며, 이는 GPU SerDes line speed를 앞으로 밀어붙이는 원동력이 되어왔다.

Scale-up domain의 크기가 증가하고 scale-up interconnect의 속도도 증가함에 따라, scale-up interconnect (그리고 결국 scale-up CPO)의 TAM이 이미 scale-out networking의 TAM을 상당히 압도했다는 것을 인식하는 것이 중요하다. CPO TAM은 scale-out networking application보다는 scale-up에 의해 지배될 가능성이 높다.

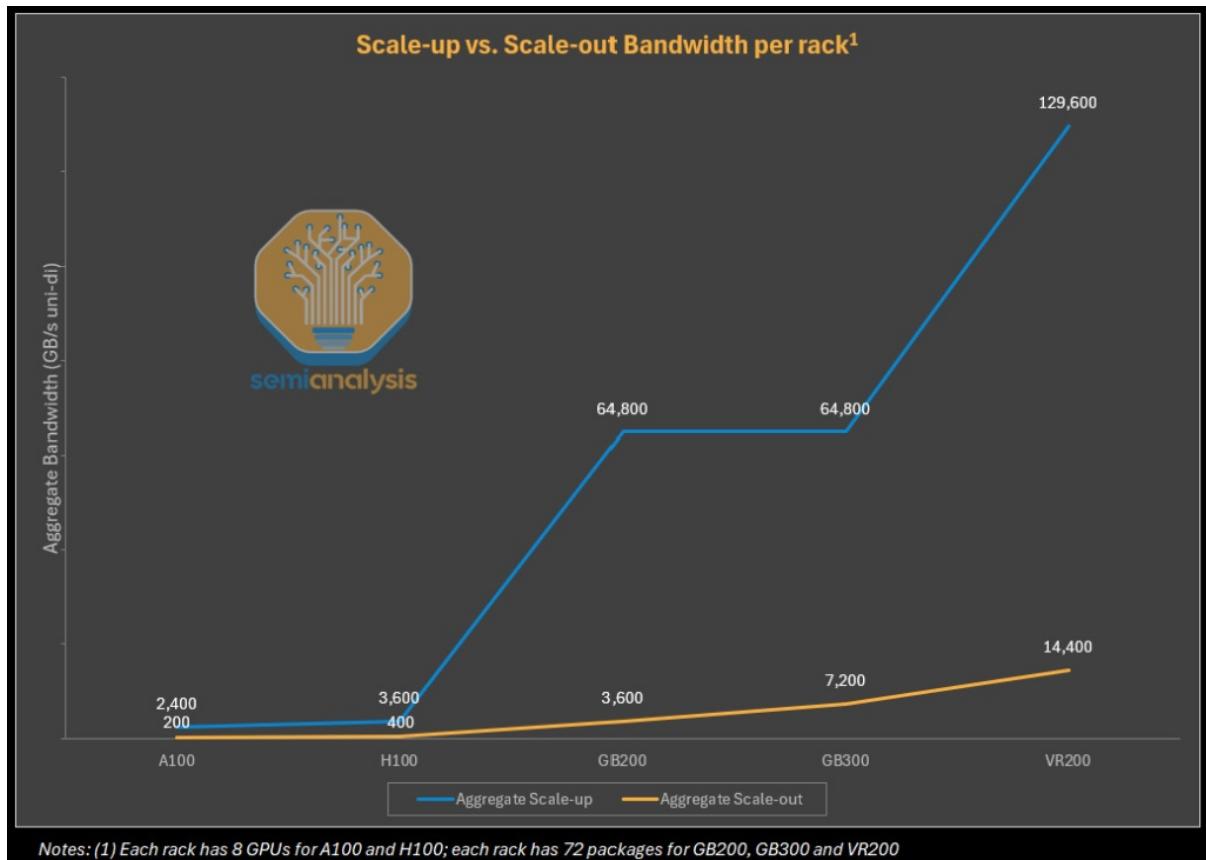


Figure 9: Source: SemiAnalysis

1.4.2 Copper vs Optics for Scale-Up: World Sizes, Density and Reach

현재 scale-up network는 좋은 이유로 전적으로 **Copper**에서 실행된다. 현재 pluggable paradigm에서 optical transceiver로 NVLink bandwidth를 맞추는 것은 비용과 전력 측면에서 엄청나게 비쌀 뿐만 아니라 원치 않는 latency를 도입할 것이다. 또한 compute tray의 face-plate 공간이 이러한 모든 transceiver를 장착하기에 충분하지 않을 수도 있다. Copper는 이러한 low-latency, high-throughput connection에서 탁월하다. 그러나 위에서 언급했듯이, copper의 제한된 도달 거리는 world size (단일 scale-up domain 내에서 연결할 수 있는 GPU 수)를 제한한다.

기술 용어: World Size와 그 중요성

World Size: 단일 scale-up domain 내에서 직접 연결될 수 있는 GPU의 총 개수를 의미합니다.

중요성:

- 더 큰 world size → 더 정교한 collective communication 가능
- 추론 기반 모델 스케일링과 test-time compute에서 핵심적
- GB200이 8 GPU → 72 GPU로 확장하여 획기적인 성능 향상 달성
- Copper 한계: 2m 도달거리로 인해 단일 rack 내로 제한 (144 GPU까지)
- CPO로 여러 rack에 걸친 확장 가능 → world size를 수천 GPU로 증가

Copper Cable의 2m 도달거리 제한 이유:

고속 electrical signal (112 Gbps SerDes)이 copper cable을 통과할 때:

- **Signal attenuation (신호 감쇠):** Copper의 저항으로 인해 거리에 따라 신호 세기가 지수적으로 감소. 고주파수 신호일수록 감쇠가 급격히 증가하며, 112 Gbps에서는 2m 이상에서 신호가 noise level 이하로 떨어짐
- **Skin effect:** 고주파수에서 전류가 도체 표면에만 흐르면서 유효 단면적 감소 → 저항 증가 → 더 큰 신호 손실
- **Dispersion (분산):** 서로 다른 주파수 성분이 다른 속도로 전파되어 신호 파형이 왜곡되고 bit error rate 급증
- **Crosstalk (누화):** 인접한 copper wire 간 전자기 간섭으로 신호 무결성을 저하. 밀집된 cable bundle에서 특히 심각
- **Impedance mismatch:** Cable, connector, PCB 간 impedance 불일치로 signal reflection 발생 → 신호 품질 저하

대조: Optical fiber는 빛으로 신호를 전송하므로 전자기 간섭, 저항, skin effect 등의 문제가 없어 수 km까지도 신호 무결성을 유지할 수 있습니다. 이것이 long-reach 연결에 optics가 필수적인 물리적 이유입니다.

World size 증가는 inference workload에서 더 많은 compute, memory capacity, memory bandwidth를 제공하여 모델 성능을 극적으로 향상시킵니다.

Scale-up world size를 증가시키는 것은 compute scaling의 매우 중요한 방법이다. 단일 scale-up domain에서 더 많은 compute, memory capacity 및 memory bandwidth를 추가하는 것은 inference 기반 model scaling 및 test time compute의 오늘날 체제에서 점점 더 중요해지고 있다. NVIDIA의 GB200 시스템은 world size를 단지 8개의 상호 연결된 GPU에서 all-to-all topology로 72개의 상호 연결된 GPU로 가져왔기 때문에 엄청난 성능 향상을 제공했다. 그 결과 scale-out network에서는 실현 가능하지 않은 더 정교한 collective communication 기술을 구현함으로써 엄청난 throughput 향상이 가능해졌다.

Copper에서 이것은 단일 rack의 footprint 내에서만 수행될 수 있어 power delivery, thermal management 및 manufacturability에 대한 엄청난 수요를 만들어냈다. 이 시스템의 복잡성으로 인해 downstream supply chain은 여전히 capacity를 늘리는 데 어려움을 겪고 있다.

NVIDIA는 계속해서 copper를 고집할 것이다. 그들은 또한 자체 scale-up network로 따라잡고 있는 AMD 및 hyperscaler와 같은 경쟁자보다 앞서 나가기 위해 scale-up world size를 더욱 높게 밀어붙여야 한다. 따라서 NVIDIA는 단일 rack 내에서 scale-up domain을 확장하기 위해 극단적인 조치를 취해야 한다. GTC 2025에서 보여진 Rubin Ultra를 위한 NVIDIA의 극단적인 Kyber rack architecture는 144 GPU package (576 GPU die)까지 확장할 수 있다. 이 rack은 이미 밀도가 높은

GB200/300 NVL72 rack보다 4배 더 밀도가 높다. GB200이 이미 제조 및 배포하기에 매우 복잡한 상황에서, Kyber는 이를 다음 단계로 끌어올린다.

Optics는 반대 접근 방식을 가능하게 하며, power delivery 및 thermal density에 어려움이 있는 밀도 높은 footprint에 더 많은 accelerator를 넣는 대신 여러 rack에 걸쳐 확장하여 world size를 증가시킨다. 이것은 오늘날 pluggable transceiver로 가능하지만, 다시 optical transceiver의 비용과 높은 전력 소비가 이를 비실용적으로 만든다.



Figure 10: Source: SemiAnalysis

1.4.3 Copper vs Optics for Scale-Up: Scaling Bandwidth

Copper에서 bandwidth를 확장하는 것도 점점 더 어려워지고 있다. Rubin을 통해 NVIDIA는 novel bi-directional SerDes 기술을 구현하여 bandwidth의 두 배를 달성하고 있으며, 여기서 transmit 와 receive operation이 동일한 channel을 공유하여 channel당 224Gbit/s transmit + 224Gbit/s receive로 full-duplex communication을 가능하게 한다. Copper에서 lane당 진정한 448G를 달성하는 것은 불확실한 시장 출시 시기를 가진 또 다른 도전적인 업적으로 남아 있다.

핵심 인사이트: Copper vs Optics - Bandwidth Scaling

Copper의 한계:

- Lane당 200Gbit/s에서 signal integrity 문제
- Bi-directional SerDes로 224Gbit/s Tx + 224Gbit/s Rx 달성 (Rubin)
- 진정한 448Gbit/s 달성을 불확실한 timeline
- 도달거리 2m로 제한 (200Gbit/s lane 기준)
- Bandwidth 증가 방법이 제한적

CPO/Optics의 Bandwidth Scaling 방법:

- **Baud Rate 증가:** Symbol rate를 높여 bandwidth 증가
- **DWDM (Dense Wavelength Division Multiplexing):** 단일 fiber에 여러 파장 사용
- **Additional Fiber Pair:** 더 많은 fiber로 병렬 전송
- **Advanced Modulation:** PAM4, QAM 등으로 bit/symbol 증가

Optics는 copper보다 훨씬 더 많은 bandwidth scaling 경로를 제공하여 미래의 요구사항을 충족할 수 있습니다.

이와 대조적으로, CPO는 bandwidth를 확장하기 위한 여러 방법을 제시한다: Baud rate, DWDM, Additional Fiber Pair 및 Modulation 이 모든 것은 이 문서의 후반부에서 자세히 논의될 것이다.

1.5 When will CPO be ready for Primetime?

그렇다면, CPO가 솔루션이라면 왜 NVIDIA는 Rubin Ultra에 대해 이것을 추구하지 않고 먼저 scale-out switch에 대해서만 추구하는가? 이것은 supply chain 미성숙, manufacturing 과정 및 배포에 대한 고객 주저로 돌아간다. Quantum 및 Spectrum CPO switch는 supply chain을 늘리고 datacenter에서 신뢰성과 serviceability에 대한 더 많은 실제 데이터를 얻는 데 도움을 주기 위해 도입되었다.

그 사이에, ECOC 주변에 발표된 Meta의 CPO 신뢰성 데이터는 몇 가지 유용한 정보를 제공한다. Meta는 이 연구를 위해 Broadcom과 협력했으며, Broadcom도 유용한 슬라이드를 제시했다. 이 연구에서 Meta는 15개의 Bailly 51.2T CPO Switch에 걸쳐 최대 1,049k 400G port device hour에 걸친 상당한 규모의 테스트 실행을 수행하고 최대 non-zero KP4 Forward Error Correction (FEC) bin을 발표했다.

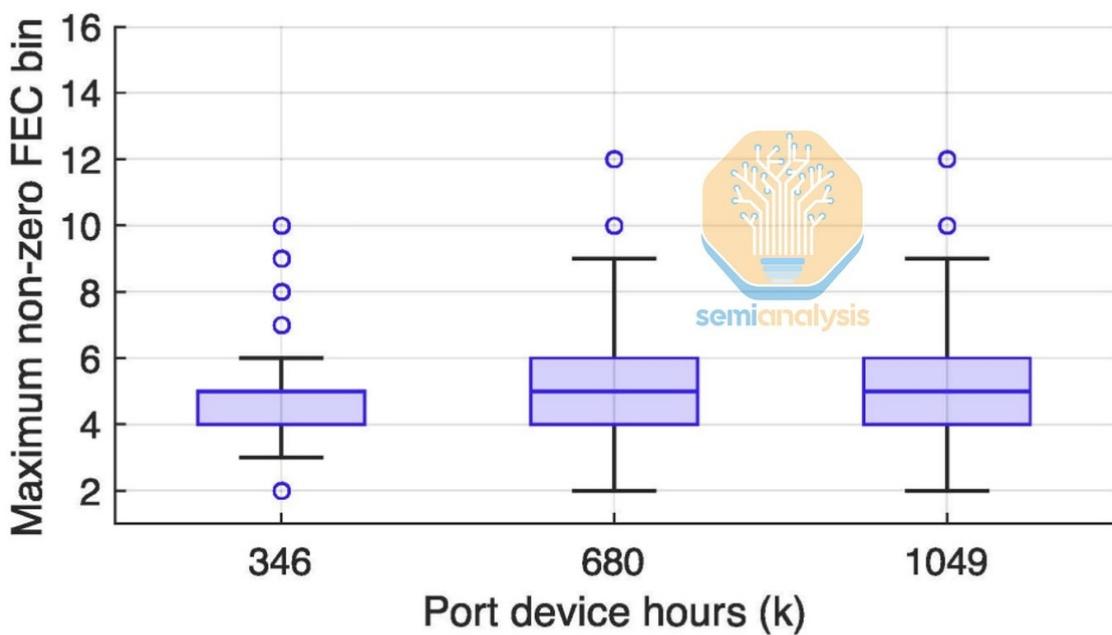


Fig. 5: Maximum non-zero FEC bin vs. 400Gbps-port device hours of operation at 40°C ambient temperature.

Figure 11: Source: Meta

논문은 또한 테스트 기간 동안 link에서 failure 또는 Uncorrectable Code Word (UCW) 가 관찰되지 않았으며, 전체 테스트 기간에 걸쳐 최대 1,049k 400G port device hour까지 FEC bin > 10의 단 하나의 사례만 관찰되었다고 설명했다.

기술 용어: 신뢰성 측정 지표

FEC (Forward Error Correction): 데이터 전송 중 발생하는 오류를 자동으로 수정하는 기술. FEC bin은 오류 수준을 나타내며, bin > 10은 높은 오류율을 의미합니다.

UCW (Uncorrectable Code Word): FEC로도 수정할 수 없는 심각한 오류로, 데이터 손실을 초래합니다. UCW 발생 시 link failure로 이어질 수 있습니다.

MTBF (Mean Time Before Failure): 평균 고장 간격 시간으로, 시스템의 신뢰성을 나타내는 핵심 지표입니다.

- 기존 2xFR4 transceiver: 0.5-1M device hour MTBF
- CPO 시스템: 2.6M device hour MTBF (2.6-5배 향상)

Meta의 테스트 결과는 CPO가 기존 트랜시버보다 더 높은 신뢰성을 가질 수 있음을 시사하지만, 대규모 production 환경에서의 검증이 여전히 필요합니다.

그러나 Meta는 거기서 멈추지 않았다. ECOC에서 동일한 논문을 발표하는 강연에서, 그들은 최대 15M 400G port-device hour에 대한 확장된 결과를 발표했다. 이러한 결과는 첫 4M 400G port device hour 동안 UCW가 없었음을 보여주었으며, 그들은 또한 400G 2xFR4 transceiver (전 세계적으로 2xFR4의 경우 550k)에 대해 0.5-1M device hour Mean Time Before Failure

(MTBF) vs CPO에 대해 2.6M device hour MTBF를 보여주었다.

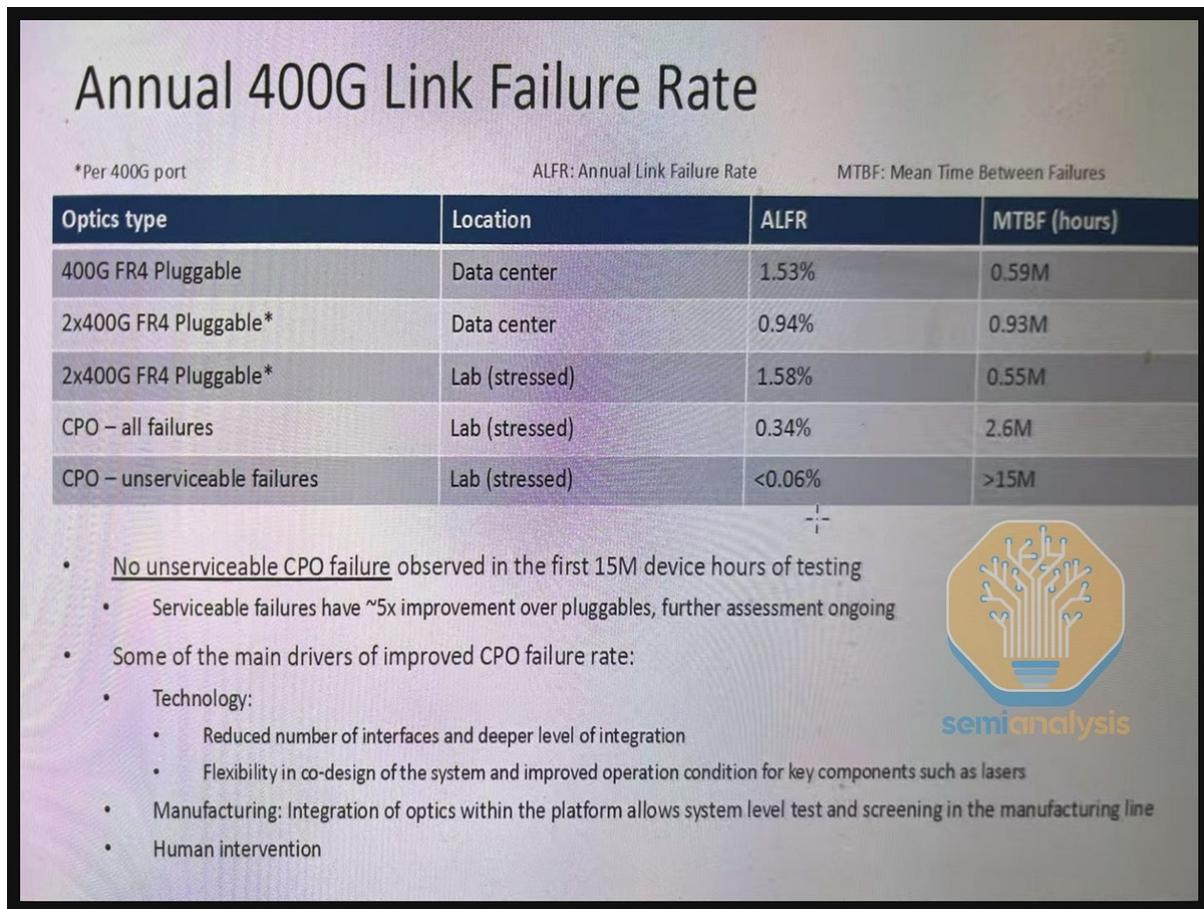


Figure 12: Source: Meta

15M port device hour가 큰 숫자처럼 들릴 수 있지만, 이것은 400G port hour 단위이다. 그래서 하나의 51.2T switch를 1 hour 동안 작동시키면 128 400G port hour를 의미한다. 15개의 51.2T switch에 걸쳐 15M 400G Port hour는 7,812 wall-clock hour 또는 약 325일을 의미한다. 실제로 이 15M hour 숫자는 종종 단순히 **hour** 또는 **device hour**로 인용되며 **port** 부분이 생략된다. 4M port device hour까지 zero failure 및 zero UCW 통계는 매우 유용하지만, 산업계가 CPO scale-out switching으로 전환하고 이 기술에 수십억 달러를 투입하기 전에 lab 환경에서 11개월 동안 테스트된 단지 15개의 CPO switch보다 훨씬 더 많은 것이 필요하다.

동적인 현장 환경에서 수천 개의 scale-out switch를 작동하는 것은 전혀 다른 도전이며, 이러한 switch가 production 환경에서 어떻게 작동할지는 아직 두고 봐야 한다. Production 환경에서의 temperature 변동은 lab보다 높을 수 있으며, 이는 component 성능이나 endurance에서 예상치 못한 변동을 초래할 수 있다. [Meta의 자체 Llama 3 논문은 datacenter에서 1-2% temperature 변동을 인용했으며](#) 이는 전력 소비 변동에 악영향을 미쳤다. 이러한 변동이 예상하기 어려운 방식으로 전체 network fabric에 영향을 미칠 수 있을까?

One interesting observation is the impact of environmental factors on training performance at scale. For Llama 3 405B, we noted a diurnal 1-2% throughput variation based on time-of-day. This fluctuation is the result of higher mid-day temperatures impacting GPU dynamic voltage and frequency scaling.



During training, tens of thousands of GPUs may increase or decrease power consumption at the same time, for example, due to all GPUs waiting for checkpointing or collective communications to finish, or the startup or shutdown of the entire training job. When this happens, it can result in instant fluctuations of power consumption across the data center on the order of tens of megawatts, stretching the limits of the power grid. This is an ongoing challenge for us as we scale training for future, even larger Llama models.

Figure 13: Source: [Meta](#)

Datacenter의 먼지와 같은 평범하게 들리는 문제조차도 fiber 끝을 청소하는 데 상당한 시간을 보낼 수 있는 지원 기술자의 골칫거리이다. 물론 CPO switch는 LC 또는 MPO type front pluggable connector를 가지고 있지만, CPO switch chassis 내부의 먼지는 어떨까? 0.06%의 unserviceable failure rate는 매력적으로 들리지만, 그러한 failure는 64 800G port의 blast radius를 가진다. 이 논문은 또한 FR optics 기반 CPO switch에 초점을 맞추고 있지만, 다음 세대의 CPO switch는 DR optics를 기반으로 할 것이다. 이것들은 단지 몇 가지 알려진 미지수이며, field testing에서 나올 수 있는 더 많은 알려지지 않은 미지수가 잠재적으로 있다.

실제로 이러한 결과는 tangible한 신뢰성 데이터를 제공함으로써 업계의 사람들을 설득하는 측면에서 영향력이 있었다. 여기서 우리의 요점은 **Fear, Uncertainty 또는 Doubt (FUD)**를 만들기 위한 것이 아니라, 업계가 예상치 못한 문제를 신속하게 이해하고 해결하여 더 광범위한 CPO 채택, 특히 scale-up networking을 위한 길을 밟을 수 있도록 훨씬 더 대규모의 field testing을 요구하기 위한 것이다. 결국, NVIDIA의 scale-out CPO 제품 출시는 실제 high-volume 배포를 위한 practice run 및 pipe-cleaner 역할을 하고 있다. 우리는 scale-out vs scale-up에 대한 훨씬 더 설득력 있는 TCO 및 Performance/TCO 이점을 고려할 때 scale-up에 대한 배포가 훨씬 더 크고 영향력이 있을 것이라고 생각한다.

더욱이 scale-out CPO와 관련하여, Rubin Ultra는 2027년 출시를 목표로 하고 있지만 (우리는 그것이 결국 2027년 후반으로 예상함) supply chain은 GPU 수요를 지원하기 위해 수천만 개의 이러한 CPO endpoint를 출하할 준비가 되지 않을 것이다. 이 timeline조차 NVIDIA에게는 너무 야심적이다. 이것이 Feynman 세대가 NVIDIA ecosystem에 CPO injection의 초점이 되는 것처럼 보이는 이유이다.

핵심 인사이트: CPO 배포 전략 - Scale-out이 먼저인 이유

핵심 질문: CPO가 그렇게 좋다면, 왜 NVIDIA는 Rubin Ultra GPU (scale-up)에 먼저 적용하지 않고 datacenter switch (scale-out)부터 시작하는가?

1. 물량의 차이 (Scale 문제)

Scale-out Switch (먼저 시작):

- 필요 수량: 대형 datacenter 하나당 수천 개의 switch
- CPO endpoint 수: 수천 ~ 수만 개 수준
- 공급망 부담: 상대적으로 관리 가능
- 실패 시 영향: 제한적 - 일부 고객의 일부 switch만 영향

Scale-up GPU (진짜 목표, 나중):

- 필요 수량: 전 세계 모든 AI datacenter의 모든 GPU
- CPO endpoint 수: 수천만 ~ 수억 개 수준 (Rubin Ultra 기준)

- **공급망 부담:** 엄청난 manufacturing capacity 필요
- **실패 시 영향:** 치명적 - 전체 GPU 출하 지연 또는 중단

구체적 예시:

- 200K GPU cluster: ~3,000개의 switch 필요 (scale-out)
- 동일 cluster: 200,000개의 GPU 각각 CPO 필요 (scale-up)
- **비율:** GPU가 switch보다 60배 이상 많음!

2. "Pipe Cleaner" 전략 (공급망 준비)

*Pipe cleaner*란? 파이프의 막힌 부분을 먼저 청소하는 도구처럼, scale-out CPO가 supply chain의 문제점을 먼저 발견하고 해결하는 역할을 합니다.

Supply chain 단계적 확장:

- **2025-2026:** 수천 개 switch로 시작 → 공급업체들이 소량 생산 경험 축적
- **2026-2027:** 수만 개 수준으로 확대 → Manufacturing process 최적화
- **2027-2028:** 수천만 개 GPU용으로 대규모 확장 → 완전한 mass production

3. Real-World Validation (현장 검증)

*Lab vs Production*의 차이:

Meta의 Lab 테스트 (2024-2025):

- 15개의 CPO switch
- 325일 동안 테스트
- 통제된 환경: 일정한 온도, 깨끗한 공기, 안정적인 전력
- 결과: 매우 좋음 (zero failures)

Production 환경 (2026-2028):

- 수천 개의 CPO switch가 여러 datacenter에 분산
- 수년간 24/7 가동
- **예상치 못한 도전:**
 - 온도 변동 (Meta 논문: 1-2% 변동이 전력 소비에 영향)
 - 먼지와 오염 물질 (fiber 끝 청소 필요)
 - 습도, 진동, 전력 불안정성
 - 다양한 운영 팀의 실수와 예상 외 사용 패턴
- **핵심 데이터 수집:**
 - Failure rate와 failure mode
 - Blast radius (한 switch 고장 시 영향 범위: 64 ports)
 - Serviceability (수리 가능성과 소요 시간)
 - 실제 MTBF (Mean Time Before Failure)

4. Timeline과 현실

시기	Scale-out (Switch)	Scale-up (GPU)
2025-2026	제한적 상용화 시작 (수백 ~ 수천 개)	아직 없음
2027	점진적 확대 (수만 개)	Rubin Ultra 출시 예정 (하지만 CPO 없이)
2028+	대규모 배포	Feynman 세대에서 CPO 본격 적용 예상

왜 Rubin Ultra (2027)도 CPO가 없는가?

- 2027년까지 supply chain이 수천만 개 CPO를 생산할 준비 안 됨
- Scale-out의 field data가 충분히 축적되지 않음
- 너무 야심적인 timeline → 리스크가 너무 큼
- 보수적 접근: Feynman (2028-2029)까지 기다려서 확실하게 적용

5. 경제적/전략적 이유

TCO (Total Cost of Ownership) 이점:

- Scale-out CPO: 23% networking 전력 절감 (좋지만 제한적)
- **Scale-up CPO:** 훨씬 더 큰 이점 (copper 제거, world size 증가)
- 진짜 큰 돈은 GPU에서 나옴 → GPU CPO가 진짜 목표

최종 목표는 scale-up CPO입니다. 하지만 그곳에 도달하려면 scale-out으로 먼저 길을 닦아야 합니다. 이는 기술적 필연성이자 비즈니스적 현명함입니다.

이제 CPO가 무엇인지, 기술적 고려사항, 과제 및 오늘날 ecosystem의 상태에 대해 심도 있게 이야기해보자.

2 Part 2: CPO Introduction and Implementation

2.1 What is CPO about and why is everyone so excited?

CPO는 optical engine을 고성능 computing 또는 networking ASIC과 동일한 package 또는 module 내에 직접 통합한다. 이러한 optical engine은 electrical signal을 optical signal로 변환하여 optical link를 통한 고속 데이터 전송을 가능하게 한다. Optical link는 몇 meter 이상의 거리에서 데이터 통신을 위해 사용되어야 하는데, copper를 통한 고속 electrical communication은 몇 meter 이상으로 도달할 수 없기 때문이다.

오늘날 대부분의 electrical에서 optical로의 변환은 pluggable optical transceiver를 통해 발생한다. 이러한 경우, electrical signal은 switch 또는 processing chip에서 **Printed Circuit Board (PCB)**를 통해 chassis의 front plate 또는 back plate에 있는 물리적 transceiver cage까지 수십 centimeter 이상을 이동한다. Pluggable optical transceiver는 그 cage에 위치한다. Transceiver는 optical **Digital Signal Processor (DSP)** chip에 의해 재조정되는 electrical signal을 수신한 다음 electrical signal을 optical signal로 변환하는 optical engine component로 전송된다. Optical signal은 그런 다음 optical fiber를 통해 link의 반대편으로 전송될 수 있으며, 여기서 다른 transceiver가 이 과정을 역으로 수행하여 optical signal을 destination silicon까지 다시 electrical signal로 변환한다.

이 과정에서 electrical signal은 optical link에 도달하기 전에 (적어도 copper의 경우) 상대적으로 긴 거리를 여러 transition point를 거쳐 통과한다. 이는 electrical signal이 저하되며 이를 구동하고 복구하기 위해 많은 전력과 복잡한 회로 (SerDes)가 필요하다. 이를 개선하려면 electrical signal이 이동해야 하는 거리를 단축해야 한다. 이것이 **Co-Packaged Optics**라는 아이디어로 이어지는데, 여기서 pluggable transceiver에 있던 optical engine이 대신 host chip과 co-package 된다. Optical engine이 XPU 또는 Switch ASIC에 훨씬 더 가깝기 때문에 electrical trace 길이가 수십 centimeter에서 수십 mm로 줄어든다. 이는 electrical interconnect 거리를 최소화하고 signal integrity 문제를 완화함으로써 전력 소비를 크게 줄이고, bandwidth density를 향상시키며, latency를 낮춘다. 아래 schematic은 CPO 구현을 보여주며, 여기서 optical engine이 compute 또는 switch chip과 동일한 package에 위치한다. Optical engine은 처음에는 substrate에 있을 것이고, 미래에는 OE가 interposer에 배치될 것이다.

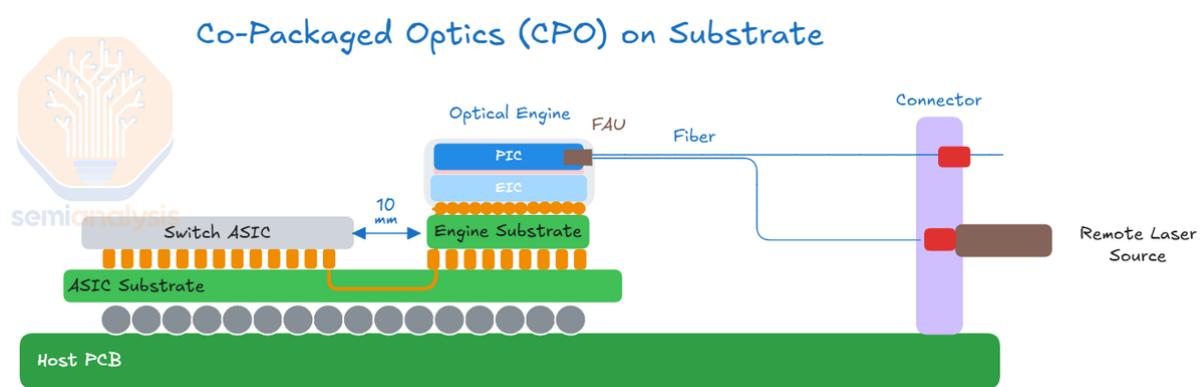


Figure 14: Source: SemiAnalysis

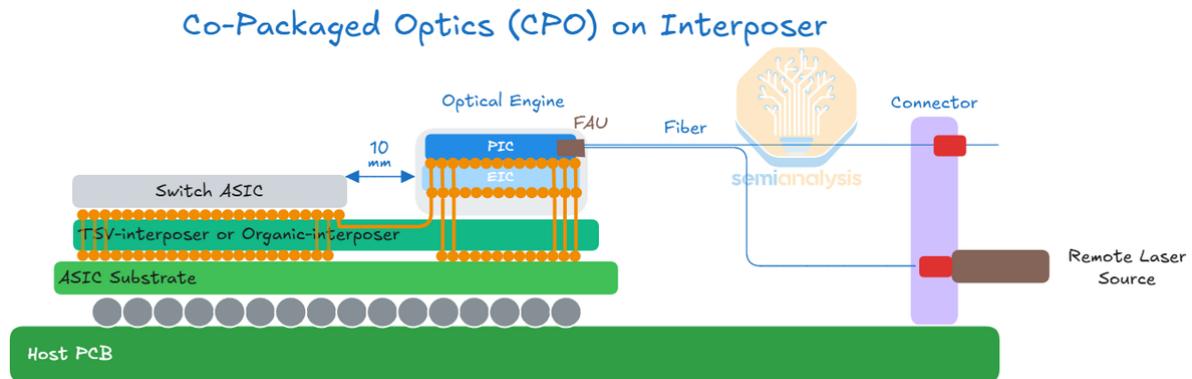


Figure 15: Source: SemiAnalysis

핵심 인사이트: CPO의 본질 - Signal Path 혁명

1. Traditional Pluggable Optics의 Signal Journey

Electrical signal의 긴 여정:

- 출발: ASIC chip의 SerDes
- 경로 1: Package substrate (1-2 cm)
- 경로 2: PCB trace 또는 flyover cable (15-30 cm)
- 경로 3: Connector 및 transceiver PCB (1-2 cm)
- 도착: Transceiver 내부의 optical engine
- 총 거리: 약 20-35 cm

이 여정의 대가:

- Signal degradation: 고주파수 성분이 exponentially 감쇠
- LR SerDes 필요: 50-70 mW/lane의 높은 전력
- DSP chip 필요: Signal reconditioning에 5-14W
- 낮은 bandwidth density: 많은 회로가 signal recovery에 사용됨

2. CPO의 혁명적 변화

Electrical signal의 짧은 여정:

- 출발: ASIC chip의 SerDes 또는 wide I/O
- 경로: Package substrate 또는 interposer (**2-5 mm**)
- 도착: 동일 package 내의 optical engine
- 총 거리: 2-5 mm

거리 단축의 마법:

- Signal degradation 최소화: mm 거리에서는 거의 감쇠 없음
- SR SerDes 또는 wide I/O: 20-30 mW/lane (50% 절감)
- DSP 제거
- 높은 bandwidth density: Signal recovery 회로 불필요

3. 구체적 수치 비교

항목	Traditional	CPO	절감률
Signal path 거리	20-35 cm	2-5 mm	100배↓
SerDes 전력/lane	50-70 mW	20-30 mW	50-70%
DSP 전력	5-14W	0W	100%
총 전력 (800G)	16-17W	4-5W	73%
Bandwidth density	0.4 Tbps/mm	1.8-10 Tbps/mm	5-25배↑

4. 왜 거리가 이렇게 중요한가?

물리 법칙: Insertion loss는 거리와 주파수에 exponentially 비례

구체적 예시 (112 Gbps, 56 GHz):

- 5 mm (CPO): 약 2-3 dB loss → 신호 50-70% 유지
- 10 cm (short PCB): 약 20-25 dB loss → 신호 1-3% 유지
- 30 cm (long PCB): 약 40-50 dB loss → 신호 0.01% 유지

5mm에서는 신호가 대부분 유지되므로: - Minimal equalization (전력 절감) - DSP 불필요 (전력 절감) - 낮은 transmit power (전력 절감)

30cm에서는 신호가 거의 완전히 소멸하므로: - 강력한 equalization 필요 (전력 소비) - DSP로 signal reconstruction 필요 (전력 소비) - 높은 transmit power 필요 (전력 소비)

핵심 통찰: CPO의 본질은 "optical engine을 chip에 붙이는 것"이 아니라 "electrical signal path를 극단적으로 단축하는 것"입니다.

오늘날 아래 다이어그램에 설명된 것처럼 front pluggable optics 솔루션이 보편적이다. 이 다이어그램에서 주요 요점은 electrical signal이 transceiver의 optical engine에 도달하기 전에 copper trace 또는 flyover cable을 통해 긴 거리 (15-30cm)를 통과해야 함을 보여주는 것이다. 위에서 논의한 바와 같이, 이는 또한 pluggable module로 구동하기 위한 long-reach (LR) SerDes의 필요성을 필요로 한다.

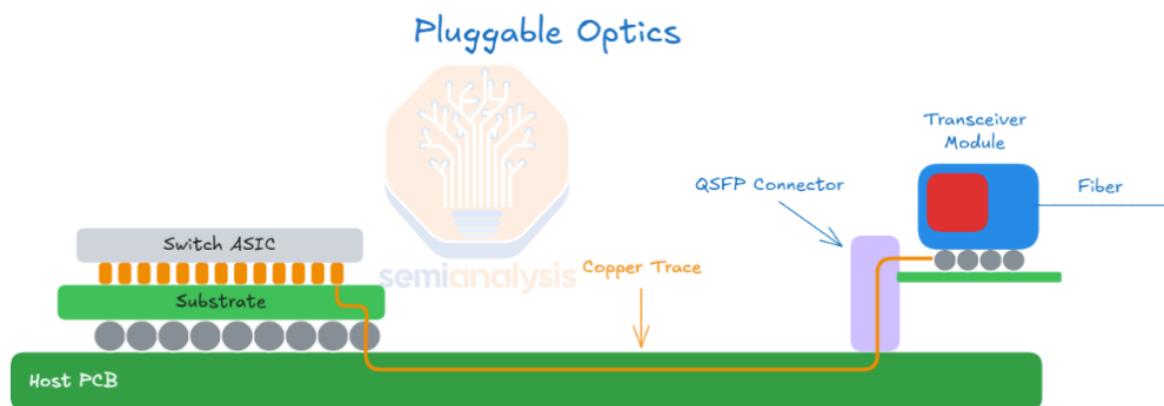


Figure 16: Source: SemiAnalysis

기술 용어: SerDes, PCB Trace, Flyover Cable

SerDes (Serializer/Deserializer):

- 정의: 병렬 데이터를 직렬로 변환(전송) 또는 직렬을 병렬로 변환(수신)하는 회로
- 필요성: Chip 내부는 넓은 병렬 bus(예: 256 bits), 외부는 좁은 직렬 lane(예: 1 bit)
- Serializer: 256 bits를 받아서 1 bit씩 순차적으로 고속 전송
- Deserializer: 1 bit씩 받아서 256 bits로 재구성

Long-Reach (LR) SerDes vs Short-Reach (SR) SerDes:

특성	LR SerDes	SR SerDes
목표 거리	15-30 cm	2-5 mm
전력 (per lane)	50-70 mW	20-30 mW
회로 복잡도	매우 높음	중간
Equalization	강력 (FFE, DFE)	최소
사용 예	Pluggable optics	CPO, chiplet

왜 LR SerDes가 더 많은 전력을 쓰는가?

- Pre-emphasis: 전송 전 고주파수 성분 증폭 (signal degradation 예상)
- Equalization: 수신 후 주파수 왜곡 보정
- CDR (Clock Data Recovery): 저하된 신호에서 clock 추출
- 더 높은 swing voltage: 긴 거리로 인한 loss 보상

PCB (Printed Circuit Board) Trace:

- 정의: PCB 기판 위/내부에 etching된 copper 경로
- CCL (Copper-Clad Laminate): PCB의 기본 재료
 - 절연체(laminate)에 copper를 입힌 구조
 - 고주파수에서 dielectric loss가 큼
- 문제: 112 Gbps 이상에서는 수 cm만 이동해도 신호 품질 심각하게 저하
- Insertion loss: 고주파수일수록 exponentially 증가

Flyover Cable:

- 정의: PCB trace를 우회하여 공중으로 연결하는 고품질 cable
- 장점:
 - PCB trace보다 훨씬 낮은 insertion loss
 - 더 나은 shielding으로 crosstalk 감소
 - 고주파수 신호에 최적화된 재료
- 단점:
 - 여전히 15-30cm 거리 필요 (LR SerDes 필수)
 - 복잡한 설치 (수작업으로 routing)
 - 물리적 공간 점유
 - 높은 비용
- 사용처: High-end server/switch에서 PCB trace 대체

추가로, CPO와 전통적인 front-pluggable optics 사이에 있는 중간 구현이 있는데, **Near-Packaged Optics (NPO)** 및 **On-Board Optics (OBO)**와 같은 것들이다. 최근 몇 년 동안 NPO는 CPO로 가는 중간 단계로 등장했다. NPO는 여러 정의를 가지고 있다. NPO는 OE가 ASIC의 substrate에 직접 위치하지 않고 다른 substrate에 co-package되는 경우이다. Optical engine은 socketable한 상태로 유지되며 substrate에서 분리될 수 있다. Electrical signal은 여전히 XPU package의 SerDes에서 일부 copper channel을 통해 Optical Engine으로 이동한다.

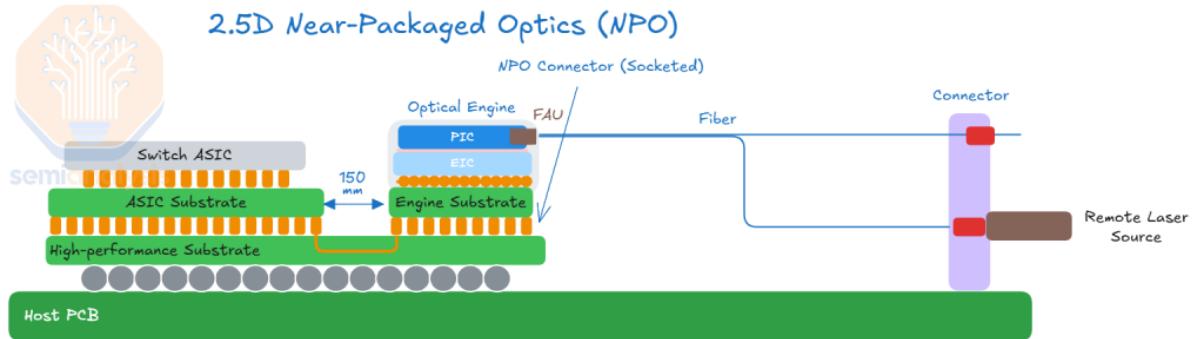


Figure 17: Source: SemiAnalysis

OBO는 optical engine을 chassis 내부의 system PCB에 통합하여 host ASIC에 더 가깝게 배치한다. 그러나 OBO는 CPO의 많은 과제를 상속하면서 bandwidth density 및 전력 절감 측면에서 더 적은 이점을 제공한다. 우리는 OBO를 **worst of both worlds**로 보는데, CPO의 복잡성을 결합하면서 front-pluggable optics의 일부 제한을 상속하기 때문이다.

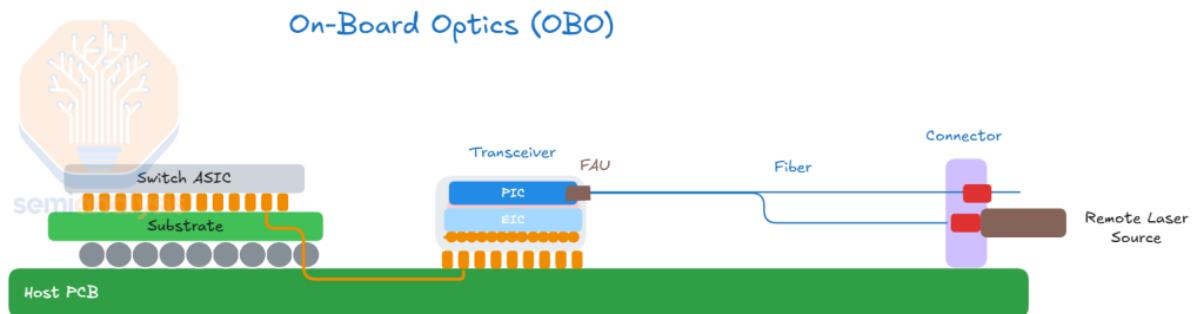


Figure 18: Source: SemiAnalysis

기술 용어: NPO, OBO - CPO로 가는 중간 단계들

Optical Integration의 진화 스펙트럼

특성	Pluggable	OBO	NPO	CPO
OE 위치	Cage	PCB	Near pkg	In pkg
Signal 거리	15-30cm	5-10cm	2-5cm	2-5mm
Serviceability	최고	중간	낮음	최저
전력 효율	낮음	중간	중간-높음	최고
복잡도	낮음	중간	높음	최고

1. OBO (On-Board Optics)

개념: Optical engine을 system PCB 위에 직접 배치 (chassis 내부, ASIC package 외부)

장점:

- Signal path가 pluggable 대비 짧아짐 (30cm → 5-10cm)
- 일부 전력 절감 가능
- Pluggable보다 높은 bandwidth density

단점:

- "**Worst of Both Worlds**": CPO의 복잡도 + Pluggable의 제한
- 여전히 LR SerDes 필요 (전력 소비 높음)
- PCB에 고정되어 serviceability 어려움
- CPO만큼의 bandwidth density나 전력 효율 달성 못함
- Thermal management 복잡

결론: 이득은 적고 복잡도는 높아서 업계에서 선호하지 않음

2. NPO (Near-Packaged Optics)

개념: Optical engine을 별도 substrate에 배치하되 ASIC package 매우 가까이 위치

주요 특징:

- OE가 ASIC substrate에 직접 붙지 않고 **별도 substrate**에 위치
- **Socketable**: OE substrate를 분리/교체 가능 (일부 serviceability 유지)
- Signal은 여전히 일부 copper channel 통과 (2-5cm)
- CPO보다 긴 electrical path이지만 pluggable보다는 짧음

장점:

- CPO로 가는 중간 단계로 supply chain 준비 가능
- 일부 serviceability 유지 (OE substrate 교체)
- Pluggable 대비 상당한 전력 절감
- CPO보다 packaging 복잡도 낮음

단점:

- CPO만큼의 극적인 전력/공간 절감은 못함
- 여전히 일부 copper channel 통과로 SerDes 전력 소비
- 표준화 부족

시장 포지셔닝: CPO가 완전히 성숙하기 전까지의 과도기 솔루션

3. 왜 CPO가 최종 목표인가?

- **물리적 한계**: Electrical signal은 거리가 짧을수록 exponentially하게 좋아짐
- **경제성**: 중간 단계들은 "일부 복잡도 + 일부 이득"으로 ROI가 불분명
- **표준화**: Pluggable은 표준화됨, CPO도 표준화 진행 중, 하지만 중간 단계들은 표준화 어려움

핵심 통찰: OBO와 NPO는 CPO로 가는 정검다리 역할을 할 수 있지만, 업계는 대부분 "조금 나은 pluggable"보다 "확실히 다른 CPO"를 기다리는 쪽을 선택하고 있습니다.

2.2 Co-packaged Copper

CPO의 또 다른 대안은 **Co-Packaged Copper (CPC)**이다. CPC는 substrate의 connector에서 직접 나오는 copper cabling을 사용한다. CPC에 사용되는 cable은 flyover와 동일한 cable이며 동일한 목적 즉, PCB trace를 우회하는 것이다. CPC는 flyover cable을 더 나아가게 하여 socket이 package substrate 자체에서 시작한다. 사용되는 cable은 cross-talk을 줄이기 위해 잘 절연된 twin-axial cable (Twinax cable)이며, 기존 electrical trace와 비교하여 insertion loss가 상당히 낮다. 이 솔루션은 여전히 copper를 사용하지만 signal integrity에서 핵심 이점을 제공한다. CPC는 448G SerDes를 배포하여 off-package interconnect의 또 다른 scaling을 허용하는 실용적인 경로를 제공할 수 있다.

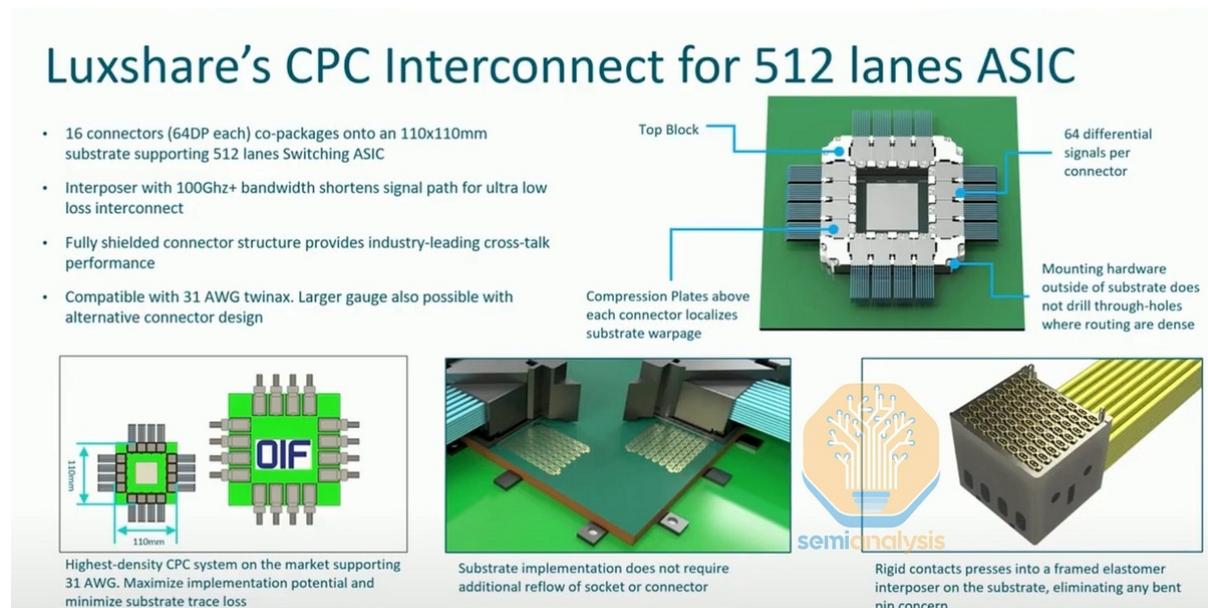


Figure 19: Source: LuxShare

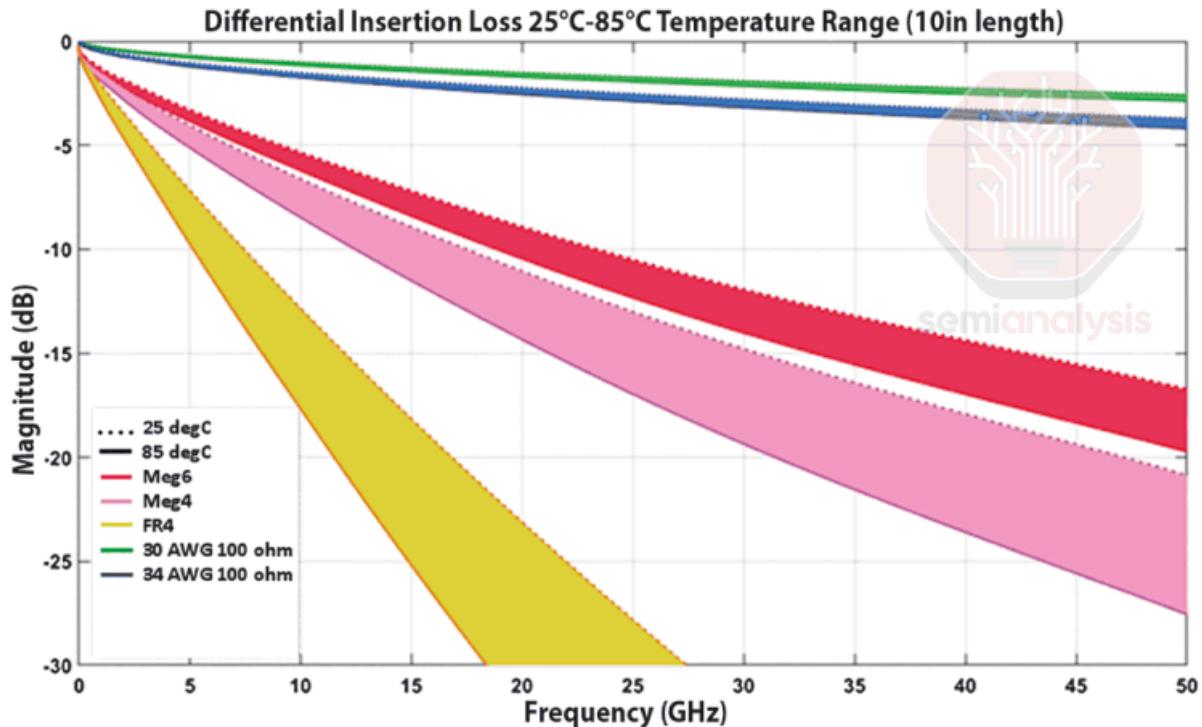


Figure 20: Source: Luxshare

CPC의 과제는 package substrate의 추가된 복잡성에 있다. Substrate는 수천 개의 이러한 cable에 전력과 신호를 routing해야 한다. 이러한 과제에도 불구하고 CPC는 여전히 supply chain의 여러 부분에서 많은 manufacturing 장애물을 극복해야 하는 CPO보다 훨씬 더 간단하다. 우리는 CPC를 in-rack scale-up connectivity와 같은 일부 short-reach application에 특히 매력적인 것으로 보며, 이는 아래에서 살펴볼 것이다. 손실이 많은 CCL trace를 우회함으로써, CPC는 448G line speed를 가능하게 하는 기술이 될 수 있다. CPC는 또한 448G를 가능하게 하기 위해 매우 광범위하게 검토되고 있는데, 이 bandwidth의 signal은 PCB를 통해 실행될 때 받아들일 수 없는 attenuation을 경험하기 때문이다.

기술 용어: Package Substrate, Twinax Cable, Cross-talk

1. Package Substrate

정의: ASIC chip과 PCB 사이의 중간층으로, chip의 미세한 bump를 PCB의 큰 solder ball로 연결하는 다층 구조

역할:

- **Fan-out:** Chip의 좁은 bump pitch (예: 100 μm) \rightarrow PCB의 넓은 ball pitch (예: 1 mm)로 확장
- **Signal redistribution:** Chip의 I/O를 적절한 위치로 재배치
- **Power delivery:** 깨끗한 전원을 chip에 공급 (decoupling capacitor 내장)
- **Thermal management:** 열을 PCB로 전달

재료와 특성:

- **Organic substrate:** FR-4 기반, 저렴하지만 electrical 성능 제한적
- **HDI substrate:** High-Density Interconnect, 미세 pitch 지원

- Advanced substrate: ABF (Ajinomoto Build-up Film), 고주파수 특성 우수
- 일반적으로 4-12 metal layer

CPC에서의 역할:

- Package substrate 표면에 **cable connector** 직접 장착
- PCB trace를 완전히 우회 → Signal degradation 최소화
- 가장 짧은 electrical path (chip → substrate → cable)

2. Twinax Cable (Twin-Axial Cable)

정의: 두 개의 copper conductor가 함께 차폐(shielding)된 고성능 cable

구조 (안쪽에서 바깥쪽으로):

- **두 개의 conductor:** 신호를 differential pair로 전송
 - Signal+ 와 Signal- 가 함께 전송
 - Differential signaling으로 noise immunity 향상
- **절연체 (dielectric):** 두 conductor 사이와 주변을 절연
- **Shield (차폐):** Aluminum foil 또는 braided copper
 - 외부 electromagnetic interference (EMI) 차단
 - Cross-talk 방지
- **Jacket (외피):** 기계적 보호

Twinax vs Coaxial:

특성	Coaxial	Twinax
Conductor 수	1개 (중심)	2개 (parallel)
Signaling	Single-ended	Differential
Noise immunity	중간	높음
용도	RF, video	High-speed data

Twinax의 장점:

- 낮은 insertion loss: PCB trace 대비 10배 이상 낮음
- 우수한 signal integrity: 차폐로 external noise 차단
- 낮은 cross-talk: 차폐로 인접 cable 간 간섭 최소
- 고주파수 지원: 448G SerDes 가능 (2m 이내)
- Flexible: 손쉬운 routing

단점:

- PCB trace 대비 높은 비용
- 물리적 공간 점유
- 거리 제한 (copper 물리 한계)

3. Cross-talk (누화)

정의: 인접한 신호선 간에 발생하는 원치 않는 전자기 간섭 (electromagnetic coupling)

발생 원리:

- **Capacitive coupling:** 인접 conductor 간 기생 capacitance
 - 고주파수 신호가 AC coupling으로 누설

- 거리가 가까울수록 심함
- **Inductive coupling:** 전류가 흐르면 자기장 생성
 - 인접 conductor에 유도 전류 발생
 - 전류가 클수록 심함

Cross-talk의 영향:

- **Signal integrity 저하:** 원래 신호에 noise 추가
- **Bit error rate (BER) 증가:** 인접 bit 간섭
- **Timing margin 감소:** Signal eye diagram 닫힘
- 고속 신호(224G+)에서 특히 치명적

PCB Trace에서 Cross-talk가 심한 이유:

- 수백-수천 개의 trace가 좁은 공간에 밀집
- Trace 간 거리: 수십-수백 μm (매우 가까움)
- 차폐 없음 → Electromagnetic field가 자유롭게 전파
- 고주파수 신호 → Coupling 증가

Twinax Cable이 Cross-talk를 막는 방법:

- **Shield (차폐):**
 - Electromagnetic field를 shield 내부에 가둠
 - 외부로 field 방출 차단
 - 외부 field 침입 차단
- **Differential signaling:**
 - Signal+ 와 Signal- 가 동일한 noise를 받음
 - Receiver가 차이만 측정 → Noise 상쇄 (common-mode rejection)
- **Physical separation:** Cable 간 물리적 간격 유지 가능

핵심 인사이트: CPC - Copper의 마지막 진화

핵심 질문: CPO가 미래라면, 왜 CPC(Co-Packaged Copper)도 논의되는가?

1. CPC의 전략

Traditional copper의 문제 진단:

- ASIC → PCB trace (손실 많은 CCL) → Connector → Cable
- **주법:** PCB trace의 높은 insertion loss
- 112G 이상에서 수 cm만 가도 신호 품질 급격히 저하

CPC의 해법:

- ASIC → Package substrate → 직접 twinax cable 연결
- **PCB trace 완전 우회!**
- Flyover cable과 유사하지만 시작점이 package substrate

2. CPC의 장점과 과제

장점:

- 448G SerDes 가능: PCB trace 우회로 448G line speed 달성
- CPO보다 훨씬 간단: Optical engine, laser, fiber coupling 모두 불필요
- Short-reach에 최적: In-rack scale-up connectivity (1-2m)
- Lower cost: CPO 대비 단순한 구조
- Proven technology: Copper cable은 성숙한 기술

과제:

- Substrate 복잡도: 수천 개의 cable connector를 substrate에 배치
- Power/Signal routing: 수천 개 cable에 전력과 신호 routing
- 기계적 스트레스: 많은 cable이 package에 연결되어 물리적 부담
- 거리 제한: 여전히 copper 물리 한계 (2m)

3. CPC vs CPO: 언제 무엇을 쓸 것인가?

CPC가 유리한 경우:

- In-rack scale-up: 같은 rack 내 GPU 간 연결 (1-2m)
 - 예: GB200 NVL72, Kyber rack
 - 모든 GPU가 하나의 rack에 밀집
 - 2m 이내 거리면 충분
- 448G adoption: CPO가 성숙하기 전 448G를 구현하는 practical path
- Cost-sensitive applications: CPO의 높은 비용 부담스러운 경우

CPO가 필수인 경우:

- Multi-rack scale-up: Rack 간 연결 (2m 이상)
- Scale-out networking: Datacenter switch fabric
- 극한의 bandwidth density: 물리적 공간 제약이 심한 경우
- 최고 전력 효율: 전력이 critical한 대규모 deployment

4. 현실적 미래: Hybrid 접근

업계는 Co-Packaged Copper + Co-Packaged Optics 병행 사용할 가능성 높음:

- In-rack: CPC로 copper 사용 (저렴하고 충분)
- Across-rack: CPO로 optics 사용 (필수)
- NVIDIA도 이 hybrid 전략 추구 중

구체적 예시:

- Rack 내부 72 GPU 연결: CPC (NVLink over copper)
- Rack 간 scale-up: CPO (NVLink over optics)
- Datacenter switch fabric: CPO (scale-out networking)

핵심 통찰: "할 수 있는 곳에서는 copper를 사용하고, 해야 할 때는 optics를 사용하라." CPC는 이 격언의 현대적 구현입니다. CPO가 모든 것을 대체하는 것이 아니라, copper와 optics가 각자 최적인 영역에서 co-exist하는 미래가 더 현실적입니다.

2.3 Past obstacles to CPO market readiness: Why only now?

기술적 우월성에도 불구하고 CPO는 비용을 증가시키는 여러 과제로 인해 실제 채택이 매우 제한적이었다. 여기에는 packaging (OE 자체보다 비용이 더 많이 드는)과 manufacturing의 복잡성, 신뢰성 및 yield 우려, 그리고 긴밀하게 통합된 optical 및 electrical component에서 발생하는 thermal management 문제가 포함된다. 또 다른 장애물은 산업 전반의 표준화 부족이다. 추가로, 고객들은 serviceability에 대해 우려하고 있으며 이는 전통적인 pluggable optics에서 CPO 솔루션으로의 전환을 방해한다.

또 다른 주요 고객 불안은 CPO를 채택함으로써 비용을 통제할 수 있는 능력을 잠재적으로 포기한다는 것이다. 더 적은 수의 switch vendor보다 더 많은 수의 transceiver 회사에 비용 압박을 가하는 것이 훨씬 쉽다. 한편, CPO가 대체할 기존 기술인 pluggable optics는 계속 개선되고 있으며 여전히 훨씬 적은 최종 사용자 불안으로 거의 모든 application에 충분히 좋은 성능을 제공한다. Part 2의 나머지 부분에서는 CPO 채택의 동기에 대해 더 깊이 탐구할 것이다. SerDes scaling이 어떻게 plateau에 도달하고 있는지 설명하는 것으로 시작하여, Wide I/O와 결합된 CPO와 같은 다른 interface type을 필요로 하게 만들고, 그 다음 manufacturing 고려사항과 시장 출시로 들어갈 것이다. Optical Engine, Fiber Coupling, External Laser Source 및 Modulator와 같은 주요 개별 CPO component를 논의할 것이다. 마지막으로, CPO에서 bandwidth를 확장하기 위한 roadmap을 다룰 것이다.

2.4 Evolving beyond DSP-based Transceivers: From LPO to CPO

DSP Transceiver는 optical signal의 transmission과 reception을 모두 처리하며 electro-optical conversion을 담당하는 **Optical Engine (OE)**을 포함한다. OE는 optical signal을 전송하기 위한 **driver (DRV)** 및 **modulator (MOD)**, 그리고 optical signal을 수신하기 위한 **transimpedance amplifier (TIA)** 및 **photodetector (PD)**로 구성된다. 또 다른 중요한 component는 optical DSP chip인데, 이는 때때로 Driver 및/또는 TIA를 하나의 package로 통합한다. Host switching 또는 processing chip에서 전송되는 high frequency electrical signal은 server chassis의 front에 있는 transceiver에 도달하기 위해 손실이 많은 copper trace를 통해 상대적으로 긴 거리를 이동해야 한다. DSP는 이 signal의 retiming 및 reconditioning을 담당한다. Signal이 switch 또는 ASIC silicon에서 substrate 또는 다른 transmission medium를 통과할 때 electrical signal degradation과 attenuation을 보상하기 위해 error correction 및 clock/data recovery를 수행한다. Modulation의 경우, **PAM4 Modulation (Pulse Amplitude Modulation with 4 Levels)**의 경우, DSP는 signal당 bit 수를 증가시키기 위해 binary signal을 4개의 distinct amplitude level로 매핑하여 더 높은 bitrate와 더 많은 bandwidth를 허용한다.

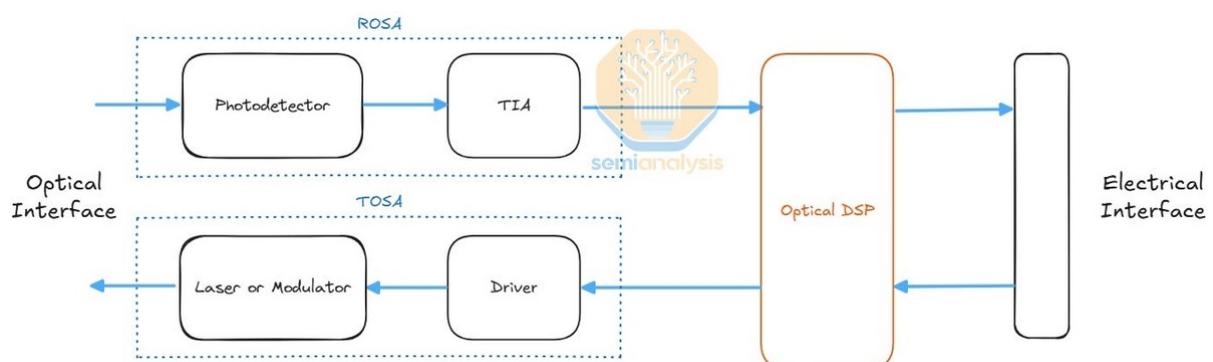


Figure 21: Source: SemiAnalysis

기술 용어: DSP Transceiver의 구조와 역할

1. Optical Engine (OE)의 구성요소

Optical Engine은 전기-광 변환을 담당하는 핵심 component:

송신 (TX) 측:

- **Driver (DRV)**: Electrical signal을 증폭하여 modulator 구동
- **Modulator (MOD)**: Laser 빛을 electrical signal에 따라 변조
 - 빛의 세기나 위상을 바꿔서 데이터 encoding
 - 예: PAM4는 4개의 다른 세기 level 사용

수신 (RX) 측:

- **Photodetector (PD)**: 광 신호를 전기 신호로 변환
 - 빛을 받으면 전류 생성
 - 매우 약한 신호 (micro-ampere 수준)
- **Transimpedance Amplifier (TIA)**: PD의 약한 전류를 증폭
 - 전류 신호를 전압 신호로 변환하면서 증폭
 - 신호 대 잡음비 향상

2. Optical DSP Chip의 역할

왜 DSP가 필요한가? 긴 copper path(15-30cm)를 거치면서 electrical signal이 심하게 degradation되기 때문

주요 기능:

- **Signal Reconditioning**: 저하된 신호를 원래 모양으로 복원
- **Retiming**: Clock 신호를 재생성하여 timing jitter 제거
- **Error Correction (FEC)**: Forward Error Correction으로 bit error 수정
- **Clock/Data Recovery (CDR)**: Signal에서 clock 추출하고 데이터 복구
- **Equalization**: 주파수 dependent loss 보상
 - 고주파수 성분이 저주파수보다 더 감쇠됨
 - 고주파수를 선택적으로 증폭하여 균형 맞춤

3. PAM4 Modulation

기본 개념:

- **PAM (Pulse Amplitude Modulation)**: 신호의 진폭(세기)으로 데이터 encoding
- **PAM2 (NRZ)**: 2개 level (0, 1) → 1 bit per symbol
- **PAM4**: 4개 level (00, 01, 10, 11) → 2 bits per symbol

PAM4의 장점:

- 동일한 baud rate에서 2배 bandwidth
- 예: 56 Gbaud × 2 bits = 112 Gbps
- Symbol rate 낮춰서 frequency 관련 loss 완화

PAM4의 단점:

- Level 간 간격이 좁아져 noise margin 감소

- 더 복잡한 DSP 필요 (level 구분)
- 더 높은 전력 소비

미래 (448G 시대):

- **PAM6**: 6 level → 약 2.58 bits per symbol
- **PAM8**: 8 level → 3 bits per symbol
- 하지만 noise margin이 더욱 줄어들어 구현 매우 어려움

4. DSP의 통합 형태

DSP chip은 종종 다른 component와 통합:

- **DSP + Driver**: 송신 측 통합
- **DSP + TIA**: 수신 측 통합
- **DSP + Driver + TIA**: 완전 통합 (single package)

핵심 통찰: DSP transceiver는 "signal degradation을 극복하기 위한 정교한 복구 시스템"입니다. 하지만 이 복구 시스템 자체가 엄청난 전력을 소비합니다. CPO의 본질은 "signal degradation 자체를 예방"하여 이 복구 시스템을 아예 제거하는 것입니다.

DSP chip은 transceiver 내에서 가장 전력을 많이 소비하고 비싼 component 중 하나이다. 800G SR8 Transceiver의 경우 – DSP는 module의 총 전력 소비의 거의 ~50%를 차지하며, 이것이 DSP를 제거하는 데 그토록 많은 초점이 맞춰진 이유이다.

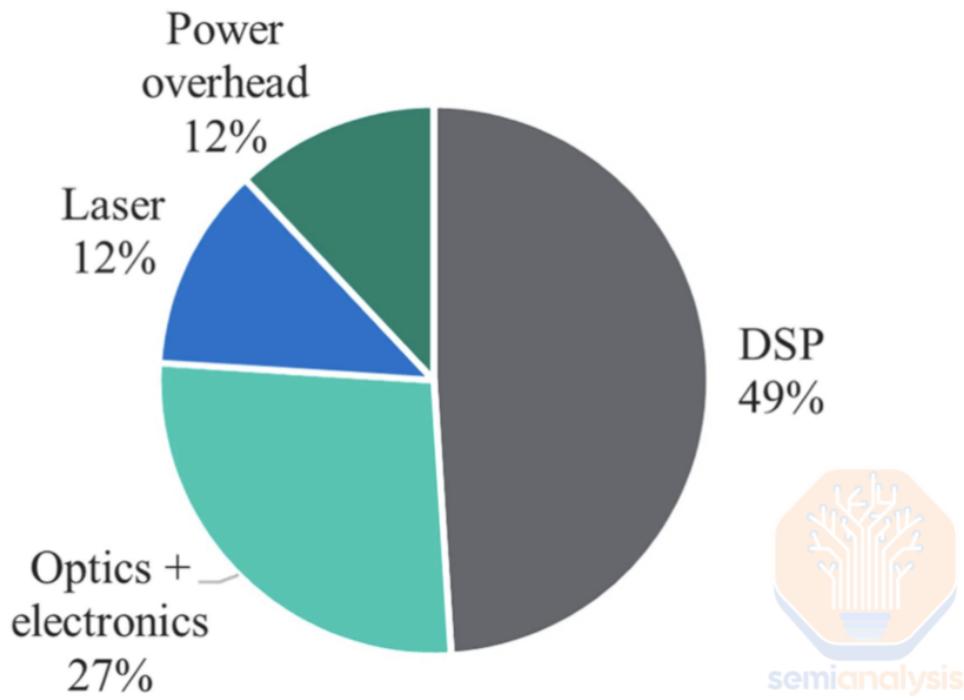


Figure 22: Source: Dr. Radha Nagarajan et al.: Recent Advances in Low-Power Digital Signal Processing Technologies for Data Center Applications

2-layer InfiniBand network로 구축된 18K GB300 Cluster는 18,432개의 800G DR4 transceiver와 27,648개의 1.6T DR8 transceiver를 필요로 할 것이다. DSP 사용으로 인한 추가 비용과 전력

요구사항은 total cost of ownership에 상당히 추가될 수 있다. 800G DSP에 6-7W, 1.6T DSP에 12-14W를 예산으로 잡으면, 이 전체 cluster의 back-end network만을 위해 480kW의 DSP 전력이 추가되며, 이는 server rack당 약 1.8kW이다. 프리미엄 브랜드 공급업체에서 조달할 때, transceiver는 cluster의 total cost of ownership의 거의 10%를 차지할 수 있다. 따라서, 일반적인 transceiver의 전력 소비의 50%와 BoM의 20-30%를 차지하는 것을 고려할 때 일부는 DSP를 비용 및 전력 효율성의 public enemy number one으로 간주한다.

GB300 NVL72 Cluster Cost and Power Budget Breakdown, per Rack-Scale 72-GPU Server						
Item	2-Layer Network (up to 10,368 GPUs)		3-Layer Network (up to 746,496 GPUs)		4-Layer Network (up to ~53.7M GPUs)	
	Cost	Power (W)	Cost	Power (W)	Cost	Power (W)
Server	\$4,357,943	142,000	\$4,357,943	142,000	\$4,357,943	142,000
Optical Transceivers	\$363,125	4,600	\$499,925	6,184	\$568,325	6,976
Switches	\$208,681	5,214	\$312,681	8,014	\$416,681	11,606
Fiber, Cables, Software, Others	\$17,723	0	\$24,338	0	\$30,953	0
Networking	\$589,529	9,814	\$836,944	14,198	\$1,015,959	18,582
All Others	\$269,152	281	\$269,152	281	\$269,152	281
Total Power	\$5,216,624	152,095	\$5,464,039	156,479	\$5,643,054	160,863

Power for a Neocloud Giant InfiniBand cluster using X800-Q3400 Back-end switches with 144 ports of 800G
Source: SemiAnalysis Networking Model

Figure 23: Source: SemiAnalysis AI Networking Model

기술 용어: DR4, DR8 - Optical Transceiver 규격

1. DR의 의미

DR = Dense Wavelength Division Multiplexing (DWDM) Reach

- D (Dense): 좁은 파장 간격으로 여러 wavelength 사용
- R (Reach): 전송 거리 category
 - SR (Short Reach): 100m 이하, multimode fiber
 - DR (DWDM Reach): 500m-2km, single-mode fiber, DWDM 지원
 - FR (Far Reach): 2km, single-mode fiber
 - LR (Long Reach): 10km, single-mode fiber
- DR은 datacenter 내부 연결에 최적화된 규격

2. 숫자의 의미 (Lane 수)

DR 뒤의 숫자 = Optical lane의 개수

DR4:

- 4개의 optical lane (4개의 laser + 4개의 photodetector)
- 각 lane이 독립적으로 데이터 전송
- 800G DR4: 4 lanes × 200G/lane = 800 Gbps
- 400G DR4: 4 lanes × 100G/lane = 400 Gbps

DR8:

- 8개의 optical lane
- 1.6T DR8: 8 lanes × 200G/lane = 1.6 Tbps
- 800G DR8: 8 lanes × 100G/lane = 800 Gbps

3. Lane 수와 Form Factor

규격	Lane 수	Fiber 수	Form Factor
400G DR4	4 TX + 4 RX	8 (4 pair)	QSFP-DD
800G DR4	4 TX + 4 RX	8 (4 pair)	QSFP-DD
800G DR8	8 TX + 8 RX	16 (8 pair)	OSFP
1.6T DR8	8 TX + 8 RX	16 (8 pair)	OSFP

왜 더 많은 lane을 사용하는가?

- Lane당 속도 제한: 단일 lane의 속도를 무한정 올리기 어려움
 - 현재: 100G-200G per lane이 주류
 - 미래: 400G per lane 목표 (매우 어려움)
- Parallel transmission: 여러 lane으로 병렬 전송하여 총 bandwidth 증가
- Redundancy: 일부 lane 고장 시에도 동작 가능 (일부 제품)

4. 구체적 예시

800G DR4 Transceiver:

- 4개의 1310nm laser (CWDM4 wavelength)
- 각 laser가 200 Gbps 전송 (PAM4 modulation)
- 4개의 fiber로 TX, 4개의 fiber로 RX
- Single-mode fiber, 최대 500m
- 전력: 약 16-17W

1.6T DR8 Transceiver:

- 8개의 laser (CWDM8 wavelength)
- 각 laser가 200 Gbps 전송
- 8개의 fiber로 TX, 8개의 fiber로 RX
- OSFP form factor (QSFP-DD보다 큼)
- 전력: 약 30-35W (DR4의 약 2배)

5. 왜 18K GPU Cluster에서 DR4와 DR8을 혼용하는가?

- GPU-to-switch: 800G DR4
 - 각 GPU는 800G로 충분
 - QSFP-DD form factor로 공간 효율적
 - 비용 효율적
- Switch-to-switch: 1.6T DR8
 - Switch 간 연결은 많은 traffic 집약 (aggregation)
 - Higher bandwidth 필요
 - Port 수 절감 (1.6T 1개 = 800G 2개)

Trade-off:

- 더 많은 lane = 더 높은 총 bandwidth, 하지만 더 많은 fiber, 더 큰 form factor, 더 높은 전력
- 적은 lane = 더 간단, 저렴, 작은 form factor, 하지만 낮은 bandwidth

핵심 인사이트: DSP - Public Enemy Number One

18K GB300 Cluster의 DSP 전력 비용

DSP 전력 계산:

- $(18,432 \times 6.5W) + (27,648 \times 13W) = 480 \text{ kW}$
- Rack당 (256 racks): **1.8 kW per rack**

DSP 점유율:

- Transceiver 전력의 **50%**
- Transceiver BoM의 **20-30%**
- Cluster TCO의 **2-3%**

"공공의 적 1호"로 불리는 이유

- Transceiver 전력/비용의 절반 차지하는 단일 component
- GPU 수가 늘수록 exponentially 증가
- **근본적 비효율**: Signal degradation을 "사후 복구" (원인 미해결)
- 하지만 **기술적으로 제거 가능** (LPO, CPO)

DSP 제거 시 절감 효과:

- **전력**: 480 kW 제거 → 10년간 \$4.2M 절감 (cluster당)
- **비용**: \$9-14M 절감 (cluster당)
- **간접**: 전력 예산 확보, cooling 절감, reliability 향상

LPO vs CPO 접근:

특성	LPO	CPO
DSP 제거	O	O
Form factor	Pluggable 유지	제거
Copper path	긴 경로 유지	극단적 단축
결과	제한적 성공	근본적 해결

2.4.1 The Crusade Against DSP

DSP가 차지하는 높은 비용과 전력 비율은 업계가 DSP를 중개할 수 있는 기술을 찾도록 동기를 부여했다. DSP에 대한 첫 번째 wave의 공격은 **Linear Pluggable Optics (LPO)**였는데, 이는 DSP를 완전히 제거하고 switch의 SerDes가 transceiver의 TX 및 RX optical element를 직접 구동하도록 시도한다. 그러나 LPO는 [DSP Diviner Loi Nguyen이 2023년 우리와의 인터뷰에서 올바르게 예측한 것처럼](#) 아직 이룩하지 못했다.

CPO는 optical engine을 compute 또는 switch chip과 동일한 package에 배치함으로써 LPO 개념을 다음 단계로 가져간다. CPO의 핵심 이점은 host와 optical engine 사이의 거리가 매우 짧기 때문에 transceiver에 있던 DSP가 더 이상 필요하지 않다는 것이다. CPO는 또한 전력 및 면적을 많이 소비하는 LR SerDes를 제거하고 shorter reach SerDes 또는 wide I/O interface의 경우 clock forwarded wide D2D SerDes를 선호함으로써 훨씬 더 큰 chip shoreline density를 가능하게 하기 때문에 LPO보다 더 나아간다.

자주 인용되는 표현은 CPO가 지난 20년 동안 바로 곧 나올 것이라고 했지만, 왜 그렇게 오랫동안 이룩하지 못했는가? 왜 업계는 pluggable DSP transceiver를 고수하는 것을 선호했는가?

Pluggable transceiver의 주요 이점 중 하나는 높은 상호운용성이다. OSFP 및 QSFP-DD와 같은 표준 form factor와 OIF standard를 준수함으로써, 고객은 일반적으로 switch 및 server vendor 와 독립적으로 transceiver vendor를 선택할 수 있어 조달 유연성과 더 강한 협상력을 누린다.

또 다른 큰 이점은 field serviceability이다. Transceiver를 설치하고 교체하는 것은 remote hand 한 쪽으로 switch 또는 server chassis에서 플러그를 뽑을 수 있기 때문에 간단하다. 이와 대조적으로 CPO의 경우 optical engine의 어떤 failure도 전체 switch를 사용할 수 없게 만들 수 있다. 서비스 가능한 failure 조차도 troubleshoot하고 수정하기가 복잡할 수 있다. 종종 laser가 가장 일반적인 failure point이며, 대부분의 CPO 구현은 이제 더 나은 serviceability 및 replaceability를 위해 pluggable external laser source를 사용하지만, 다른 non-pluggable CPO component의 failure에 대한 불안은 여전히 남아 있다.

2.5 Why CPO? The I/O challenge, BW density, and bottlenecks

전력을 많이 소비하고 비용이 많이 드는 DSP를 제거하고 LR SerDes의 사용을 최소화하거나 제거하는 것 외에도, CPO 채택의 또 다른 큰 이점은 에너지 소비 대비 더 큰 interconnect bandwidth density이다. Bandwidth density는 단위 면적 또는 channel당 전송되는 데이터의 양을 측정하며, 제한된 공간이 고속 데이터 전송을 위해 얼마나 효과적으로 활용될 수 있는지를 반영한다. Energy efficiency는 데이터 단위를 전송하는 데 필요한 에너지를 정량화한다.

따라서 에너지 소비 대비 interconnect bandwidth density는 주어진 interconnect의 객관적 품질을 결정할 때 매우 중요한 **Figure of Merit (FoM)**이다. 물론 최적의 interconnect는 거리 및 비용 parameter 내에 맞는 것이기도 하다. 아래 차트를 검토할 때 명확한 추세가 나타난다. 이 figure of merit는 거리가 증가함에 따라 electrical link의 경우 exponentially하게 저하된다. 또한 순수한 electrical interface에서 optical-electrical conversion이 필요한 interface로 이동하면 효율성이 상당히 떨어진다. 잠재적으로 magnitude의 order만큼. 이러한 drop은 chip에서 transceiver 가 있는 front-panel까지 일정 거리로 signal을 구동하는 데 에너지가 필요하기 때문에 발생한다. Optical DSP에 전력을 공급하는 데 더 많은 에너지가 필요하다. CPO 기반 communication을 위한 figure of merit curve는 pluggable보다 확실히 위에 있다. 아래 차트에 표시된 바와 같이, CPO는 동일한 거리 범위에서 소비된 에너지당 면적당 더 많은 bandwidth density를 제공하여 객관적으로 더 나은 interconnect를 만든다.

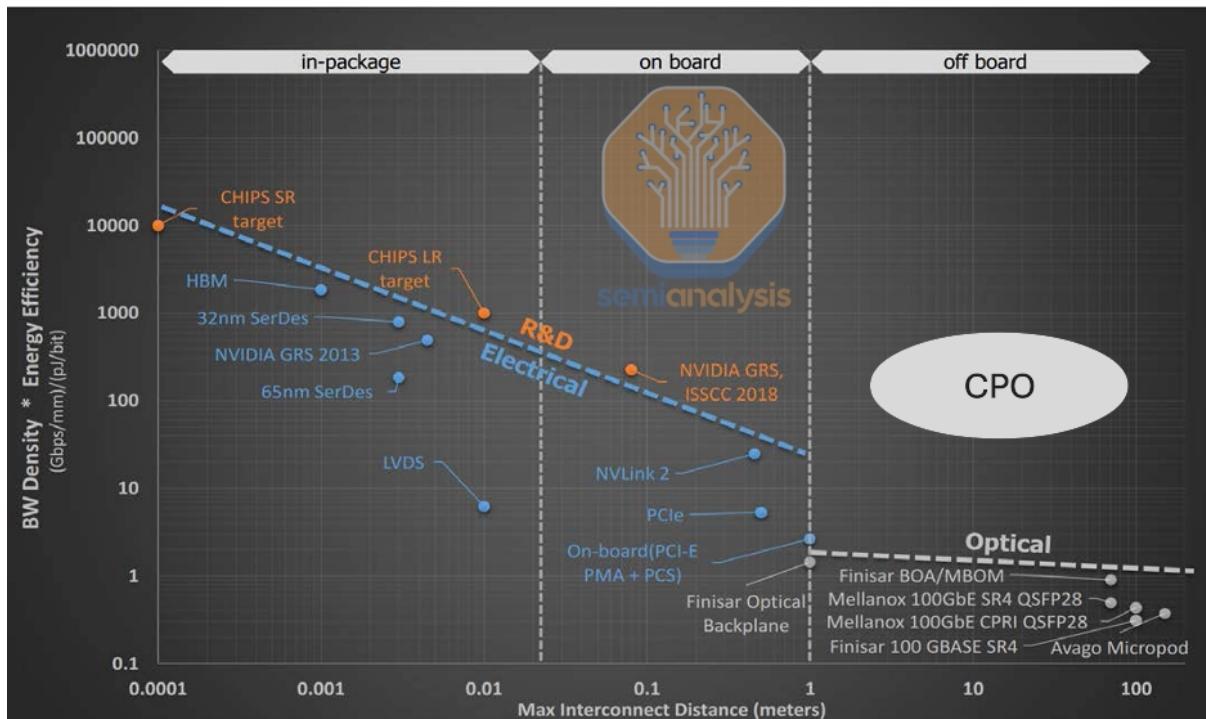


Figure 24: Source: G Keeler, DARPA 2019, SemiAnalysis

핵심 인사이트: Figure of Merit (FoM) - Interconnect의 성적표

1. Figure of Merit (FoM)란?

정의: Interconnect의 객관적 품질을 나타내는 지표

FoM 공식:

$$FoM = \frac{\text{Bandwidth Density (Gbps/mm)}}{\text{Energy Efficiency (pJ/bit)}}$$

더 높은 FoM = 더 좋은 interconnect:

- 단위 면적(mm)당 더 많은 데이터 전송
- 비트당 더 적은 에너지 소비
- 즉, 공간 효율 × 전력 효율

2. 구성 요소 분해

Bandwidth Density (대역폭 밀도):

- 단위: Gbps/mm (또는 Tbps/mm)
- Chip shoreline 1mm당 전송 가능한 데이터량
- Reticle-sized chip ($\approx 858\text{mm}^2$, perimeter $\approx 120\text{mm}$)에서 중요
- 예: $0.4 \text{Tbps/mm} \times 120\text{mm} = 48 \text{Tbps total bandwidth}$

Energy Efficiency (에너지 효율):

- 단위: pJ/bit (pico-Joule per bit)
- 1 bit 전송에 소비되는 에너지
- 낮을수록 좋음
- 예: $10 \text{ pJ/bit} \times 1 \text{Tbps} = 10\text{W}$

3. 차트 해석: 거리에 따른 FoM 변화

Pure Electrical (Copper):

- Short reach (<1cm): 최고 FoM
 - Die-to-die, chiplet interconnect
 - 예: PCIe, 2.5D/3D packaging
 - 이유: 신호 감쇠 최소, DSP 불필요
- Medium reach (1-10cm): FoM 급격히 저하
 - Package-to-package interconnect
 - 여전히 copper 가능하지만 LR SerDes 필요
- Long reach (>10cm): FoM 매우 낮음
 - PCB trace를 통한 연결
 - 높은 전력 소비, 낮은 bandwidth density

Pluggable Optics:

- Electrical-Optical transition에서 FoM 급락
 - E/O, O/E conversion loss
 - DSP chip의 높은 전력
 - LR SerDes 전력
- 10cm-10m 범위: FoM이 낮지만 일정
 - Optical signal은 거리에 둔감
 - 대부분의 전력은 DSP와 E/O conversion
 - 거리가 늘어나도 추가 전력 거의 없음

CPO (Co-Packaged Optics):

- Pluggable보다 확실히 위: 1-10m 범위에서
 - DSP 제거로 전력 대폭 절감
 - SR SerDes 또는 wide I/O로 전력 효율 향상
 - 더 높은 bandwidth density (shoreline)
- Electrical과 optical 사이의 "sweet spot"
 - Electrical의 energy efficiency 상속 (짧은 전기 경로)
 - Optical의 reach 상속 (긴 광 경로)

4. "할 수 있는 곳에서는 Copper, 해야 할 때는 Optical"

i) 격언을 FoM 차트가 증명:

- <1cm: Pure copper가 왕 (chiplet, interposer)
- 1-2m: Copper 가능하지만 FoM 떨어짐 (in-rack)
- >2m: Optics 필수 (rack-to-rack)
- CPO의 역할: Optics 영역(>2m)에서 FoM을 최대한 끌어올림

핵심 통찰: FoM 차트는 interconnect의 "물리 법칙"을 보여줍니다. Copper는 짧은 거리에서 exponentially 좋지만, 거리가 늘어나면 exponentially 나빠집니다. Optics는 전환 비용이 크지만(E/O conversion), 일단 optical domain에 들어가면 거리에 둔감합니다. CPO는 "전환 비용을 최소화"하여 optics의 최적 FoM을 달성합니다.

이 차트는 또한 “할 수 있는 곳에서는 copper를 사용하고 해야 할 때는 optical을 사용하라”는 격언을 보여준다. Copper를 통한 shorter reach communication은 사용 가능할 때 우수하다. NVIDIA는 copper를 통해 네트워크로 연결될 수 있는 GPU 수를 최대화하기 위해 intra-rack density의 한계를 밀어붙이는 목적으로만 설계된 rack-scale GPU architecture로 이 mantra를 받아들인다. 이것이 [GB200 NVL72](#)에 사용되는 scale-up network architecture 뒤의 rationale이며, NVIDIA는 [Kyber rack](#)에서 이 아이디어를 더욱 발전시키고 있다. 그러나 – CPO의 성숙도가 scale-up을 위한 FoM curve의 일부에 접근하는 것을 실행 가능하게 만들고 performance per TCO 관점에서 가치 있게 만드는 것은 시간 문제일 뿐이다.

2.5.1 Input/Output (I/O) Speedbumps and Roadblocks

Transistor density와 compute (FLOP로 표현되는)는 잘 확장되었지만, I/O는 훨씬 더 느리게 확장되어 전체 시스템 성능에서 bottleneck을 만들었다. Off-chip I/O를 위해 사용 가능한 shoreline은 매우 제한적인데, off-chip로 나가는 데이터가 organic package substrate의 제한된 수의 I/O를 통해 escape해야 하기 때문이다. 추가로, 각 개별 I/O의 signaling speed를 증가시키는 것이 점점 더 어려워지고 전력 집약적이 되어, 데이터 이동을 더욱 제약한다. 이것이 interconnect bandwidth가 다른 computing trend에 비해 지난 수십 년 동안 그렇게 형편없이 확장된 주요 이유이다.

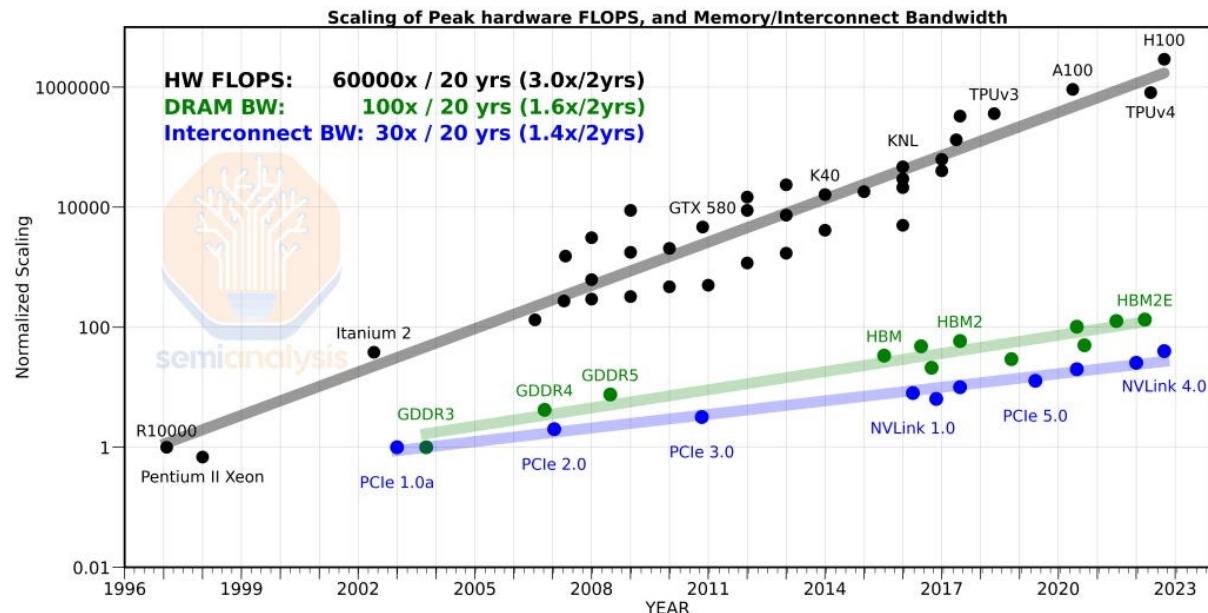


Figure 25: Source: Amir Gholami

HPC application을 위한 off-package I/O density는 단일 flip-chip BGA package의 bump 수 제한으로 인해 plateau에 도달했다. 이것은 escape bandwidth를 확장하는 데 대한 제약이다.

Interconnect Scaling for Higher Bandwidth



Figure 26: Source: TSMC

핵심 인사이트: I/O Wall - AI 시대의 새로운 Bottleneck

1. The Great Imbalance

Transistor Density의 폭발적 증가:

- Moore's Law: Transistor 수 18-24개월마다 2배
- 지난 20년: 약 $2^{10} = 1,000$ 배 증가
- FLOP (compute) 성능도 유사하게 증가

I/O Bandwidth의 정체:

- 같은 기간 I/O bandwidth는 약 10-20배만 증가
- 100배 격차 발생!
- 차트에서 명확히 보임: Compute curve vs I/O curve 급격히 벌어짐

2. 왜 I/O가 느리게 증가하는가?

Shoreline의 물리적 한계:

- Shoreline: Chip 외곽 둘레 (perimeter)
- Die size는 reticle limit (858mm^2)로 제한
- Perimeter: 약 $4 \times \sqrt{858} \approx 120\text{mm}$
- I/O는 shoreline을 통해서만 나갈 수 있음
- Bump pitch는 물리적 한계로 더 줄이기 어려움

Package Substrate의 한계:

- Organic substrate의 escape bandwidth 제한
- 수천-수만 개의 bump를 모두 routing하기 어려움
- Layer 추가는 비용과 복잡도 급증
- HPC용 flip-chip BGA package는 plateau 도달

SerDes Speed의 한계:

- 각 I/O의 속도(Gbps)를 올리는 것도 한계 도달 중
- 224G SerDes는 이미 매우 어려움
- 448G는 더욱 어려움 (뒤에서 상세 설명)

3. AI가 I/O를 더욱 악화시킴

Traditional HPC vs AI:

- Traditional HPC:** Compute-bound (I/O 덜 중요)
- AI Training/Inference:** Communication-intensive
 - All-reduce, all-gather 등 collective communication 빈번
 - Model parameter 교환으로 enormous 데이터 이동
 - Transformer architecture: Attention mechanism으로 communication 유발

4. I/O Wall의 영향

시스템 레벨:

- GPU utilization 저하 (50-70%로 떨어질 수 있음)
- Training time 증가
- TCO 악화 (비싼 GPU를 놀림)

Architecture 레벨:

- Model parallelism strategy 제약
- Scalability limit (GPU 수 늘려도 성능 향상 둔화)
- 새로운 architecture 탐색 필요 (예: sparse model)

5. CPO가 해결하는 방법

Shoreline Density 극대화:

- Wide I/O (PCIe 등) 사용 가능: 10 Tbps/mm
- SR SerDes의 면적/전력 효율 향상
- 0.4 Tbps/mm → 10 Tbps/mm = **25배 향상**

실제 예시 (가상):

- CPO + Wide I/O로 120mm shoreline 활용
- $10 \text{ Tbps/mm} \times 120\text{mm} = 1.2 \text{ Pbps}$ (1,200 Tbps!)
- Blackwell 대비 50배 이상 I/O bandwidth
- Arithmetic intensity를 workload에 맞출 수 있음

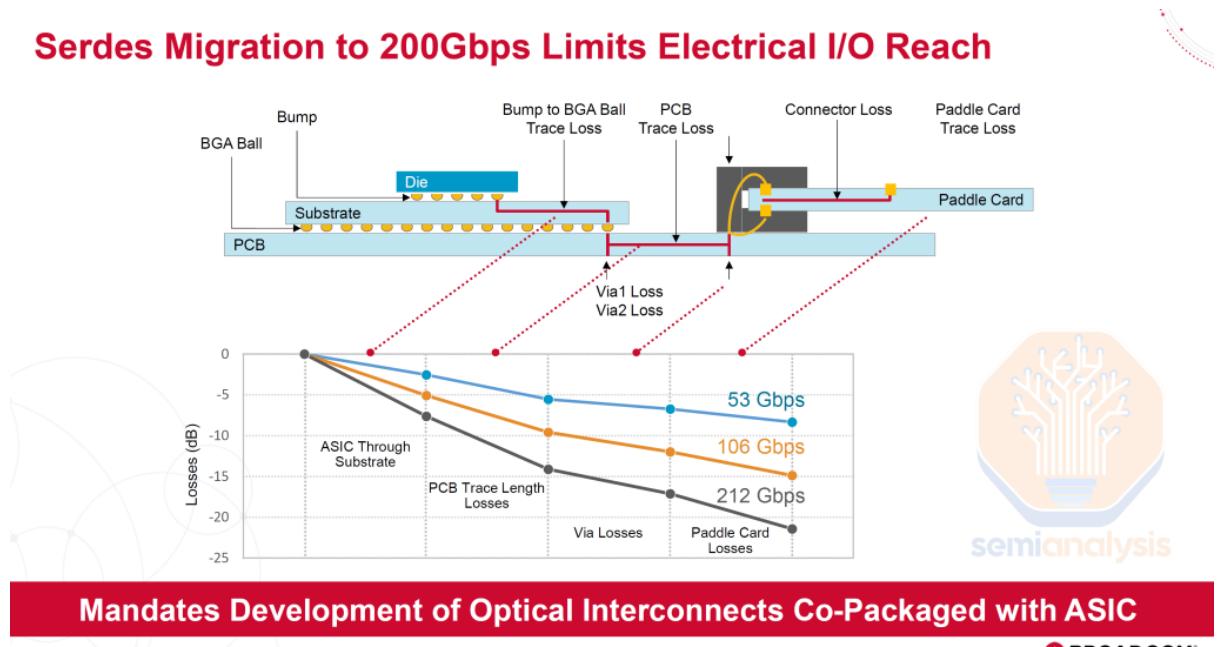
핵심 통찰: "I/O Wall"은 AI 시대의 진짜 bottleneck입니다. Moore's Law가 계속되어도 I/O가 따라가지 못하면 GPU는 "굶주리게" 됩니다. CPO + Wide I/O는 이 wall을 허무는 가장 유망한 해법입니다. 이것이 CPO가 단순히 "전력 절감 기술"이 아니라 "AI scaling의 미래를 결정하는 핵심 기술"인 이유입니다.

2.5.2 Electrical SerDes Scaling Bottlenecks

제한된 수의 I/O로 더 많은 escape bandwidth를 실현하는 방법은 각 I/O가 signaling하는 frequency를 밀어붙이는 것이다. 오늘날 NVIDIA와 Broadcom은 SerDes IP의 선두에 있다. NVIDIA

는 Blackwell에서 224G SerDes를 출하하고 있으며, 이것이 그들의 blazing fast NVLink를 가능하게 한다. 마찬가지로 Broadcom은 2024년 후반부터 optical DSP에서 224G SerDes를 sampling하고 있다. 업계에서 가장 많은 AI FLOP을 출하하는 두 회사가 high-speed SerDes IP에서도 선두를 달리는 것은 우연이 아니다. 이는 AI 성능과 throughput 사이의 근본적인 연결을 강화하며, 여기서 데이터 이동 효율성을 최대화하는 것은 raw compute power를 제공하는 것만큼 중요하다.

그러나 바람직한 reach에서 더 높은 line speed를 제공하는 것이 점점 더 어려워지고 있다. frequency가 증가함에 따라 insertion loss가 증가하며, 아래 차트에 표시된 대로이다. Signal path가 길어짐에 따라 특히 더 높은 SerDes signaling speed에서 loss가 증가하는 것을 볼 수 있다.



Mandates Development of Optical Interconnects Co-Packaged with ASIC

3 | Broadcom Proprietary and Confidential. Copyright © 2023 Broadcom. All Rights Reserved. The term "Broadcom" refers to Broadcom Inc. and/or its subsidiaries.



Figure 27: Source: Broadcom

SerDes scaling은 plateau에 접근하고 있다. 더 높은 속도는 추가 signal recovery component 없이는 매우 짧은 거리에서만 유지될 수 있으며, 이는 차례로 복잡성, 비용, latency 및 전력 소비를 증가시킨다. 224G SerDes에 도달하는 것이 어려웠다.

448G SerDes를 앞서 보면, 단지 몇 centimeter 이상을 구동하는 feasibility는 더 불확실하다. NVIDIA는 bi-directional SerDes 기술을 사용하여 Rubin에서 electrical channel당 448G connectivity를 제공하고 있다. True 448G uni-directional SerDes를 달성하는 것은 추가 개발이 필요할 것이다. 56G SerDes 시대 이후 널리 퍼진 PAM4 modulation 대신 PAM6 또는 PAM8과 같은 더 높은 order의 modulation으로 이동해야 할 수도 있다. Signal당 2 bit를 encoding하는 PAM4를 사용하여 448G에 도달하려면 과도한 전력 소비와 insertion loss로 인해 untenable할 가능성성이 있는 244Gbaud의 baud rate가 필요할 것이다.

2.5.3 SerDes Scaling Plateau as a Roadblock for Scaling NVLink

NVLink protocol에서 NVLink 5.0의 bandwidth는 NVLink 1.0과 비교하여 11배 이상 증가했다. 그러나 이 성장은 lane 수의 상당한 증가에서 나온 것이 아니며, NVLink 1.0의 32 lane에서 NVLink 5.0의 36 lane으로 약간만 증가했다. Scaling의 핵심 driver는 SerDes lane speed의 10배 증가였으며, 20G에서 200G로 증가했다. 그러나 NVLink 6.0에서 NVIDIA는 200G SerDes에 머물 것으로 예상되며, 이는 lane 수를 두 배로 늘려야 함을 의미한다. 동일한 수의 물리적 copper wire를 사

용하면서 lane 수를 효과적으로 두 배로 늘리기 위해 bi-directional SerDes를 사용함으로써 이를 영리하게 제공한다. 이를 넘어서면 SerDes speed를 확장하거나 더 많은 lane을 맞추기 위한 제한된 shoreline 가용성을 극복하는 것이 점점 더 어려워질 것이며 total escape bandwidth는 막히게 될 것이다.

Escape bandwidth를 확장하는 것은 throughput이 차별화 요소인 선두 기업에게 중요하다. NVLink scale-up fabric이 중요한 moat인 NVIDIA의 경우, 이 roadblock은 AMD 및 hyperscaler와 같은 경쟁자가 따라잡기 쉽게 만들 수 있다.

NVLink Speeds								
	Calculation	NVLink 1.0	NVLink 2.0	NVLink 3.0	NVLink 4.0	NVLink 5.0	NVLink 6.0	NVLink 6.0 vs. NVLink 1.0 Multiple
Year		2014	2017	2020	2022	2024	2026	
Link Number	(a)	4	6	12	18	18	18	4.5x
Lane Number per Link	(b)	8	8	4	2	2	4	0.5x
Total Lane	(a)*(b)=(c)	32	48	48	36	36	72	2.3x
Data Rate per Lane (Gb/s)	(d)	20	25	50	100	200	200	10.0x
Total Bidirectional Bandwidth (GB/s)	((c)*(d))/8	160	300	600	900	1,800	3,600	22.5x

Figure 28: Source: NVIDIA, SemiAnalysis

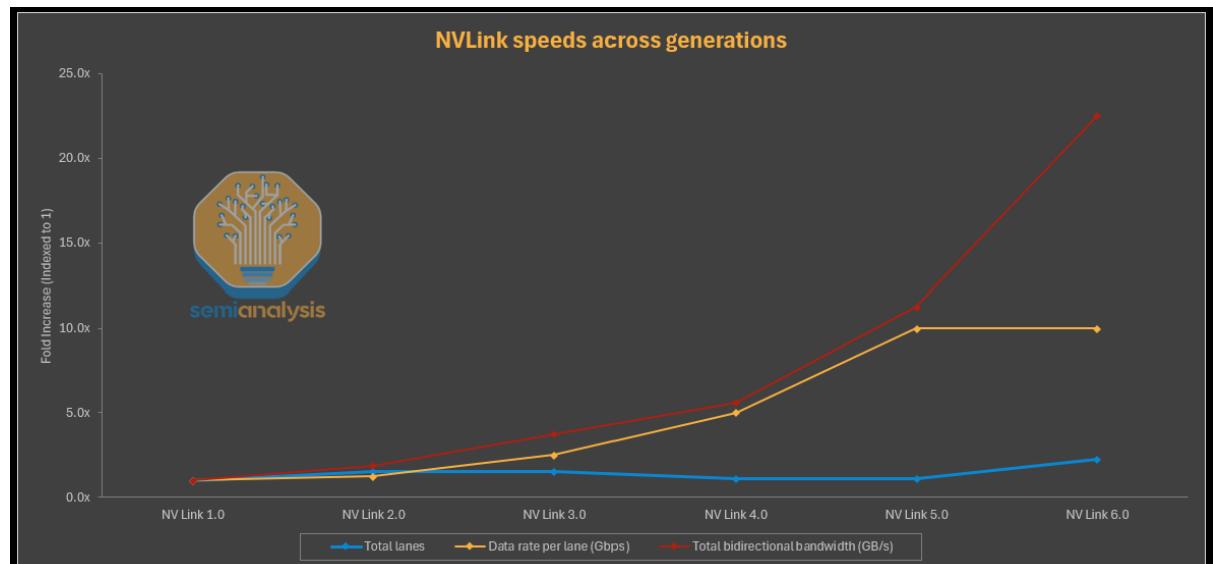


Figure 29: Source: NVIDIA, SemiAnalysis

기술 용어: SerDes Scaling의 물리적 한계

1. Insertion Loss의 exponential 증가

Insertion Loss란?:

- Signal이 medium(copper)을 통과하며 잃는 에너지
- 단위: dB (decibel) - 높을수록 더 많은 loss
- Frequency dependent: 고주파수일수록 loss 커짐
- Distance dependent: 거리가 길수록 loss 커짐

왜 고주파수에서 loss가 커지는가?:

- Skin effect: 전류가 표면에만 흐르면서 저항 증가
- Dielectric loss: PCB 절연체가 고주파수를 흡수

- **Radiation loss:** 고주파수 신호가 전자기파로 방출

구체적 수치 (차트 기반):

- 56G SerDes (28 GHz): 10cm에서 약 20 dB loss
- 112G SerDes (56 GHz): 10cm에서 약 35 dB loss
- **224G SerDes (112 GHz): 10cm에서 약 50+ dB loss**
- 448G SerDes (224 GHz): 10cm에서 추정 80+ dB loss (거의 불가능)

2. 왜 224G SerDes가 어려운가?

PAM4 with 224G:

- $224 \text{ Gbps} \div 2 \text{ bits/symbol} = 112 \text{ Gbaud}$
- $112 \text{ Gbaud} = 112 \text{ GHz fundamental frequency}$
- PCB trace를 통과하면 50+ dB insertion loss
- **50 dB = 신호가 1/100,000로 감소!**

Recovery의 어려움:

- **Equalization:** 고주파수 성분을 선택적으로 증폭
 - Transmitter equalization (pre-emphasis)
 - Receiver equalization (CTLE, DFE)
 - 매우 복잡한 회로, 높은 전력 소비
- **CDR (Clock Data Recovery):** 저하된 신호에서 clock 추출 극도로 어려움
- **Noise margin:** PAM4의 4개 level이 noise에 묻힘

전력 소비 폭발:

- 56G SerDes: 약 20-30 mW/lane
- 112G SerDes: 약 40-50 mW/lane
- **224G SerDes: 추정 80-100+ mW/lane**
- 448G SerDes: 예상 150-200 mW/lane (untenable)

3. 448G SerDes의 딜레마

PAM4 approach (현재 표준):

- $448 \text{ Gbps} \div 2 \text{ bits/symbol} = \mathbf{224 \text{ Gbaud}}$
- 224 GHz frequency → 거의 불가능한 insertion loss
- 전력 소비가 untenable 수준 (150-200 mW/lane)

Higher-order modulation (PAM6, PAM8):

- **PAM6:** 6 levels → $\log_2(6) = 2.58 \text{ bits/symbol}$
 - $448 \text{ Gbps} \div 2.58 = 174 \text{ Gbaud}$ (PAM4보다 나음)
 - 하지만 level 간 margin 더 좁아짐
- **PAM8:** 8 levels → 3 bits/symbol
 - $448 \text{ Gbps} \div 3 = 149 \text{ Gbaud}$ (더 나음)
 - 하지만 noise margin 극도로 작아짐 (level 간 1/8)
 - DSP 복잡도와 전력 급증

NVIDIA의 해법: Bi-directional SerDes:

- 하나의 wire에서 양방향 동시 전송
- $2 \times 224\text{G} = 448\text{G}$ effective
- 영리하지만 복잡도 증가, echo cancellation 필요

4. NVLink Scaling의 위기

NVLink 역사 (차트 참고):

- NVLink 1.0: $20\text{G} \times 32 \text{ lanes} = 640 \text{ Gbps}$
- NVLink 2.0: $25\text{G} \times 32 \text{ lanes} = 800 \text{ Gbps}$
- NVLink 3.0: $50\text{G} \times 32 \text{ lanes} = 1.6 \text{ Tbps}$
- NVLink 4.0: $100\text{G} \times 36 \text{ lanes} = 3.6 \text{ Tbps}$
- NVLink 5.0: $200\text{G} \times 36 \text{ lanes} = 7.2 \text{ Tbps}$

핵심 관찰:

- Lane 수: $32 \rightarrow 36$ (거의 변화 없음)
- SerDes speed: $20\text{G} \rightarrow 200\text{G}$ (**10배** 증가)
- Bandwidth 증가는 대부분 SerDes speed에서 나옴

NVLink 6.0의 문제 (Rubin):

- NVIDIA는 200G SerDes에 머물 것으로 예상
- 대신 bi-directional SerDes로 448G 달성
- 하지만 이것도 한계: Lane 수를 늘리기 어려움 (shoreline 제한)
- Beyond NVLink 6.0: Scaling path가 불투명

경쟁자들의 기회:

- NVIDIA의 NVLink가 plateau → moat 약화
- AMD, Intel, hyperscaler들이 따라잡기 쉬워짐
- NVIDIA에게는 existential threat

핵심 통찰: SerDes scaling의 물리적 한계는 단순히 "기술적 도전"이 아니라 "AI 업계의 구조를 바꿀 수 있는 전환점"입니다. NVIDIA의 NVLink moat가 SerDes plateau로 무너질 수 있습니다. CPO는 이 위기에서 NVIDIA를 구할 수 있는 핵심 기술입니다.

이 딜레마에 대한 솔루션 또는 다른 방식으로 표현하면 필요한 타협은 electrical I/O를 가능한 한 짧게 하고 더 높은 bandwidth를 달성하기 위해 host ASIC에 가능한 한 가까운 곳에서 optical link로 데이터 전송을 offload하는 것이다. 이것이 CPO가 interconnect의 **holy grail**로 간주되는 이유이다. CPO는 substrate를 통해서는 interposer를 통해서는 ASIC package에서 optical communication이 발생하도록 허용한다. Electrical signal은 손실이 많은 **Copper-Clad Laminate (CCL)**를 통해 수십 centimeter가 아니라 package substrate를 통해 몇 millimeter만 이동하면 되거나, 이상적으로는 더 높은 품질의 interposer를 통해 훨씬 더 짧은 거리를 이동하면 된다.

SerDes는 대신 shorter reach에 최적화될 수 있으며, 이는 동등한 long reach SerDes보다 훨씬 적은 회로가 필요하다. 이는 설계를 더 쉽게 만들면서 더 적은 전력과 silicon 면적을 소비한다. 이러한 단순화는 higher-speed SerDes를 구현하기 더 쉽게 만들고 SerDes scaling roadmap을 확장한다. 그럼에도 불구하고 우리는 전통적인 bandwidth model에 의해 제약을 받으며, 여기서 bandwidth

density는 SerDes speed에 비례하여 계속 확장된다.

훨씬 더 높은 B/W density를 달성하기 위해, wide I/O PHY는 극도로 짧은 거리에서 더 나은 옵션이며, SerDes interface보다 소비된 전력당 더 나은 bandwidth density를 제공한다. wide I/O는 또한 훨씬 더 advanced한 package의 비용을 수반한다. 그러나 CPO의 경우 이것은 무의미한 point이다. Packaging이 이미 매우 advanced하므로 wide I/O PHY를 통합하는 것은 추가 packaging 복잡성을 거의 또는 전혀 추가하지 않는다.

2.6 Wide I/O vs SerDes

Electrical signal을 상대적으로 긴 거리로 구동할 필요가 더 이상 없으면, 짧은 거리에서 훨씬 더 나은 shoreline density를 제공하는 wide interface를 사용하여 serialized interface를 완전히 escape 할 수 있다.

그러한 예 중 하나는 UCIe interface이다. UCIe-A는 advanced package를 위해 설계된 최대 ~10 Tbit/s/mm의 shoreline density를 제공할 수 있다 (즉, sub-2mm reach의 interposer를 통해 interfacing하는 chiplet). Reticle sized chip의 long edge에서 이것은 최대 330 Tbit/s (41TByte/s) 의 off-package bandwidth이다. 이것은 양쪽 edge에서 660 Tbit/s의 bi-directional bandwidth이다. 이것은 23.6 Tbit/s의 off-package BW만 가진 Blackwell과 비교되며, 이는 약 0.4 Tbit/s/mm 의 shoreline density에 해당하며, 이는 큰 차이이다.

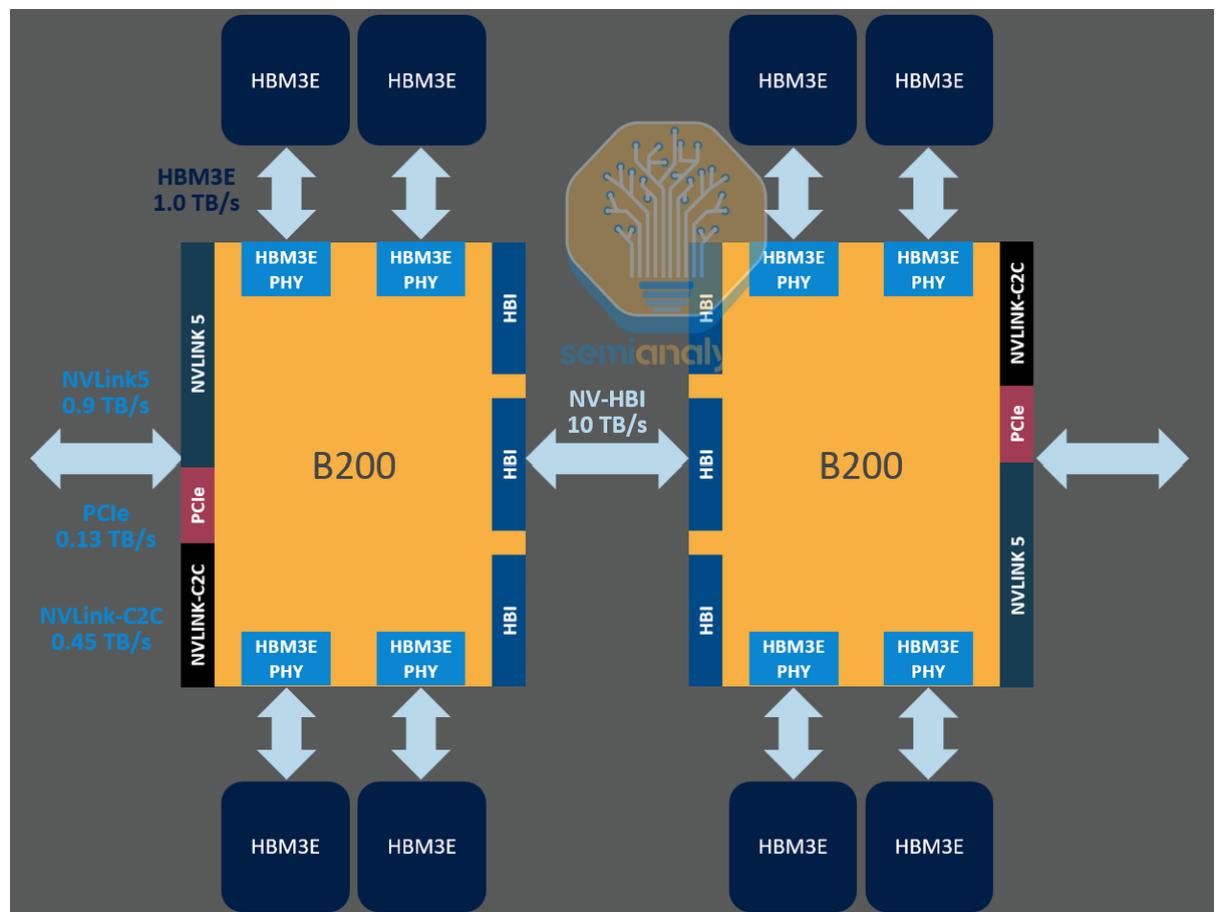


Figure 30: Source: SemiAnalysis

Shoreline Density Comparison					
Location	PHY Interface	Interface Type	Edge Dimensions (mm)	Bi-directional B/W (TB/s)	B/W Density (Tbps/mm)
 North+South	HBM3E PHY	Wide	101.2	8.0	0.6
	NVLink5			1.8	
	NVLink-C2C			0.9	
	PCIe Gen 6			0.3	
	Total off-package B/W		65.0	3.0	0.4
East + West	Nvidia HBI	Wide	32.5	10.0	2.5

Figure 31: Source: SemiAnalysis

기술 용어: Wide I/O vs SerDes - Paradigm Shift

1. SerDes vs Wide I/O의 근본적 차이

SerDes (Serializer/Deserializer):

- 개념: 병렬 데이터를 직렬로 변환하여 소수의 고속 lane으로 전송
- 예시: 32 bit 데이터를 1개의 224G lane으로 serialize
- 장점: 적은 wire 수, 긴 거리 지원
- 단점: 복잡한 회로, 높은 전력, 고주파수 신호

Wide I/O (Parallel Interface):

- 개념: 다수의 저속 lane으로 병렬 전송
- 예시: 32 bit 데이터를 32개의 저속 lane으로 동시에 전송
- 장점: 간단한 회로, 낮은 전력, 저주파수 신호
- 단점: 많은 wire 필요, 짧은 거리만 지원

2. UCIE (Universal Chiplet Interconnect Express)

UCIE란?:

- Chiplet 간 interconnect를 위한 industry standard
- Wide I/O 기반 (parallel interface)
- Intel, AMD, ARM, TSMC 등 주요 업체 참여

UCIE-A (Advanced Package):

- 목표: Interposer 기반 (2.5D/3D packaging)
- 거리: Sub-2mm (매우 짧음)
- Bandwidth density: 최대 10 Tbps/mm
- 전력 효율: 매우 우수 (저주파수)

UCIE-S (Standard Package):

- 목표: Organic substrate 기반
- 거리: 수 mm ~ 수 cm
- Bandwidth density: 약 1.8 Tbps/mm
- 전력 효율: 여전히 SerDes보다 좋음

3. Bandwidth Density 비교

Reticle-sized chip의 shoreline (약 120mm):

Technology	Density	Total BW
224G SerDes (LR)	0.4 Tbps/mm	48 Tbps
UCIE-S (substrate)	1.8 Tbps/mm	216 Tbps
UCIE-A (interposer)	10 Tbps/mm	1,200 Tbps

Blackwell 실제:

- Off-package BW: 23.6 Tbps (NVLink 5.0)
- Shoreline density: 약 0.4 Tbps/mm
- UCIE-A 대비 25배 차이!

4. 왜 Wide I/O가 CPO에 최적인가?

거리가 문제가 아니므로:

- CPO에서 electrical path는 2-5mm (매우 짧음)
- Wide I/O의 단점(짧은 거리)이 문제 없음
- 장점만 취함: 높은 bandwidth density + 낮은 전력

SerDes의 불필요한 복잡도 제거:

- Serialization/Deserialization 회로 제거
- Clock recovery 회로 단순화 (clock forwarding 사용)
- Equalization 최소화 (신호 깨끗함)

전력 효율 극대화:

- 저주파수 신호 → 낮은 전력
- 간단한 회로 → 작은 면적
- 예: UCIE-A는 224G SerDes 대비 전력 50-70% 절감 추정

5. 왜 Broadcom/NVIDIA는 SerDes를 고집하는가?

Hybrid 전략:

- In-rack copper: 여전히 필요 (CPC)
 - Copper는 1-2m 거리에서 SerDes 필요
 - Wide I/O는 copper에 부적합
- CPO/Optics: Wide I/O 사용
- 동일 silicon에 둘 다 지원: 여러 tape-out 피하기

점진적 migration:

- CPO 채택이 느릴 것으로 예상
- 당분간 SerDes 기반 copper 계속 사용
- Wide I/O 전용 chip은 niche market

SerDes IP 활용:

- NVIDIA/Broadcom은 세계 최고 SerDes IP 보유
- 기존 투자 활용

6. 미래: Wide I/O의 승리?

Short term (2025-2027):

- SerDes 기반 CPO가 주류
- UCIe-S (substrate) 일부 사용

Medium term (2027-2030):

- UCIe-S 더 널리 사용
- Interposer 기반 CPO 등장 시작

Long term (2030+):

- UCIe-A (interposer) 기반 CPO가 ultimate goal
- 10 Tbps/mm bandwidth density
- SerDes는 copper-only application으로 제한

물론 이것은 like-for-like 비교가 아닌데, 이러한 off-package PHY는 긴 거리를 구동하는 데 필요하기 때문이다. 어떤 것이든 이것이 설명되고 있는 바로 그 point이다: CPO를 사용하면 signal이 electrically하게 긴 거리로 구동되지 않기 때문에 reach는 더 이상 고려 사항이 아니다. 10 Tbit/s/mm의 bandwidth density에서 bottleneck은 더 이상 electrical interface에 있지 않고 link의 다른 부분에 있으며, 즉 반대편의 fiber에서 얼마나 많은 bandwidth가 escape할 수 있는지에 있다.

이 제약에 도달하는 것은 오늘날의 현실에서 매우 먼 end state이며 OE는 host와 interposer를 공유해야 할 것이다. Interposer 자체에 CPO를 통합하는 것은 substrate에 OE를 안정적으로 통합하는 것보다 roadmap에서 훨씬 더 멀리 있다. Substrate의 PHY 성능은 물론 열등하며 UCIe-S는 약 1.8Tbit/s/mm의 shoreline density를 제공한다. 이것은 여전히 우리가 믿는 224G SerDes가 ~0.4Tbit/s/mm에서 제공하는 것에 비해 상당한 향상이다.

그러나 Broadcom과 NVIDIA는 wide interface가 제공하는 이점에도 불구하고 roadmap에서 electrical SerDes를 고집하고 있다. 주요 이유는 그들이 여전히 SerDes를 확장할 수 있다고 믿고 optics의 채택이 느리기 때문에 특히 copper를 위해 설계해야 한다고 믿기 때문이다. 또한 hybrid co-packaged copper와 co-packaged optics 솔루션이 계속 유지될 가능성이 더 높아 보이며, 이는 그들이 둘 다를 최적화해야 함을 요구한다. 이 접근 방식은 서로 다른 솔루션에 대한 여러 tape-out의 필요성을 제거하기 위해 취해진다.

2.7 Link Resiliency

Link resiliency와 reliability는 CPO 기술의 또 다른 매우 중요한 추진 요인이다. 대규모 AI cluster에서 link downtime은 전체 cluster availability에 상당한 기여를 하며, link availability와 stability의 작은 개선조차도 infrastructure 투자에 대한 큰 수익을 제공한다.

오늘날 pluggable module을 사용하는 1M link에 접근하는 대규모 AI cluster에서는 하루에 수십 건의 link interruption이 발생할 수 있다. 이 중 일부는 component failure 또는 hardware 품질의 결과로 인한 **hard failure**이며, 많은 것들은 pluggable 기반 솔루션의 고유한 복잡성과 가변성에서 비롯되는 다양한 root cause의 결과로 인한 **soft failure**이다. Signal integrity 문제 및 variation, connector 및 wirebond 품질, component 및 pin의 contamination, noise injection 및 기타 transient effect를 포함하되 이에 국한되지 않는 긴 failure mode tail이 있다. Component failure와의 correlation은 거의 없다. link failure로 인해 반품되는 optical module의 80%는 **no trouble found**이다. CPO는 다음을 통해 대규모 AI network에서 high-speed signal path의 고유한 복잡성과 가변성을 크게 줄인다:

- Optical interface의 component 수를 크게 줄인다. photonic level과 chip/package level 모두

에서의 높은 수준의 통합은 중요한 high-speed assembly의 복잡성을 줄이고 system level에서 reliability와 yield를 향상시킨다. E/O interface의 수도 감소하여 각 interface에서 발생하는 power loss를 최소화한다

- Host ASIC (예: switch)과 optical engine 사이의 host electrical interface의 signal integrity를 크게 개선한다. insertion loss, reflection 및 기타 non-linear impairment는 매우 잘 정의되고 결정적인 design rule 및 manufacturing tolerance를 가진 first level package에 optical engine을 packaging함으로써 크게 감소한다
- Switch 전체에서 high-speed signal path의 port-to-port variation을 줄여 DSP signal processing, host 및 module equalization, host 및 module firmware 및 link optimization algorithm에 overhead와 복잡성을 추가한다. 모든 pluggable module 솔루션과 host SerDes는 복잡성과 failure point로 이어지는 port별 성능의 이러한 variation을 수용하도록 설계되어야 한다

Optical link provisioning의 human 요소를 제거한다. CPO switch 또는 optical engine은 공장에서 완전히 조립되고 테스트된 '**known-good**' 상태로 나오며 switch의 optics를 provision하기 위한 상당한 현장 작업이 필요하지 않으며, 이는 설치 variation, 손상, contamination 및 system과 optical module 간의 compatibility 문제로 이어질 수 있다.

3 Part 3: Bringing CPO to Market and deployment challenges

3.1 CPO optical engine manufacturing considerations and go to market

CPO는 아직 광범위한 채택에 상응하는 수량으로 제조되지 않았다. Broadcom은 Bailly 및 Hum-boldt switch를 포함하는 CPO를 특징으로 하는 production system을 출하한 유일한 vendor였지만, 이제 NVIDIA도 합류하고 있다. 이것들은 매우 낮은 volume으로 출하되었다. CPO는 많은 새로운 manufacturing process와 상당한 manufacturability 과제를 도입한다. 당연히 supply chain의 미성숙과 reliability에 대한 데이터 부족을 고려할 때, 고객들도 기술 채택에 뛰어드는 것을 꺼린다.

CPO가 견인력을 얻으려면 업계 leader가 이러한 제품을 출하하는 데 투자하여 supply chain이 scalable한 manufacturing 및 testing process를 개발하도록 추진해야 한다. NVIDIA가 뛰어들고 있으며 그들의 의도는 supply chain을 준비시키고, 문제를 식별하고 해결하며, 우리가 생각하기에 killer application이 될 것을 위해 datacenter operator를 준비시키는 것이다. Scale-up networking, CPO와 관련하여 집중해야 할 몇 가지 핵심 component와 고려사항이 있으며, 다음의 모든 것이 성능과 manufacturability에 영향을 미친다.

1. Host and Optical Engine packaging
2. Fibers and Fiber coupling
3. Laser Sources and Wavelength Multiplexing
4. Modulator Type

3.2 Host and Optical Engine Packaging

기술 용어: CPO Packaging의 핵심 구성요소

PIC (Photonic Integrated Circuit): Photodetector와 modulator를 포함하는 optical component가 집적된 회로. Silicon photonics process를 통해 제조되며, 빛의 생성, 변조, 전송을 담당한다.

EIC (Electric Integrated Circuit): Driver와 Transimpedance Amplifier (TIA)를 포함하는 electrical 회로. CMOS process로 제조되며, electrical signal을 optical signal로 변환하거나 그 역을 수행한다.

Monolithic Integration: PIC와 EIC를 동일한 silicon wafer에 제조하는 방식. Parasitic과 latency가 가장 낮지만, 약 35nm 미만의 geometry에서 한계를 가진다.

Heterogeneous Integration: SiPho process로 PIC를 제조하고 advanced packaging을 통해 CMOS wafer의 EIC와 통합하는 방식. 3D integration을 통해 최고의 성능 제공.

Hybrid Bonding: Die-to-wafer 또는 wafer-to-wafer bonding 기술로, bump 없이 직접 연결하여 trace 길이를 최소화하고 parasitic을 감소시킨다. TSMC SoIC가 대표적.

FAU (Fiber Array Unit): 여러 optical fiber를 정밀하게 정렬하여 optical engine에 연결하는 unit. Coupling efficiency와 manufacturing yield에 중요한 영향을 미친다.

이름에서 알 수 있듯이 **Co-packaged optics**는 근본적으로 packaging 및 assembly 과제이다. optical engine은 optical 및 electrical component를 모두 가지고 있다. Photodetector와 modulator는 **Photonic Integrated Circuit (PIC)**에 포함된 optical component이다. Driver와 Transimpedance Amplifier는 **Electric Integrated Circuit (EIC)**에 포함된 electrical 회로이다. OE

가 작동하려면 PIC와 EIC가 통합되어야 한다. 이 PIC-EIC 통합을 달성하기 위해 여러 packaging 방법이 존재한다.

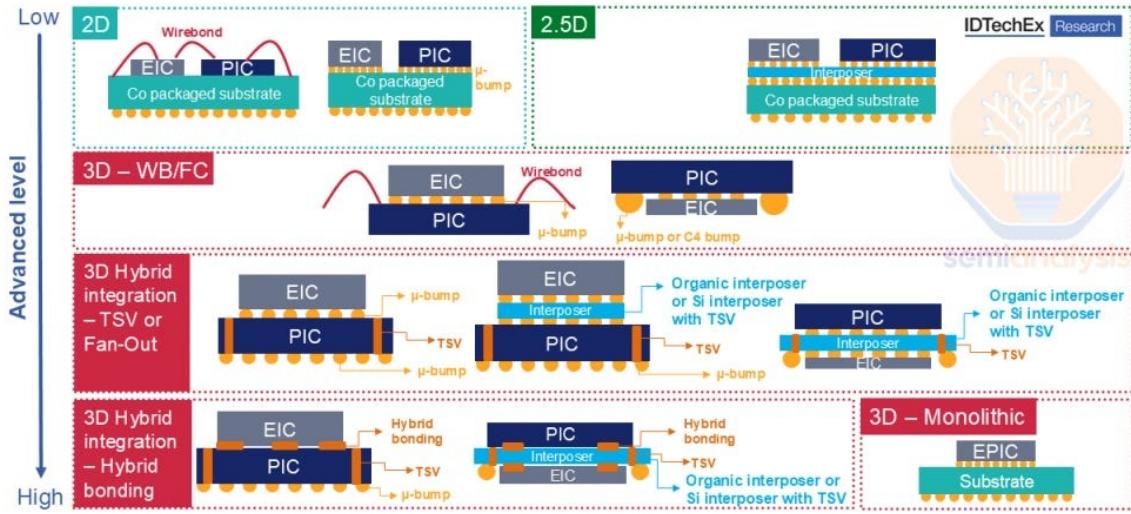


Figure 32: Source: ID TechEx

Optical engine은 PIC와 EIC를 동일한 silicon wafer에 제조함으로써 monolithic일 수 있다. Monolithic integration은 parasitic, latency 및 전력 측면에서 가장 우아한 접근 방식이다. 이것 이 Ayar Labs가 2세대 TeraPHY chiplet에 대해 취한 접근 방식이다 (하지만 그들의 next-gen chiplet은 TSMC COUPE로 pivot한다). GlobalFoundries, Tower 및 Advanced Micro Foundry는 monolithic CMOS 및 SiPho process를 제공할 수 있는 foundry이다. 그러나 photonics process는 전통적인 CMOS처럼 scale할 수 없기 때문에 monolithic process는 약 35nm 미만의 geometry에서 멈춘다. 이는 특히 CPO 시스템에서 예상되는 더 높은 lane speed를 고려할 때 EIC의 capability를 제한한다. 고유한 단순성과 우아함에도 불구하고 이는 monolithic integration을 scaling에 대한 dealbreaker로 만든다. 이것이 Ayar Labs도 추가 scaling을 허용하기 위해 heterogeneously integrated OE로 roadmap을 이동시키고 있는 이유이다.

Heterogenous integration은 SiPho process를 사용하여 PIC를 제조하고 advanced packaging을 통해 CMOS wafer의 EIC와 통합하는 것을 포함하는 주류 접근 방식이 되고 있다. 다양한 packaging 솔루션이 존재하며, 더 advanced한 packaging 솔루션이 더 높은 성능을 제공한다. 그 중에서 3D integration은 최고의 bandwidth와 energy efficiency를 제공한다. EIC와 PIC 통신에 있어서 큰 문제는 성능을 저하시키는 parasitic이다. Trace 길이를 줄이면 parasitic이 크게 감소하고 따라서 coupling efficiency가 증가한다. 3D integration은 bandwidth와 전력 관점에서 CPO의 성능 목표에 도달하는 유일한 방법이다.

3.3 TSMC COUPE is emerging as the integration option of choice

핵심 인사이트: TSMC COUPE - CPO Ecosystem의 새로운 표준

TSMC가 역사적으로 silicon photonics에서 제한적 존재감을 가졌음에도 불구하고 COUPE (COmplex Universal Photonic Engine)가 업계 표준으로 급부상하고 있다. 이는 세 가지 핵심 요인에 기인한다:

- 기술적 우위:** TSMC는 유일하게 합리적 scale에서 die-to-wafer hybrid bonding을 성공적으로 입증한 foundry이다. SoIC (System-on-Integrated-Chips) bumpless interface는

monolithic 통합 다음으로 가장 짧은 trace 길이를 제공하며, iso-power에서 bump 기반 통합보다 23배 이상의 bandwidth density를 달성한다.

2. 통합 ecosystem: COUPE는 단순한 manufacturing process가 아닌 완전한 솔루션이다. EIC (N7 node), PIC (SOI N65 node), 3D stacking, optical I/O design, 완전한 PDK 및 EDA workflow까지 포함한다. 이는 설계자가 fragmented supply chain을 관리할 필요 없이 end-to-end 솔루션을 얻을 수 있음을 의미한다.

3. Vendor lock-in 효과: COUPE를 채택하면 사실상 TSMC 제조 PIC를 사용하기로 commit하는 것이다. TSMC는 다른 foundry의 SiPho wafer를 package하지 않기 때문이다. 이는 TSMC가 value chain의 더 많은 부분을 포착할 수 있게 한다.

결과적으로 NVIDIA, Broadcom, Ayar Labs 등 주요 player들이 기존 solution에서 COUPE로 pivot하고 있다. 특히 Broadcom은 SPIL FOWLP 접근 방식이 parasitic capacitance로 인해 lane당 100G를 넘어 scaling할 수 없다는 것을 인식하고 COUPE로 전환했다. 이는 단순히 하나의 기술 옵션이 아닌, CPO 산업의 **de facto standard**가 되어가고 있음을 시사한다.

TSMC는 fabless giant와 startup 모두를 위한 차세대 OE의 foundry 파트너로 선택되어 앞서 나가고 있다. CPO endpoint를 특징으로 하는 첫 번째 high-volume 제품은 **COnpact Universal Photonic Engine (COUPE)** 이름으로 소개되고 있다. 여기에는 EIC 및 PIC의 제조와 TSMC의 COUPE 솔루션에 따른 heterogeneous integration이 포함된다. NVIDIA는 GTC 2025에서 COUPE optical engine을 자랑스럽게 전시했으며, 이것들이 출하될 첫 번째 COUPE 제품이 될 것이다. Broadcom도 다른 supply chain 파트너와 함께 기존 세대의 OE를 가지고 있음에도 불구하고 미래 roadmap을 위해 COUPE를 채택하고 있다. 앞서 언급한 것처럼 이전에 monolithic optical engine을 위해 Global Foundries의 Fotonix platform에 의존했던 Ayar Labs도 이제 roadmap에 COUPE를 가지고 있다.

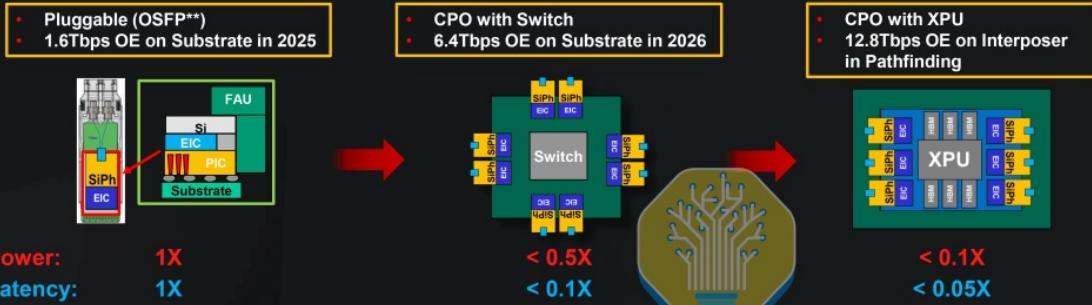
전통적인 CMOS logic에서의 지배력과 달리 TSMC는 이전에 silicon photonics에서 제한적인 존재감을 가졌으며, 여기서 Global Foundries와 Tower Semi가 선호되는 foundry 파트너였다. 그러나 최근 몇 년 동안 TSMC는 photonic capability와 관련하여 빠르게 따라잡고 있다. TSMC는 또한 EIC component를 위한 leading edge CMOS logic에서의 의심할 여지 없는 강점과 leading packaging capability를 가져온다. TSMC는 합리적인 scale에서 die-to-wafer hybrid bonding capability를 성공적으로 입증한 유일한 foundry이며, volume으로 다양한 AMD hybrid bonded chip을 출하했다. hybrid bonding은 PIC와 EIC를 bonding하는 더 성능이 좋은 접근 방식이지만, 상당히 높은 비용이 수반된다. Intel은 유사한 capability를 개발하기 위해 노력하고 있지만 이 기술을 개척하는 데 상당한 과제에 직면했다.

전반적으로 TSMC는 이전에 더 약했던 standalone SiPho capability에도 불구하고 이제 CPO에서 매우 핵심적인 player가 되었다. 다른 주요 player와 마찬가지로 TSMC는 가능한 한 많은 value chain을 포착하는 것을 목표로 한다. TSMC의 COUPE 솔루션을 채택함으로써 고객은 사실상 TSMC 제조 PIC를 사용하기로 commit하는 것인데, TSMC는 다른 foundry의 SiPho wafer를 package하지 않기 때문이다. 많은 CPO 중심 기업들은 실제로 향후 몇 년 동안 TSMC의 COUPE를 go to market 솔루션의 일부로 만드는 쪽으로 결정적으로 pivot했다.

3D Optical Engine (OE) for Next-Gen Communication



- Optics is crucial for rapid and reliable data transmission and lowering network power consumption for AI
- EIC-on-PIC stacked using the SoIC-X process (COUPE™*) offers unparalleled interconnect density while maintaining optimal system power
- We will enable COUPE in pluggable in '25, followed by COUPE on substrate in a CoWoS CPO with a 2x reduction in power and a 10x reduction in latency in '26
- COUPE on CoWoS interposer for another 5x reduction in power and 2x reduction in latency is being explored



* COUPE: Compact Universal Photonic Engine

** OSFP: Octal Small Form Factor Pluggable

© 2024 TSMC, Ltd

9

TSMC Property

semianalysis

Figure 33: Source: TSMC

Die fabrication: TSMC는 die fabrication을 위한 포괄적인 솔루션 suite를 제공한다. EIC는 high-speed optical modulator driver와 TIA를 통합하는 N7 node에서 제조된다. 또한 wavelength stabilization과 같은 기능을 가능하게 하는 heater controller를 통합한다. 반면 PIC는 SOI N65 node에서 제조되며, TSMC는 photonic circuit design, photonic layout design 및 verification, 그리고 photonic circuit의 simulation 및 modeling (RF, noise 및 multi-wavelength와 같은 측면을 다룸)에 대한 광범위한 지원을 제공한다.

EIC와 PIC는 TSMC-SoIC-bond process를 사용하여 bonding된다. 앞서 언급했듯이 더 긴 trace 길이는 더 많은 parasitic을 의미하며, 이는 성능을 저하시킨다. TSMC의 SoIC는 monolithic 하지 않으면서 가능한 가장 짧은 trace 길이를 제공하는 bumpless interface이므로 EIC와 PIC를 heterogeneously하게 통합하는 가장 성능이 좋은 가능한 방법이다. 아래에 표시된 것처럼 iso-power에서 SoIC 기반 OE는 bump로 통합된 OE보다 23배 이상의 bandwidth density를 제공한다.

TABLE I. COMPARISON OF 3D INTEGRATION OF MULTIPLE WAVELENGTH SYSTEM WITH μ BUMP AND SOIC



		μ bump	SoIC
Pitch		1x	< 0.25x
Bond array per MRM unit		4 x 3	5 x 4
MRM density		1x	> 10.7x
Iso-power	Speed	1x	2.2x
	BWD	1x	> 23x

Figure 34: Source: TSMC

COUPE는 전체 optical engine design 및 integration process를 지원한다. optical I/O의 경우, wafer 또는 chip level에서 micro-lens의 통합을 가능하게 하는 μ Lens design을 지원하며, mirror, μ Lens, grating coupler (GC) 및 reflector를 다루는 optical I/O path simulation을 지원한다. 3D stacking의 경우, 3D floorplanning, SoIC-X/TDV/C4 bump layout implementation, interface physical checking, 그리고 high-frequency channel model extraction 및 simulation을 지원한다. seamless한 개발을 보장하기 위해 회사는 COUPE design 및 verification을 위한 완전한 PDK 및 EDA workflow를 제공하여 designer가 기술을 효율적으로 구현할 수 있도록 한다.

Coupling: 나중에 더 자세히 설명하겠지만 두 가지 주요 coupling 방법이 있다 – grating coupling (GC)과 edge coupling (EC). COUPE는 GC와 EC 모두에 대해 PIC bumpless stacking 구조에 하나의 공통 EIC를 사용한다. 그러나 COUPE-GC 구조는 Silicon lens (Si lens)와 MR (metal reflector)를 독특하게 사용하는 반면, COUPE-EC는 고유하게 EC facet (fiber로 EC를 종료하기 위한)를 가질 것이다. GC의 경우 Si lens는 770 μ m silicon carrier (Si-carrier)에 설계되고 MR은 optical 성능에 필요한 최적화 dielectric layer와 함께 GC 바로 아래에 배치된다. 그런 다음 Si-carrier는 CoW (chip-on-wafer) wafer에 WoW (wafer-on-wafer) bonding된다.

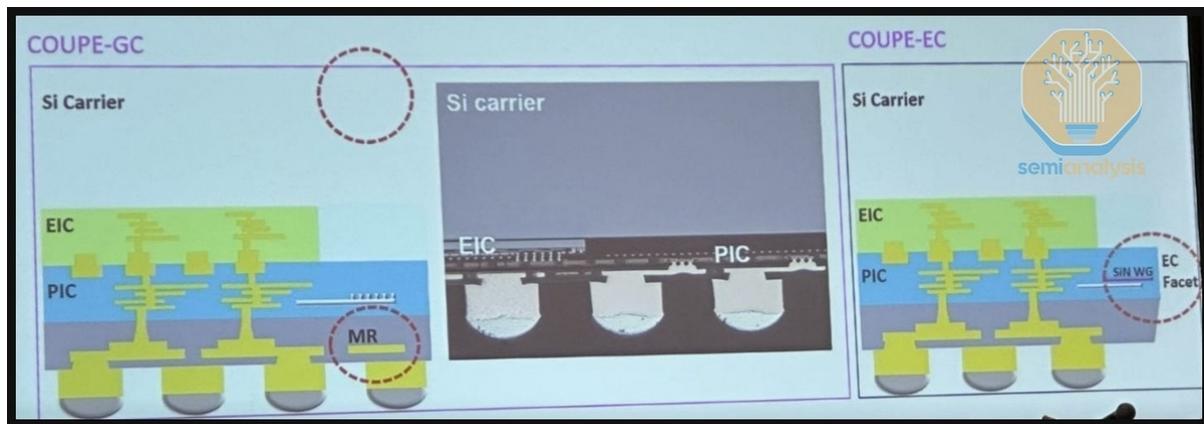


Figure 35: Source: TSMC

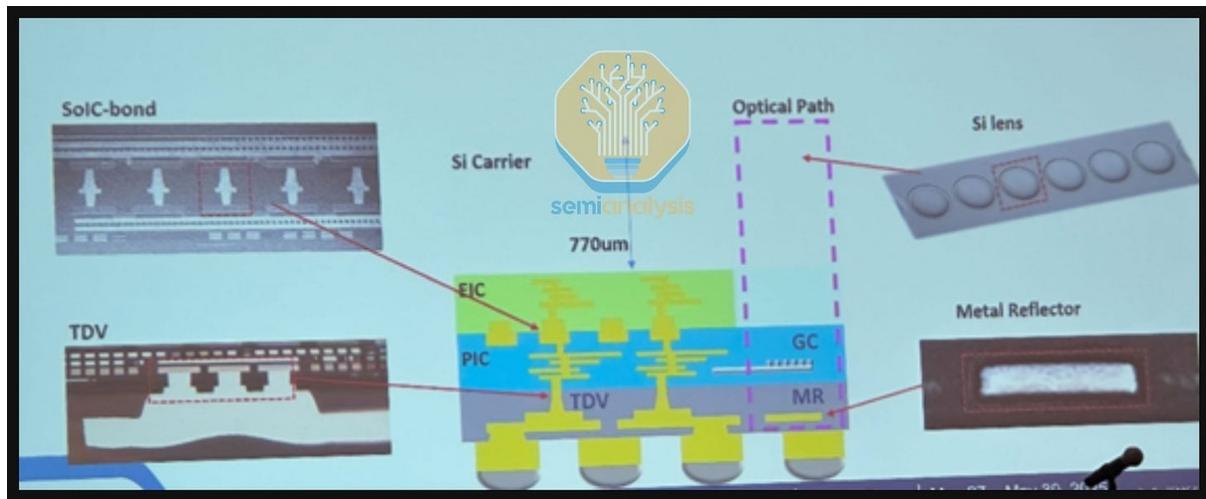


Figure 36: Source: TSMC

Fiber Attach Unit (FAU): FAU는 COUPE의 optical path에 따라 co-design되어야 한다. FAU의 목적은 Si lens에서 optical fiber로 빛을 낮은 insertion loss로 coupling하는 것이다. I/O 수가 증가함에 따라 manufacturing 난이도가 증가하지만, 산업이 특정 표준을 준수하면 개발 시간과 비용이 감소한다. 전반적으로 각 component는 최상의 optical 성능을 달성하기 위해 최적화된 design이 필요하다.

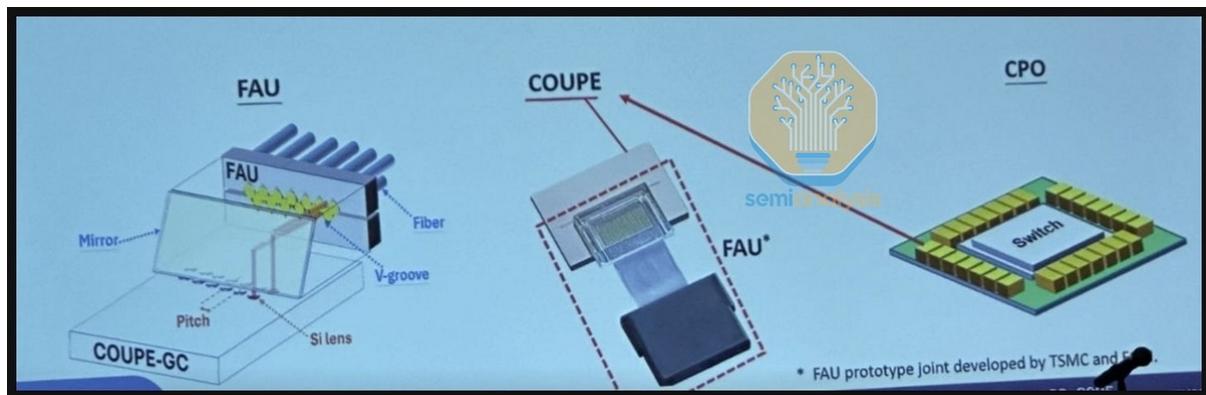


Figure 37: Source: TSMC

Product roadmap: COUPE의 첫 번째 iteration은 substrate에 있는 optical engine이 될 것이며 궁극적인 목표는 interposer에 OE를 배치할 수 있는 것이다. Interposer는 훨씬 더 많은 I/O density를 제공하므로 OE와 ASIC PHY 사이에 더 큰 bandwidth를 가능하게 하며, 개별 OE가 각각 최대 12.8Tbit/s의 bandwidth를 가질 가능성이 있어 약 4 Tbit/s/mm로 환산된다. Interposer를 통합하는 데 있어 과제는 OE를 수용하기 위해 interposer 크기 (package substrate보다 비쌈)를 확장하는 것이다.

이것이 Broadcom이 SPIL에 의해 개발된 **Fan-Out Wafer-Level Packaging (FOWLP)** 접근 방식을 사용하여 여러 세대의 CPO를 iterate했음에도 불구하고 CPO 솔루션을 위해 TSMC COUPE로 전환하고 있는 이유이다. 특히 Broadcom은 미래 switch 및 customer accelerator roadmap을 위해 COUPE에 commit했다. 우리가 이해하기로는 FOWLP 접근 방식은 과도한 parasitic capacitance로 인해 lane당 100G를 넘어 scaling할 수 없는데, electrical signal이 EIC에 도달하기 위해 **Through-Mold Via (TMV)**를 통과해야 하기 때문이다. 경쟁력 있는 roadmap을 유지하기 위해 Broadcom은 우수한 성능과 scalability를 제공하는 COUPE로 전환해야 한다. 이는 TSMC

의 기술적 우위를 강조하며, 역사적으로 더 약하다고 여겨졌던 optics domain에서도 승리를 확보할 수 있게 한다.

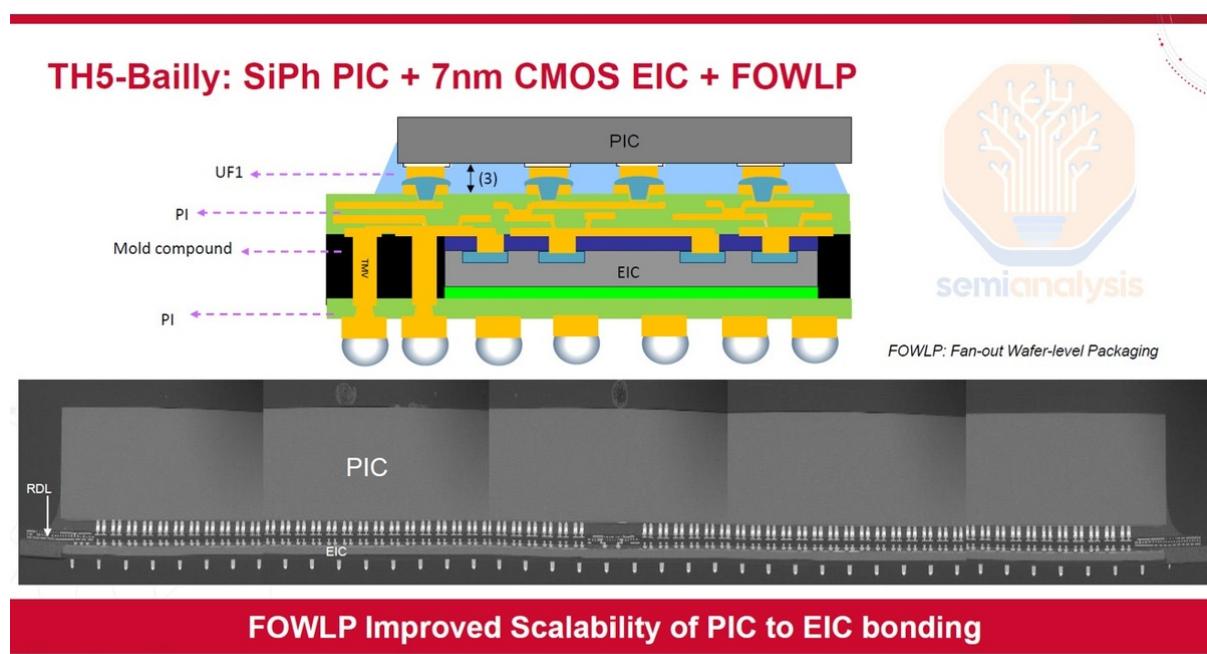


Figure 38: Source: Broadcom

FOWLP Innovation: Dual-Side Attach for Co-Packaged Optics

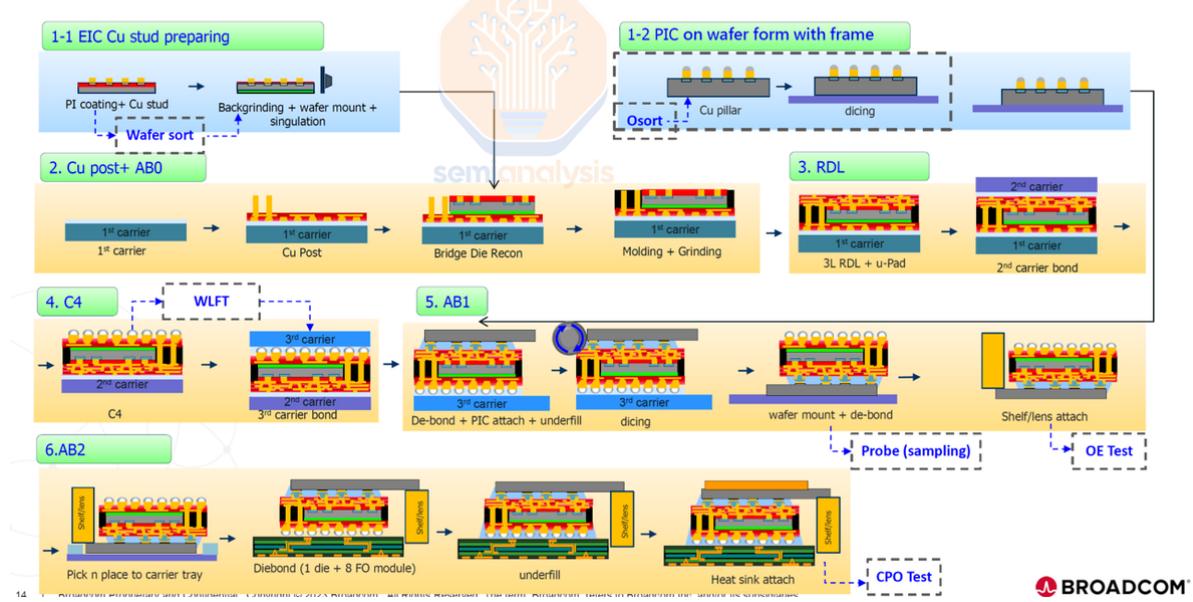


Figure 39: Source: Broadcom

3.4 Packaging the OEs with the host

OE 자체는 substrate에 배치되며, 그 후 substrate는 host package에 flip-chip bonding된다. OE를 co-packaging하기 위해서는 많은 package 면적이 필요하다. 이는 배치되는 위치에 따라 package

substrate 또는 interposer를 상당히 확대해야 함을 필요로 한다. NVIDIA의 Spectrum-X Photonics switch ASIC package의 경우 substrate는 110mm x 110mm를 측정할 것이다. 참고로 이것은 그 자체로 매우 큰 chip인 70mm x 76mm의 Blackwell package와 비교된다.

추가로 substrate에 더 많은 element를 부착하는 것은 yield 과제를 제시한다. 다시 Spectrum-X의 경우 36개의 known good OE가 먼저 substrate에 flip-chip bonding되어야 한다. CoWoS assembly를 완료하기 위한 'on Substrate' 단계를 위해 interposer module을 bonding하기 전에.

마찬가지로 interposer의 경우 훨씬 더 큰 interposer를 제조해야 하는 필요성은 비용이 많이 들며 훨씬 더 많은 element를 bonding해야 하므로 yield 과제를 제시한다. 추가로 이러한 과제는 interposer/substrate 크기가 확장됨에 따라 더욱 두드러지는 warpage 문제로 인해 악화된다.

3.5 FAUs and Fiber Coupling

기술 용어: Fiber Coupling 방식의 이해

Edge Coupling (EC): Fiber를 chip의 edge를 따라 정렬하는 방식. Fiber 끝의 microlens가 빛을 chip으로 집중시키고, waveguide taper가 점진적으로 넓어져 smooth mode transition 을 제공한다. 낮은 coupling loss (<1dB), 광범위한 wavelength 지원, polarization insensitivity의 장점을 가지지만, 1D 구조로 fiber density가 제한되고 die stacking과 호환되지 않는다.

Grating Coupling (GC): 빛이 위에서 들어오고 주기적 grating 구조가 빛을 아래로 scatter하여 waveguide로 유도하는 방식. 2D density 가능, 더 작은 footprint, 간단한 제조, interposer 배치 가능한 장점을 가지지만, 더 높은 optical loss (3-5dB)와 더 좁은 optical bandwidth를 가진다. NVIDIA와 TSMC가 선호하는 방식.

Fiber Pitch: Fiber 간의 간격. 현재 최소 pitch는 127 μm (mm당 최대 8개 fiber)이며, 업계는 80 μm pitch와 multicore fiber를 향해 발전 중이다.

Insertion Loss: Optical signal이 fiber에서 waveguide로 coupling될 때 발생하는 power loss. EC는 일반적으로 <1dB, GC는 3-5dB의 insertion loss를 가진다.

Waveguide Taper: EC에서 사용되는 점진적으로 넓어지는 waveguide 구조. Reflection 과 scattering을 줄여 coupling efficiency를 향상시킨다.

Fiber는 데이터를 전송하기 위해 OE에서 나온다. 하나의 optical lane은 2개의 fiber 또는 하나의 fiber pair (transmit plus receive)로 구성된다. **Fiber coupling** (원활하고 효율적인 빛 전송을 위해 fiber를 on-chip waveguide와 정밀하게 정렬하는 것)은 CPO에서 중요하고 도전적인 단계이며, **Fiber Array Unit (FAU)**는 그 과정을 지원하기 위해 CPO에서 널리 사용된다. 이를 수행하는 두 가지 주요 방법이 있는데, 즉 **Edge Coupling (EC)**과 **Grating Coupling (GC)**이다.

3.5.1 Edge Coupling (EC)

Edge coupling은 fiber를 chip의 edge를 따라 정렬한다. 아래 이미지에서 빛 beam이 edge coupler에 정확하게 들어가도록 보장하기 위해 fiber 끝이 chip의 polished edge와 정밀하게 정렬되어야 함을 볼 수 있다. Fiber 끝의 microlens는 빛을 chip 쪽으로 집중하고 방향을 지정하여 waveguide로의 진입을 이끈다. Waveguide taper는 점진적으로 넓어져 reflection과 scattering을 줄여 coupling efficiency를 보장하는 smooth mode transition을 허용한다. 그러한 lens와 taper가 없으면 fiber facet과 waveguide facet 사이의 interface에서 상당한 optical loss가 발생할 것이다.

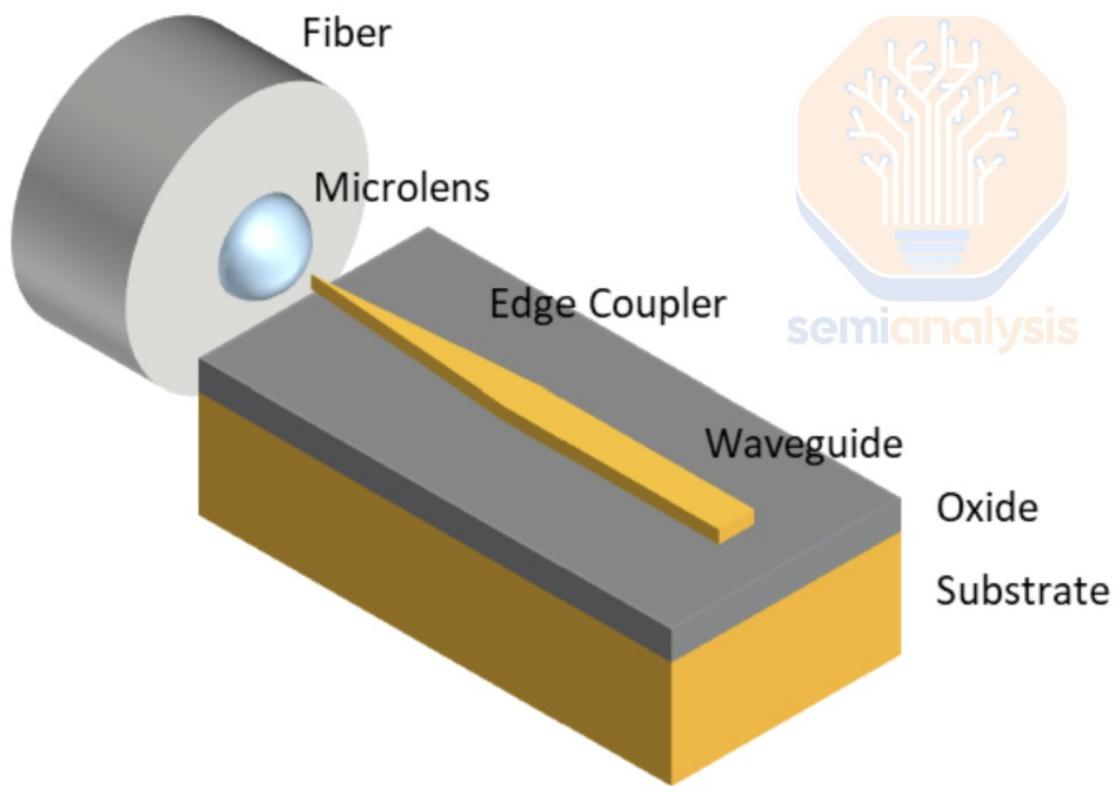


Figure 40: Source: Ansys

Edge coupling은 낮은 coupling loss, 광범위한 wavelength와 작동하는 능력, 그리고 polarization insensitivity로 인해 선호된다. 그러나 다음의 몇 가지 단점도 있다:

1. Fabrication process가 더 복잡하며 undercut과 deep etching이 필요하다;
2. 1D 구조이기 때문에 fiber density가 제한될 수 있다;
3. Die stacking과 호환되지 않는다 (TSV가 thinning을 필요로 하기 때문에);
4. Form factor의 mechanics reliability, mechanical stress, warpage 및 fiber handling에 대한 과제가 있다;
5. 더 낮은 thermal reliability를 제공한다; 그리고
6. 일반적으로 ecosystem compatibility의 부족이 있다.

Global Foundries (GFS)는 올해 VLSI conference에서 signature 45nm Fotonix platform에서 32 channel과 127μm-pitch를 가능하게 하는 monolithically integrated SiN edge coupler를 시연했다.

3.5.2 Grating Coupling (GC)

Grating Coupler (GC)에서 빛은 위에서 들어오고 fiber는 grating 위의 작은 각도로 배치된다. 빛이 grating에 도달하면 주기적 구조가 빛을 아래로 scatter하고 구부려 waveguide로 들어간다.

Grating/vertical coupling의 주요 이점은 여러 행의 fiber를 가질 수 있어 optical engine당 더 많은 fiber를 허용한다는 것이다. GC는 또한 substrate의 바닥에 배치될 필요가 없어 OE를 interposer에 배치하는 것이 가능하다. 마지막으로 GC는 극도의 정밀도로 배치될 필요가 없으며 간단한 two-step etched process로 더 쉽게 제조될 수 있다. GC의 단점은 single-polarization grating coupler가 제한된 범위의 wavelength에서만 작동하고 polarization에 매우 민감하다는 것이다.

NVIDIA는 여러 이점 때문에 GC를 선호했다. 2D density를 가능하게 하고, 더 작은 footprint를 제공하며, 제조가 더 쉽고, EC와 비교하여 더 간단한 wafer-level testing을 허용한다. 그러나 회사는 또한 GC의 여러 단점을 알고 있다. 일반적으로 더 높은 optical loss를 도입하고 EC보다 더 좁은 optical bandwidth를 가진다 (후자는 일반적으로 더 넓은 spectral range를 수용할 수 있다). TSMC도 COUPE platform에서 지원되는 GC에 대한 더 높은 선호도를 명확히 가지고 있다.

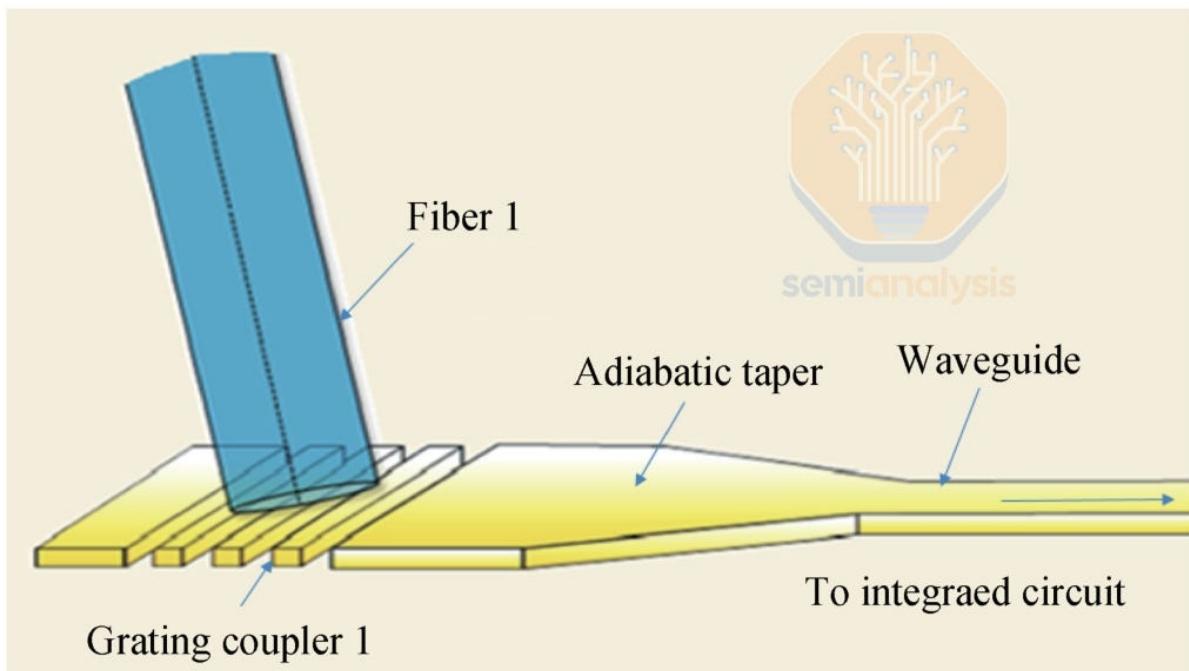


Figure 41: Source: Journal of Semiconductors

3.6 Laser Type and Wavelength Division Multiplexing (WDM)

기술 용어: Laser와 Wavelength Multiplexing

ELS (External Light Source): Laser가 fiber를 통해 optical engine에 연결된 별도의 module에 위치하는 방식. 주로 OSFP와 같은 pluggable form factor 사용. Field servicing이 간단하지만 connector loss, fiber coupling loss, modulator inefficiency로 인해 높은 전력 소비 (laser와 TEC가 전력의 70

CWDM (Coarse Wavelength Division Multiplexing): 일반적으로 20nm spacing으로 더 적은 channel을 전달하는 방식. Wider spacing으로 capacity가 제한됨.

DWDM (Dense Wavelength Division Multiplexing): 매우 tight spacing (종종 <1nm)으로 많은 lane을 pack하는 방식. 40, 80 또는 100+ channel 수용 가능. 제한된 fiber pair를 최대화하는 데 핵심적.

Lambda (λ): 개별 optical lane의 개별 wavelength를 지칭. 각 lambda는 하나의 modulator를 필요로 한다.

DFB (Distributed FeedBack) Laser: Single wavelength를 생성하는 laser type. Lumentum^o single high-power DFB와 DFB array를 제공.

Quantum-dot Mode-locked Comb: 여러 wavelength를 동시에 생성하는 laser 기술. Innolume^o 개발 중.

Pumped Nonlinear Resonant Comb: 여러 wavelength를 생성하는 또 다른 접근법. Xscape, Enlightra, Iloomina가 개발 중.

Laser를 CPO에 통합하는 두 가지 주요 방법이 있다. 첫 번째 접근 방식인 on-chip laser는 일반적으로 III-V (InP) 재료를 silicon에 bonding하여 laser와 modulator를 동일한 photonic chip에 통합한다. On-chip laser는 design을 단순화하고 insertion loss를 줄이지만 몇 가지 과제가 있을 수 있다:

1. Laser는 system 내에서 가장 failure가 발생하기 쉬운 component 중 하나로 알려져 있다 – CPO engine에 통합되면 전체 chip을 다운시킬 것이기 때문에 failure는 높은 blast radius를 가질 것이다;
2. Laser는 또한 열에 민감하며, co-packaged OE에 배치하면 system의 가장 뜨거운 부분인 host silicon에 매우 가까울 것이기 때문에 laser를 높은 열에 노출시켜 문제를 더욱 악화시킬 것이다;
3. On-chip laser는 일반적으로 충분히 높은 power output을 제공하는 데 어려움을 겪는다.

업계가 합의에 도달한 하나의 접근 방식은 **External Light Source (ELS)**를 사용하는 것이다. Laser는 fiber를 통해 optical engine에 연결된 별도의 module에 있다. 종종 이 laser는 OSFP와 같은 pluggable form factor에 있다. 이 설정은 laser failure의 상당히 일반적인 경우에 field servicing을 단순화한다.

ELS의 단점은 더 높은 전력 소비이다. 아래 다이어그램에 표시된 것처럼 ELS 기반 system에서 output power는 connector loss, fiber coupling loss 및 modulator inefficiency와 같은 다양한 요인으로 인해 여러 단계에서 손실된다. 따라서 이 system의 각 laser는 loss를 보상하고 안정적인 전송을 보장하기 위해 24.5 dBm의 optical power를 제공해야 한다. High-output laser는 더 많은 열을 생성하고 thermal stress 하에서 더 빨리 저하되며, laser와 thermo-electric cooler가 ELS 전력 소비의 ~70%를 차지한다. Laser design, packaging 및 optical path의 점진적인 개선이 도움이 되지만 laser의 높은 전력 요구사항 문제는 완전히 해결되지 않았다.

올해 VLSI conference에서 NVIDIA는 ecosystem 내의 여러 laser 파트너를 강조했다. Single high-power DFB를 위한 Lumentum, DFB array를 위한 Ayar Labs, quantum-dot mode-locked comb를 위한 Innolume, 그리고 pumped nonlinear resonant comb를 위한 Xscape, Enlightra 및 Iloomina.

NVIDIA는 또한 잠재적인 대안 laser 솔루션으로 VCSEL array를 탐색하는 것에 대해 논의했다. Fiber당 데이터 rate는 더 낮을 것이고 일부 thermal 문제가 있을 수 있지만, VCSEL은 전력 및 비용 효율성을 제공할 수 있으며 **wide-and-slow** application에 적합할 수 있다. 그렇긴 하지만 우리는 이것이 NVIDIA에게 즉각적인 우선순위라고 보지 않는다.

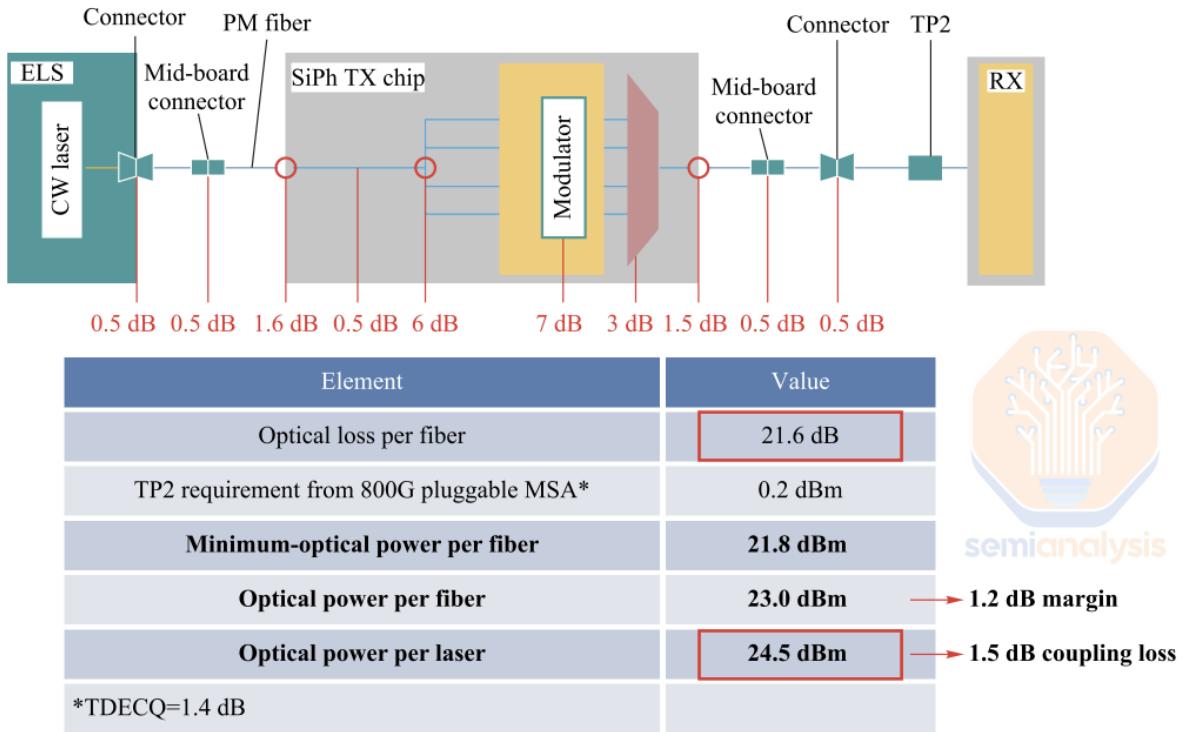


Figure 42: Source: CPO status, challenges, and solutions

Wavelength Division Multiplexing (WDM)은 여러 다른 wavelength 또는 lambda의 빛이 동일한 fiber strand를 통해 전송되는 것이다. WDM의 두 가지 일반적인 variant는 **Coarse Wavelength Division Multiplexing (CWDM)**과 **Dense Wavelength Division Multiplexing (DWDM)**이다. CWDM은 일반적으로 상대적으로 멀리 떨어진 (일반적으로 20 nm spacing) 더 적은 channel을 전달하는 반면, DWDM은 매우 tight spacing (종종 <1nm spacing)으로 많은 lane을 pack한다. CWDM의 wider channel spacing은 capacity를 제한하는 반면, DWDM의 narrower spacing은 40, 80 또는 심지어 100+ channel을 수용할 수 있다. WDM이 중요한 이유는 오늘날 제안된 대부분의 CPO 구현이 optical engine에 부착될 수 있는 fiber 수에 의해 제한되기 때문이다. 제한된 fiber pair는 각 fiber pair가 최대화되어야 함을 의미한다.

3.7 Modulator Types

핵심 인사이트: Modulator 선택 - 성능과 Scalability의 Trade-off

Modulator 선택은 CPO system의 성능, 전력 효율, scalability를 결정하는 가장 중요한 architectural decision 중 하나이다. 세 가지 주요 modulator type은 각각 명확한 trade-off를 가진다:

MZM (Mach-Zehnder Modulator) - "High Performance, Low Density": - 장점: 가장 낮은 thermal sensitivity, 200 Gbaud 달성 가능, PAM4/QAM 등 advanced modulation 지원, 낮은 chirp - 단점: 매우 큰 크기 ($12,000\text{mm}^2$), 높은 전력 소비, 높은 bias voltage 필요 - 채택: Nubis 등 일부 startup이 사용하지만 제한적

MRM (Micro-Ring Modulator) - "High Density, High Complexity": - 장점: 극도로 compact ($25\text{-}225\text{mm}^2$), WDM에 최적 (built-in mux/demux), 매우 낮은 전력, 낮은 driving voltage - 단점: 10-100배 더 temperature-sensitive, non-linear (higher-order

modulation 복잡), 정밀한 control 필요, 표준화 어려움 - 채택: NVIDIA, TSMC, Ayar Labs, Lightmatter, Ranovus가 집중 - 업계 주류 선택

EAM (Electro-Absorption Modulator) - "Thermal Resilience, Limited Ecosystem": - 장점: MRM보다 훨씬 낮은 temperature sensitivity (35°C instant shift 견디), MZM 보다 작고 낮은 전력 (250mm^2) - 단점: GeSi reliability 우려, C-band로 제한 (O-band ecosystem 활용 불가), MRM/MZI보다 높은 insertion loss (4-5dB), 별도 multiplexer 필요 - 채택: Celestial AI가 거의 유일 - XPU 아래 interposer 배치에 thermal tolerance 활용

산업적 합의: NVIDIA와 TSMC의 MRM 선택은 단순한 기술 선호가 아니다. MRM은 구현 난이도가 높지만, TSMC의 advanced CMOS expertise와 결합하여 high-precision, 훌륭한 Q-factor 달성이 가능하다. 이는 결국 **MRM의 potential bandwidth density 우위가 CPO scaling의 핵심이 될 것임을 시사한다.** 반면 EAM의 thermal advantage는 특정 niche application (co-packaged compute chiplet)에서 차별화 요소가 될 수 있다.

Laser가 PIC에 들어가면 electronic signal이 laser의 wavelength로 encoding되는 modulation 단계 (driver에 의해 구동됨)를 거친다. 이 과정에 사용되는 세 가지 주요 modulator type은 **Mach-Zehnder Modulator (MZM)**, **Micro-Ring Modulator (MRM)** 및 **Electro-Absorption Modulator (EAM)**이다. 각 개별 lambda (개별 optical lane의 개별 wavelength)는 하나의 modulator를 필요로 한다.

3.7.1 Mach-Zehnder Modulator (MZM)

MZM은 continuous-wave optical signal을 두 개의 waveguide arm으로 분할하여 데이터를 encoding하며, 이들의 refractive index는 applied voltage에 의해 변화한다. Arm이 재결합될 때 그들의 interference pattern의 signal의 intensity 또는 phase를 modulate한다.

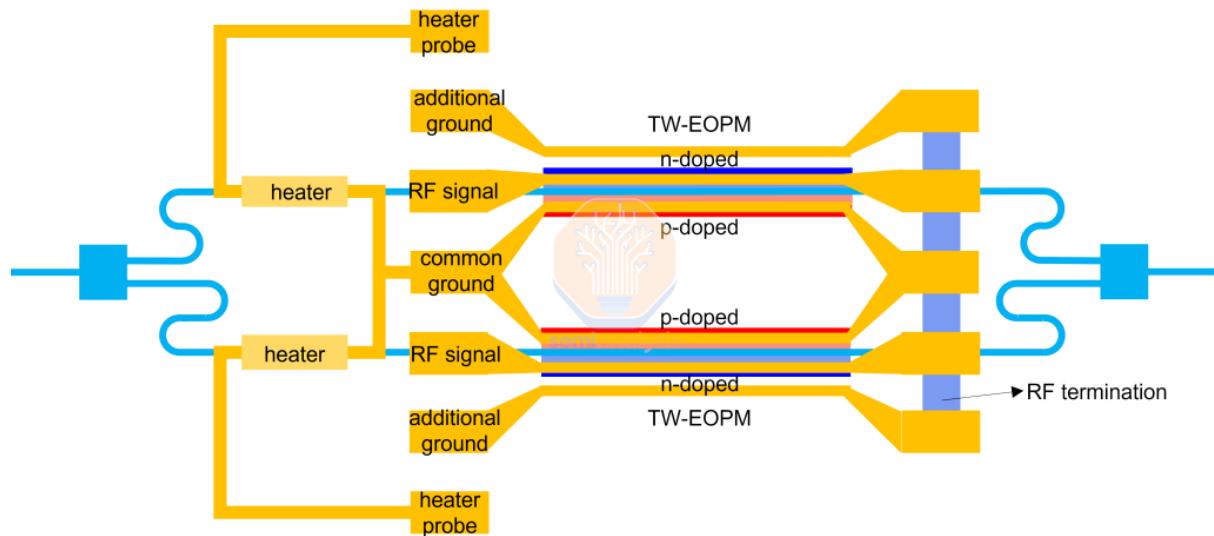


Figure 43: Source: Luceda Academy

MZM은 세 가지 중 구현하기 가장 쉽고 낮은 thermal sensitivity를 가지고 있어 정밀한 temperature control의 필요성을 줄인다. 그들의 높은 linearity는 PAM4 및 coherent QAM (QAM은 HPC/AI workload에 적합하지 않지만)과 같은 advanced modulation format을 지원한다. MZM의 낮은 chirp는 higher-order modulation 및 long-distance transmission을 위한 signal integrity

를 향상시킨다. MZM은 또한 channel당 더 높은 bandwidth를 가능하게 한다: lane당 200G가 작동하는 것으로 입증되었으며, non-coherent PAM modulation으로 lane당 400G가 가능할 것으로 믿어진다.

그러나 MZM의 단점은 다음과 같다:

- Millimeter scale로 측정되는 dimension을 가진 큰 form factor (micron scale의 MRM과 비교), 두 개의 waveguide arm과 combining region이 필요하기 때문에 더 많은 chip 면적을 소비하고 OE PIC에 포함된 modulator (따라서 channel)의 density를 제한한다. MZM 크기는 ~12,000mm² 정도이고, EAM은 약 250mm² (5x50mm)이며 MRM은 25mm²에서 225mm² 사이 (diameter 5-15mm²)이다. 이것은 scaling을 제한할 수 있는 MZM의 한 가지 중요한 단점이다. 그러나 modulator 주변의 driver 및 optical/electrical control circuitry를 포함한 전체 PIC/EIC 조합의 크기를 고려하면 MZM의 크기 단점은 덜 두드러질 수 있다,
- 높은 전력 소비, phase-shifting process가 상당한 에너지를 요구하기 때문이다. 또한 sub-voltage에서 작동하는 MRM보다 더 높은 bias condition (기본적으로 initiating voltage)을 가진다. 그러나 Nubis와 같은 기업은 MZM의 전력 단점을 개선하기 위해 영리한 design을 개발하려고 노력하고 있다.

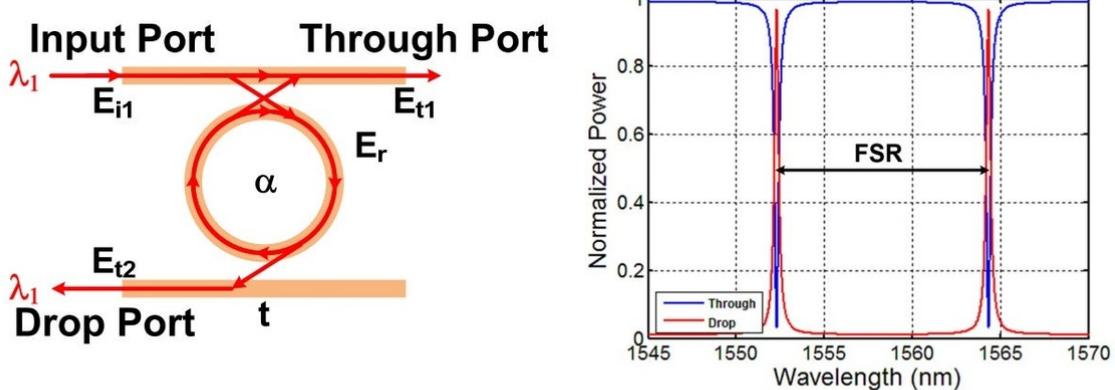
startup ecosystem에서 Nubis는 scale-up CPO 솔루션에 주로 MZM을 활용하는 기업 중 하나이다. MZM은 큰 form factor와 제한된 수의 lambda로 인해 startup ecosystem에서 널리 선택되지 않는다.

3.7.2 Micro-Ring Modulators (MRMs)

MRM은 하나 이상의 straight waveguide에 결합된 compact ring waveguide를 사용한다. Electrical signal은 ring의 refractive index를 변경하여 resonant wavelength를 이동시킨다. Resonance를 input light와 align하거나 misalign하도록 tuning함으로써 MRM은 optical signal의 intensity 또는 phase를 modulate하여 데이터를 encoding한다.

Light source는 input port에서 ring으로 전달된다. 대부분의 wavelength의 빛에 대해 ring에 resonance가 없어 빛이 device를 통과하여 input port에서 through port로 전달된다. Wavelength 가 resonance condition을 만족하면 빛은 ring에서 constructively하게 interfere하고 대신 drop port로 당겨진다. 아래 normalized power graph에 표시된 것처럼 특정 wavelength의 빛은 drop port에서 transmission power의 sharp peak를 야기하고 through port에서 transmission의 해당 drop 을 야기한다. 이 효과는 modulation에 사용될 수 있다.

Ring Resonator Filter



- Ring resonators display a high- Q notch filter response at the through port and a band-pass response at the drop port
- This response repeats over a free spectral range (FSR)

7

Figure 44: Source: Sam Palermo, Texas A&M University

Optical Engine은 일반적으로 여러 MRM을 사용하며, 이러한 각 ring은 다른 wavelength로 tuning될 수 있어 WDM을 달성하기 위해 추가 device set를 필요로 하는 것과 반대로 ring 자체를 사용하여 **Wavelength Division Multiplexing (WDM)**을 가능하게 한다. MRM은 다음의 몇 가지 주요 이점을 가지고 있다:

- 극도로 compact하며 (수십 micron scale), MZM보다 훨씬 더 높은 modulator density를 허용한다. MZM 크기는 ~12,000mm² 정도이고, EAM은 약 250mm² (5x50mm)이며 MRM은 25mm²에서 225mm² 사이 (diameter 5-15mm²)이다.
- Ring은 WDM application (8 또는 16 wavelength를 가진 DWDM 포함)에 매우 적합하며, built-in mux/demux 기능을 가지고 있다.
- MRM은 매우 에너지 효율적일 수 있다 (bit당 더 낮은 전력).
- 마지막으로 ring은 낮은 chirp를 가지며, 이는 signal 품질을 향상시킨다.

그러나 MRM은 또한 아래의 몇 가지 과제를 동반한다:

- MRM은 MZM 및 EAM보다 10-100배 더 temperature-sensitive할 수 있으며, design 및 manufacturing이 어려운 매우 정밀한 control system을 필요로 한다.
- 그들은 non-linear하여 PAM4/6/8과 같은 higher-order modulation을 복잡하게 만든다.

- MRM의 sensitivity와 tight temperature control tolerance는 각 design이 정밀한 요구사항을 가지기 때문에 표준화를 어렵게 만들 수 있다.

솔루션 제공업체 중에서 NVIDIA는 MRM에 대한 명확한 선호를 가지고 있다. 그들은 CPO system에서 MRM을 design하고 넣은 최초라고 주장한다. 회사는 MRM의 핵심 이점이 전력 소비를 줄이는 데 도움이 되는 compact size와 낮은 driving voltage라고 믿는다. 그러나 MRM 기술은 또한 control하기 어려운 것으로 알려져 있어 성공적인 구현을 위해 design precision이 중요하다. 이것은 실제로 NVIDIA의 강점이다.

Fabrication 측면에서 TSMC의 advanced CMOS expertise는 high-precision과 훌륭한 Q-factor로 MRM을 fabrication하는 데 매우 적합하다. 추가로 Tower도 photonics node에 강력한 fabrication capability를 가져온다. MRM은 구현하기 어렵지만 확실히 실행 가능하다. 그들은 잠재적으로 MZM보다 더 높은 bandwidth density를 가능하게 할 수 있다. 그래서 TSMC, NVIDIA 및 Ayar Labs, Lightmatter 및 Ranovus와 같은 많은 CPO 기업이 이 기술 roadmap에 집중한다.

3.7.3 Electro-Absorption Modulators (EAM)

EAM은 signal을 modulate하여 applied voltage에 기반하여 빛을 흡수하는 능력을 변경한다. 더 구체적으로 EAM에 낮거나 voltage가 applied되지 않으면 device는 들어오는 laser 빛의 대부분을 통과시켜 transparent하거나 open하게 보이게 한다. 더 높은 voltage가 applied되면 GeSi modulator의 band gap이 high C-band 범위 (1500nm 이상)를 커버하도록 이동하여 해당 wavelength에 대한 absorption coefficient를 증가시키고 nearby waveguide를 통과하는 optical signal을 attenuating closing한다. 이것은 Franz-Keldysh effect로 알려져 있다. 이 open과 close state 사이의 switching은 빛의 intensity를 modulate하여 optical signal에 데이터를 효과적으로 encoding한다.

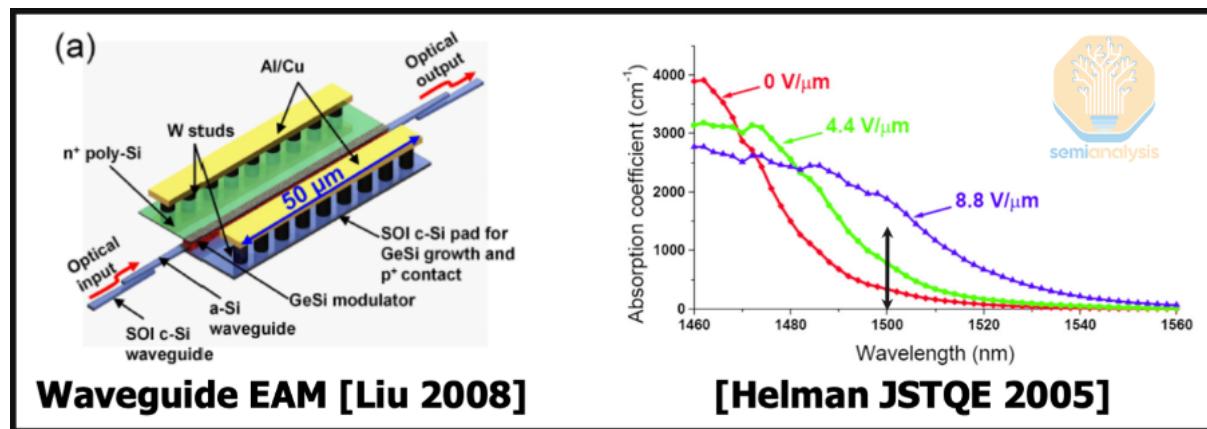


Figure 45: Source: Texas A&M University, Liu 2008, Helman 2005

동일한 원리가 오늘날 modulation을 위해 **Electro-Absorption Modulated Laser (EML)**를 사용하는 transceiver에서 사용된다. **Continuous Wave (CW)**, **Distributed FeedBack (DFB)** laser와 InP 기반 EAM이 함께 결합되어 하나의 lane을 modulate할 수 있는 단일 discrete **Electro-Absorption Modulated Laser (EML)**을 구축한다. 예를 들어 800G DR8 transceiver는 8개의 개별 fiber lane에 걸쳐 8개의 EML을 사용하며, 각각 PAM4 modulation (2 bits/signal)을 사용하고 ~56 GBaud에서 signaling한다. GeSi 기반 modulator와 달리 InP modulator의 band gap은 모든 Datacom DR optics에서 사용되는 표준 wavelength인 O-band (1310nm)에 해당하여 높은 수준의 상호운용성을 허용한다. InP modulator는 CPO에서 사용하기에 이상적이지 않은 몇 가지 단점을 가지고 있다. InP wafer는 작은 경향이 있으며 (3" 또는 6") 낮은 yield에 시달린다.

두 요인 모두 8" 또는 12" process에서 구축될 수 있는 Silicon과 비교할 때 InP 기반 device의 unit cost를 증가시킨다. InP를 Silicon에 coupling하는 것도 GeSi를 다른 silicon device에 coupling하는 것보다 훨씬 더 어렵다.

EAM은 MRM 및 MZI와 비교하여 여러 이점을 가지고 있다:

- 분명히 EAM과 MRM 모두 temperature variation에 대해 둘 다를 안정화시키기 위해 작동하는 control logic과 heater를 가지고 있지만, EAM은 근본적으로 temperature에 대한 sensitivity가 더 낮다. MRM과 비교하여 EAM은 50°C 이상에서 훨씬 더 나은 thermal stability를 가지는 반면 MRM은 temperature에 매우 민감하다. MRM의 일반적인 stability인 70-90 pm/C는 2°C variation으로 resonance를 0.14nm만큼 이동시킨다는 것을 의미하며, 이는 MRM 성능이 붕괴되는 0.1nm resonance shift를 훨씬 넘는다.

이와 대조적으로 EAM은 최대 35°C의 순간 temperature shift를 견딜 수 있다. 이 tolerance는 특히 Celestial AI의 접근 방식에서 중요한데, 그들의 EAM modulator가 수백 watt의 전력을 dissipate하는 high-XPU power compute engine 아래의 interposer 내에 위치하기 때문이다. EAM은 또한 약 80°C의 높은 ambient temperature 범위를 견딜 수 있으며, 이는 XPU 옆에 있고 그 아래에 있지 않은 chiplet application에 적용될 수 있다.

- MZI와 비교하여 EAM은 크기가 훨씬 작고 전력을 덜 소비하는데, MZI의 상대적으로 큰 크기가 high voltage swing을 필요로 하여 0-5V의 swing을 달성하기 위해 SerDes를 증폭하기 때문이다. Mach Zender Modulator (MZM)는 ~12,000mm² 정도이고, EAM은 약 250mm² (5x50mm)이며 MRM은 25mm²에서 225mm² 사이 (diameter 5-15mm²)이다. MZI는 또한 그러한 큰 device를 desired bias로 유지하는 데 필요한 heater에 대해 더 많은 전력 사용을 필요로 한다.

반면에 CPO를 위해 GeSi EAM을 사용하는 데 몇 가지 단점이 있다:

- MRM 및 MZI와 같은 Silicon 또는 Silicon Nitride에 구축된 물리적 modulator 구조는 GeSi 기반 device보다 훨씬 더 큰 endurance와 reliability를 가진 것으로 인식되어 왔다. 실제로 많은 이들이 Germanium 기반 device를 다루고 통합하는 어려움을 고려할 때 GeSi 기반 device의 reliability에 대해 걱정하지만, Celestial은 본질적으로 Photodetector의 역인 GeSi 기반 EAM은 오늘날 transceiver에서 photodetector의 ubiquity를 고려할 때 reliability와 관련하여 알려진 양이라고 주장한다.
- GeSi modulator의 band edge는 자연적으로 C-band (즉, 1530nm-1565nm)에 있다. 이것을 O-band (즉, 1260-1360nm)로 이동시키기 위해 quantum well을 design하는 것은 매우 어려운 engineering 문제이다. 이는 GeSi 기반 EAM이 book-ended CPO system의 일부를 형성할 가능성이 높으며 open chiplet 기반 ecosystem에 참여하는 데 쉽게 사용될 수 없음을 의미한다.
- C-band laser source 주변의 laser ecosystem을 구축하는 것은 O-band CW laser source 주변의 잘 개발된 ecosystem을 사용하는 것과 비교할 때 scale의 diseconomy를 가질 수 있다. 대부분의 datacom laser는 O-band를 위해 구축되지만 Celestial은 상당한 volume의 1577nm XGS-PON laser가 제조된다고 지적한다. 이것들은 일반적으로 consumer fiber to the home 및 business connectivity application에 사용된다.
- SiGe EAM은 MRM과 MZI 모두에 대한 3-5dB와 비교하여 약 4-5dB의 insertion loss를 가진다. MRM은 다른 wavelength를 직접 multiplex하는 데 사용될 수 있는 반면, EAM은 CWDM

또는 DWDM을 구현하기 위해 별도의 multiplexer를 필요로 하여 잠재적 loss budget에 약간 추가한다.

전반적으로 EAM은 현재 CPO 구현에서 널리 사용되지 않으며, Celestial AI가 이 접근 방식을 적극적으로 추구하는 몇 안 되는 기업 중 하나로 두드러진다.

3.8 OE roadmap – scaling OEs

오늘날 사용 가능한 Optical Engine은 일반적으로 1.6T에서 3.2T 사이의 aggregate bandwidth를 제공한다. NVIDIA의 Quantum CPO는 1.6T engine을 포함하며, Spectrum용 3.2T 버전이 계획되어 있다. Broadcom은 Bailly용 6.4T OE를 선보였지만 form factor가 매우 크고 (NVIDIA의 2-3배 너비) 두 개의 FAU를 필요로 하므로 bandwidth density는 NVIDIA의 제공과 유사할 수 있다. Marvell의 6.4T OE도 동일한 경우로 2개의 FAU를 필요로 하므로 큰 footprint를 차지한다. Marvell의 OE는 우리가 아는 한 곧 production system에 들어가지도 않는다.

Optical Engine Specifications					
Name	Company	Year	Total number of lambdas	Lambda Speed - unidirectional (Gbps)	Bandwidth - unidirectional (Tbps)
Odin 8	Ranovus	2021	8	50	0.4
Odin 32	Ranovus	2022	32	50	1.6
Tomahawk 3.2T OE	Broadcom	2022	32	100	3.2
XT1600	Nubis	2023	16	100	1.6
XT3200	Nubis	2025	16	200	3.2
TeraPHY (1st Gen)	Ayar Labs	2023	64	32	2.0
OCI chiplet	Intel	2024	64	32	2.0
Bailly 6.4T OE	Broadcom	2024	64	100	6.4
3D SiPho Engine	Marvell	2024	32	200	6.4
TeraPHY (2nd Gen)	Ayar Labs	2025	64	64	4.0
Nvidia 1.6T OE	Nvidia	2025	8	200	1.6
Nvidia 3.2T OE	Nvidia	2026	16	200	3.2

Figure 46: Source: SemiAnalysis

논의한 바와 같이 NVIDIA의 Spectrum-X photonics switch의 3.2T OE 구현은 long reach SerDes에 의해 구동되는 pluggable보다 더 많은 shoreline bandwidth density를 제공하지 않는다. 다시 말해 optical engine density는 설득력 있는 성능 이점을 제공하고 고객 채택을 추진하기 위해 여러 번 scale해야 한다. 이는 host silicon과 OE EIC 사이의 electrical interface를 scaling하는 것뿐만 아니라 fiber에서 나오는 bandwidth 양을 scaling하는 것을 의미한다.

하지만 만약 우리가 free hand로 차세대 interconnect를 design할 수 있다면 이 세대와 그 이후를 위해 더 큰 bandwidth를 가능하게 하는 접근 방식은 무엇일까?

3.9 Key approaches for scaling bandwidth

기술 용어: Bandwidth Scaling의 핵심 Parameter

Baud Rate: 초당 전송되는 symbol 수. 현재 advanced system은 100 Gbaud에서 작동하며, 업계는 200 Gbaud를 추진 중. Higher baud rate는 modulator가 더 높은 frequency에서 switch하도록 요구.

Modulation Scheme: - NRZ (Non-Return-to-Zero): Symbol당 1 bit 전송 - PAM4 (4-level Pulse Amplitude Modulation): 4개의 다른 amplitude로 symbol당 2 bit 전송

- **PAM6**: Symbol당 2.6 bits 전송 (연구 단계) - **PAM8**: Symbol당 3 bits 전송 (연구 단계)
- **DP-16QAM**: 두 orthogonal plane, 각각 4 amplitude \times 4 phase = 총 256 가능한 signal (signal당 8 bit)
- Lane Speed 계산**: Baud Rate \times Bits per Symbol = Lane Speed - 예: 100 Gbaud \times 2 bits (PAM4) = 200 Gbit/s per lane
- Fast and Narrow**: FAU당 더 적은 수의 fiber (높은 두 자릿수)와 각 fiber pair에서 빠른 link. Higher baud rate와 advanced modulation에 의존.
- Slow and Wide**: 훨씬 더 많은 수의 fiber pair (finer pitch)와 개별 fiber pair당 느린 bandwidth. Fiber density와 WDM에 의존.
- Fiber Pitch**: Fiber 간격. 현재 최소 127μm (mm당 8 fiber), 목표는 80μm pitch와 multicore fiber.

Co-packaged optical engine에서 bandwidth를 scaling하기 위한 주요 접근 방식을 논의해보자:

1. Electrical SerDes 기반 PHY를 계속 사용: long reach SerDes 대신 **Short Reach (SR)** SerDes를 사용하여 더 간단한 design 구현, 감소된 면적 및 더 낮은 전력을 활용한다. 이것은 결국 runway가 부족한 electrical interface에서 SerDes speed에 의해 여전히 제한될 것이다. 여기서 아이디어는 silicon designer가 I/O를 재설계할 필요가 없도록 interim 솔루션을 사용하는 것이다. 추가로 electrical SerDes를 사용하면 동일한 silicon으로 기존 pluggable optics 및/또는 copper를 사용할 수 있는 유연성을 제공한다.
2. NRZ modulation과 함께 56G와 같은 더 낮은 Baud rate에서 작동하는 PCIe와 같은 wide I/O PHY를 사용한다. 이것은 optical engine의 EIC에 덜 요구되며 낮은 속도에서 parasitic이 덜 문제가 되므로 [expensive hybrid bonding](#)의 필요성을 제거할 수도 있다. 그러나 낮은 signal rate를 사용하면 optical engine을 떠나는 fiber의 수가 더 빨리 bottleneck이 될 수 있다. wavelength-division multiplexing은 각 fiber가 parallel로 여러 data stream을 전달할 수 있게 함으로써 이를 해결하는 데 도움이 된다.
3. PCIe와 같은 wide I/O PHY를 사용한 다음 EIC가 더 적은 수의 optical fiber lane으로 serialize하도록 한다. 각 optical lane의 속도를 최대화하기 위해 PAM4 modulation과 함께 high Baud rate를 계속 사용하고, WDM scheme을 사용하여 필요한 경우 여러 wavelength를 추가하여 bandwidth를 더욱 증가시키기 위해 fiber pair당 여러 lambda를 허용한다.

Electrical 측면이 해결되면 다음 과제는 fiber를 통해 얼마나 많은 escape bandwidth를 전달할 수 있는가이다. 전체 fiber bandwidth는 세 가지 주요 요소에 따라 달라진다: 1) fiber의 수 (optical lane을 정의함) 2) lane당 속도, 그리고 3) fiber당 wavelength 수 - 각각은 scaling을 위한 vector를 나타낸다.

최근 업계는 개념을 두 가지 주요 접근 방식으로 나누었다: **Fast and Narrow vs Slow and Wide**. **Fast and Narrow**는 FAU당 더 적은 수의 fiber (기껏해야 높은 두 자릿수)와 각 fiber pair에서 빠른 link를 구상하는 반면, **Slow and Wide**는 훨씬 더 많은 수의 fiber pair (훨씬 더 finer pitch를 가진)와 개별 fiber pair당 훨씬 더 느린 bandwidth의 아이디어에 기반한다.

1. **More fiber pairs**: Fiber density는 fiber pitch에 의해 제한되며 단일 FAU 내의 전체 fiber 수는 yield가 문제가 되기 전에 manufacturing 가능한 것에 의해 제한된다. 현재 fiber의 최소 pitch는 127 micron (μm)이며, mm당 최대 8개의 fiber를 의미한다. 업계는 일정 면적이 수용 할 수 있는 fiber 수를 더욱 scaling하기 위해 80- μm pitch와 multicore fiber를 향해 작업하고 있다. 그러나 더 많은 fiber를 부착하는 것은 manufacturability 과제를 가져온다:

A) 여전히 많은 manual process를 포함하는 fiber를 정렬하는 것은 yield loss에 취약하며, FAU yield는 정렬해야 하는 각 successive fiber와 함께 저하된다; Ficontec과 같은 회사가 제공하는 automation tool이 있지만 여전히 낮은 throughput에 시달린다,

B) Coupling 선택도 중요하다: Edge coupling은 fiber array를 단일 row로 제한하는 반면, grating coupling은 여러 row를 지원할 수 있다. 현재 우리가 본 가장 sizeable한 fiber array는 36 fiber를 가진 Nubis의 2D FAU이다.

2. **Speed per lane:** Lane speed에 영향을 미칠 수 있는 두 가지 dimension이 있다:

A) **Baud rate:** 초당 전송되는 symbol 수를 정의한다; 오늘날의 advanced system은 100 Gbaud에서 작동하는 반면, 업계는 200 Gbaud를 추진하고 있다. 그러나 더 높은 baud rate는 modulator가 더 높은 frequency에서 switch하도록 더 높은 요구사항을 부과한다; 다양한 type 중에서 MZM은 이 metric에서 가장 capable하며 200 Gbaud를 달성하는 상대적으로 명확한 경로를 가지고 있다.

B) **Modulation:** Symbol당 전달되는 bit 수를 정의한다. NRZ (symbol당 1 bit) 및 PAM4 (4개의 다른 amplitude를 통해 symbol당 2 bit)는 오늘날 널리 채택되고 있다. 연구는 PAM6 (~2.6 bits per symbol) 및 PAM8 (symbol당 3 bit)로 확장되고 있다. higher order modulation scheme은 여러 amplitude level 외에도 빛의 다른 phase를 사용하여 signaling함으로써 접근 할 수 있다. DP-16QAM은 각각 4개의 다른 amplitude, 4개의 다른 phase를 가진 두 개의 orthogonal plane을 가능하게 하여 총 256개의 가능한 signal – signal당 8 bit를 전달한다.

3. **Wavelength Division Multiplexing (WDM):** Optical fiber는 여러 wavelength의 빛을 동시에 전달할 수 있다. 예를 들어 각각 200Gbit/s로 데이터를 전달하는 8 wavelength를 가진 fiber는 1.6 Tbit/s의 aggregate capacity를 전송할 수 있다. 오늘날 상업적으로 사용 가능한 DWDM 솔루션은 일반적으로 8-lambda 또는 16-lambda configuration을 제공한다. 연구자들은 또한 lambda 수를 증가시키기 위해 broad-spectrum, band multiplexing 및 interlaying 기술을 탐색하고 있다. wavelength 수를 scaling하는 데 있어 한 가지 주요 과제는 여러 lane의 빛을 안정적이고 효율적으로 생성할 수 있는 reliable laser source를 개발하는 것이다. Ayar Labs의 Supernova light source는 16 wavelength가 가능한 laser를 가지고 있다 (laser는 Sivers가 공급). Scintil의 wafer-scale InP laser는 마찬가지로 최대 16 wavelength를 제공하며, Xscape Photonics는 최대 64 wavelength를 가진 tuneable comb laser를 개발하기 위해 노력하고 있다. modulator 중에서 MRM은 여러 wavelength를 처리하는 데 가장 적합하며 built in multiplexing (mux) 및 demultiplexing (demux) 기능을 가지고 있다.

아래 표는 optical engine을 12.8T 이상으로 scaling하기 위한 여러 접근 방식을 설명한다.

Approaches to Scaling Optical Engines to 12.8T and Beyond																		
Description	Fiber pitch	# fibers / mm per row	# fibers per row	Implied FAU edge (mm)	Number of rows	Total fibers in FAU	# Optical lanes	# laser fibers	Modulation scheme	Lane Speed (Gbps) unidirectional	# Wavelengths /lambda per channel	Bandwidth per optical lane (Gbps)	Total # modulators i.e. lambdas	Bandwidth density (Gbps/mm)	Bandwidth density (Tops/mm)	Total OE BW (Tbps)		
		(a)		(b)	(c) = (a) x (b)	(d)	(e) = (d) / 4			(f)	(g)	(h) = (f)x(g)		(i) = (d) x (g)		(j) = (d) x (h)		
Micro Ring Modulator																		
Scaling fibers (50G MRM)	0.127	7.9	9	1.1	1	9	4	1	NRZ	50	1	50	4	200	0.2	0.2		
	0.127	7.9	18	2.3	2	36	16	4		50	1	50	16	800	0.8	0.8		
	0.127	7.9	27	3.4	3	81	36	9		50	1	50	36	1,800	1.8	1.8		
	0.127	7.9	36	4.6	4	144	64	16		50	1	50	64	3,200	3.2	3.2		
Scaling fibers (200G MRM)	0.127	7.9	9	1.1	1	9	4	1	PAM4	200	1	200	4	800	0.8	0.8		
	0.127	7.9	9	1.1	2	18	8	2		200	1	200	8	1,600	1.6	1.6		
	0.127	7.9	18	2.3	1	18	8	2		200	1	200	8	1,600	1.6	1.6		
	0.127	7.9	18	2.3	1	18	8	2		50	4	200	32	1,600	1.6	1.6		
Scaling colors (50G MRM)	0.127	7.9	18	2.3	1	18	8	2	NRZ	50	8	400	64	3,200	3.2	3.2		
	0.127	7.9	18	2.3	1	18	8	2		50	8	400	64	3,200	3.2	3.2		
	0.127	7.9	18	2.3	1	18	8	2		50	16	800	128	6,400	6.4	6.4		
	0.127	7.9	18	2.3	1	18	8	2		200	4	800	32	6,400	6.4	6.4		
Scaling colors (200G MRM)	0.127	7.9	18	2.3	1	18	8	2	PAM4	200	8	1,600	64	12,800	12.8	12.8		
	0.127	7.9	18	2.3	1	18	8	2		50	4	200	64	3,200	3.2	3.2		
	0.127	7.9	18	2.3	2	36	16	4		50	8	400	128	6,400	6.4	6.4		
	0.127	7.9	18	2.3	2	36	16	4		50	16	800	256	12,800	12.8	12.8		
Scaling colors and fibers (50G MRM)	0.127	7.9	18	2.3	2	36	16	4	PAM4	200	4	800	64	12,800	12.8	12.8		
	0.127	7.9	18	2.3	2	36	16	4		200	8	1,600	128	25,600	25.6	25.6		
	0.127	7.9	18	2.3	2	36	16	4		200	1	200	36	3,500	3.5	7.1		
	0.127	7.9	16	2.0	5	80	36	9		400	1	400	5	1,400	1.4	2.1		
Scaling fibers (200G MZM)	0.127	7.9	9	1.1	1	9	4	1	PAM4	200	1	200	4	700	0.7	0.8		
	0.127	7.9	12	1.5	2	24	11	3		200	1	200	11	1,400	1.4	2.1		
	0.127	7.9	12	1.5	3	36	16	4		200	1	200	16	2,100	2.1	3.2		
	0.127	7.9	16	2.0	5	80	36	9		200	1	200	36	3,500	3.5	7.1		
Scaling fibers (400G MZM)	0.127	7.9	12	1.5	1	12	5	1	PAM4	400	1	400	5	1,400	1.4	2.1		
	0.127	7.9	12	1.5	2	24	11	3		400	1	400	11	2,800	2.8	4.3		
	0.127	7.9	16	2.0	3	48	21	5		400	1	400	21	4,199	4.2	8.5		
	0.127	7.9	16	2.0	5	80	36	9		400	1	400	36	6,999	7.0	14.2		
	0.127	7.9	16	2.0	8	128	57	14		400	1	400	57	11,199	11.2	22.8		

Figure 47: Source: SemiAnalysis

3.10 CPO adoption pace and deployment challenges

NVIDIA의 첫 번째 CPO 제품은 2H 2025에 사용 가능한 InfiniBand CPO와 H2 2026에 사용 가능한 Ethernet CPO switch를 갖춘 backend scale-out switch 용이 될 것이다. 우리는 이 초기 단계가 주로 market test와 supply chain 성숙도를 높이기 위한 preparatory phase 역할을 할 것이라고 생각한다. 우리는 2026년 total shipment volume이 10-15k unit 범위가 될 것으로 예상한다.

배포가 더 나아가고 더 빠르게 진행되어 진정으로 ubiquitous하게 되려면 CPO 채택에 대한 더 설득력 있는 rationale이 있어야 할 것이다. 두 가지 가능성은 CPO 채택으로 인한 극적인 total cost of ownership 이점이거나, switch ASIC에서 switch box의 front panel로 signal을 구동하는 데 필요한 electrical LR SerDes가 hard speed 또는 range wall에 부딪히는 경우일 수 있다. TCO 이점을 상쇄하는 것은 datacenter operator가 CPO 기반 system 배포에 대해 싫어하는 두 가지 주요 사항이다: interoperability의 부족과 serviceability 과제.

CPO의 과제는 package 외부와 전체 system으로 확장된다. Fiber 관리, front plate density, external laser는 모두 과제인 필수 부분이다. CPO를 가능하게 하려면 chip 회사가 고객이 배포할 수 있는 end-to-end 솔루션을 제공해야 한다. 이것은 특히 성능을 확장하기 위한 system design에 집중하고 있는 NVIDIA에서 우리가 보고 있는 추세의 연속이다.

3.11 Proprietary solutions vs standards

핵심 인사이트: CPO의 근본적 딜레마 - Interoperability vs Innovation Speed

CPO는 pluggable optics의 성공 model과 근본적으로 다른 paradigm을 요구하며, 이는 업계에 심각한 딜레마를 제시한다.

Pluggable의 성공 공식: - Electrical: OIF가 표준 관리 - Optical: IEEE 802.3 PMD layer 가 modulation, lane speed, wavelength 등 정의 - Mechanical: MSA가 form factor와 multi-vendor interoperability 보장 → 결과: 완벽한 plug-and-play, multi-vendor ecosystem, 경쟁적 가격

CPO의 현실: CPO는 단순히 optics를 package 내부로 이동시킨 것이 아니다. 전체 system-level 문제가 된다: - Fiber management가 chassis 내부로 이동 - Front plate density 최적화 필요 - Modulator architecture 선택이 wavelength band 결정 (O-band vs C-band) - External laser module 통합 - FAU coupling precision과 thermal management

이러한 복잡성 때문에 NVIDIA와 같은 vendor는 **end-to-end 솔루션**을 제공해야 한다. 고객이 여러 vendor의 component를 mix-and-match할 수 없다.

딜레마의 본질: 1. **Standards-first 접근:** Interoperability 보장, multi-vendor 경쟁, 낮은 가격 → 하지만 innovation이 느려지고, 누가 초기 R&D 비용을 감당하나? 2. **Proprietary-first 접근:** 빠른 innovation, vendor가 investment 회수 가능 → 하지만 customer lock-in, 제한된 채택

현재 상황: CPO는 "wild west" 단계. NVIDIA Spectrum-X, Broadcom Bailly/Humboldt 는 각각 완전히 다른 proprietary solution. OIF의 새로운 high-density interconnect 노력 (CPX paradigm)이 이를 해결하려 하지만, 성공 여부는 미지수.

산업적 함의: 역설적이게도, CPO의 광범위한 채택은 초기에 proprietary solution의 성공에 달려 있을 수 있다. NVIDIA와 같은 major vendor가 volume을 창출하고 supply chain을 성숙시켜야만, 그 다음에 표준화와 interoperability 논의가 의미를 가질 것이다. Pluggable이 즉시 표준화된 것이 아니라 GBIC → SFP → SFP+ → QSFP → QSFP-DD로 진화했듯이, CPO도 유사한 경로를 밟을 가능성성이 높다.

CPO 채택에 대한 한 가지 주요 과제는 잘 확립되고 매우 상호운용 가능한 pluggable optics model에 대한 업계의 뿌리 깊은 의존을 극복하면서 interoperability를 달성하는 것이다. Interoperability의 세 가지 주요 flavor가 있다: (1) electrical, (2) optical, 그리고 (3) mechanical Pluggable의 경우 interoperability는 다음과 같은 특성이 있다:

1. 일반적으로 **Optical Internetworking Forum (OIF)**에 의해 처리된다,
2. 일반적으로 IEEE (때로는 OIF)에 의해 처리된다. IEEE는 Ethernet **Physical Medium Dependent (PMD)** layer를 정의하는 IEEE 802.3 standard를 통해 중심 역할을 한다. 이러한 specification은 modulation format, lane speed, lane count, reach 및 media type, 그리고 optical signal을 위한 wavelength와 같은 주요 parameter를 다룬다. 이러한 표준화된 PMD를 준수함으로써 다른 vendor의 transceiver는 상호 교환 가능하게 작동하여 multi-vendor ecosystem 전체에서 진정한 plug-and-play compatibility를 보장한다,
3. 그리고 일반적으로 **Multi-Source Agreement (MSA)**에 의해 처리된다. MSA는 specialized 솔루션을 정의하고 공식 IEEE standard 외부에서 multi-vendor interoperability를 보장한다.

OIF, IEEE standard 및 MSA의 조합을 통해 pluggable transceiver는 광범위한 interoperability와 강력한 multi-vendor ecosystem을 달성한다. CPO의 경우:

1. CPO module이 electrically하게 compliant한 것이 critical한데, 그렇지 않으면 state-of-the-art SerDes와 통신할 수 없기 때문이다.
2. Optical에 대한 compatibility는 유용한데, cluster의 다른 곳에서 standard pluggable과 호환될 수 있기 때문이다.
3. CPO가 **wild west** 단계에 있으며, 일부 솔루션과 architectural decision이 완전히 proprietary form factor로 이어지고 있다는 것을 이해하는 것이 중요하다. 이것이 새로운 OIF high-density interconnect 노력 (CPX paradigm과 같은)이 해결하려고 하는 것이다.

(1)+(2)+(3)가 충족되면 CPO는 운영상 pluggable과 매우 유사해질 수 있으며, 이는 광범위한 채택을 가능하게 하는 데 도움이 될 것이다.

그러나 현재 CPO는 아직 pluggable만큼 standard를 수용하지 않고 있으며 optical transceiver가 할 수 있는 정도로 interoperability를 보장할 수 없다. 과제의 일부는 vendor가 box maker에게 silicon만을 판매하는 것이 아니라 system level 솔루션을 추진하고 있다는 것이다. 이는 CPO의 과제가 package 외부와 전체 system으로 확장되기 때문이다. fiber 관리, front plate density, modulator architecture 및 external laser는 모두 과제인 필수 부분이다. CPO 채택을 bootstrap하기 위해 NVIDIA와 같은 기업은 end-to-end 솔루션을 제공하는 것으로 시작해야 한다.

이를 위한 한 가지 접근 방식은 co-packaged OE가 wavelength, speed 및 modulation에 대한 Ethernet standard 또는 MSA와 align하는 photonic component—laser, modulator 및 photodiode와 같은—와 함께 표준화된 fiber interface를 따르는 component level에서 표준화된 솔루션을 채택하는 것일 수 있다. 이는 진정한 interoperability를 가능하게 하여 고객이 한 vendor에서 모든 장비를 조달할 필요 없이 다양한 vendor의 제품을 mix and match할 수 있게 한다. OE가 socketable하다면 operator는 failure의 경우 쉽게 교체할 수 있다. 그러한 표준화는 또한 고객을 위한 비용을 낮출 수 있는 더 경쟁적이고 강력한 multi-vendor 시장을 만들 것이다. 또한 고객이 한 vendor에 얹매이지 않고 lead time, 가격 또는 고유한 기능을 기반으로 다양한 vendor를 선택할 수 있게 한다.

3.12 Serviceability and reliability

핵심 인사이트: CPO의 아킬레스건 - Serviceability의 현실

NVIDIA가 CPO의 reliability 이점을 강조하지만, serviceability는 datacenter operator의 가장 큰 우려사항이다. 이는 단순한 engineering 문제가 아닌 operational paradigm shift이다.

Pluggable의 운영 모델 (현재): 1. Front panel에서 고장난 transceiver 감지 2. Hot-swap으로 새 transceiver 교체 (< 5분) 3. 전문 technician 불필요, downtime 최소화 → "Finger-replaceable unit" 수준의 serviceability

CPO의 운영 현실 (Quantum-X CPO switch 예시): 1. 36개 OE가 substrate에 flip-chip bonding 2. 각 OE에서 fiber ribbon이 fiber cassette를 통해 routing 3. FAU는 detachable 하지만 chassis 내부 깊숙이 위치 4. 고장 발생 시: - Chassis를 열어야 함 - 다른 fiber를 방해하지 않고 고장난 FAU 제거 - Cassette를 통해 새 FAU를 정밀하게 재부착 - Sub-micron accuracy로 submicron waveguide에 정렬 - Cramped하고 thermally active한 환경에서 작업
비용 합의: - 숙련된 technician 필요 (일반 datacenter operator 불가) - 장시간 downtime (시간 단위) - Collateral damage risk (인접 fiber 손상 가능성) - Preventive maintenance complexity 증가

Reliability vs Serviceability Paradox: CPO는 이론적으로 더 reliable 할 수 있다 (fewer connector, shorter electrical path). 하지만 실제 datacenter 운영에서 중요한 것은: - MTBF (Mean Time Between Failures) < MTTR (Mean Time To Repair) × Operational Complexity

현재 CPO는 MTBF가 더 높을 수 있지만, MTTR × Complexity가 훨씬 크다. 100,000개의 port를 운영하는 hyperscaler에게는 "How fast can I fix it?"이 "How often does it fail?"보다 더 중요할 수 있다.

산업적 합의: CPO가 mainstream이 되려면, serviceability model의 근본적 혁신이 필요하다: - Socketable OE with known-good-die testing - Tool-less FAU replacement mechanism - Automated fiber alignment system - 또는 MTBF가 MTTR × Complexity를 압도할 정도로 높아야 함

이것이 해결되지 않으면, CPO는 tier-1 hyperscaler의 mission-critical deployment가 아닌,

niche high-performance application으로 제한될 수 있다.

CPO가 광범위하게 채택되기 전에 해결되어야 하는 문제는 fiber coupling에서 나온다. Pluggable optical module에서 silicon photonics device에 fiber를 연결하는 것은 상대적으로 간단한 반면, CPO를 위해 optical engine에 fiber를 coupling하는 것은 훨씬 더 어렵다. CPO system에서 fiber는 chip의 매우 작은 waveguide (종종 width와 height 모두 submicron)로 빛을 couple하기 위해 sub-micron accuracy로 정밀하게 정렬되어야 하는 반면, pluggable은 pre-aligned, 표준화된 connector를 사용하기 때문에 덜 번거롭다. 또한 CPO 기반 system에서 fiber coupling은 cramped하고 thermally active한 switch chassis 내에서 발생하는 반면, pluggable의 경우 fiber coupling이 device의 main package에서 멀리 떨어진 곳에서 발생한다.

Copper 기반 electrical communication은 optical과 비교할 때 열등한 insertion loss와 signal integrity로 고통받지만, copper는 일반적으로 다른 방식으로 reliable하다. optical device는 본질적으로 temperature sensitive하다. Operating temperature의 변화는 laser wavelength를 변경하고, component efficiency를 감소시키며, reliability에 부정적인 영향을 미칠 수 있으며, 종종 specialized temperature control 또는 calibration mechanism을 필요로 한다. 추가로 photonic component는 또한 시간이 지남에 따라 자연스럽게 저하되며, aging, contamination 또는 mechanical stress로 인해 점진적으로 optical efficiency를 잃어 장기 reliability를 더욱 어렵게 만든다. Dust, humidity 및 mechanical disturbance와 같은 environmental factor는 copper 기반 electrical link보다 더 취약한 optical system에 불균형적으로 영향을 미친다.

이러한 environmental 및 operational 과제는 물리적 fiber 자체로 직접 확장된다. Fiber 성능은 특히 bending과 같은 물리적 disturbance에 매우 민감하며, 이는 optical insertion loss를 증가시킬 뿐만 아니라 breakage와 failure를 가속화한다. 여러 fiber array를 포함하는 chassis에서(일반적으로 engine당 2개, connector용 하나와 light source용 ELS의 하나), 각 array는 엄격한 topological 고려사항과 함께 routing되어야 한다. 모든 개별 fiber link는 faceplate에서 각 optical engine까지의 거리 variation과 인접한 array에 의해 부과된 routing constraint를 고려하여 고유한 길이를 필요로 한다.

아래 Quantum-X CPO switch의 close-up 이미지에서 OE에서 나오는 fiber ribbon이 fiber를 적절하게 관리하기 위해 fiber cassette를 통해 routing되어야 함을 볼 수 있다. FAU는 교체가 필요한 breakage를 provision하기 위해 detachable하다. 그러나 switch 내에서 fiber의 더 복잡한 routing은 fiber/FAU 교체가 face plate의 front에서 단순히 hot swap하는 고장난 pluggable transceiver를 교체하는 것보다 훨씬 더 부담스럽다는 것을 의미한다. CPO switch의 경우 engineer는 box/chassis 내부로 들어가 고장난 FAU를 제거한 다음 cassette를 통해 새 FAU를 적절히 재부착해야 한다. 이것은 다른 fiber를 방해하지 않고 수행되어야 한다. NVIDIA가 CPO의 reliability 이점을 강조해 왔지만, serviceability는 더 길게 논의할 가치가 있는 또 다른 요소이다.

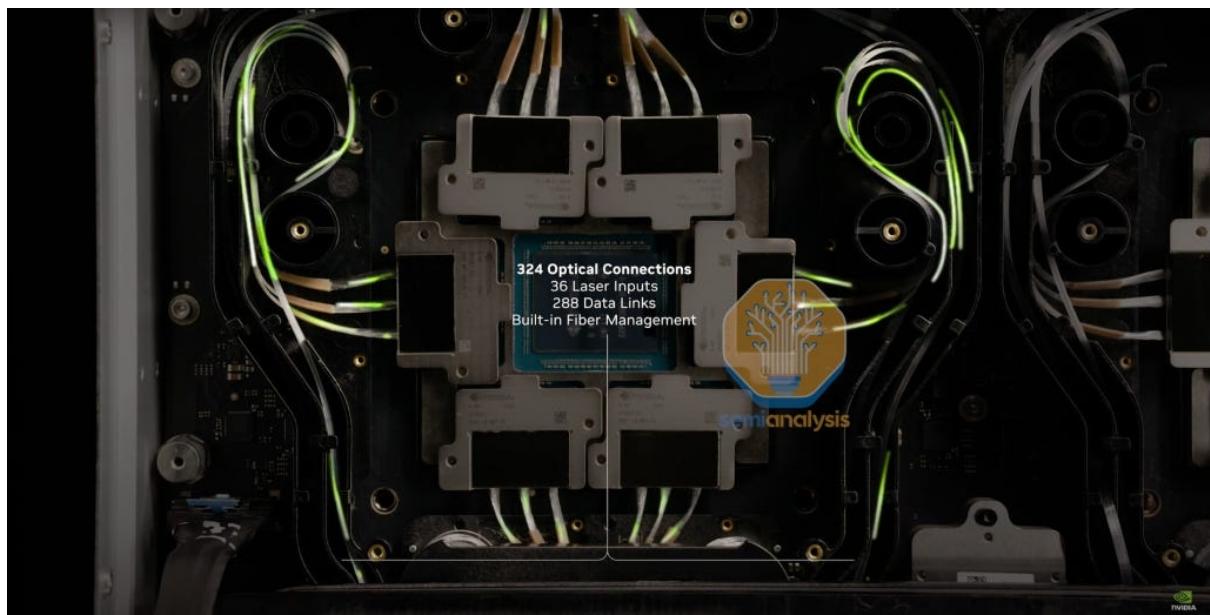


Figure 48: Source: NVIDIA

4 Part 4: CPO Products of Today and Tomorrow

이 part에서 우리는 오늘날 시장에 있거나 곧 시장에 출시될 CPO 제품을 소개하는 것으로 시작하여 NVIDIA와 Broadcom의 portfolio로 시작하고 다양한 CPO 중심 기업의 offering을 설명하는 것으로 이동할 것이다. Intel CPO, MediaTek의 CPO 작업, Ayar Labs, Nubis, Celestial AI, Lightmatter, Xscape Photonics, Ranovus 및 Scintil을 다루며, 각 제공업체의 솔루션을 상세히 설명하고 각 기업의 접근 방식에 대한 중요한 장단점을 평가한다. 마지막으로 Optical Engine 및 External Laser Source와 같은 주요 CPO component의 manufacturing, testing 및 assembly를 다루는 supply chain을 논의하기 위해 다시 돌아올 것이다.

4.1 NVIDIA CPO

GTC 2025에서 NVIDIA는 scale-out network를 위한 첫 번째 CPO 기반 switch를 선보였다. 세 가지 다른 CPO 기반 switch가 발표되었다. 우리는 각각을 차례로 살펴볼 것이지만, 먼저 모든 중요한 specification을 집계하는 깔끔한 표를 제시한다:

Nvidia CPO Roadmap			
Switch Model	Quantum 3450 CPO	Spectrum 6810 CPO	Spectrum 6800 CPO
Launch Date	2H 2025	2H 2026	
Networking Standard	InfiniBand	Ethernet	
Switch ASIC	Quantum-3	Spectrum-6	
Throughput per Package	28.8 Tbps	102.4 Tbps	
Number of Switch Packages	4	1	4
Switch Aggregate Bandwidth	115.2 Tbps (not all-to-all)	102.4 Tbps	409.6 Tbps (not all-to-all)
SerDes speed (Gb/s uni-di)	200 Gbps	200 Gbps	
Optical Connectivity	DR Optics	DR Optics	
Physical MPO Ports	144	128	512
Bandwidth and Logical Port Configurations Available	144 Ports of 800G	512 Ports of 200G 256 Ports of 400G 128 Ports of 800G	512 Ports of 800G
Bandwidth per Optical Engine (OE)	1.6 Tbps	3.2 Tbps	
Number of OEs	72	32	128
External Light Sources (ELSS)	18	16	64
<i>For Spectrum CPO there are 36 OEs on the package, but only 32 OEs are enabled</i>			

Figure 49: Source: SemiAnalysis

4.1.1 Quantum-X Photonics

2H 2025까지 시장에 나올 첫 번째 CPO switch는 Quantum X800-Q3450이 될 것이다. 144개의 Physical MPO port를 특징으로 하며 이는 800G의 144 logical port 또는 1.6T의 72 logical port를 가능하게 하여 115.2T의 aggregate bandwidth를 제공한다. spaghetti monster와의 유사성은 저자들의 배를 고프게 한다.

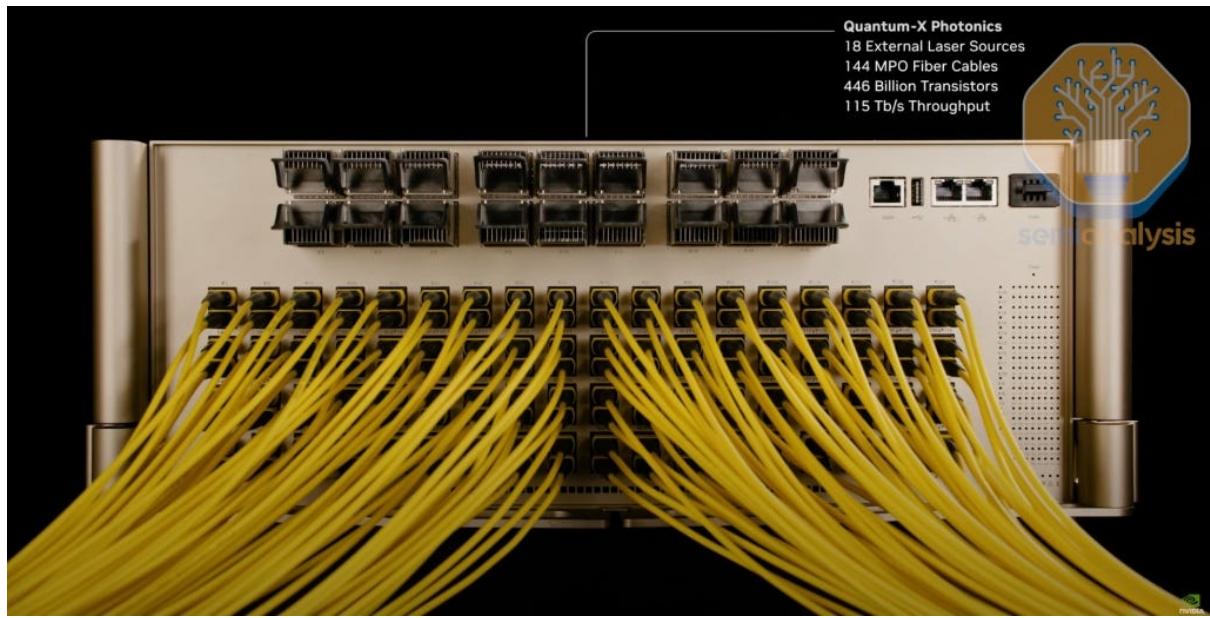


Figure 50: Source: NVIDIA

Quantum X800-Q3450은 multi-plane configuration에서 각각 28.8 Tbit/s bandwidth를 가진 4개의 Quantum-X800 ASIC chip을 사용하여 이 높은 radix와 높은 aggregate bandwidth를 달성 한다. 이 multi-plane configuration에서 각 physical port는 4개의 switch ASIC 각각에 연결되어 4 개의 다른 switch ASIC를 통해 모든 4개의 200G lane에 걸쳐 데이터를 spraying하여 모든 physical port가 다른 port와 통신할 수 있게 한다.

3-layer network의 maximum cluster size와 관련하여 이는 이론적으로 200G logical port size를 가진 28.8T switch box의 4배를 사용하는 것과 동일한 최종 결과를 제공한다. 둘 다 746,496 GPU 의 maximum cluster size를 허용한다. 차이점은 X800-Q3400 switch를 사용할 때 shuffle \diamond switch box 내부에서 깔끔하게 발생하는 반면, discrete 28.8T switch box로 동일한 network를 설정하려면 훨씬 더 많은 수의 destination으로 가는 훨씬 더 많은 개별 fiber cable \diamond 필요하다는 것이다.

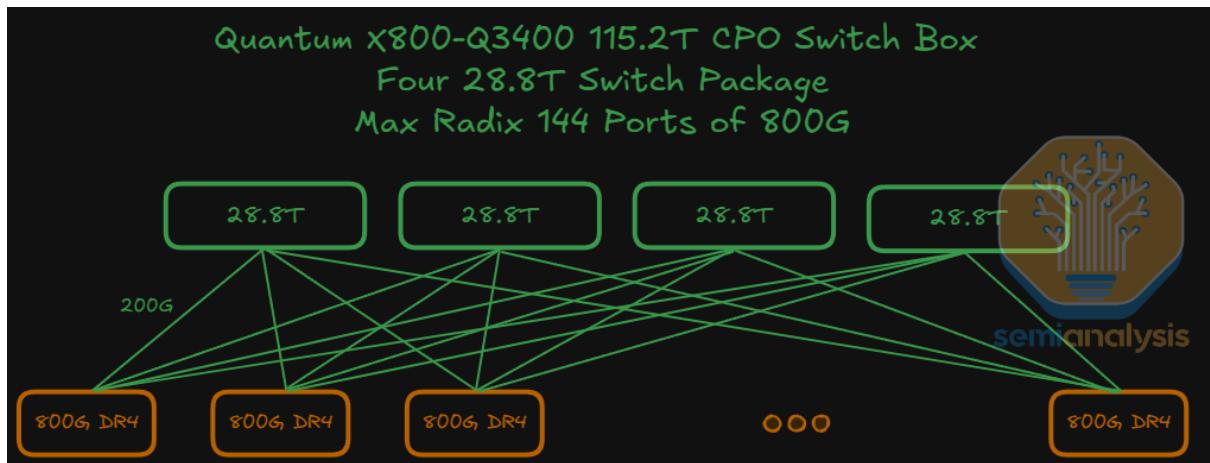


Figure 51: Source: SemiAnalysis

Quantum-X800-Q3450의 각 ASIC는 6개의 detachable optical sub-assembly로 둘러싸여 있으며, 각 sub-assembly는 3개의 optical engine을 수용한다. 각 optical engine은 1.6 Tbit/s의 bandwidth를 제공하여 ASIC당 총 18개의 optical engine과 ASIC당 28.8 Tbit/s의 aggregate optical

bandwidth를 생성한다. 이러한 sub-assembly는 detachable하므로 purist는 이것을 엄격히 **CPO**가 아니라 기술적으로 **NPO**로 간주할 수 있다. detachable OE와 관련된 약간의 extra signal loss가 있지만, 실제로 우리는 이것이 성능에 상당한 영향을 미치지 않을 것이라고 믿는다.

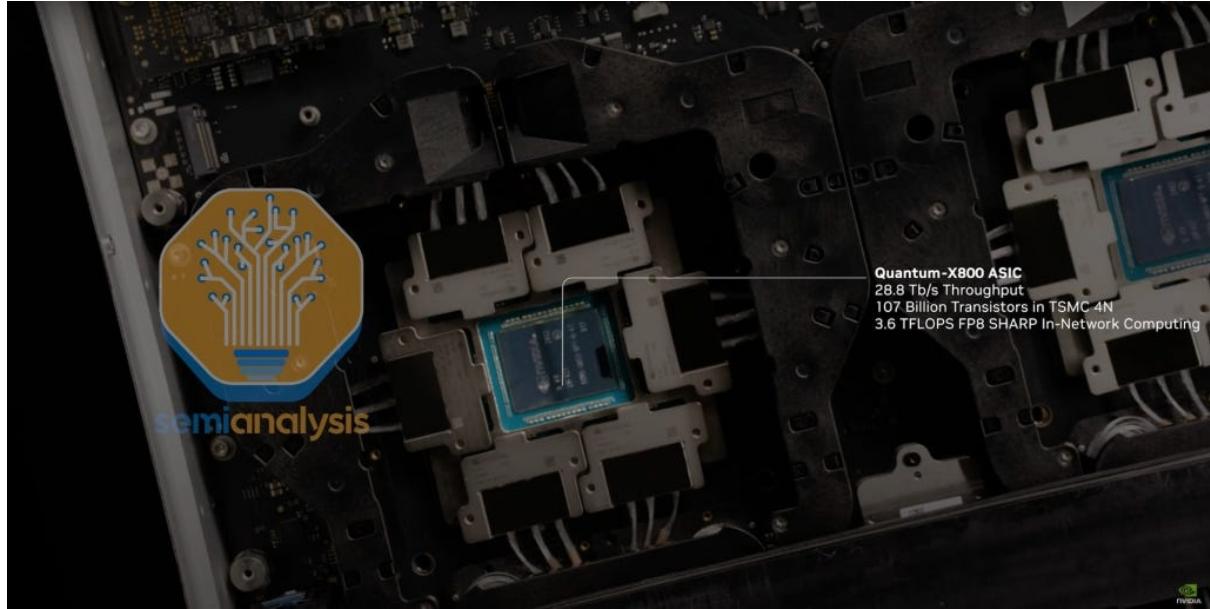


Figure 52: Source: NVIDIA

각 engine은 8개의 electrical 및 optical channel로 작동하며, electrical 측면에서 200G PAM4 SerDes에 의해 구동되고 optical 측면에서 8개의 **Micro-Ring Modulator (MRM)**가 modulator당 200G를 달성하기 위해 PAM4 modulation을 사용한다. 이 design 선택은 발표의 큰 takeaway 중 하나였다. 즉, NVIDIA와 TSMC가 production에서 200G MRM을 출하할 수 있다는 것. 이것은 오늘날 가장 빠른 MZM과 일치하며 MRM이 NRZ modulation으로 제한된다는 업계 notion을 반증한다. 이 milestone에 도달한 것은 NVIDIA의 매우 인상적인 engineering 성과이다.

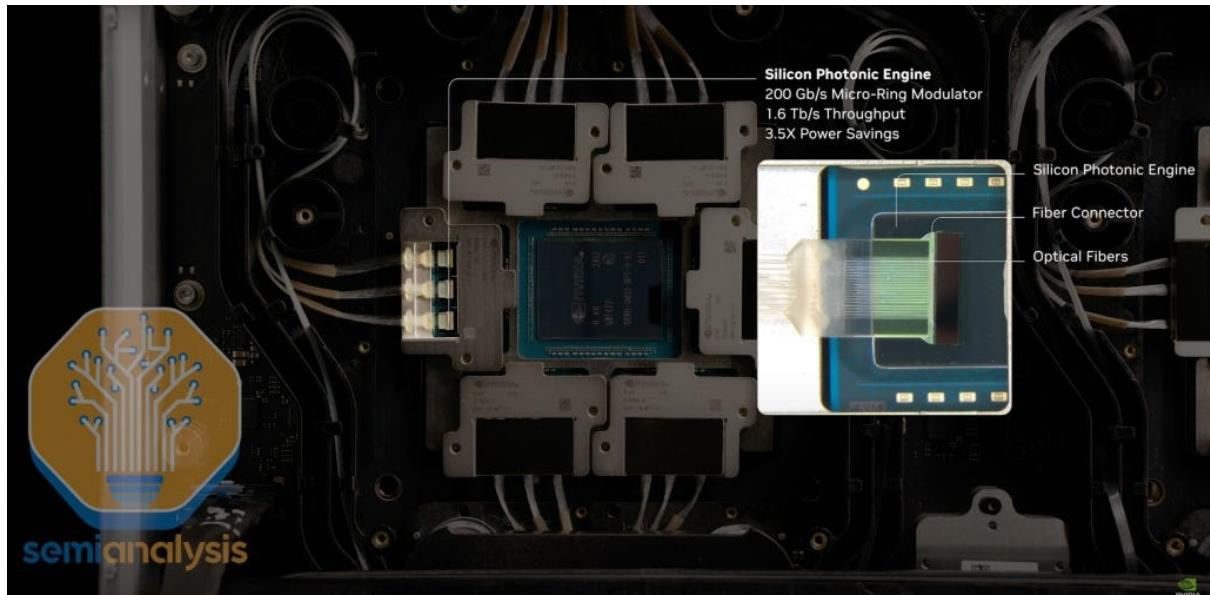


Figure 53: Source: NVIDIA

각 optical engine은 mature N65 process node에 구축된 **Photonic Integrated Circuit (PIC)**과 advanced N6 node에서 fabrication된 **Electronic Integrated Circuit (EIC)**를 통합한다. PIC는 modulator, waveguide 및 detector와 같은 optical component를 포함하기 때문에 older node를 활용한다 (Scaling으로 이점을 얻지 못하고 종종 더 큰 geometry에서 더 나은 성능을 발휘하는 device) 이와 대조적으로 EIC는 advanced node가 가능하게 하는 더 높은 transistor density와 향상된 power efficiency로부터 상당한 이점을 얻는 driver, TIA 및 control logic을 포함한다. 이 두 die는 그런 다음 TSMC의 COUPE platform을 사용하여 hybrid bonding되어 photonic과 electronic domain 사이에 ultra-short, high-bandwidth interconnect를 가능하게 한다.

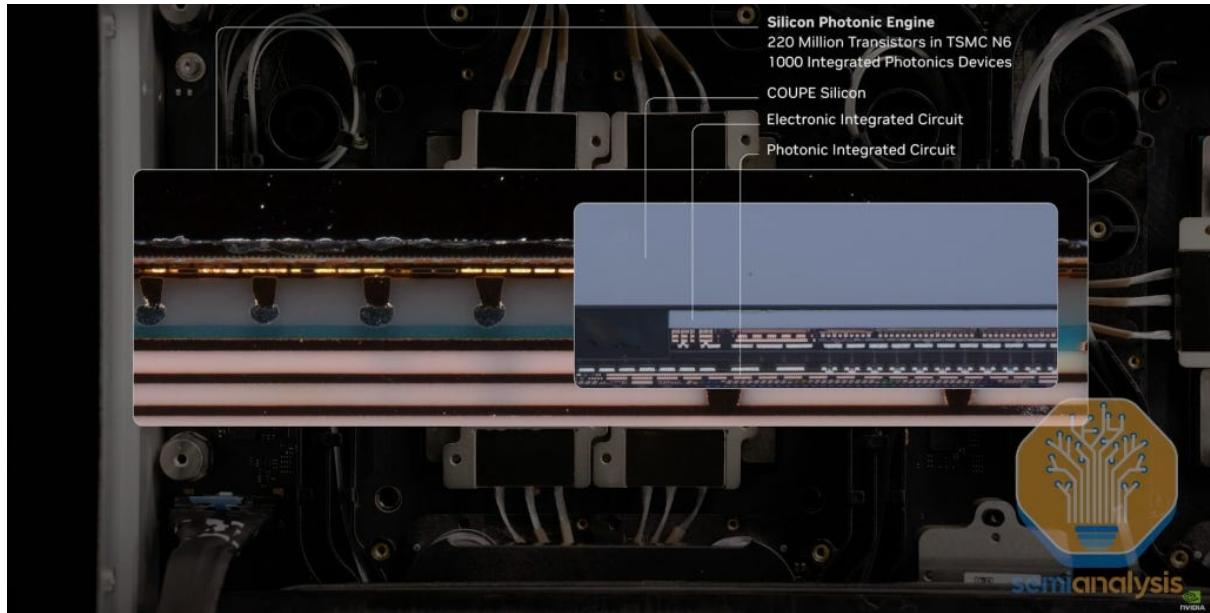


Figure 54: Source: NVIDIA

Quantum-X800-Q3450의 ASIC 위에 2개의 copper cold plate가 각 switch ASIC에서 열을 효율적으로 dissipate하는 closed-loop liquid cooling system의 일부로 위치한다. Cold plate에 연결된 black tubing은 coolant fluid를 순환시켜 thermal stability를 유지하는 데 도움을 준다. 이 cooling system은 ASIC뿐만 아니라 인접한 temperature-sensitive co-packaged optics의 thermal stability를 유지하는 데 필수적이다.

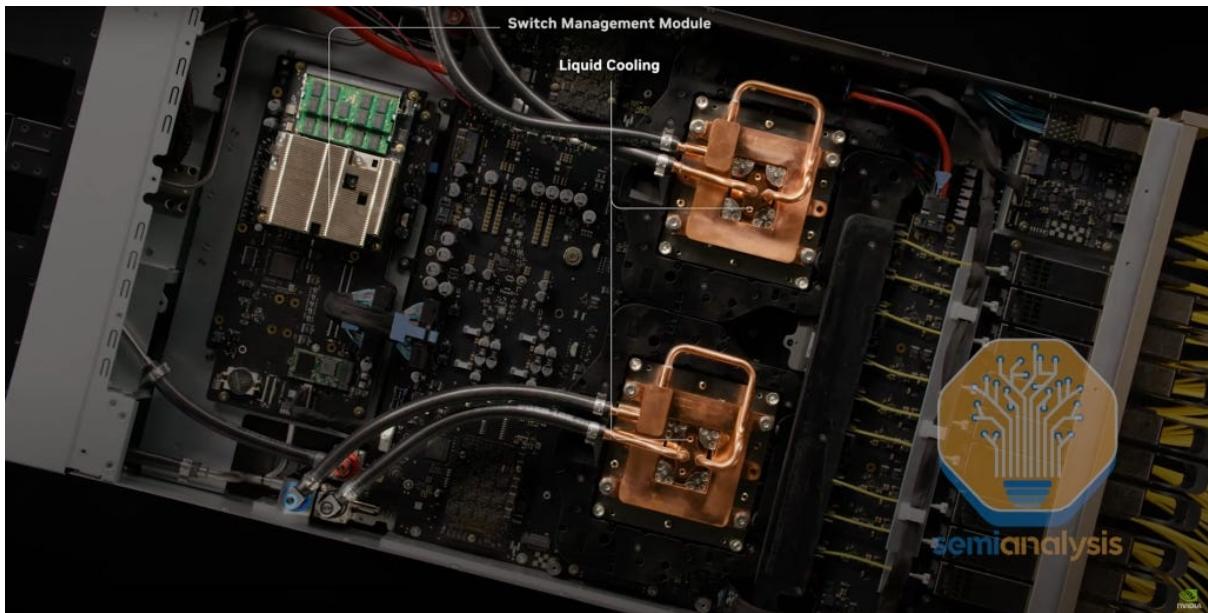


Figure 55: Source: NVIDIA

4.1.2 Spectrum-X Photonics

Spectrum-X Photonics는 2026년 후반에 출시될 예정이며, 102.4T aggregate bandwidth를 가진 X800-Q3450 CPO switch의 Ethernet Spectrum-X variant, 102.4T aggregate bandwidth를 제공하는 Spectrum 6810, 그리고 4개의 discrete Spectrum-6 Multi-Chip Module (MCM)을 사용하여 409.6T aggregate bandwidth를 제공하는 더 큰 cousin Spectrum 6800의 두 가지 별도 switch configuration 제품이 출시될 것이다.

Quantum X800-Q3450 CPO switch는 multi-plane configuration에서 physical에 연결된 4개의 discrete switch package를 활용하며, 각 switch package는 required SerDes 및 기타 electrical component와 함께 28.8T switch ASIC를 포함하는 monolithic die이다. 이와 대조적으로 Spectrum-X Photonics switch silicon은 중앙에 훨씬 더 큰 reticle size 102.4T switch ASIC가 있고 각 측면에 2개씩 8개의 224G SerDes I/O chiplet로 둘러싸인 MCM이다.

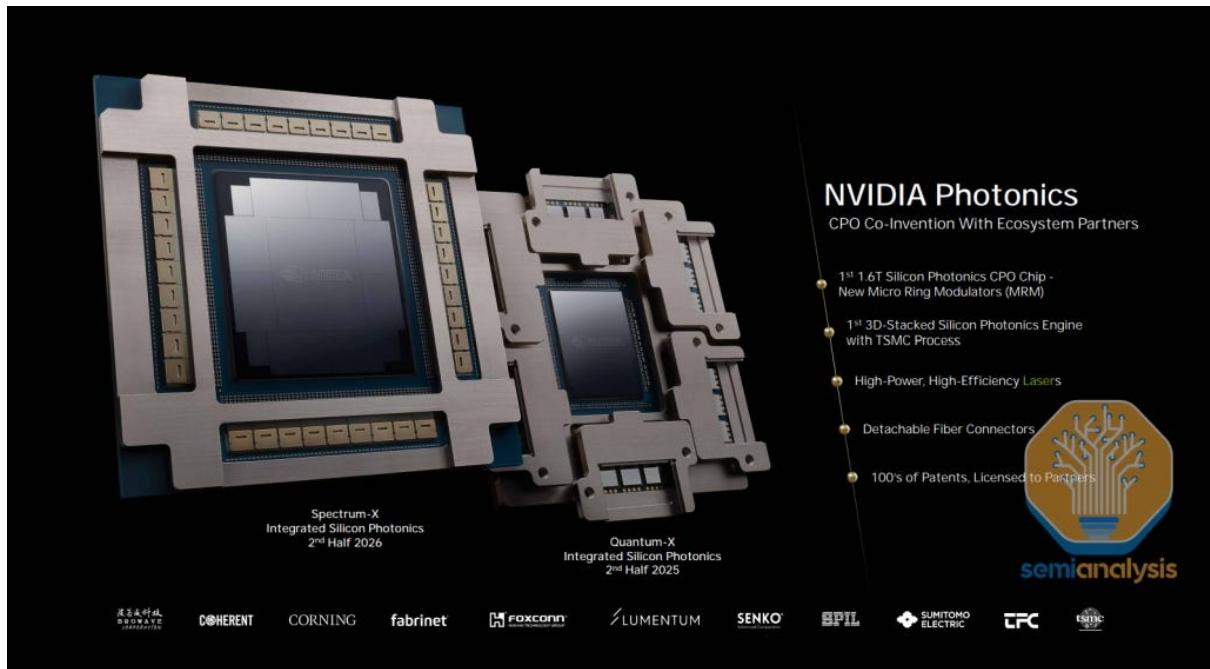


Figure 56: Source: NVIDIA

각 Spectrum-X photonics multi-chip module switch package는 단일 102.4T switch package에 36개의 optical engine을 가질 것이다. 이 package는 각각 200G의 16 optical lane을 가진 NVIDIA의 2세대 optical engine을 3.2T bandwidth로 사용할 것이다. 32개의 optical engine만 active하며, OE가 실패할 경우를 대비한 redundancy 목적으로 추가 4개가 있다는 점에 유의하라. 이는 OE가 substrate에 soldering되어 쉽게 교체할 수 없기 때문이다.

각 I/O chiplet은 64 SerDes lane으로 구성된 총 12.8T의 unidirectional bandwidth를 제공하며 각각 4개의 OE와 interface한다. 이것이 Spectrum-X가 SerDes를 위한 훨씬 더 많은 shoreline과 면적으로 Quantum-X Photonics보다 훨씬 더 많은 aggregate bandwidth를 제공할 수 있게 하는 것이다.

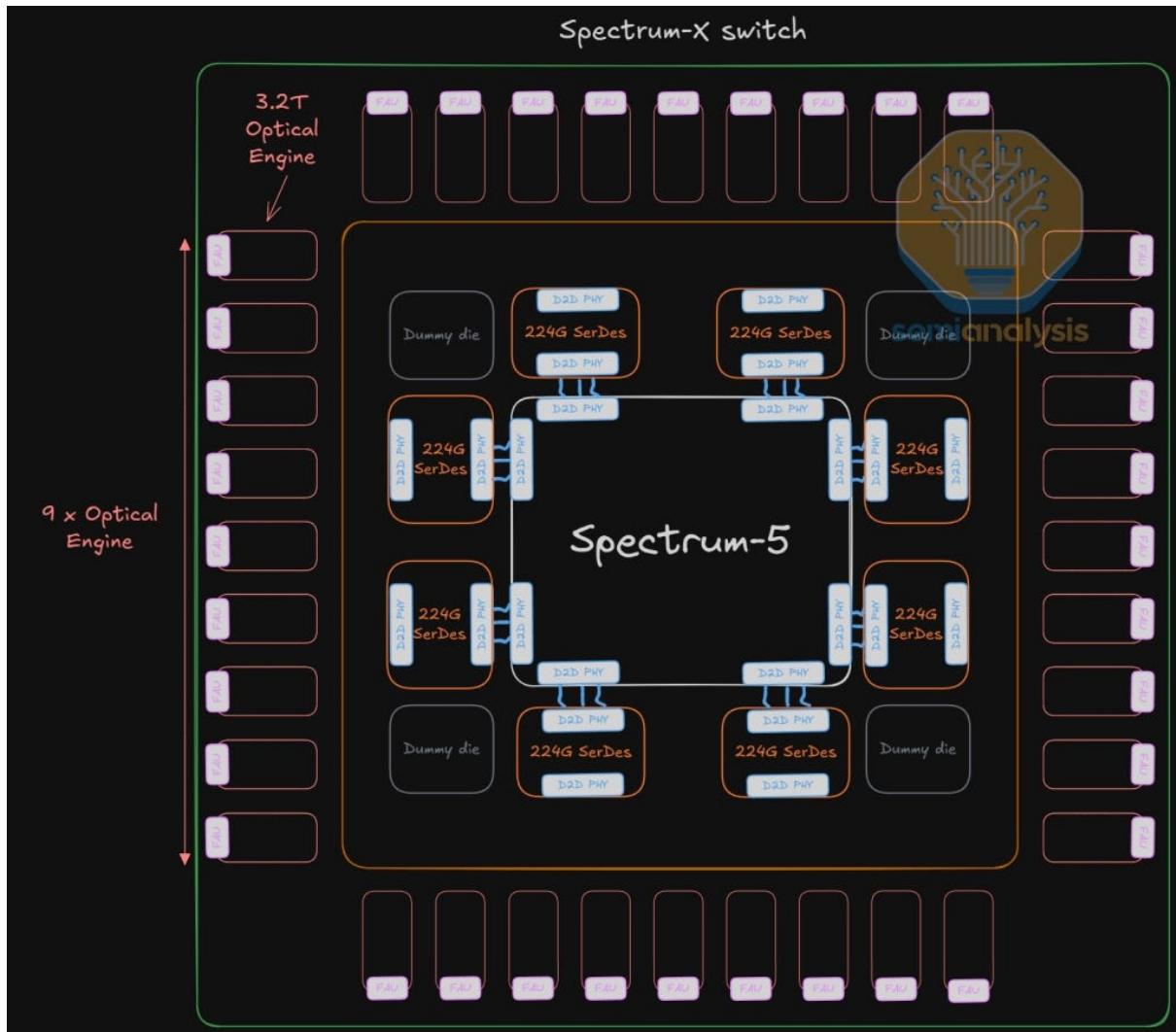


Figure 57: Source: SemiAnalysis

Spectrum-X 6810 Switch Box는 위의 switch package 1개 unit을 사용하여 102.4T의 aggregate bandwidth를 제공한다. 더 큰 Spectrum-X 6800 Switch Box SKU는 multi-plane configuration에서 external physical port에도 연결된 위의 Spectrum-X switch package 4개를 활용하여 달성된 409.6T의 aggregate bandwidth를 가진 high-density chassis이다.

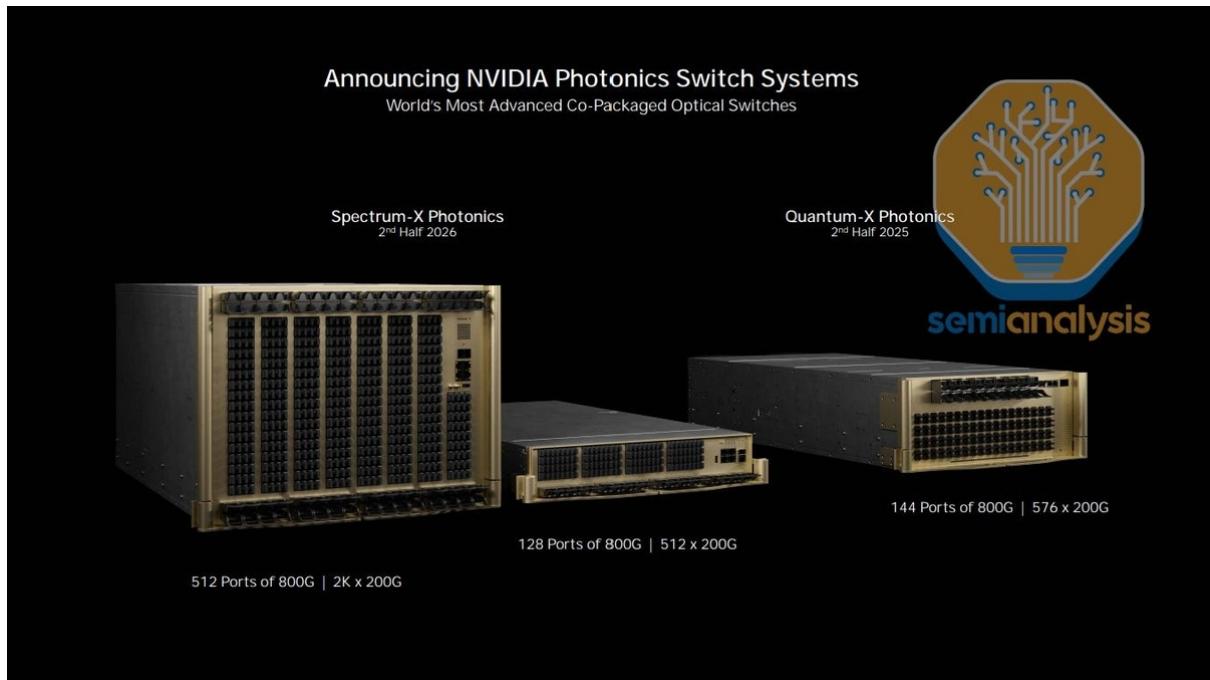


Figure 58: Source: NVIDIA

4개의 ASIC 115.2T Quantum X800-Q3450과 마찬가지로 Spectrum-X 6800은 각 port를 모든 4개의 ASIC에 물리적으로 연결하기 위해 internal breakout을 사용한다.



Figure 59: Source: SemiAnalysis

4.2 The Broadcom CPO Switch Portfolio

Broadcom CPO Roadmap			
Switch Model	Humbolt	Bailly	Davisson
Launch Date	2022	2024	2026
Networking Standard	Ethernet	Ethernet	Ethernet
Scale-out or Scale-up	Scale-out	Scale-out	Scale-out
Switch ASIC	Tomahawk 4	Tomahawk 5	Tomahawk 6
Switch ASICs per Package	1	1	1
Throughput per Package	25.6 Tbps	51.2 Tbps	102.4 Tbps
Number of switch packages	1	1	1
Switch Aggregate Bandwidth	25.6 Tbps (half-electrical)	51.2 Tbps	102.4 Tbps
SerDes speed (Gb/s uni-di)	100G	100G	200G
Optical Connectivity	DR Optics	FR4 Optics	DR4 Optics
Bandwidth and Logical Port Configurations Available	256 Ports of 100G 128 Ports of 200G 64 Ports of 400G	128 Ports of 400G 64 Ports of 800G	128 Ports of 800G 64 Ports of 1.6T
Bandwidth per Optical Engine (OE)	3.2 Tbps	6.4 Tbps	6.4 Tbps
Number of OEs	4	8	16

Figure 60: Source: SemiAnalysis

Broadcom은 실제 CPO 지원 system을 제공한 최초 기업 중 하나이며 따라서 CPO의 leader로 간주된다. Broadcom의 1세대 CPO device인 Humboldt는 주로 proof of concept 역할을 했다. **TH4-Humboldt**로 불리는 이것은 총 capacity를 전통적인 electrical connection과 CPO 사이에 균등하게 나누는 25.6Tbit/s Ethernet switch이다. 그 중 12.8Tbit/s는 각각 100 Gbit/s의 32 lane 을 제공하는 4개의 3.2 Tbit/s optical engine에 의해 처리된다. 이러한 copper와 optics의 hybrid design에는 몇 가지 두드러진 use case가 있다. 한 시나리오에서 **Top-of-Rack (ToR)** switch는 nearby server로의 short-distance copper connection을 위해 electrical interface에 의존하는 반면, optical port는 다음 switching tier로 uplink한다. 다른 시나리오에서 aggregation layer에서 electrical port는 rack 내의 다양한 switch를 interconnect하고, optical link는 해당 layer 위 또는 아래의 switching tier로 확장된다.

TH4-Humboldt: First Generation System



Product Features:

- 25.6T Ethernet Switch
- Half CPO, Half Electrical connectivity
- Four 3.2T optical engines (32x100Gbps DR connectivity)
- Optical engine is a PIC bonded to a SiGe EIC
- Each optical engine has ~ 250 optical components



SiGe dissipates additional 3 pJ/bit power consumption compared to CMOS solutions

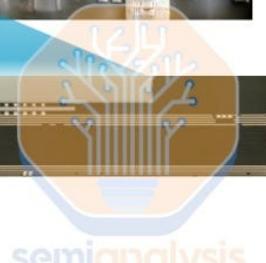
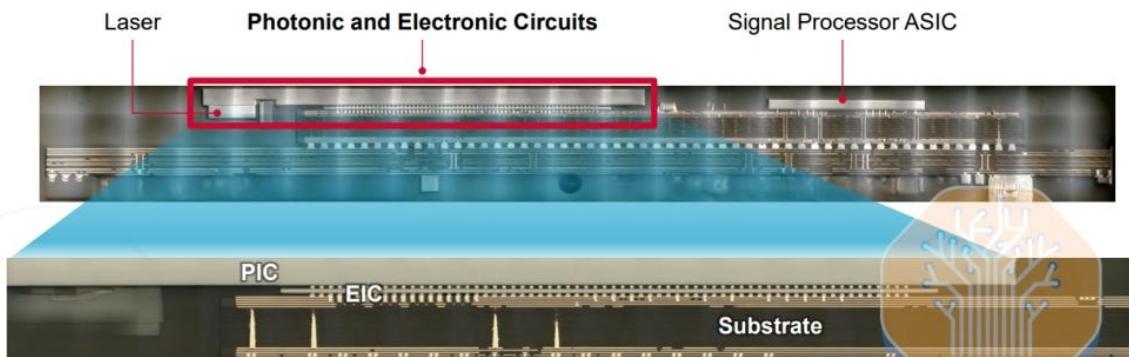
10 | Broadcom Proprietary and Confidential. Copyright © 2023 Broadcom. All Rights Reserved. The term "Broadcom" refers to Broadcom Inc. and/or its subsidiaries.



Figure 61: Source: Broadcom

o] design에서 Broadcom은 Silicon Germanium (SiGe) EIC를 사용했지만 다음 세대 (즉, Bailly)에서는 CMOS로 전환했다.

TH4-Humboldt: SiPh PIC + SiGe EIC + TSV



11 | Broadcom Proprietary and Confidential. Copyright © 2023 Broadcom. All Rights Reserved. The term "Broadcom" refers to Broadcom Inc. and/or its subsidiaries.

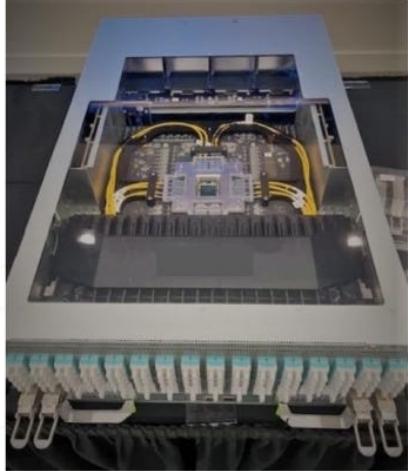


Figure 62: Source: Broadcom

Broadcom의 2세대 CPO device인 Bailly는 half-optical predecessor와 달리 전적으로 optical I/O에 의존하는 51.2 Tbit/s Ethernet switch이다. 각각 100 Gbit/s의 64 lane을 제공하는 8개의 6.4Tbit/s optical engine으로 구성된다. 또 다른 주목할 만한 변화는 SiGe EIC를 사용하는 대신 이제 7nm CMOS EIC를 사용한다는 것이다. CMOS EIC로 이동하면 추가 control logic이 있는 더

복잡하고 통합된 design이 가능해졌으며, 이는 차례로 더 높은 lane count를 가능하게 했다 (이전 32 lane에서 새로운 optical engine에서 64 lane까지 scaling)

TH5-Bailly: Second Generation System



Product Features:

- 51.2T Ethernet Switch
- All Optical CPO connectivity
- Eight 6.4T optical engines (64x100Gbps FR4 connectivity)
- Optical engine is a PIC bonded to a CMOS EIC
- Each optical engine has ~ 1000 optical components



semianalysis

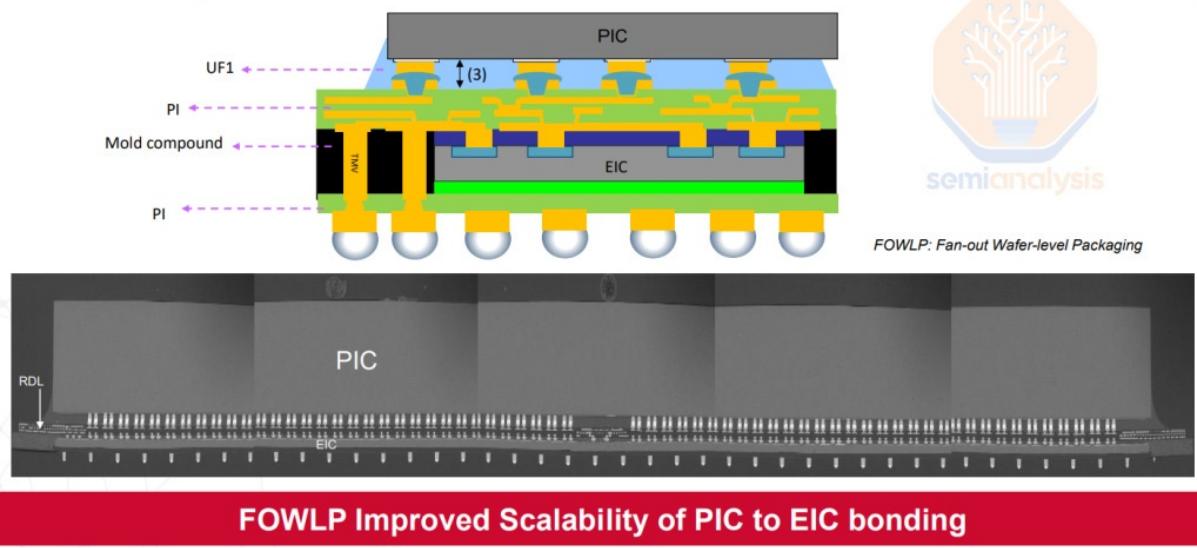
 **BROADCOM**

12 | Broadcom Proprietary and Confidential. Copyright © 2023 Broadcom. All Rights Reserved. The term "Broadcom" refers to Broadcom Inc. and/or its subsidiaries.

Figure 63: Source: Broadcom

1세대에서 2세대로의 또 다른 주목할 만한 shift는 TSV process에서 **Fan-Out Wafer-Level Packaging (FOWLP)**으로의 전환이다. 이 design에서 EIC는 PIC까지 signal을 routing하기 위해 textbfThrough-Mold Via (TMV)를 활용하는 반면 copper pillar bump는 substrate에 연결한다. FOWLP를 채택하는 주요 이유는 mobile handset 시장에서 이미 입증되었고 OSAT에 의해 널리 지원되어 기술에 더 큰 scalability를 제공하기 때문이다. ASE/SPIL이 이 FOWLP process의 OSAT 파트너였다.

TH5-Bailly: SiPh PIC + 7nm CMOS EIC + FOWLP



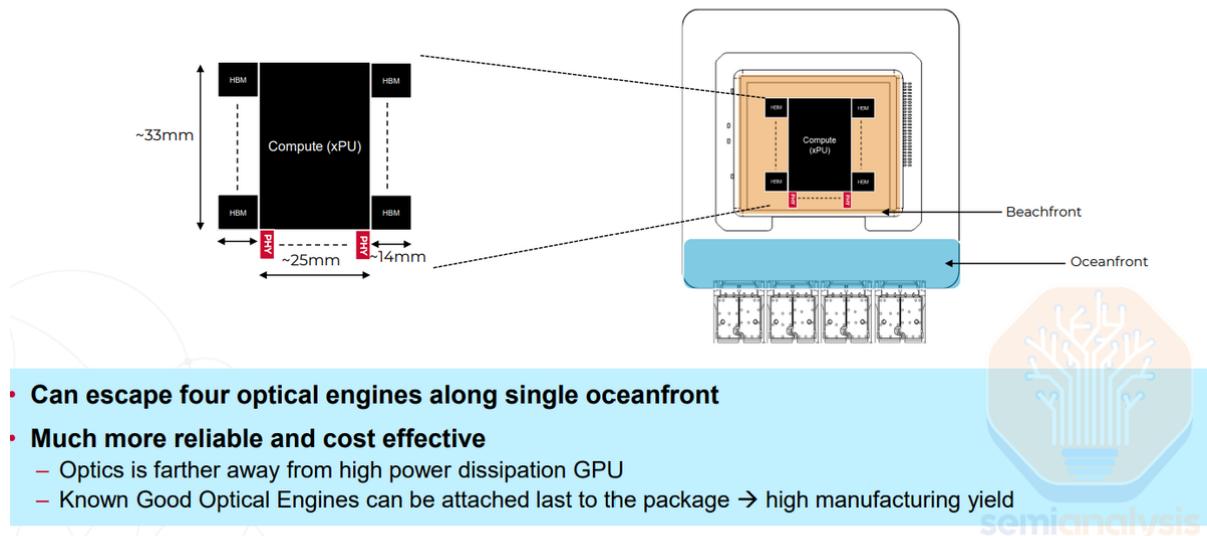
13 | Broadcom Proprietary and Confidential. Copyright © 2023 Broadcom. All Rights Reserved. The term "Broadcom" refers to Broadcom Inc. and/or its subsidiaries.

BROADCOM

Figure 64: Source: Broadcom

Broadcom은 Hot Chips 2024에서 하나의 logic die, 두 개의 HBM stack 및 SerDes tile과 함께 package에 6.4 Tbit/s optical engine을 통합하는 experimental design을 공개했다. 그들은 substrate의 east 및 west edge에 HBM을 배치하는 fan-out 접근 방식을 제안하여 동일한 package에 2개의 optical engine을 위한 공간을 허용했다. CoWoS-S에서 CoWoS-L로 이동하면 edge에서 100 mm를 초과하는 substrate로 이동한다. 따라서 최대 4개의 optical engine을 수용하고 51.2 Tbit/s의 bandwidth를 달성할 수 있을 것이다.

Beachfront vs Oceanfront: Utilizing Fan-Out



24 | Broadcom Proprietary and Confidential. Copyright © 2023 Broadcom. All Rights Reserved. The term "Broadcom" refers to Broadcom Inc. and/or its subsidiaries.

BROADCOM

Figure 65: Source: Broadcom

Scale-Up Optical Oceanfront Density Roadmap

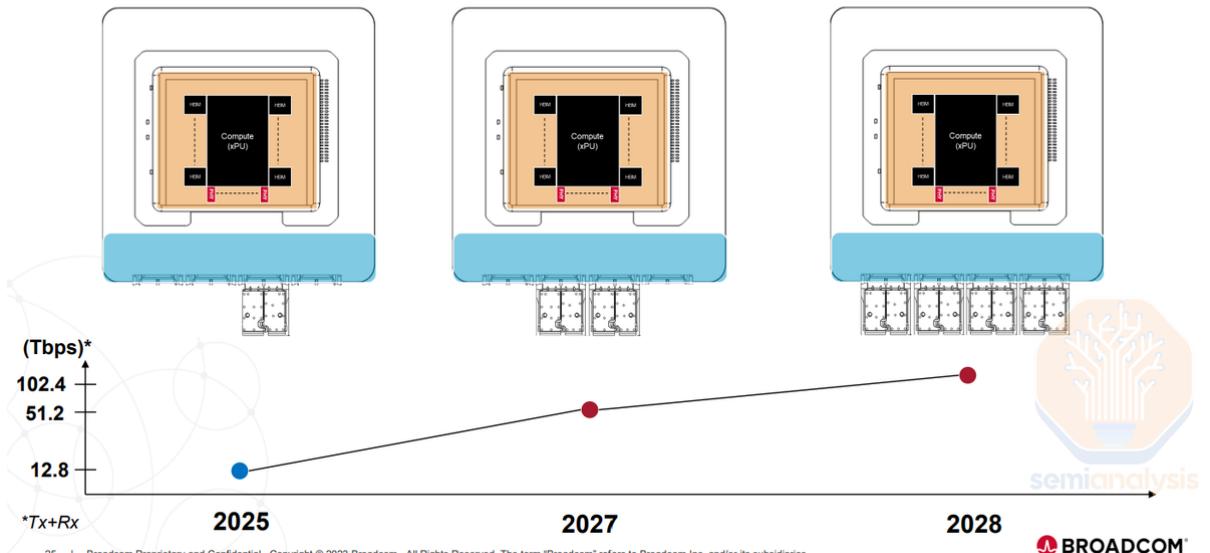


Figure 66: Source: Broadcom

올해 Broadcom은 16개의 6.4T OE를 통합하는 Tomahawk 6 기반 Davisson CPO switch를 출시하고 있다. Switch ASIC는 TSMC의 N3 process node를 사용하여 fabrication되며 package 당 102.4 Tbit/s의 bandwidth를 제공한다. Broadcom은 box assembly를 위해 Micas 및 Celestica 와 같은 **Contract Manufacturer (CM)**를 사용한다. 추가로 NTT Corp (Japan)은 Broadcom 의 TH6 bare die를 구매하고 Broadcom에서 조달하지 않은 proprietary OE 및 optical 솔루션을 사용하여 자체 CPO system을 구축하고 있는 것으로 알려져 있다. 이 접근 방식은 TH6 기반 CPO system에 대한 잠재적 business 기회를 확대하고 더 open한 vendor ecosystem을 장려한다.



Figure 67: Source: SemiAnalysis

Scale-up fabric에서 CPO에 대한 더 큰 가치를 보기 때문에 Broadcom이 제공하는 첫 번째

mass-produced CPO system은 고객의 AI ASIC에 있을 것이라고 믿는다. CPO에 대한 Broadcom의 경험은 중기적으로 ASIC roadmap에서 CPO를 보는 고객에게 매력적인 design 파트너로 만든다. 우리는 이것이 OpenAI가 Broadcom을 선택하게 만든 주요 요인이었다고 이해한다. 흥미롭게도 Broadcom의 가장 큰 ASIC 고객인 Google은 datacenter에 CPO를 배포하는 것을 가장 주저하는 hyperscaler이다. Google의 infrastructure philosophy는 절대 성능보다 reliability에 더 중점을 두며, CPO의 reliability는 그들에게 deal breaker이다. 우리는 Google이 곧 CPO를 채택할 것으로 기대하지 않는다.

Broadcom CPO endpoint의 미래 세대도 TSMC의 COUPE platform으로 이동하고 있다. COUPE가 제공하는 기능이 bandwidth scaling에 대한 경로를 제공한다는 명확한 signal이며, 이것은 OE를 package하는 방식의 변화일 뿐만 아니라 Broadcom의 이전 세대가 edge coupling과 MZM을 사용했다는 것이다. 이 두 가지 선택은 구현 관점에서 더 간단했지만 위에서 논의한 것처럼 더 scalable하기도 했다. COUPE는 grating coupling과 MRM에 bias되어 있으며 이는 기존 접근 방식과 극적인 변화이다. Broadcom이 가장 많은 CPO 경험을 가지고 있음에도 불구하고 이러한 technical approach의 변화는 Broadcom이 기술의 일부 측면에서 본질적으로 새로 시작해야 함을 의미한다. 문제는 TSMC가 Broadcom을 위해 design을 더 쉽게 만들기 위해 얼마나 많은 도움을 제공할 수 있는가이다.

4.3 Intel's CPO Roadmap

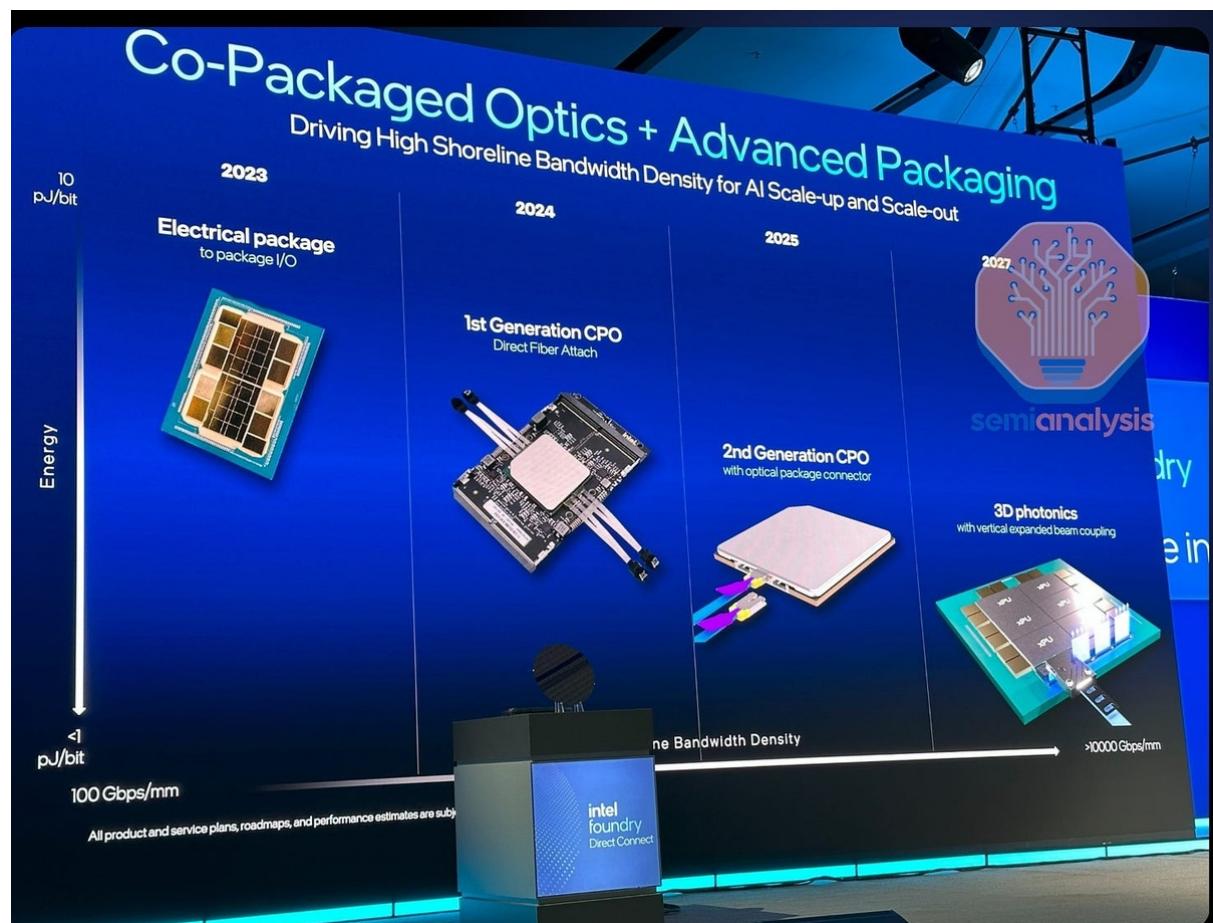


Figure 68: Source: Intel

Intel은 올해 Intel Foundry Direct Connect에서 CPO roadmap을 공개하고 4단계 개발 roadmap을 설명했다

2023: Intel은 optical integration의 precursor로 advanced electrical package-to-package I/O connectivity에 대한 concept을 설명했다. 이 milestone은 multi-die system을 지원하기 위해 chip package 간 (전통적인 PCB trace를 우회하는) high-bandwidth, short-reach electrical link를 가능하게 하는 데 초점을 맞췄다. 나중에 optical channel로 augment될 수 있는 package-level I/O infrastructure를 확립함으로써 photonics를 통합하기 위한 stage를 설정했다.

2024: Intel은 direct fiber attach를 특징으로 하는 1세대 CPO 솔루션을 시연했다. 이 접근 방식에서 optical engine chiplet은 external connector 없이 optical fiber에 직접 coupled되어 link를 단순화한다. OFC 2024에서 Intel은 concept Xeon CPU와 co-package된 4 Tbit/s (bi-directional) textbf{Optical Compute Interconnect (OCI)} chiplet을 선보이며 single-mode fiber link를 통해 error-free 데이터를 실행하고 각각 32 Gbit/s에서 64 lane을 제공했다. optical interface는 ~5 pJ/bit에서 뛰어난 efficiency를 달성했다.

2025: Intel의 2세대 CPO 솔루션은 permanent fiber pigtail 대신 detachable optical package connector를 통합한다. Intel은 embedded 3D waveguide와 on-package photonics를 standard fiber connector와 interface하는 mechanical alignment feature를 포함하는 package 측면에 slot되는 glass optical bridge를 개발했다. 이 optical package connector design은 modular assembly를 가능하게 하여 더 connectorized되고 serviceable한 form factor로의 전환을 표시한다.

2027: Intel은 3D-integrated photonics의 breakthrough를 목표로 하고 있다 (Vertical expanded beam coupling을 사용하여 photonic component를 vertically stacking). 이 구상된 3세대 design에서 optical I/O는 short free-space 또는 in-glass optical path를 통해 die layer 사이 (예를 들어 photonic interposer layer와 logic die 사이)를 vertically하게 routing될 것이다. package를 통해 빛을 vertically하게 coupling함으로써 Intel은 electrical bottleneck을 더욱 줄이고 10년 후반에 ultra-high bandwidth chiplet fabric을 가능하게 하는 것을 목표로 한다.

4.4 MediaTek CPO plans

Custom ASIC design house인 MediaTek는 design platform에 CPO capability를 통합하기 위해 노력하고 있다. 그들은 custom accelerator와 seamless하게 작동할 수 있는 PIC/EIC design을 제공하는 것을 목표로 한다. 그들은 200G-per-lane 세대에서 **Near-Packaged Copper (NPC)**가 fiber pitch >900 μm로 효과적인 솔루션이 될 수 있다고 생각한다. 데이터 rate가 200-300G 범위로 증가함에 따라 더 dense한 pitch >400 μm를 가진 CPC가 더 선호될 수 있다. 그러나 속도가 400G-per-lane 이상에 도달하면 ~130 μm의 더 dense한 fiber pitch와 더 compact한 interconnect IP를 가진 CPO architecture로 이동하는 것이 필요할 것이다.

4.5 CPO Focused Companies

NVIDIA, Broadcom 및 Marvell이 자체 proprietary 솔루션을 만들면서 자체 경로를 개척하고 있는 동안, 여러 CPO 중심 기업은 또 다른 접근 방식 set를 탐색하고 있다. 이러한 기업의 질문은 주요 switch silicon 및 GPU/ASIC provider와 어떻게 경쟁할 것인가이다. 특히 이러한 incumbent 대부분이 이미 proprietary 솔루션을 발표하거나 시연했기 때문이다. AMD는 예외로 남아 있다. 그들은 어떤 offering도 선보이지 않았지만 내부적으로 photonic IP를 개발하고 있는 것으로 알려져 있다.

Ayar Labs, Lightmatter, Celestial AI, Nubis 및 Ranovus와 같은 OE chiplet provider의 경우 과제는 established player를 overtake하고 통합하기에 충분히 설득력 있는 솔루션을 제공하는 것이다. Ayar Labs, Celestial AI 및 Ranovus는 완전한 "bookended" system을 공급하며, 이는 고객이

완전한 end-to-end 솔루션을 선택해야 함을 의미한다. 반면 Nubis는 구현을 streamline하고 선택을 더 간단하게 만드는 것을 목표로 더 open하고 standards-based 솔루션에 집중하고 있다. 다른 한편으로 일부 기업의 제품 roadmap의 중요한 부분을 형성하는 더 radical한 접근 방식이 있다 (Lightmatter의 optical interposer와 Celestial AI의 photonic bridge) 이러한 솔루션은 전체 잠재력을 실현하기 위해 package와 host silicon design 모두의 근본적인 재고를 필요로 한다. 그러나 이러한 접근 방식에는 elevated cost와 상당한 불확실성, 특히 CMOS 기반 silicon 및 high-volume manufacturing과의 seamless한 통합과 관련하여 수반된다.

이러한 각 기업의 architecture와 go to market plan을 통한 tour를 시작해보자.

4.5.1 Ayar Labs

Ayar Labs의 제품은 XPU, switch ASIC 또는 memory에 package될 수 있는 TeraPHY optical engine chiplet이다. 1세대 TeraPHY는 단지 10W의 전력을 사용하면서 2Tbit/s의 uni-directional bandwidth를 제공할 수 있다. 2세대 TeraPHY는 4 Tbit/s의 unidirectional bandwidth를 제공한다. 이것은 세계 최초의 PCIe optical retimer chiplet으로 host signal을 optically하게 전송하기 위해 chiplet 내에서 E/O conversion을 수행한다. PCIe의 선택은 host chip에 쉽게 구현될 수 있는 표준화된 interface를 가지고 있기 때문에 고객에게 매력적으로 만들어야 한다.

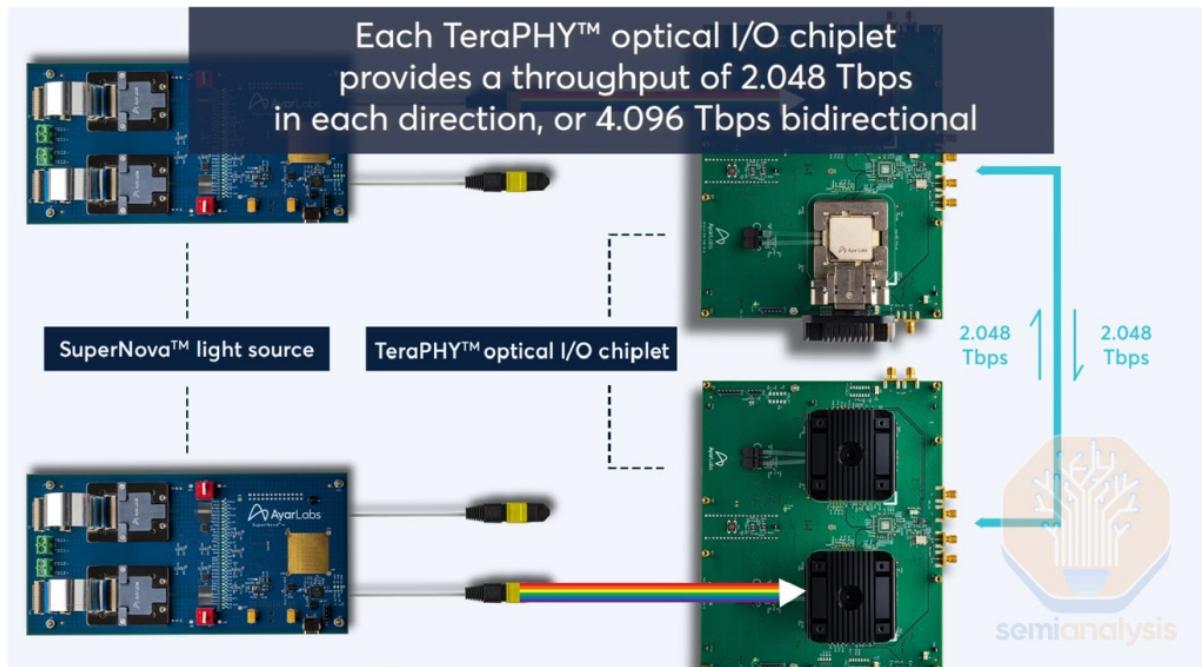


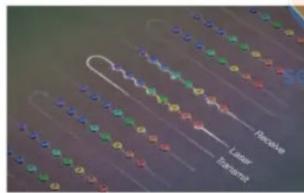
Figure 69: Source: Ayar Labs

Ayar Labs는 TeraPHY의 처음 두 세대를 GlobalFoundries의 45 nm process에서 electronic과 silicon photonics를 모두 통합하는 monolithic 솔루션으로 제조했으며, 3세대 TeraPHY는 대신 TSMC COUPE를 선택한다. Ring modulator, waveguide 및 control circuitry의 이러한 긴밀한 통합은 electrical loss를 줄이는 데 도움이 된다. 그러나 처음 두 세대에서 사용된 mature monolithic node는 EIC의 성능을 제약하며 이것이 TeraPHY의 처음 몇 세대가 낮은 modulation rate를 사용한 이유이다.

Making Optical I/O a Reality



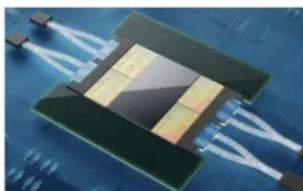
Dense optics: Micro-ring modulators



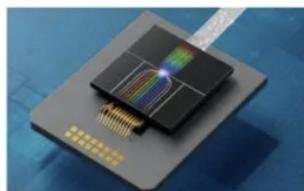
Monolithic integration with electronics



High volume scalable manufacturing



Advanced packaging and fiber attach techniques



Disaggregated multi-wavelength laser sources



Ecosystem and standardization:
electrical I/O, wavelengths, data rates

Figure 70: Source: Ayar Labs

4 Tbit/s unidirectional 2세대 code-named **Eagle**에서 TeraPHY는 각각 MRM에 의해 modulate된 32 Gbit/s NRZ x 16-wavelength architecture로 구동되는 8개의 512 Gbit/s I/O port를 통합한다. SuperNova라고 불리는 external laser source는 Swedish 회사 Sivers가 공급한다. Laser는 DWDM을 사용하여 16 lambda ("color")를 하나의 fiber로 결합한다. 그런 다음 각 port는 transmit (Tx) 및 receive (Rx)를 위해 하나의 single-mode fiber pair를 사용하며, 이는 각 4T chiplet이 총 24 fiber에 연결됨을 의미한다 (Rx/Tx용 16개와 laser input용 8개). 회사는 packaging process에서 edge coupling (EC)을 사용하지만 grating coupling (GC)도 지원할 수 있다.

Chiplet당 bandwidth를 scaling하기 위해 회사는 fiber density (현재 chiplet당 24개)가 connector 기술이 발전함에 따라향후 몇 년 동안 realistically하게 두 배가 될 수 있다고 언급했다. 추가로 port/fiber당 bandwidth도 per-wavelength 데이터 rate를 증가시켜 두 배가 될 수 있으며, 가까운 미래 roadmap에서 전체 4배 bandwidth 확장에 기여한다.

SuperNova laser는 **Multi-Source Agreement (MSA)** compliant하여 다른 CW-WDM standard optical component와 상호운용할 수 있다.



Figure 3. Ayar Labs' 16-wavelength SuperNova™ light source

Figure 71: Souce: Ayar Labs

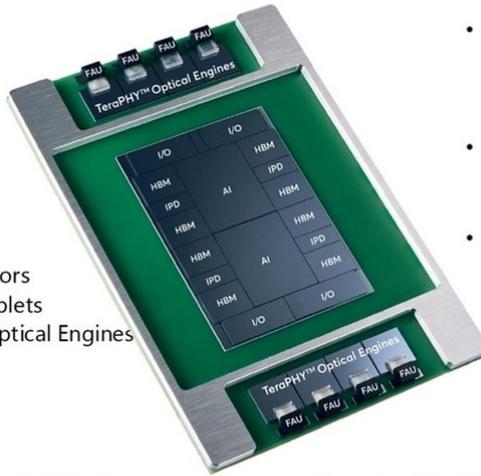
Ayar의 3세대 TeraPHY는 TSMC COUPE를 사용하는 것으로 pivot하며 각 optical engine당 13.5 Tbit/s uni-directional 이상을 제공할 수 있으며, 8개의 optical engine이 아래 ~108Tbit/s의 total package scale-up bandwidth를 제공한다. 이 ~13.5+Tbit/s는 PAM4 Modulation을 사용하여 lambda당 ~200Gbit/s의 bandwidth를 사용하여 달성된다.

Ayar Labs가 정확한 port architecture (즉, DWDM wavelength 수, FAU당 fiber 등)를 공개하지 않았지만, bi-directional optical link의 사용은 Tx와 Rx를 위해 최대 ~64 fiber strand가 필요하고 external laser source에 연결하기 위해 최대 수십 개가 더 필요함을 의미한다. 그러나 Ayar의 전략은 항상 WDM에 집중해 왔으며, 이는 FAU당 total fiber count가 총 32개 정도로 낮을 수 있음을 의미한다. 처음 두 세대와 마찬가지로 3세대 TeraPHY는 optical chiplet이 작게 유지되면서 미래 bandwidth scaling을 위한 vector로 CWDM 또는 DWDM을 가능하게 하기 위해 Microring Modulator를 계속 사용한다.

New XPU: Future Collaboration in AI Accelerator



- 2 Full Reticle AI Accelerators
- 4 Protocol Converter Chiplets
- 8 Ayar Labs TeraPHY™ Optical Engines
- 8 HBM
- IPD



- Optics brought directly on-package
 - High bandwidth, high radix
 - Low latency, low end-to-end energy
- I/O protocol converter chiplets
 - UCIe-A to UCIe-S
 - Scale-up protocol endpoint
- IPD – Integrated Passive Device
 - Improving package step response
 - Customize capacitors

Ayar Labs Optical Engines on a common substrate with Alchip's solutions

Silicon Heart of AI

Figure 72: Source: Ayar Labs, Alchip

Ayar Labs는 또한 Alchip 및 GUC와 협력하여 chiplet을 Alchip 및 GUC의 XPU 솔루션에 통합할 수 있도록 했다. 위의 예는 2개의 reticle size compute die와 8개의 TeraPHY optical engine을 가진 XPU를 보여주며, 이는 최대 108 Tbit/s uni-directional의 bandwidth를 가능하게 할 수 있다.

Hot Chips 2025에서 [Ayar Labs는 결과를 공유했다](#). Slow thermal cycling link test에서 약 5°C/min의 rate로 4시간 이상의 thermal cycling을 보여주며 전체에 걸쳐 강력한 link BER을 입증했다.

System EVT: Thermal Cycling Link Test

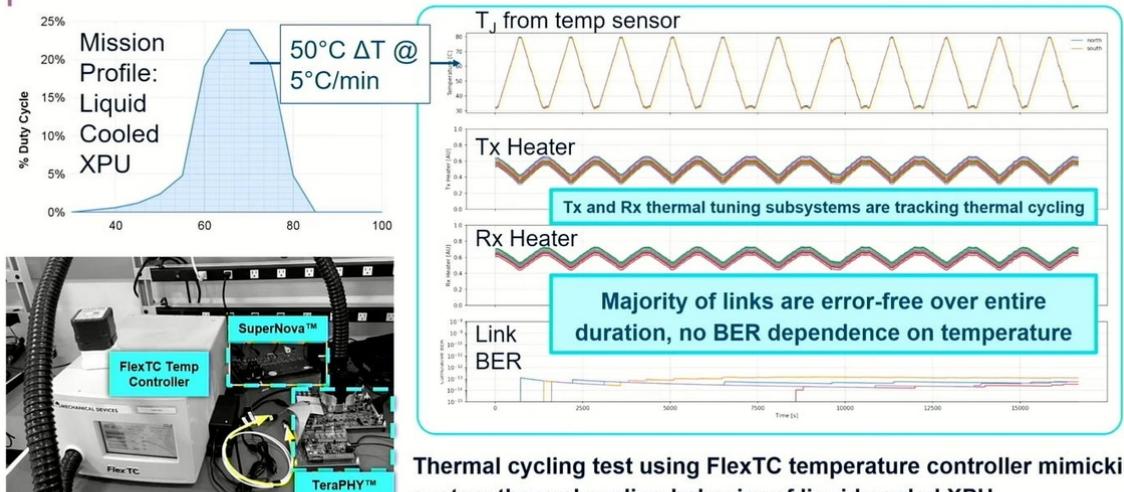
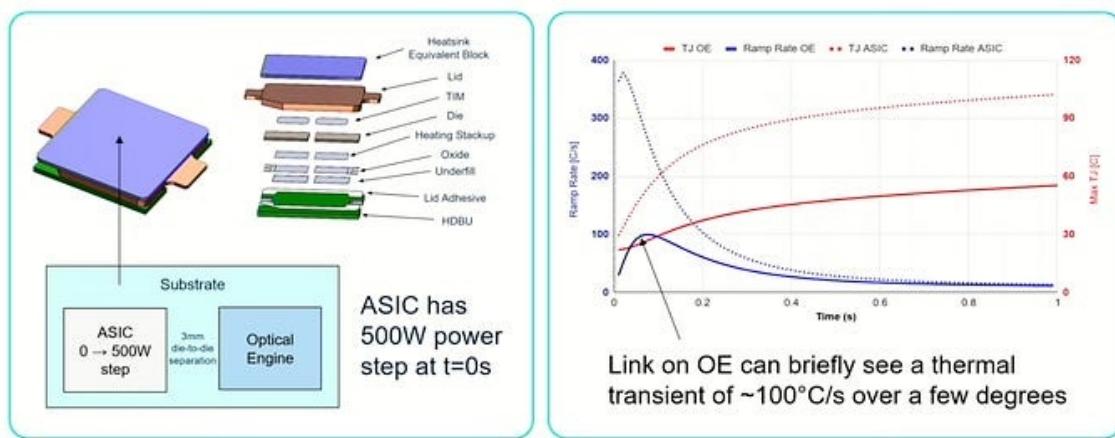


Figure 73: Source: [Ayar Labs](#)

그러나 temperature의 빠른 변화에 대한 MRM의 resilience를 연구하는 것은 긴 기간 동안 넓은 temperature 범위에 걸쳐 link의 stability를 입증하는 것만큼 중요하다. 동일한 Hot Chips 강연에서 Ayar는 실제로 0에서 500W step을 수행할 수 있는 on-package ASIC를 갖는 대신 laser wavelength를 sweeping하여 fast temperature ramp를 emulate하기로 선택한 방법을 설명했다. Control circuit은 ring resonance가 drift하는지 감지한다. 이는 incoming laser가 wavelength를 변경하거나 ring temperature의 변화로 인해 발생할 수 있으므로 temperature의 동등한 변화에 해당하는 rate로 laser wavelength를 sweep한다. 예를 들어 20nm/s sweep은 0.2 second에 걸쳐 64C 변화를 simulate하여 320 C/s에 해당한다. 이 연구는 temperature 변화의 800C/s까지 bit error가 없음을 보여주었다.

Application Study: Co-packaged 500W ASIC Link Thermal Transient



16

Figure 74: Source: Ayar Labs

Ayar Labs는 GlobalFoundries, Intel Capital, NVIDIA, AMD, TSMC, Lockheed Martin, Applied Materials 및 Downing을 포함한 광범위한 strategic backer를 보유하고 있다.

4.5.2 Nubis

Nubis는 최근 2025년 10월 Ciena에 의해 인수되었다. Ayar와 유사하게 Nubis는 고객 host silicon과 통합하기 위한 optical engine chiplet을 제공하지만 single wavelength connection에 중점을 둔다. Nubis는 interoperability – protocol과 mechanical (즉, pluggable) 모두 – 에 집중해 왔으며 이는 그들의 기술 선택을 결정했다. Nubis는 또한 일반적으로 I/O wall을 해결하기 위한 더 광범위한 mission을 가지고 있으며, 그들의 솔루션은 optics와 copper를 모두 포함한다.

기존 optical engine 제품은 Vesta 100 1.6T NPX optical engine이다. 이것은 100G의 16 lane으로 1.6T의 bi-directional bandwidth를 제공하는 socketable module이다. module은 6x7mm의 footprint를 가지고 있다. Nubis는 다른 기업과 달리 modulator의 interoperability, reliability 및 maturity로 인해 주로 MZM을 사용하고 있다. 다른 주요 design 선택은 Nubis가 대부분의 ASIC developer가 이러한 기술을 계속 활용할 것이라고 믿기 때문에 IEEE/OIF standard-compliant electrical interface와 호환되도록 design되었다는 것이다.

Nubis의 주요 차별화 point는 fiber를 couple하는 방법이다. Nubis는 PIC의 표면을 couple하

며, 특히 fiber를 routing하고 정렬하는 데 도움이 되는 얇은 glass piece를 사용한다. optical fiber가 chip의 edge에 연결되는 edge coupling과 달리 Nubis의 2D fiber array 접근 방식은 silicon photonics die의 top에서 optical fiber를 연결하는 것을 포함한다. 아래 다이어그램을 보면: PIC (bottom의 green)는 modulator, photo detector 및 waveguide를 포함하며 top에 EIC가 mounted되어 있다. red pole은 optical fiber이고, optical fiber를 포함하는 block은 fiber holder로 사용되는 glass block (FAU)이다. FAU는 정확한 fiber positioning을 보장하기 위해 block의 top에 laser-drilled hole을 가지고 있다. 2D fiber array를 사용함으로써 그들은 36개의 optical fiber (transmit 용 16개, receive용 16개, laser용 4개)를 PIC에 연결하고 더 적은 fiber에 더 많은 lambda를 얻기 위해 WDM의 필요성을 피할 수 있다. 이것은 Nubis FAU를 현재 shipping 중인 가장 dense한 것 중 하나로 만든다.

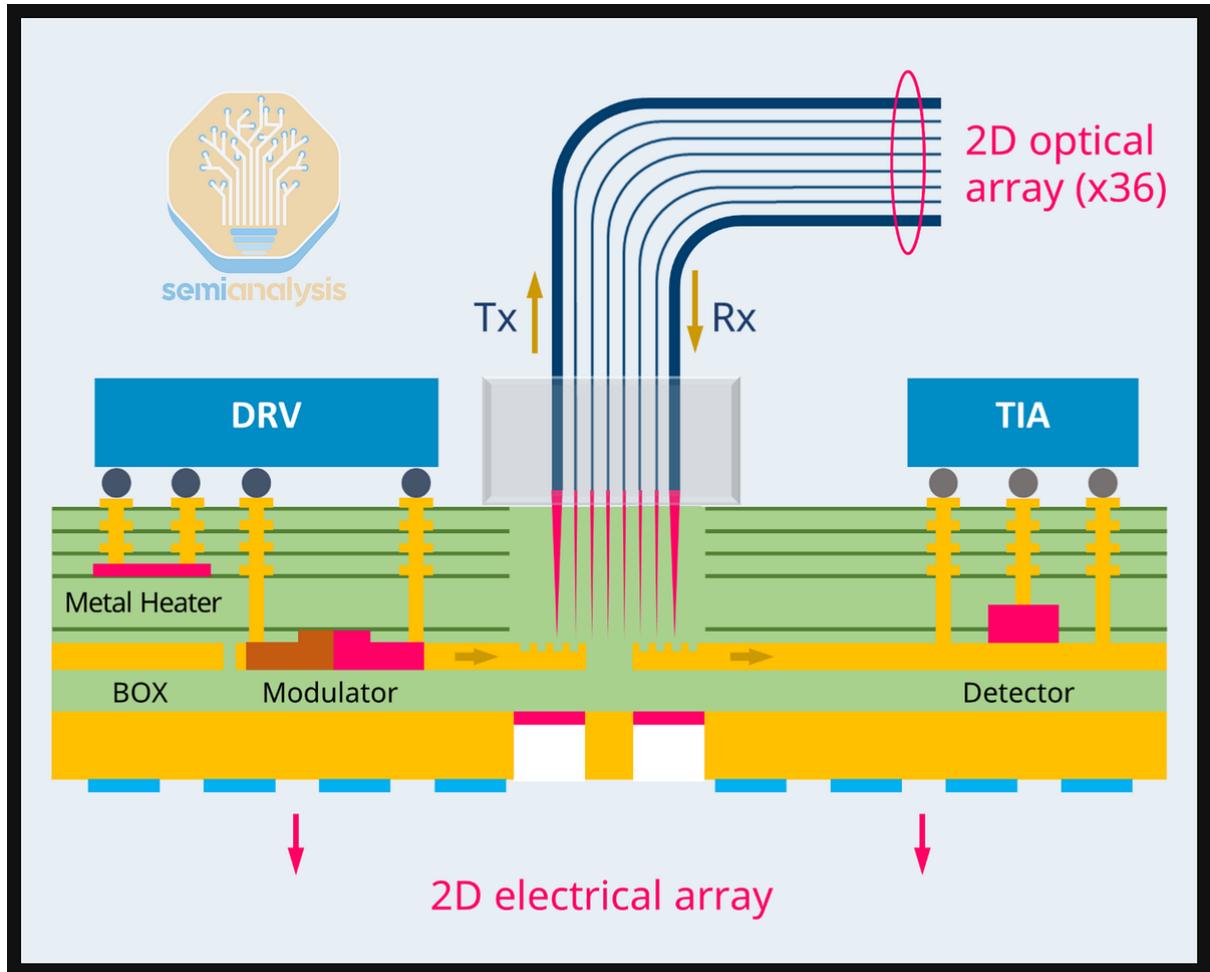


Figure 75: Source: Nubis

2D fiber array는 TSMC와 같은 기업의 roadmap에 있고 vertical coupling의 주요 이점이지만, Nubis 외에는 아무도 아직 이것을 shipping하지 않으며 이것이 그들을 구별하지만, 다른 기업들은 나중에 2D array로 이동할 계획이다. Optical fiber는 위로 올라가고 Sumitomo Electric이 개발한 **FlexBeamGuidE**라고 하는 특수 optical fiber를 사용하여 90-degree angle로 구부러지면서 높은 reliability와 낮은 loss를 나타낼 수 있어 sideways로 구부러진다.

Edge coupling이 아닌 2D array를 사용하는 또 다른 이점은 연결할 수 있는 fiber 수에 의해 물리적으로 덜 제한된다는 것이다. 아래 다이어그램에서 볼 수 있듯이 Nubis의 2D Fiber array 구조를 사용하면 package가 허용하는 경우 여러 row의 optical engine을 ASIC 주위에 배치하여

bandwidth density를 증가시킬 수 있다.

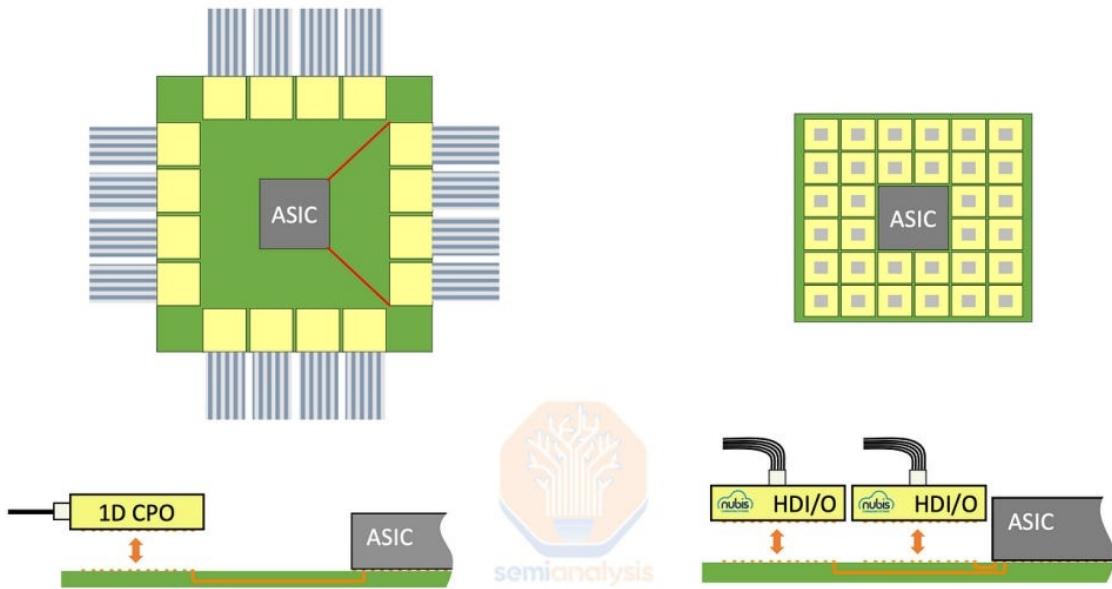


Figure 76: Source: Nubis

2025년 4월 Nubis는 0.5Tbps/mm의 unidirectional beachfront density (electrical host interface density와 일치)를 가진 16 x 200G per lane silicon photonics IC인 차세대 PIC의 availability를 발표했다. 추가로 Nubis는 Samtec Si-Fly HD Co-Packaged copper connector와 snap-in 호환 가능한 32x 200G (6.4T) optical module을 sample할 Samtec과의 partnership를 발표했다. 대안 CPO 접근 방식과 비교하여 이 접근 방식은 공통 copper 및 optical footprint를 가능하게 한다; 이것은 또한 시간이 지남에 따라 CPO를 배포하기 위한 open pluggable ecosystem을 만들 수 있다.

마지막으로 copper에서 Nubis는 OFC에서 200G over copper의 reach를 several meter로 확장할 수 있는 **Active Copper Cable (ACC)**용 linear redriver chip인 Nitro를 발표하고 시연했다. 이것은 Nitro linear redriver를 기반으로 ACC를 구축할 Amphenol과의 partnership으로 수행된다.

4.5.3 Celestial AI

Celestial AI는 AI scale-up network를 위한 optical interconnect 솔루션을 전문으로 하는 IP, 제품 및 system 회사이다. 회사 기술의 주요 목표는 photonic device (modulator, PD, waveguide 등)를 외부 세계와의 interface (FAU를 가진 GC)와 결합된 interposer에 구축하는 것이다. 아래 다이어그램은 Celestial AI가 **Photonic Fabric TM (PF)**라고 부르는 photonics 기반 interconnect 솔루션 suite의 좋은 표현이다.

Celestial AI product portfolio

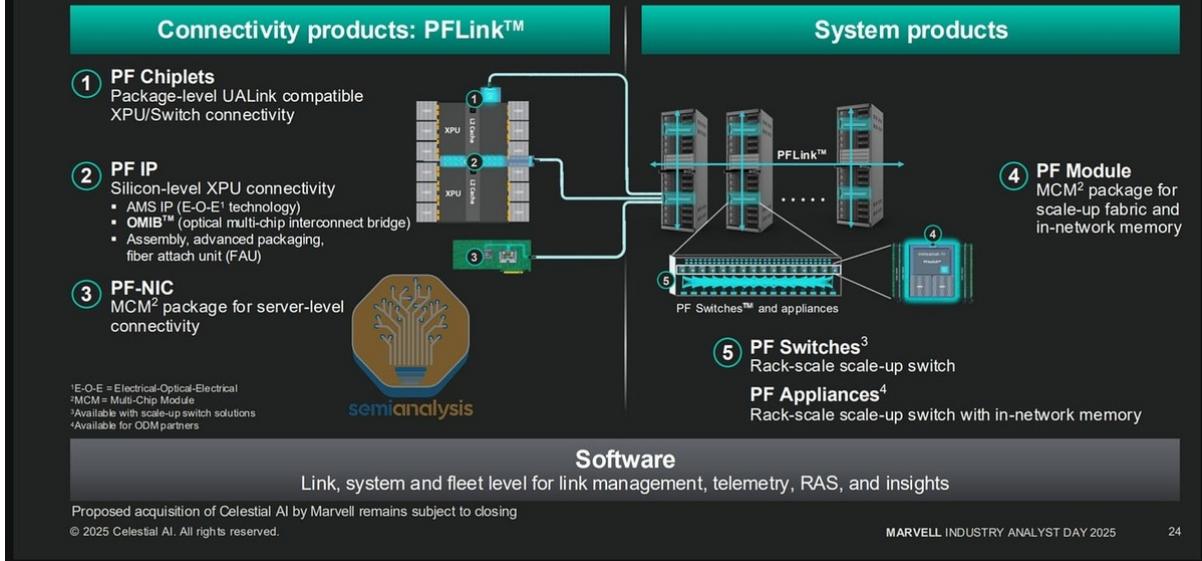


Figure 77: Source: Celestial AI

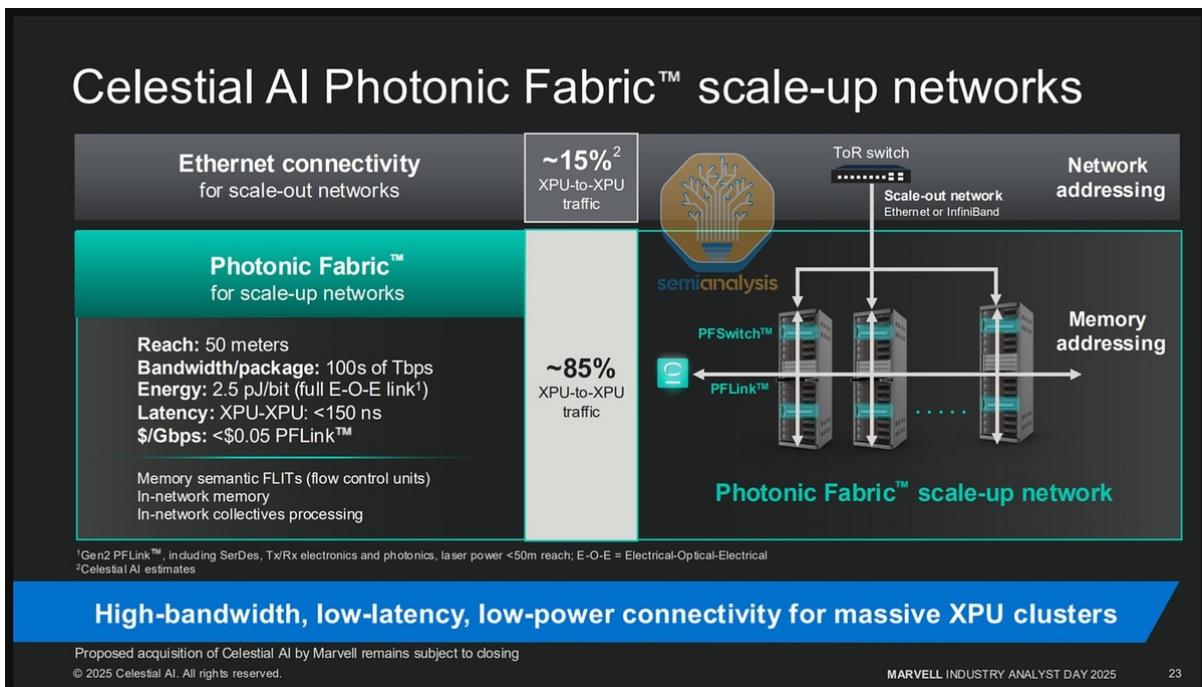


Figure 78: Source: Celestial AI

PF Chiplet은 Universal Chiplet Interconnect Express (UCIE) 및 MAX PHY와 같은 die-to-die interface를 통합하는 TSMC 5nm chiplet으로 XPU-to-XPU, XPU-to-Switch 및 XPU-to-Memory connectivity를 가능하게 한다. 고객이 XPU와 함께 co-package할 수 있으며, electrical SerDes interface를 기반으로 하는 CPO 제품보다 더 높은 bandwidth density와 더 낮은 전력 소비를 제공한다. Celestial AI는 특정 D2D interface 및 protocol을 수용하기 위해 customer별로 이러한 chiplet을 개발한다. 1세대 PF Chiplet은 16 Tbit/s의 bandwidth를 지원하는 반면, 2세대

는 64 Tbit/s를 제공할 것이다.

Optical chiplet은 전통적인 copper trace와 비교하여 강력한 전력 이점을 제공한다. 224G의 linear SerDes를 가진 전통적인 copper cable은 ~5 pJ/bit를 필요로 한다. 2개의 end가 필요하므로 총 전력 소비는 ~10 pJ/bit이다. Celestial AI의 솔루션은 전체 electrical-optical-electrical link에 단지 ~2.5 pJ/bit만 필요하다 (plus external laser를 위한 ~0.7 pJ/bit).

다음으로 Photonic Fabric TM **Optical Multichip Interconnect Bridge TM (OMIB TM)**은 본질적으로 CoWoS-L style 또는 EMIB style packaging 솔루션이다. Interposer의 embedded bridge에 직접 photonics를 추가하여 bridge가 consumption point로 직접 데이터를 이동할 수 있도록 한다. Beachfront constraint에 의해 제한되지 않기 때문에 PF Chiplet보다 더 높은 전체 chip bandwidth를 제공한다.

metal interconnect를 가진 전통적인 interposer 또는 substrate에서 chip의 center에 I/O를 배치하는 것은 비실용적인데, high-density signal congestion으로 인한 과도한 routing 복잡성과 심각한 crosstalk 문제를 만들기 때문이다. 그러나 OMIB optical interposer를 사용하면 Celestial AI는 interposer를 ASIC 바로 아래에 배치하여 shoreline limitation을 우회하고 최소한의 crosstalk로 더 빠르고 효율적인 데이터 이동을 가능하게 할 수 있다.

Optical interposer는 I/O를 chip의 어디에나 배치할 수 있는데, **optical** waveguide는 거리에 걸쳐 무시할 수 있는 signal degradation을 경험하여 우리가 알고 있는 shoreline의 전통적인 제약을 제거하기 때문이다. 또한 crosstalk를 제거하는데, 다른 waveguide의 빛 signal은 densely packed copper trace의 electrical signal처럼 interfere하지 않기 때문이다. 왜냐하면 cladding의 외부에 작은 evanescent field만 있는 waveguide core 내에 매우 confined되어 있기 때문이다. 이러한 I/O design 및 배치의 ground-up 재설계는 optics가 제공하는 잠재력을 완전히 활용한다.

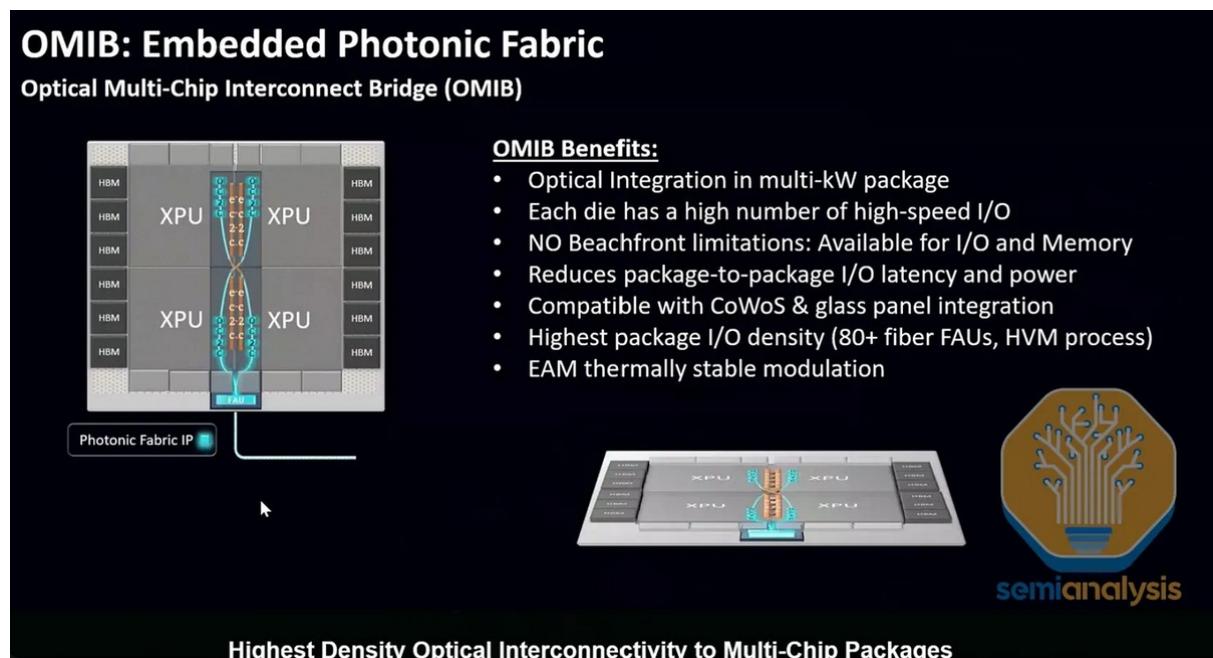


Figure 79: Source: Celestial AI

Celestial AI Photonic Fabric™ chiplet

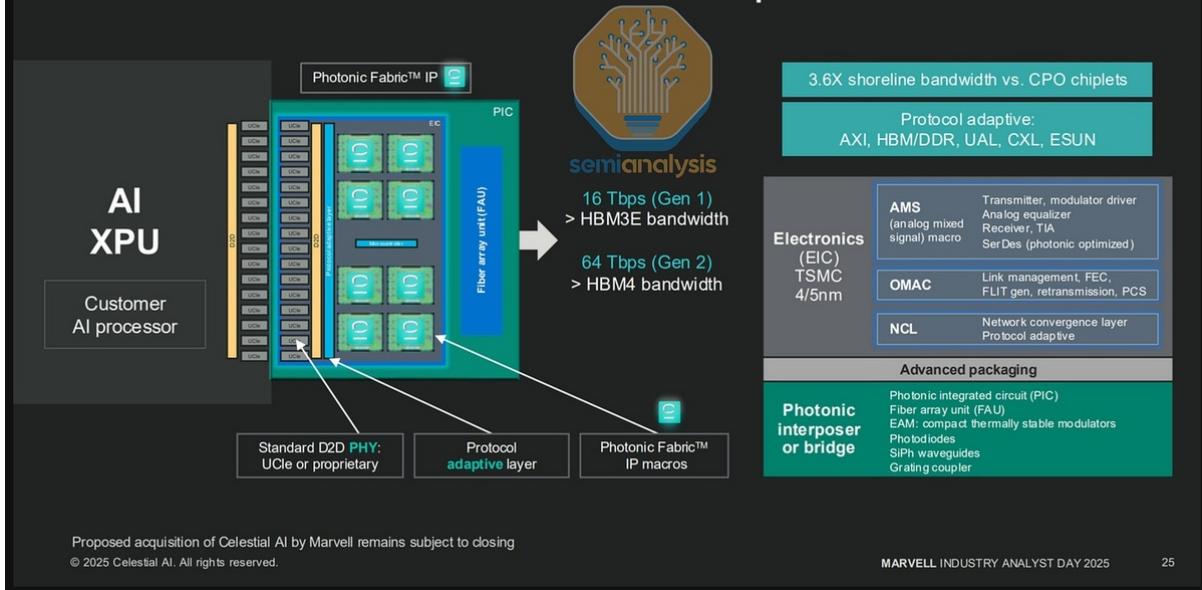


Figure 80: Source: Celestial AI, Marvell

Optical interposer의 아이디어 또는 advanced package를 통한 optical signal channeling은 둘다 logic chip 아래로 optical signal을 routing하여 shoreline constraint를 피한다는 점에서 Lightmatter의 솔루션과 몇 가지 유사성을 가지지만 몇 가지 주요 차이점이 있다. Celestial AI는 silicon bridge (CoWoS-L silicon bridge처럼 생각)와 유사한 photonic bridge를 채택하는 반면 Lightmatter는 여러 개별 chip 아래에 위치하는 큰 multi-reticle photonic interposer를 사용한다. Lightmatter의 concept는 scope에서 더 야심적이다. M1000 3D Photonic Superchip에서 4,000 mm² interposer size를 목표로 하는 동시에 interposer 내에서 optical circuit switching을 지원하고 매우 높은 114 Tbit/s의 total aggregate bandwidth를 목표로 한다.

Celestial AI Photonic Fabric™ Link: PFLink™

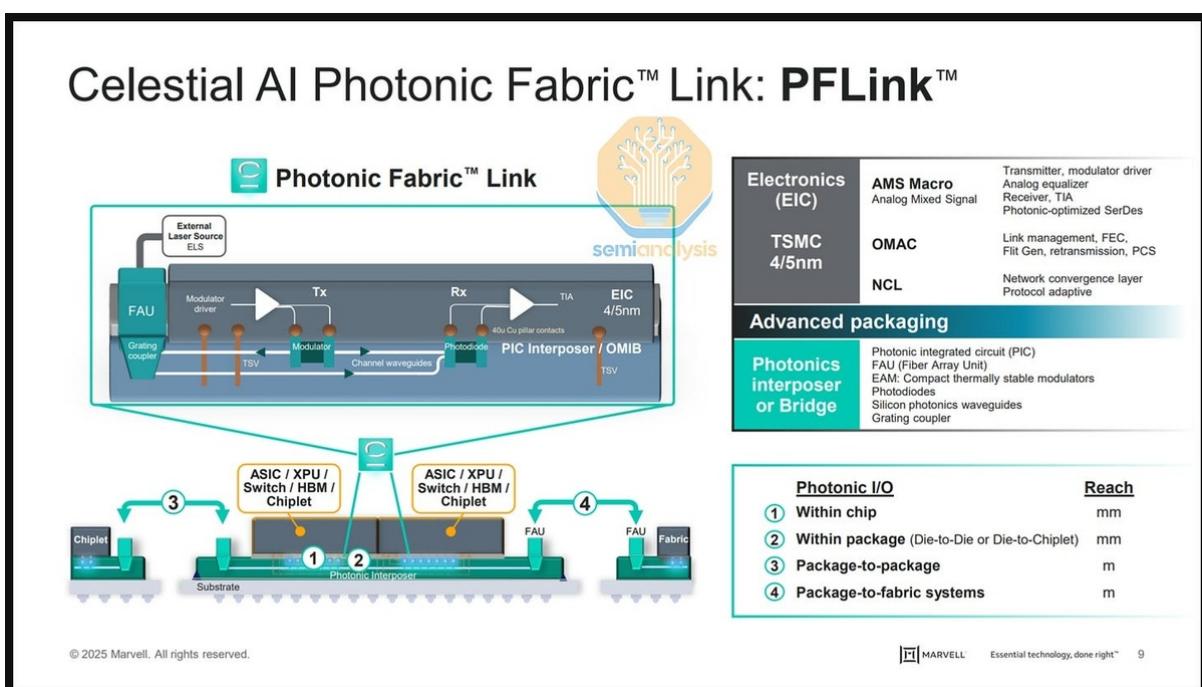


Figure 81: Source: Celestial AI, Marvell

마지막으로 Celestial AI는 각 ASIC당 7.2T scale-up bandwidth에서 16개의 ASIC를 연결 할 수 있는 115.2T total bandwidth를 가진 TSMC 5nm에 구축된 in-network memory를 가진 high-bandwidth, low-latency scale-up fabric인 **Photonic Fabric TM Memory Appliance (PFMA)**를 제공한다. 주목할 만하게도 PFMA는 chip의 center에 위치한 on-die optical I/O를 가진 세계 최초의 silicon device로 memory controller를 위한 희소한 perimeter physical I/O를 남긴다. 이것은 PFMA를 KVCache offloading을 위한 host CPU memory와 storage 사이의 **warm memory tier**로 위치시킨다.

Celestial AI 기술의 주요 차별화 요소는 **Electro Absorption Modulator (EAM)**의 사용이다. 이 문서의 Part 3은 EAM이 어떻게 작동하는지 더 자세히 설명하고 EAM과 관련된 이점과 trade-off를 논의한다. EAM의 장단점을 이해하는 것이 Celestial의 go to market을 이해하는 데 핵심이기 때문에 여기서 이 논의의 대부분을 반복한다.

EAM은 MRM 및 MZI와 비교하여 아래의 여러 이점을 가지고 있다:

- 분명히 – EAM과 MRM 모두 temperature variation에 대해 둘 다를 안정화시키기 위해 작동하는 control logic과 heater를 가지고 있지만, EAM은 근본적으로 temperature에 대한 sensitivity가 더 낮다. MRM과 비교하여 EAM은 50°C 이상에서 훨씬 더 나은 thermal stability를 가지는 반면 MRM은 temperature에 매우 민감하다. MRM의 일반적인 stability 인 70-90 pm/C는 2°C variation의 resonance를 0.14nm만큼 이동시킨다는 것을 의미하며, 이는 MRM 성능이 붕괴되는 0.1nm resonance shift를 훨씬 넘는다. 이와 대조적으로 EAM은 최대 35°C의 순간 temperature shift를 견딜 수 있다. 이 tolerance는 특히 Celestial AI의 접근 방식에서 중요한데, 그들의 EAM modulator가 수백 watt의 전력을 dissipate하는 high-XPU power compute engine 아래의 interposer 내에 위치하기 때문이다. EAM은 또한 약 80°C의 높은 ambient temperature 범위를 견딜 수 있으며, 이는 XPU 옆에 있고 그 아래에 있지 않은 chiplet application에 적용될 수 있다.
- MZI와 비교하여 EAM은 크기가 훨씬 작고 전력을 덜 소비하는데, MZI의 상대적으로 큰 크기가 high voltage swing을 필요로 하여 0-5V의 swing을 달성하기 위해 SerDes를 증폭하기

때문이다. Mach Zender Modulator (MZM)는 ~12,000mm² 정도이고, EAM은 약 250mm² (5x50mm)이며 MRM은 25mm²에서 225mm² 사이 (diameter 5-15mm²)이다. MZI는 또한 그러한 큰 device를 desired bias로 유지하는 데 필요한 heater에 대해 더 많은 전력 사용을 필요로 한다.

반면에 CPO를 위해 GeSi EAM을 사용하는 데 몇 가지 단점이 있다:

- MRM 및 MZI와 같은 Silicon 또는 Silicon Nitride에 구축된 물리적 modulator 구조는 GeSi 기반 device보다 훨씬 더 큰 endurance와 reliability를 가진 것으로 인식되어 왔다. 실제로 많은 이들이 Germanium 기반 device를 다루고 통합하는 어려움을 고려할 때 GeSi 기반 device의 reliability에 대해 걱정하지만, Celestial은 본질적으로 Photodetector의 역인 GeSi 기반 EAM은 오늘날 transceiver에서 photodetector의 ubiquity를 고려할 때 reliability와 관련하여 알려진 양이라고 주장한다.
- GeSi modulator의 band edge는 자연적으로 C-band (즉, 1530nm-1565nm)에 있다. 이것을 O-band (즉, 1260-1360nm)로 이동시키기 위해 quantum well을 design하는 것은 매우 어려운 engineering 문제이다. 이는 GeSi 기반 EAM이 book-ended CPO system의 일부를 형성할 가능성이 높으며 open chiplet 기반 ecosystem에 참여하는 데 쉽게 사용될 수 없음을 의미한다.
- C-band laser source 주변의 laser ecosystem을 구축하는 것은 O-band CW laser source 주변의 잘 개발된 ecosystem을 사용하는 것과 비교할 때 scale의 diseconomy를 가질 수 있다. 대부분의 datacom laser는 O-band를 위해 구축되지만 Celestial은 상당한 volume의 1577nm XGS-PON laser가 제조된다고 지적한다. 이것들은 일반적으로 consumer fiber to the home 및 business connectivity application에 사용된다.
- SiGe EAM은 MRM과 MZI 모두에 대한 3-5dB와 비교하여 약 4-5dB의 insertion loss를 가진다. MRM은 다른 wavelength를 직접 multiplex하는 데 사용될 수 있는 반면, EAM은 CWDM 또는 DWDM을 구현하기 위해 별도의 multiplexer를 필요로 하여 잠재적 loss budget에 약간 추가한다.

전반적으로 Celestial AI는 custom link를 혁신하기 위해 노력해 왔다. 그들은 어떤 gearbox component에도 의존하지 않으며 더 나은 latency와 power efficiency를 제공하고 다양한 type의 protocol에 adaptive하다. 앞서 언급했듯이 Celestial AI는 주로 modulation을 위해 EAM을 사용하는 유일한 주요 player이다. 한 가지 주요 implication은 그들이 EAM design을 foundry에 통합하는 데 앞으로 일부 작업이 있을 것이라는 것인 반면, 다른 CPO 기업은 MRM 및 관련 heater가 이미 PDK의 일부인 TSMC COUPE에 의존할 수 있다.

Celestial AI Silicon Photonics Differentiation

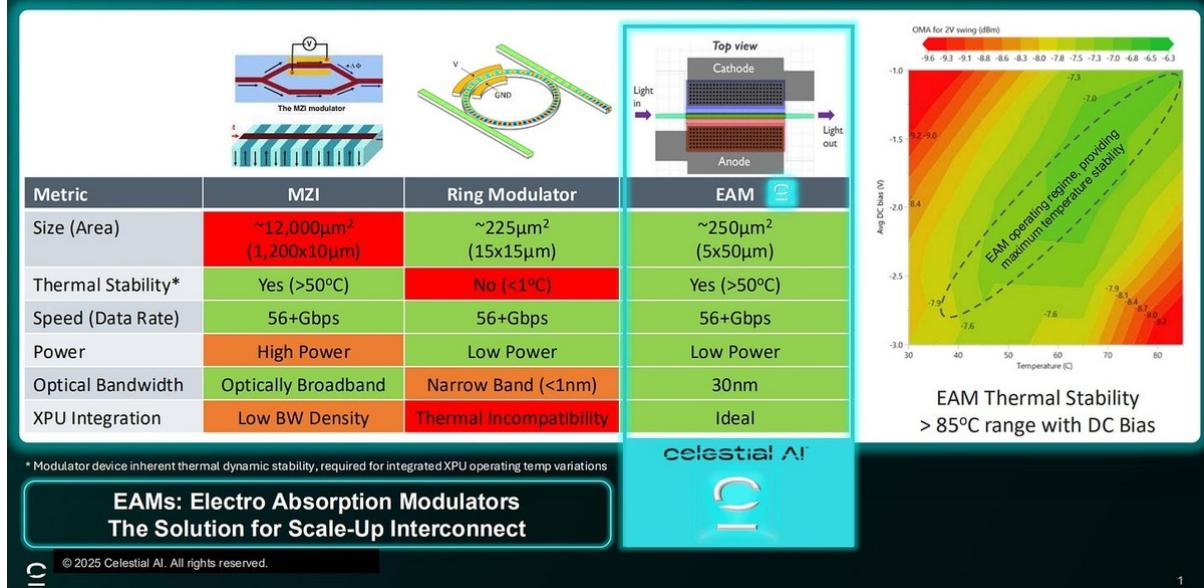


Figure 82: Source: Celestial AI

단기적으로 Celestial AI는 chiplet 출시에 대한 야심찬 timeline에 commit하고 있다. Marvell은 transaction summary에서 Marvell의 Fiscal Year 2028 말 (즉, F1/28)인 2028년 1월 말 Celestial로부터의 estimated revenue run rate가 \$500 million에 도달할 것으로 예상된다고 발표했다. Barclay's Global Technology Conference에서 그들은 이 run rate가 Calendar Year 2028 말까지 (CY28의 대부분이 2029년 1월에 끝나는 Marvell의 FY에 해당. 즉, F1/29) \$1B로 두 배가 될 것으로 예상된다고 추가했으며, 이는 제품이 commercial viability를 달성하기까지 지금부터 2027년 말까지 2년 기간을 의미한다.

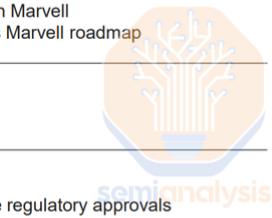
Deal term의 일부로 Celestial AI의 equity holder에 대한 추가 \$2.25B의 payout은 회사가 2029년 1월 (Marvell의 Fiscal Year 2029의 말, 즉 F1/29)까지 최소 \$2.0 billion의 cumulative revenue를 달성하는 것에 달려 있다. 전체 payout에 대한 첫 번째 milestone은 payout의 1/3를 위해 2029년 1월까지 \$500M의 cumulative revenue를 달성하는 것이다. F1/29를 exit하며 예상되는 \$1B revenue run rate는 earn-out 금액의 절반이다. Celestial이 \$2B earn-out target을 달성하기 위해 order book에 추가 고객을 추가해야 함을 의미한다.

Celestial AI transaction summary

Transaction consideration	<ul style="list-style-type: none"> \$3.25 billion payable at closing, consisting of \$1.0 billion in cash, as well as approximately 27.2 million shares of Marvell common stock, having a value of \$2.25 billion Up to \$2.25 billion in contingent consideration tied to milestone achievements. The first milestone representing one-third of the earnout consideration, will be achieved if Celestial AI reaches cumulative revenue of at least \$500 million by the end of Marvell's fiscal¹ year 2029. The full earnout would be paid if Celestial AI's cumulative revenue by the end of Marvell's fiscal¹ year 2029 exceeds \$2.0 billion.
Financial impact	<ul style="list-style-type: none"> Accretive to non-GAAP earnings 2H fiscal¹ 2028 Expect revenue to reach \$500 million annualized run rate by Q4 fiscal¹ 2028 \$1 billion annualized run rate revenue by Q4 fiscal¹ 2029 Adds high-value Photonic Fabric™ optical scale-up platform
Management and governance	<ul style="list-style-type: none"> David Lazovsky (CEO), Preet Virk (COO), Philip Winterbottom (CTO) to join Marvell Celestial AI leadership to help drive integration of Photonic Fabric™ across Marvell roadmap
Financing	<ul style="list-style-type: none"> Combination of cash and stock at close Additional contingent payments based on performance milestones Financed with cash on hand
Expected close	<ul style="list-style-type: none"> Expected to close in the first quarter of calendar 2026 Subject to satisfaction of customary closing conditions, including applicable regulatory approvals

¹Marvell's fiscal year is the 52- or 53-week period ending on the Saturday closest to January 31; as an example, FY2026 refers to the period February 1, 2025, through January 31, 2026

© 2025 Marvell. All rights reserved.



semianalysis

MARVELL Essential technology, done right™ 12

Figure 83: Source: Celestial AI, Marvell

Celestial AI 인수와 관련하여 Marvell은 2025년 12월 2일 [8-K report](#)를 제출하여 2030년 12월 31일까지 \$87.0029의 exercise price로 Amazon warrant를 발행했다. 이러한 warrant는 2030년 12월 31일까지 간접적으로 또는 직접적으로 Amazon의 Photonic Fabric 제품 구매를 기반으로 vest하며, 이것이 2027년 후반에 ramp하기 시작하는 AWS의 Trainium^o target 제품임을 강력히 시사한다. Marvell의 Industry Analyst Day에서 Celestial AI는 주요 hyperscaler가 해당 hyperscaler의 차세대 processor에서 volume production으로 이동할 advanced AI system을 위한 optical interconnectivity를 위해 그들을 선택한 방법을 논의했다. 이것은 transaction summary의 earn-out timing 및 product revenue guidance와 함께 Celestial AI가 Trainium 4 내에 솔루션을 배포하는 것을 목표로 하고 있음을 시사한다.

Item 8.01 Other Events.

On December 2, 2025, the Company announced that it entered into a definitive agreement to acquire Celestial AI Inc. In connection with this acquisition, which is subject to customary closing conditions, the Company entered into a Transaction Agreement with Amazon.com, Inc. ("Amazon", including its affiliates) and related Warrant, under which the Company issued to Amazon a warrant (the "Warrant") to acquire up to 1,045,171 shares (the "Warrant Shares") of Company common stock (the "Common Stock"). The Warrant Shares vest based on Amazon's purchases of photonic fabric products, indirectly or directly, through December 31, 2030.

Subject to certain conditions, including vesting, the Warrant may be exercised, in whole or in part and for cash or on a net exercise basis, at any time before December 2, 2031, at a purchase price per share of Common Stock equal to \$87.0029 (the "Exercise Price"). The Exercise Price and the Warrant Shares issuable are subject to customary antidilution adjustments.

The Transaction Agreement includes customary representations, warranties and covenants of the parties and sets forth certain provisions relating to Amazon's equity interest in the Company.

The Warrant and the Warrant Shares have not been registered under the Securities Act, in reliance on the exemption from registration provided by Section 4(a)(2) of the Securities Act and rules and regulations of the U.S. Securities and Exchange Commission promulgated thereunder.

Figure 84: Source: Marvell SEC Filings

chiplet에 연결될 16Tbit/s Photonics Link를 중심으로 oriented될 시장에 나올 첫 번째 scale-up 솔루션에 대해 더 자세히 설명하여 Celestial AI에 대한 논의를 마무리하자. FAU는 grating coupler를 통해 channel waveguide에 연결된다. Scale-up switch ASIC (아마도 Marvell의 115.2T scale-up ASIC)은 Photonics link와 PF chiplet을 통해 XPU에 optically하게 연결될 것이다. Celestial

은 초기 go-to-market revenue의 대부분이 chiplet에 의해 기여될 것으로 예상하지만, system 회사로 자신을 포지셔닝하며 이 첫 번째 scale-up networking 솔루션 이후 시장에 나올 수 있는 여러 optical 기반 memory expansion 솔루션을 제안했다.

optics를 사용하여 여러 switch layer를 통해 scale-up world size를 증가시키는 것은 새로운 concept이 아니지만, 물론 아직 productized되는 것에 가까이 가지 못했다. 그러한 concept는 GB200의 NVL576 concept를 mirror하는 topology를 가질 수 있으며, 여기서 2개의 switch layer가 있고 각 switch layer는 OSFP transceiver module 및 optical fiber를 통해 다른 layer에 연결된다. 여러 switch layer를 사용하는 Celestial AI의 접근 방식은 유사하지만 실제 transceiver의 사용을 skip 한다.

그러나 NVL576 concept와의 가장 큰 차이점은 scale-up ASIC가 router와 memory endpoint 모두로 double할 수 있는 반면 NVSwitch는 GPU 간 high-bandwidth link만 route한다는 것이다. 이것은 중요한 구분인데, Celestial AI의 pitch는 scale-up 솔루션이 XPU에 부착할 수 있는 HBM stack 수를 제한하는 silicon beachfront constraint를 sidestep할 수 있다는 것이기 때문이다.

이를 달성하기 위해 XPU에 부착된 HBM stack은 shared HBM pool인 Photonic Fabric에 연결되는 chiplet으로 교체된다. Shared HBM pool은 각각 하나의 port로 구성된 16개의 Photonic Fabric ASIC로 구성된 2U-rack-mountable system인 **Photonic Fabric Appliance (PFA)**이다. 각 ASIC는 2개의 36GB HBM3e memory와 8개의 external DDR5와 통합된 2.5 layer package이다.

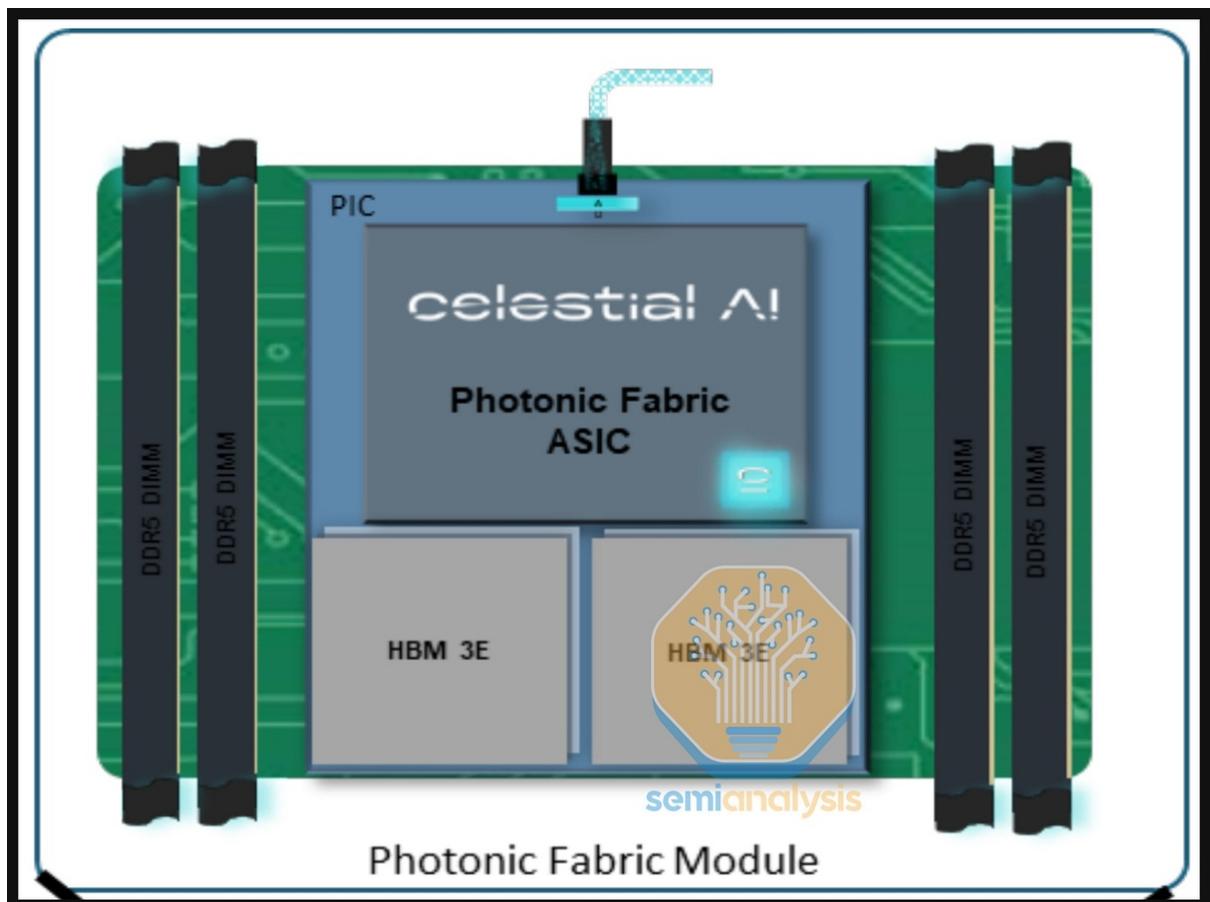


Figure 85: Source: [Celestial AI](#)

Optical I/O (Photonic Fabric IP)는 beachfront가 아닌 ASIC의 middle에 mount되어 다른 use case를 위해 shoreline을 free한다.

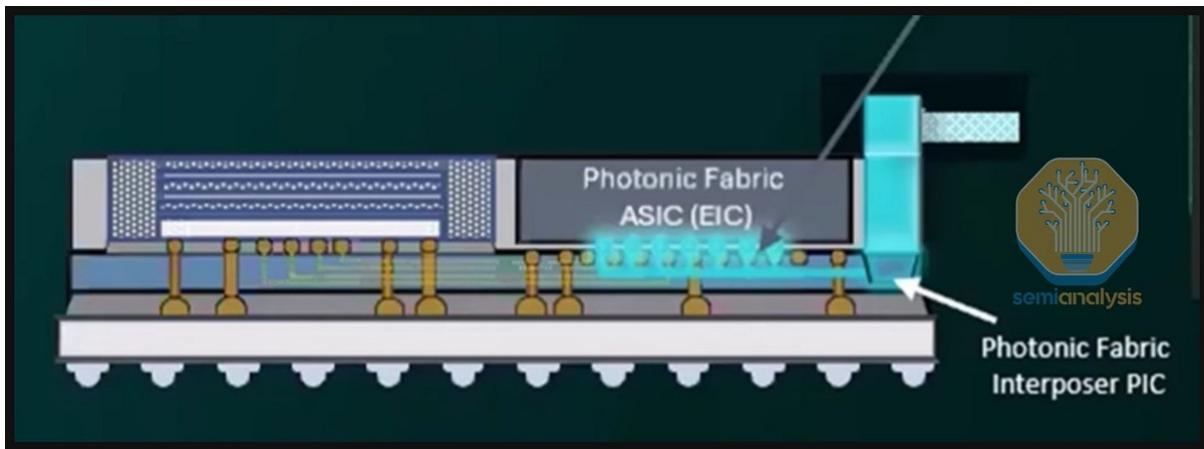


Figure 86: Source: [Celestial AI](#)

Zoom out하면 각 PFA module은 최대 16개의 XPU를 지원할 수 있는 16-radix switch이다. 각 XPU가 모든 16 port로 fan-out하는 대신 all-to-all connectivity는 switch box 내부에서 발생하며, 여기서 각 Switch ASIC에 연결된 **Fiber Attach Unit (FAU)**가 16개의 switch I/O 각각으로 fan out한다. 따라서 모든 XPU는 box 외부의 하나의 switch port에 하나의 Fiber link만 가진다.

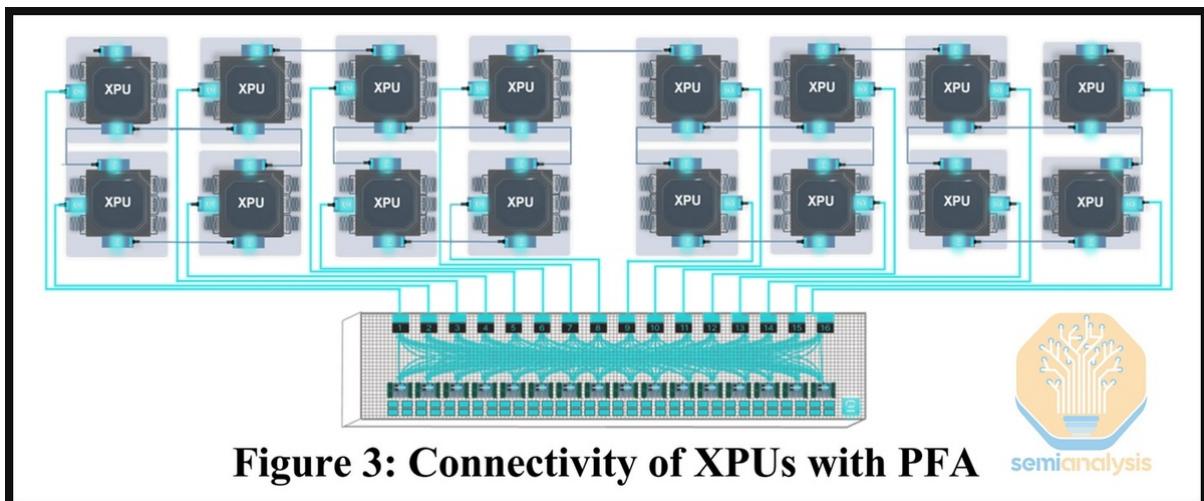


Figure 3: Connectivity of XPUs with PFA

Figure 87: Source: [Celestial AI](#)

Memory를 XPU 외부에 배치하고 shared switching interface 내에 배치함으로써 데이터가 집계되고 subsequently all-reduce communication collective에서 모든 XPU에 의해 shared memory pool에서 접근된다.

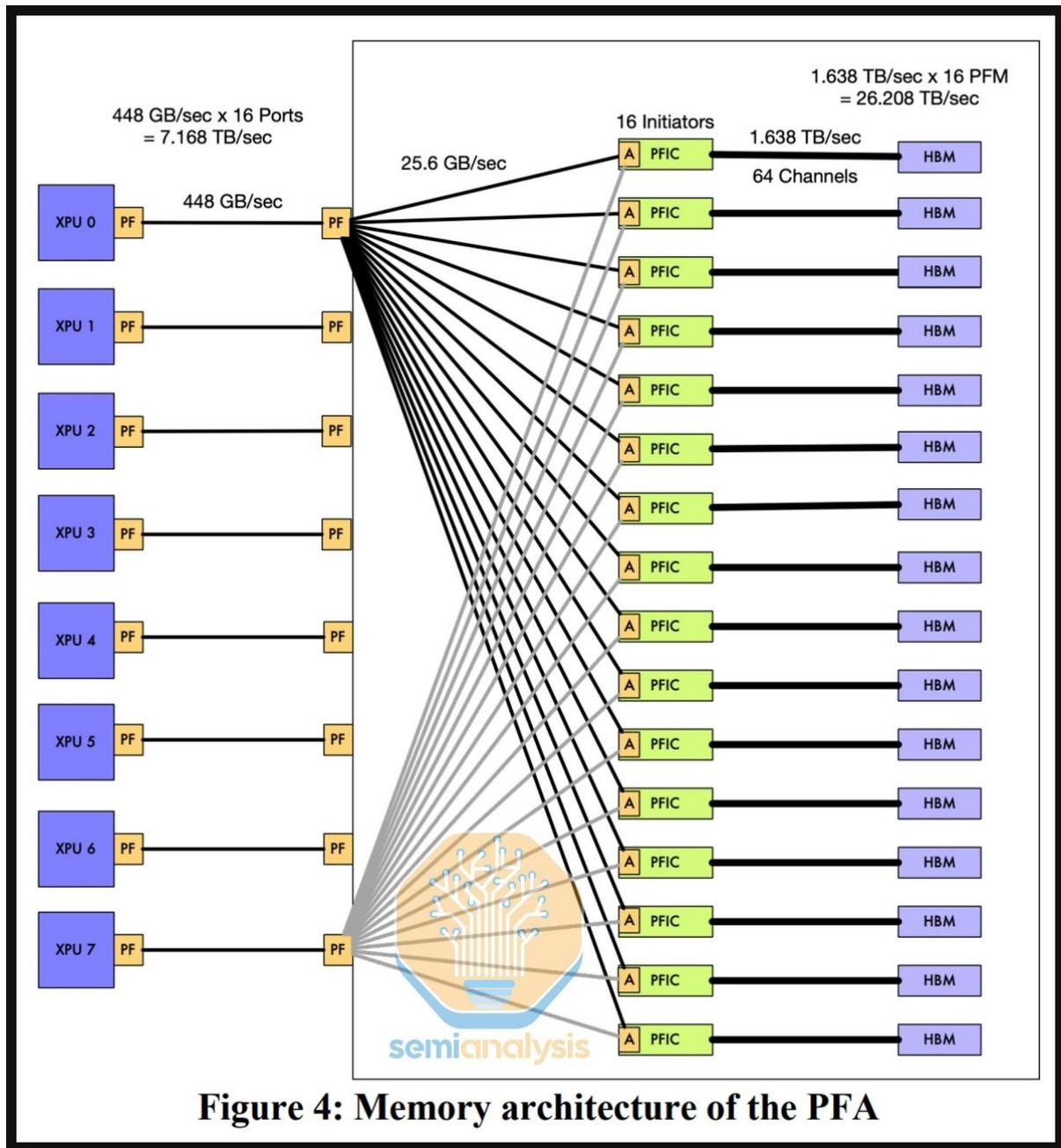


Figure 4: Memory architecture of the PFA

Figure 88: Source: [Celestial AI](#)

4.5.4 Lightmatter

Lightmatter는 Optical Interposer 제품인 Passage™ M1000 3D Photonic Superchip으로 잘 알려져 있지만 CPO roadmap의 다양한 단계에 맞는 여러 솔루션을 도입하고 있으며, 여러 chiplet이 TSMC에서 tape out되고 있다.

시장에 나올 첫 번째 솔루션은 2026/2027년 textbfNear Packaged Optics (NPO)를 위한 optical engine이 될 것이다. NPO 솔루션에서 Optical Engine은 baseboard에 soldering되며, copper가 XPU의 LR SerDes를 Optical Engine에 연결한다. Lightmatter의 optical engine은 FAU당 40 fiber strand를 가진 최대 3개의 FAU를 지원하여 총 120 fiber strand를 제공한다. NPO 전략은 CPO 채택에서 Hyperscaler의 첫 번째 단계가 먼저 NPO로 operational 경험을 얻는 것이라는 아이디어에 기반하며, 이는 Hyperscaler가 CPO에 commit할 필요가 없기 때문에 제품을 derisk한다.

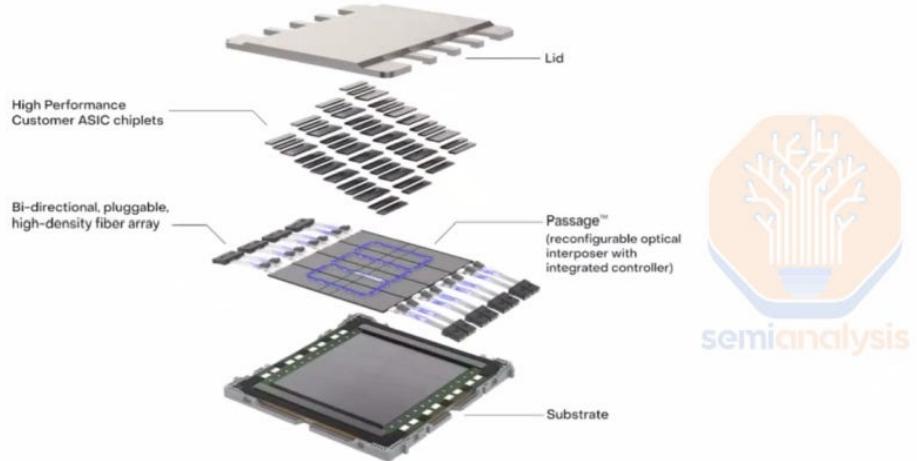
왜냐하면 궁극적으로 XPU 또는 switch의 LR SerDes와 interface하기 위해 optical 또는 copper scale-up 솔루션을 사용하도록 선택할 수 있기 때문이다.

Lightmatter의 optical engine 솔루션은 TSMC COUPE 및 GF 45nm SPCLO process를 기반으로 하므로 많은 scaling vector가 table에 있다. 100Gbaud PAM4를 통해 lane당 200Gbit/s (uni-directional)를 제공하는 것 외에도 DWDM8로 PAM4에서 200Gbit/s를 지원하거나 fiber당 3.2T를 달성하기 위해 DWDM16을 사용하여 PAM4에서 100Gbit/s를 지원할 수 있다.

일부 다른 CPO 기업은 merchant laser source ecosystem을 사용하기로 선택한 반면, Lightmatter는 현재 sampling 중인 GUIDE로 알려진 자체 external laser source를 개발했다. 다른 laser source가 InP wafer를 singulate하여 discrete laser diode를 생성하는 반면, GUIDE는 업계 최초의 **Very Large Scale Photonics (VLSP)** laser로 최대 50 Tbit/s의 bandwidth를 지원하기 위해 단일 silicon chip에 수백 개의 InP laser를 통합하는 새로운 class의 laser이다. Lightmatter는 이러한 많은 InP laser를 관리하기 위한 고유한 control 기술을 가져온다고 주장하며, 이는 또한 InP laser 수를 overprovisioning하고 여전히 작동하는 diode로 swap하여 **self repair**를 허용함으로써 전체 reliability를 증가시키는 이점을 가진다. 800G의 144 port를 특징으로 하는 NVIDIA Quantum-X CPO Switch는 18개의 ELS를 필요로 하며, Lightmatter는 2개의 GUIDE laser source가 동일한 전체 bandwidth 요구사항을 충족할 수 있다고 주장한다.

Lightmatter는 2027년과 2028년에 본격적으로 CPO 솔루션을 제공하는 COUPE roadmap과 align한 다음 2029년 이후 flagship PassageTM M1000 솔루션에 집중하는 것을 목표로 한다. Lightmatter의 M1000 3D Photonic Superchip은 host compute engine 아래에 배치되고 electrical에서 optical로 signal conversion을 처리하는 4,000 mm² optical interposer이다. M1000은 **Supercomputing 2025 (SC25)**에서 live rack-scale demonstration으로 시연되었으며, Lightmatter는 이를 reference design으로 사용 가능하게 했다. Passage는 XPU와 optical engine 사이의 electrical signal과 전력을 전달하기 위해 TSV를 사용하고 둘을 연결하기 위해 SerDes를 사용한다. ASIC를 optical interposer에 직접 배치함으로써 Passage는 크고 전력을 많이 소비하는 SerDes의 필요성을 제거한다. 대신 114Tbit/s의 total I/O bandwidth를 가능하게 하기 위해 1,024개의 compact, lower-power SerDes (기존 SerDes보다 ~8배 작음)를 활용한다 (각 SerDes는 112Gbit/s에서 작동). ASIC를 optical interposer의 top에 직접 배치함으로써 chip shoreline constraint도 완화된다.

PASSAGE™



PASSAGE ENABLES PER-CHIP PERFORMANCE SCALING.

MASSIVE RADIX, BANDWIDTH ENABLING 128K NODES + FOR NEXT-GEN LLMS

WE PARTNER WITH CHIP COMPANIES AND HYPERSCALERS.

Figure 89: Source: Lightmatter

System은 redundancy를 관리하는 built-in **Optical Circuit Switch (OCS)**를 통합한다. 하나의 communication route가 실패하면 traffic이 alternate path를 통해 reroute되어 그러한 대규모 system에서 중단 없는 operation을 보장할 수 있다. 추가로 neighboring tile은 electrically하게 함께 stitched되어 PCIe와 같은 interface를 사용하여 electronically하게 통신할 수 있다.

Passage는 각각 resistive heater와 통합된 ~15 μm의 diameter를 가진 MRM을 사용하고 56 Gbit/s NRZ modulation을 달성한다. Module은 각각 최대 16 color (wavelength)를 전달할 수 있는 16개의 horizontal bus로 구성된다. 이러한 color는 200 GHz grid에서 fiber당 16 wavelength를 제공하는 GUIDE에 의해 공급될 것이다.

Passage는 DWDM을 통해 각각 16 wavelength을 unidirectionally하게 (또는 8 wavelength를 bi-directionally하게) 전달하는 256개의 optical fiber를 활용하여 fiber당 1 Tbit/s에서 1.6 Tbit/s의 bandwidth를 제공한다. Yield를 개선하기 위해 그들은 chip에 부착된 fiber 수를 최소화하여 복잡성과 manufacturing 과정을 줄였다. 추가로 그들은 결함이 있는 fiber를 panel에서 쉽게 분리하고 교체할 수 있는 fiber attach system을 구현하여 reliability와 serviceability를 향상시켰다. 아래 표는 Passage가 현재 지원하는 다른 mode를 반영한다.

Bandwidth	Modulation	Number of Wavelengths	Transmission Type	CWDM/DWDM
56 Gbps	NRZ	16	Bi-directional	DWDM
56 Gbps	NRZ/PAM4	16	Uni-directional	DWDM
112 Gbps	PAM4	16	Uni-directional	DWDM
224 Gbps	PAM4	4	Bi-directional	CWDM
224 Gbps	PAM4	4	Uni-directional	CWDM

Figure 90: Source: [Lightmatter](#)

PASSAGE와 관련된 주요 debate 중 하나는 optical interposer가 매우 뜨거운 XPU 바로 아래에 위치한다는 점을 고려할 때 사용된 MRM의 thermal stability이다. 비교하자면, CPO에 대한 다른 접근 방식은 modulator가 XPU 바로 아래에 배치되는 것을 구상하지 않으므로 thermally하게 관리하기가 더 쉽다. 이 point에 대응하여 Lightmatter는 PASSAGE의 MRM에 사용되는 control loop이 초당 2,000C의 excursion을 처리할 수 있고 0에서 105C 사이의 temperature를 처리할 수 있다고 설명했다. 즉, 60에서 80C temperature transition^{o]} optical link를 방해하지 않고 10ms 내에 발생할 수 있다.

[SC25 demonstration](#) video는 25C에서 105C 사이의 temperature variation의 illustration을 묘사하여 광범위한 operating temperature를 보여주었지만, 이 특정한 것은 80C transition이 약 1분 정도 걸려, 상당히 낮은 1.33C per second excursion이지만 SC25에서의 별도 demonstration도 on-chip thermal aggressor를 사용하여 2,000C/s rate에 도달했으며, MRM stabilizer heater가 MRM 자체에서 훨씬 더 낮은 -2에서 +2 C/s 범위를 허용했다.

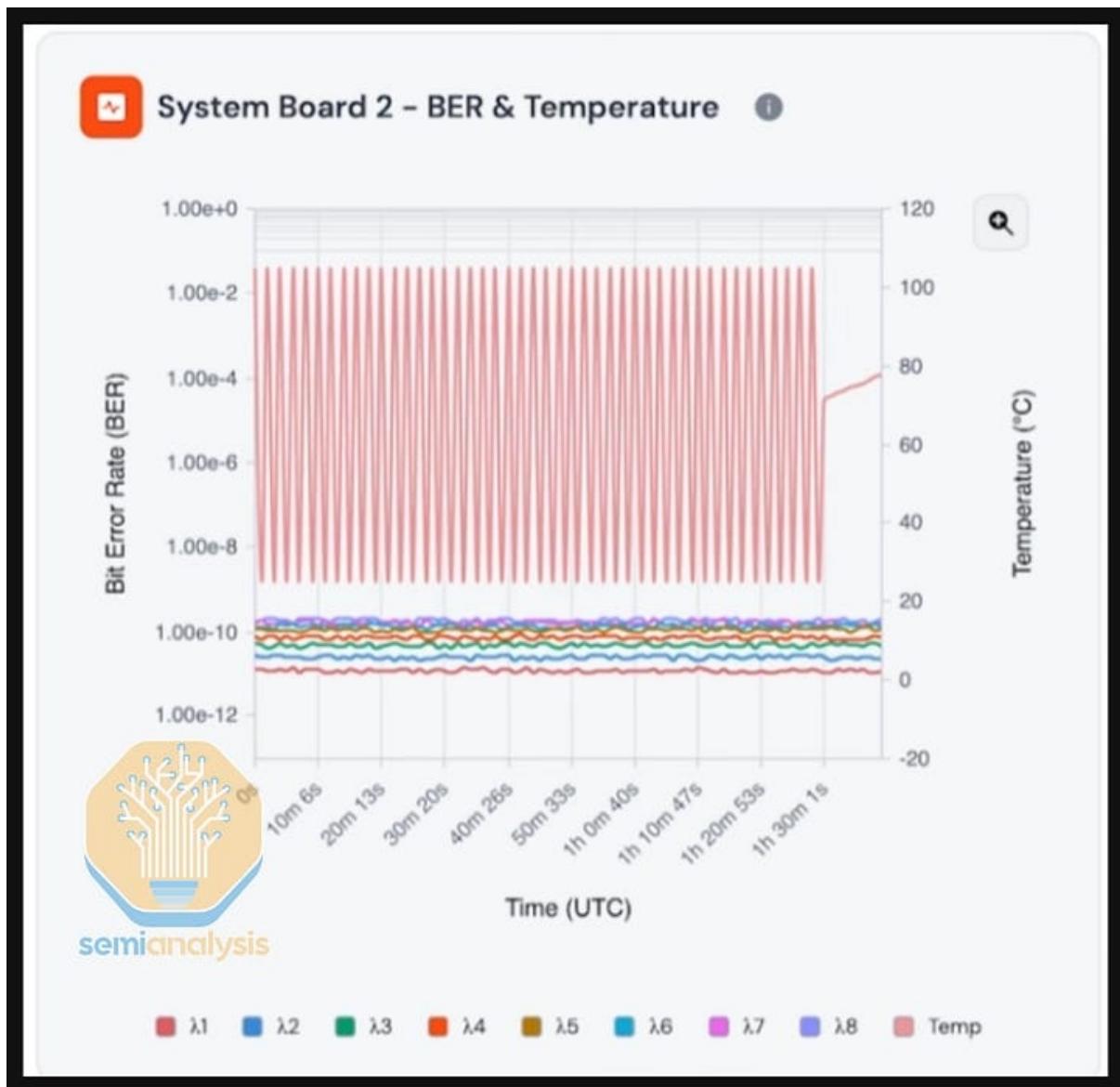


Figure 91: Source: Lightmatter

4.5.5 Xscape Photonics

Xscape Photonics는 4에서 16 wavelength를 제공하는 programmable laser인 ChromX를 작업하고 있는 혁신적인 기업이며, 미래에 최대 128 wavelength를 제공할 계획이다. 최대 128개의 다른 color를 제공함으로써 ChromX는 4에서 8 wavelength만 제공하는 기존 laser와 비교하여 상당히 더 높은 bandwidth를 달성할 수 있을 것이다. ChromX는 external III-V laser와 WDM을 위한 여러 wavelength를 생성하는 데 도움이 되는 on-chip multicolor generator에 의존한다.

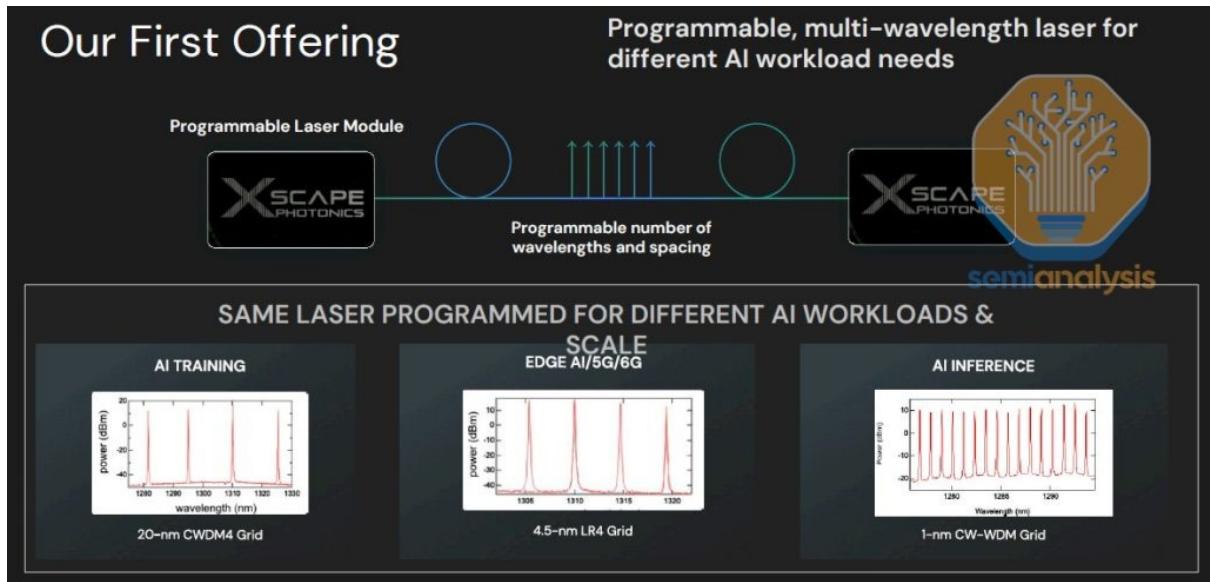


Figure 92: Source: Xscape Photonics

Laser가 programmable하다는 사실은 다른 bandwidth 및 거리 요구사항을 충족하기 위해 다양한 type의 workload에 wavelength를 제공하는 유연성을 제공한다. 흥미롭게도 그들의 솔루션은 단지 하나의 laser만 필요한 반면 기존 CPO 솔루션은 극도로 높은 전력과 전기 소비를 가진 여러 laser를 필요로 한다. 더욱이 모든 wavelength는 단일 fiber를 통해 전달되어 여러 fiber를 필요로 하는 대부분의 CPO system을 괴롭히는 복잡성을 피하여 fiber coupling 문제를 크게 줄인다.

4.5.6 Ranovus

Ranovus는 optical chiplet 기술과 laser design 및 manufacturing 모두에 집중하고 있다. 그들은 GlobalFoundries에서 monolithic CPO 제품 (GF가 AMF를 인수하기 전에 원래 AMF에서 tape out됨)과 다양한 geometry의 PIC와 EIC를 자연스럽게 통합하는 TSMC COUPE 기반 제품을 포함하여 몇 가지 다른 경로에 걸쳐 제품을 tape out했다. Odin Optical Engine은 Microring Resonator modulator를 사용하여 PAM4 modulation을 사용하여 최대 64 lane의 100Gbit/s를 제공할 수 있다.

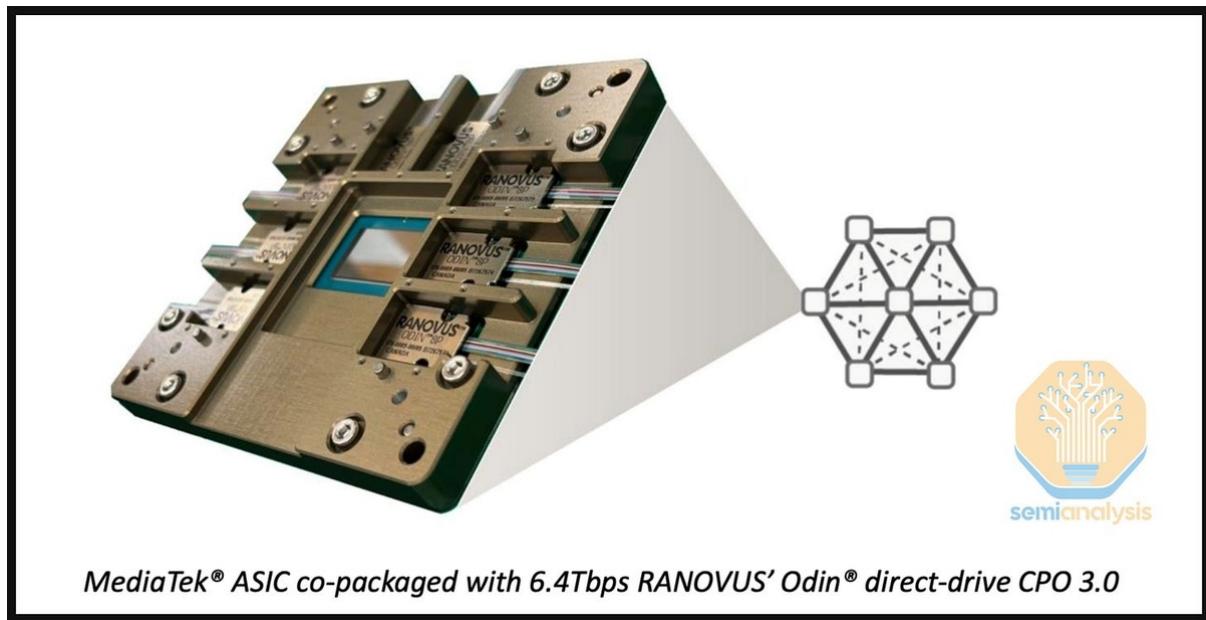


Figure 93: Source: Ranovus

Ranovus의 go to market은 고객이 원하는 interoperable 솔루션을 제공하는 것을 중심으로 한다. 현재 이것은 100G PAM4 DR optics를 의미하지만, modulation을 위한 Microring Resonator의 사용은 56 Gbaud NRZ와 같은 다른 scheme으로 pivot을 가능하게 하지만 WDM을 사용하여 4 lambda를 결합하여 fiber pair당 400G를 제공한다.

Ranovus는 800G chiplet에서 AMD와 interoperability를 시연했으며 Hyperscaler의 미래 custom silicon XPU를 위한 chiplet 솔루션으로 Odin direct-drive CPO 3.0을 제공하기 위해 MediaTek와 partnership을 맺었다.

4.5.7 Scintil

Scintil의 주요 제품은 LEAF Light로, die format (**Known Good Die (KGD)**)으로 제공되거나 module로 조립될 수 있는 **Photonic System-on-Chip (PSoC)**으로, 200 GHz 또는 100 GHz interval로 spaced된 다양한 color의 8개 또는 16개 laser를 통합하여 DWDM을 통해 단일 fiber에서 여러 wavelength를 전달할 수 있다. 그들은 temperature variation 하에서도 wavelength 간 100 GHz 또는 200 GHz spacing을 정확하게 유지할 수 있는 electronic control을 개발했다. OSFP와 유사한 **External Laser Source Form Factor Pluggable (ELSFP)** module에 대한 package reference design이 **Optical Internetworking Forum (OIF)**에 의해 정의되어 고객이 외부 레이저 소스를 통합하기 더 쉽게 만든다. Scintil의 솔루션은 ring modulator 기반 co-packaged optics와 잘 작동한다.

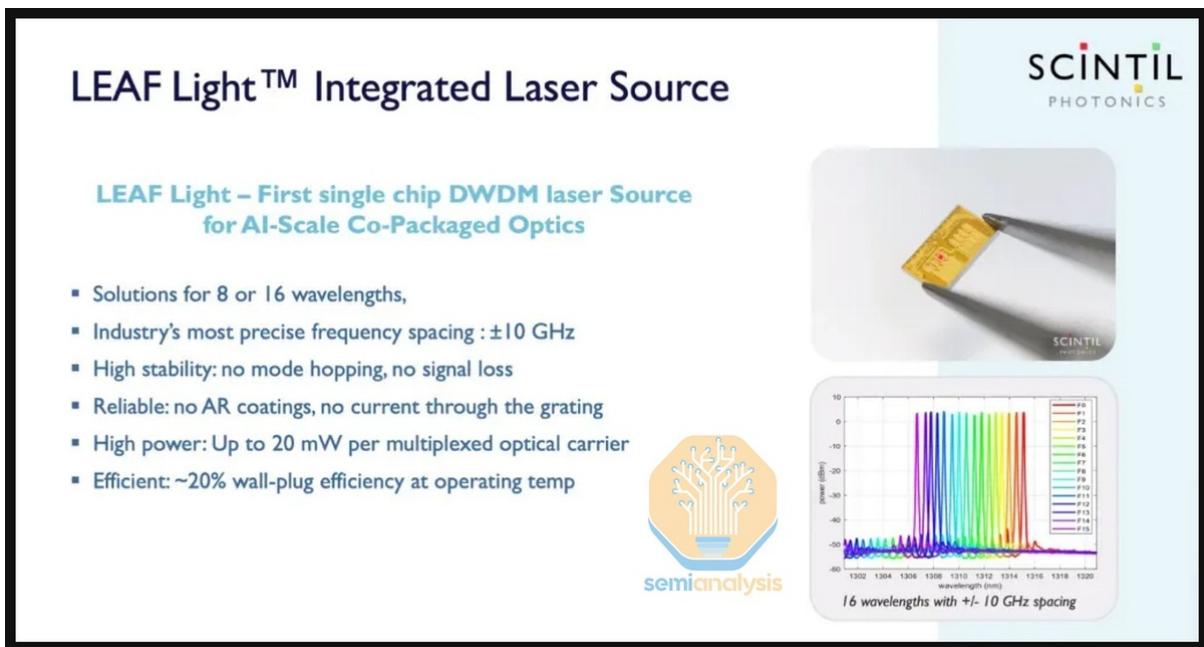


Figure 94: Source: Scintil

Scintil의 process는 **Scintil Heterogeneous Integrated Photonics (SHIP)**라고 불린다. 기술의 본질은 wafer-level process에서 standard silicon photonics에 III-V laser를 통합하는 것이다. Process는 standard silicon photonics wafer로 시작한다. Waveguide, detector 및 mux/demux가 완비된 conventional foundry flow를 사용하여 fabrication된다. 그런 다음 wafer가 flip되고 새 handle에 bonding되어 original substrate를 제거하고 buried oxide layer를 드러낼 수 있다. Unpatterned III-V material이 subsequently 이 새로 노출된 표면에 bonding된다. 그런 다음 III-V가 lithography로 patterned되고 laser를 fabrication하기 위해 etched되어 on-board laser를 가진 monolithically integrated silicon photonics chip이 생성된다. 이것은 E-beam writer를 사용하여 patterned된 전통적인 InP 기반 laser와 대조적이며, 여기서 DWDM을 위한 정밀한 wavelength control을 달성하는 것이 더 어려울 수 있어 tightly spaced channel을 지원하기 어렵게 만든다.

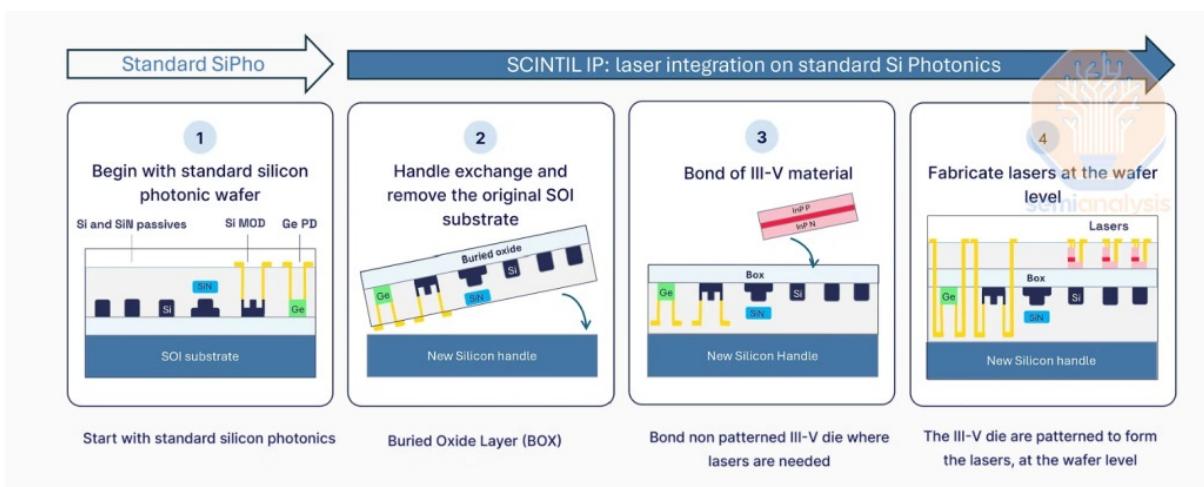


Figure 95: Source: Scintil

DWDM DFB laser array를 개발하는 것은 각 wavelength의 frequency가 정밀하게 생성되어야 하기 때문에 어렵다. 100 GHz channel separation을 달성하기 위해 silicon photonics foundry

와 photolithographic process의 advanced capability가 silicon의 grating을 정확하고 반복적으로 pattern하기 위해 활용되어야 한다. 더욱이 laser가 wafer level에서 생산되기 때문에 각 wafer에서 수백 개의 device를 fabrication할 수 있어 high-volume이고 scalable한 production을 가능하게 한다.

Scintil 솔루션의 한 가지 주요 이점은 power efficiency이다. Scintil 솔루션은 multiplexed color 당 target power를 달성하기 위해 매우 high-power laser를 필요로 하는 combiner splitter를 따라 여러 discrete laser를 사용하는 솔루션과 달리 단일 chip에서 여러 colour (8 또는 16)를 생성하고 multiplex할 수 있다. Scintil 솔루션은 우수한 power efficiency와 증가된 bandwidth density를 제공하는 동시에 전송된 각 bit에 필요한 에너지를 절반으로 줄인다. 이것은 낮은 modulation speed를 가진 여러 wavelength보다 높은 modulation speed를 가진 단일 wavelength를 사용하는 기존 co-packaged 솔루션 (NVIDIA가 현재 Q3450 CPO switch에 채택하고 있는 것 포함)과 비교된다.

5 Part 5: NVIDIA's CPO Supply Chain

CPO system에서 핵심 component가 수행하는 역할을 논의했으며, 이 섹션에서는 supply chain의 특정 기업과 component BOM 비용을 NVIDIA의 공급업체에 초점을 맞춰 논의할 것이며, **Laser Source**, **ELS Module**, **FAU**, **FAU Alight Tool**, **FAU Assembly**, **Shuffle Box**, **MPO Connector**, **MT Ferrule**, **Fiber 및 E/O Testing**에 대한 주요 공급업체를 명시한다.

5.1 Optical Engines

NVIDIA의 X800-Q3450 CPO switch는 115.2T total throughput을 특징으로 하며 scale-out network와 함께 사용하도록 설계되었다. 초기 버전은 각각 1.6Tbit/s로 작동하는 72개의 optical engine을 사용할 것이다; 이후 버전은 각각 3.2Tbit/s에서 작동하는 36개의 optical engine으로 전환할 가능성이 있으며, 단위당 (FAU 포함) ~\$1,000의 비용이 든다. 결과적으로 optical engine의 total BOM cost는 약 \$35-40k (3.2T OE 버전의 경우)에 달한다.

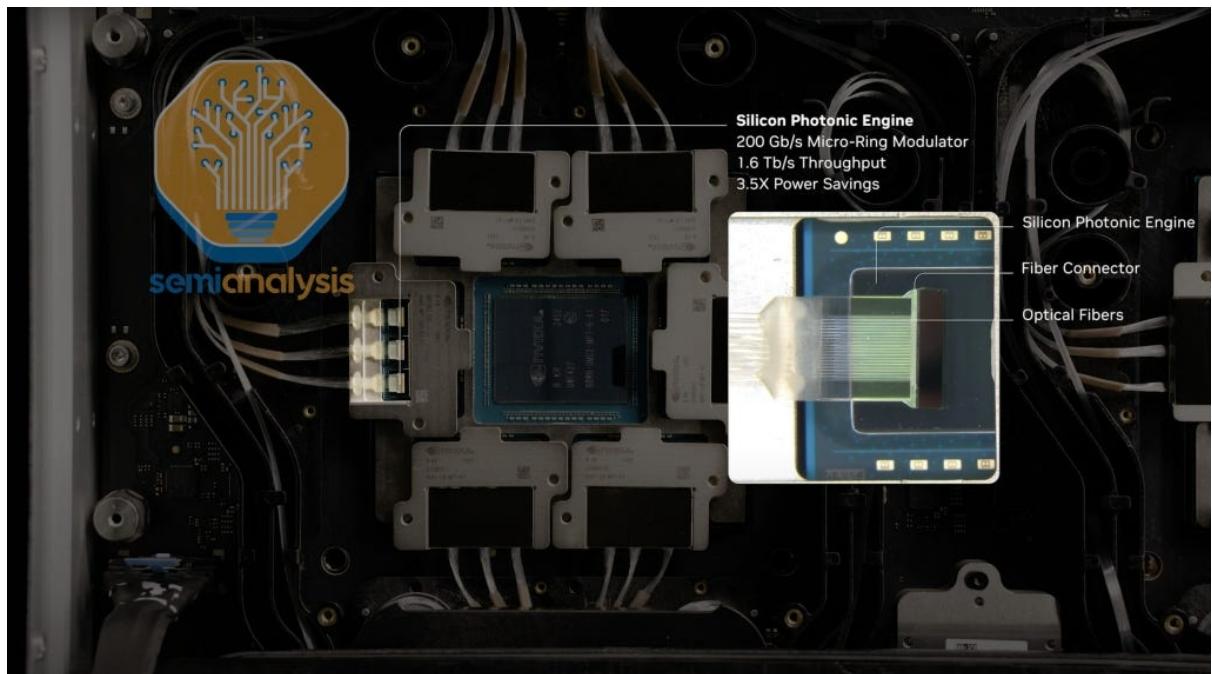


Figure 96: Source: NVIDIA

5.2 External Laser Source (ELS)

NVIDIA X800-Q3450 CPO switch는 laser source를 위해 18개의 ELS module을 사용하며, 각각 8 개의 **Continuous Wave (CW) Distributed Feedback (DFB)** laser chip을 포함한다. CPO system은 상대적으로 higher-power laser source를 사용해야 하며, 각 CWDFB chip은 ~350mW의 전력을 제공한다.

CW laser unit을 생산할 수 있는 주요 업계 player에는 Broadcom (U.S.), Furukawa (Japan), Lumentum (U.S.), Coherent (U.S.), Yuanjie (China) 및 Shijia (China)가 포함된다. Lumentum, Coherent, Furukawa 및 Broadcom은 일반적으로 Chinese vendor (Yuanjie 및 Shijia)보다 높은 가격을 부과한다. 우리는 Lumentum이 NVIDIA의 초기 batch CPO switch shipment의 sole supplier가 될 것으로 예상하며, Coherent는 2026년 후반에 두 번째 공급업체로 진입할 수 있다. Chinese manufacturer는 CW laser source가 일반적으로 상대적으로 표준화되고 commoditized된 것으로

간주되기 때문에 앞으로 기회를 볼 수 있지만, CPO application에 필요한 high power laser source를 구축하는 데는 여전히 어느 정도의 moat가 있다.

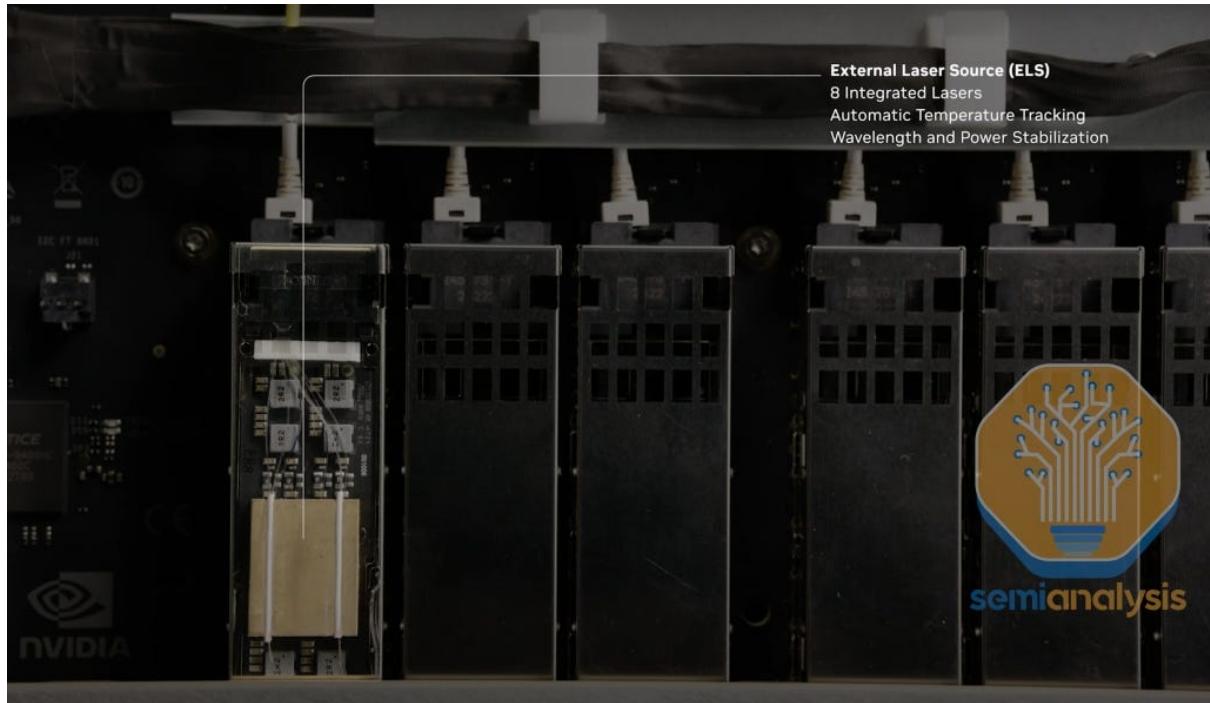


Figure 97: Source: NVIDIA

5.3 Fiber Attach Units (FAUs)

FAU는 optical fiber를 Optical Engine과 coupling하는 것을 담당하는 중요한 passive component이다. High-quality FAU와 그들의 정밀한 alignment는 최적의 optical 성능을 보장하는 데 critical하다. FAU assembly process의 현재 과제 중 하나는 coupling loss를 측정하는 데 사용되는 testing equipment가 아직 완전히 automated될 수 없다는 것이다. 결과적으로 testing process는 여전히 manual labor에 크게 의존하며, 이는 전체 production 속도를 늦추고 더 비용이 많이 듈다. Corning은 Spectrum X CPO system에 대해 FAU testing당 평균 10-15 minute를 추정했다.

material 및 component 비용 외에도 labor는 FAU 비용의 큰 부분이다. Skilled labor는 FAU unit에서 fiber의 high quality assembly 및 alignment를 보장하는 데 필수적이다. X800-Q3450의 각 1.6T OE는 20 fiber를 가진 FAU를 가지고 있다: 8 transmit, 8 receive 그리고 external laser용 4. System당 Tx/Rx용 1152를 포함하여 총 1,440 fiber이다.

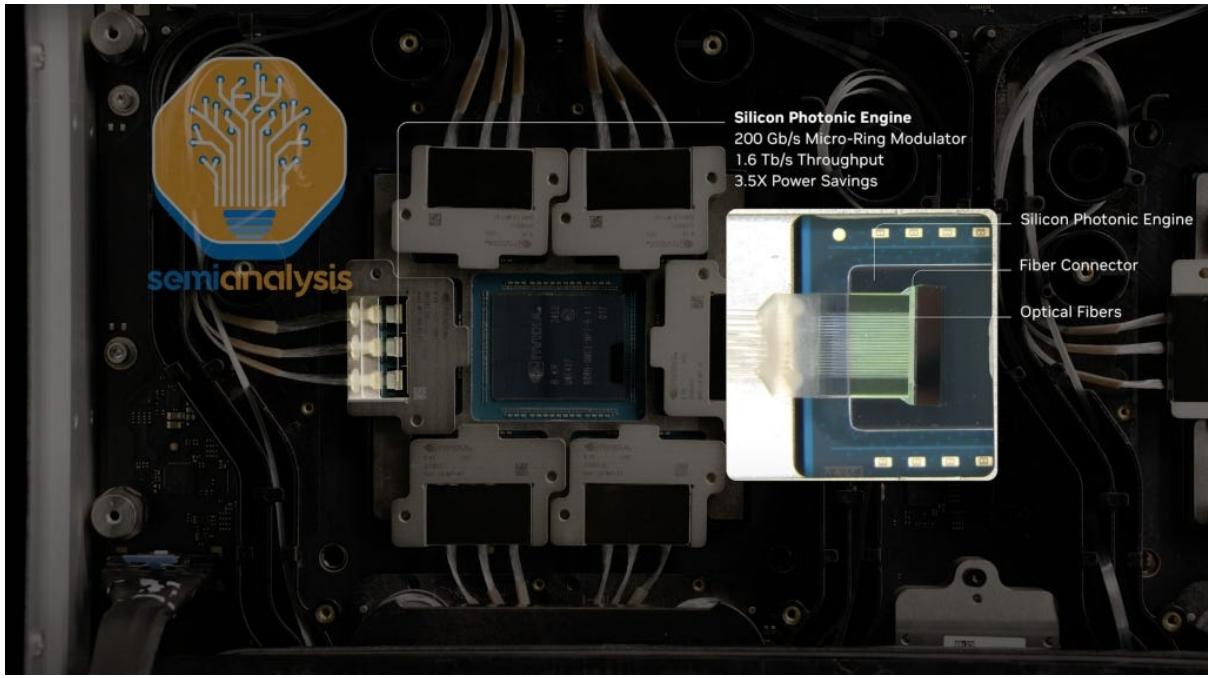


Figure 98: Source: NVIDIA

FAU의 leading 기업은 TFC Optical (300394.SH), Senko (9069.JP) 및 FOCI (3363.TW)이다. TFC는 X800-Q3450 CPO switch-용 FAU를 공급할 강력한 기회를 가지고 있는 반면, Senko는 Spectrum X CPO와 Broadcom의 Tomahawk 6 CPO system 모두에 대한 매우 가능성 있는 후보로 간주된다. 한편 FOCI는 NVIDIA의 large-up CPO 솔루션에 더 집중할 가능성이 있다.

TFC의 핵심 강점은 위 단락에서 논의한 것처럼 경쟁력을 유지하는 데 핵심 요소인 China의 강력한 manufacturing capability와 숙련되었지만 비용 효율적인 대규모 labor pool에 대한 접근에 있다. 추가로 TFC는 약 3년 전에 CPO design에 대해 NVIDIA와 협력하기 시작했으며, 그 초기 partnership은 이제 결실을 맺고 있으며, TFC는 NVIDIA의 진행 중인 CPO roll-out 노력에서 핵심 역할을 할 것으로 예상된다.

Senko는 CPO system을 위한 detachable FAU 솔루션을 제공하는 signature **SEAT (Senko Elastic Averaging Technology)** platform을 가지고 있다. 회사는 edge-coupling 기술에 대해 GFS와 긴밀하게 협력하고 있으며 Senko mirror를 wafer trench에 직접 통합하고 있다. 이 접근 방식은 edge coupling을 위한 wafer-level testing을 가능하게 하며, 이는 지금까지 grating-coupling의 주요 이점이었다.

Space의 다른 주목할 만한 player에는 Sumitomo 및 **Advanced Fiber Resources (AFR)** 등이 포함된다. AFR은 Broadcom의 supply chain 내에서 긴밀한 관계를 가지고 있는 것으로 믿어진다.

SiPho chip은 micron scale의 waveguide channel을 특징으로 하며, 각각 system에 들어오는 빛과 정밀한 alignment가 필요하다. 따라서 process 동안 exceptionally 높은 accuracy를 가진 coupling machine이 필요하다. FiconTEC (Germany)은 현재 high-precision coupling machine을 제공하는 업계를 선도하고 있다. 그들의 machine은 unit당 \$300k 이상에 판매될 수 있지만, 우수한 accuracy로 인해 여전히 고객들에게 매우 인기가 있다. 한편 All Ring Tech (Taiwan)도 automated fiber attach equipment를 제공한다. 회사는 workforce의 약 10%를 space에 투입했으며 coupling equipment revenue가 2026년부터 견인력을 얻을 것으로 예상한다. 마지막으로 GMT Global (Taiwan)은 FAU를 통한 정확한 transmission을 보장하기 위해 빛의 wavelength를 correct하는 FAU bonding, alignment 및 inspection machine을 design한다. 회사는 일반적으로 \$200–250K를 부과

하는 비교 가능한 Japanese machine보다 ~75% 낮게 machine 가격을 책정하는 것을 목표로 한다. 회사는 Q4 2025에 일부 pile run을 가질 것으로 예상한다.

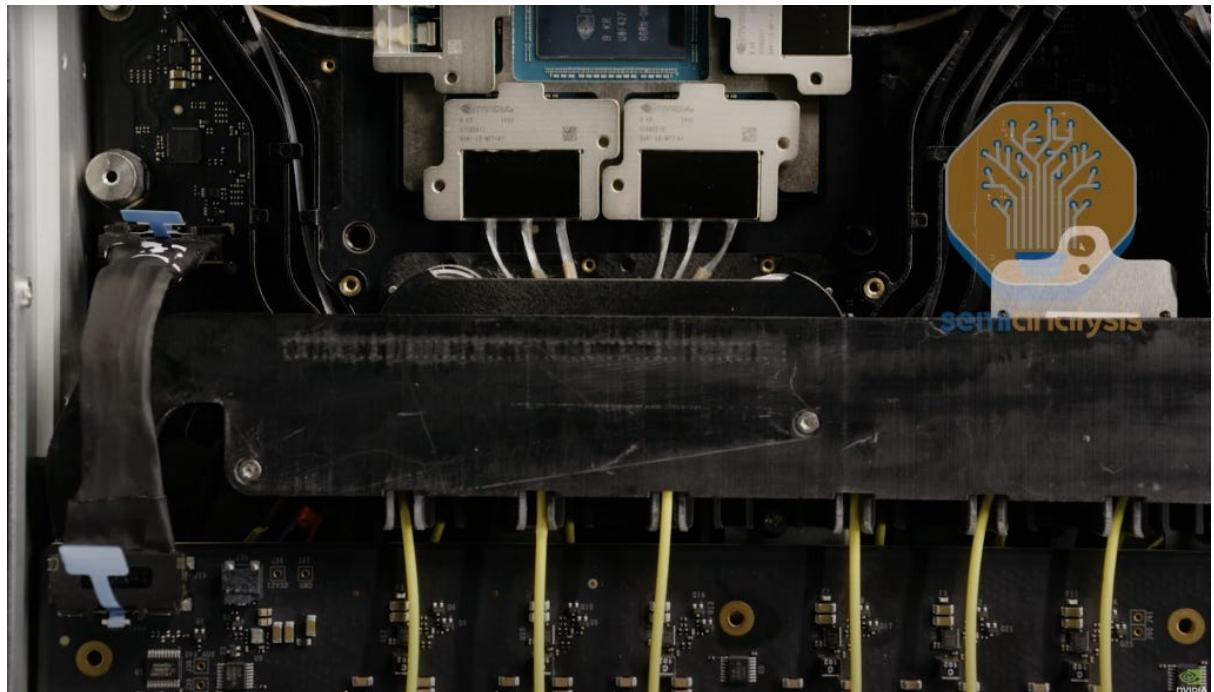


Figure 99: Source: NVIDIA

5.4 Fiber Shuffle Box

NVIDIA의 X800-Q3450 CPO switch의 경우 1,000개 이상의 optical fiber가 OE에서 나오므로 이를 정리하고 destination으로 routing하기 위해 shuffle box가 필요하다. 전통적으로 이 shuffle box alignment process는 기술자가 물리적으로 fiber를 배열하여 수동으로 수행되었다. 그러나 일부 leading 기업은 더 높은 accuracy와 efficiency로 alignment를 수행할 수 있는 automated machine을 개발했다.

Shuffle box의 가격은 일반적으로 관리하는 fiber 수와 연결된다. 예를 들어 T&S Communications는 48-fiber shuffle box를 ~\$150에, 300-fiber 버전을 ~\$1,000에, 그리고 500-fiber model 을 ~\$1,600에 판매한다. 수천 개의 fiber를 가진 X800-Q3450의 경우 shuffle box는 NVIDIA에게 조달하는 데 \$3,000 이상의 비용이 든다. Shuffle box의 주요 BOM component에는 MT ferrule과 fiber가 포함된다.

T&S Communications (300570.SH)는 shuffle box 산업의 leading player이다. 회사는 shuffle box 내에서 fiber를 정렬하기 위한 automated machine을 개발했다. 기술은 특허를 받았으며, 경쟁자가 이를 우회하는 데 시간과 추가 비용이 걸릴 수 있다. T&S의 주요 고객은 Corning이며, 두 회사는 종종 고객에게 서비스를 제공하기 위해 tandem으로 작업한다. 예를 들어 Corning은 고객 (NVIDIA, Broadcom 등)이 CPO 솔루션을 위한 fiber network를 design하는 데 도움을 주고 shuffle box 부분을 T&S에 subcontract한다.

Molex는 또 다른 주목할 만한 player이며 시장에 꽤 일찍 진입했다. 그러나 더 낮은 manufacturing efficiency로 인해 제품 가격은 일반적으로 T&S보다 ~20% 더 높다.

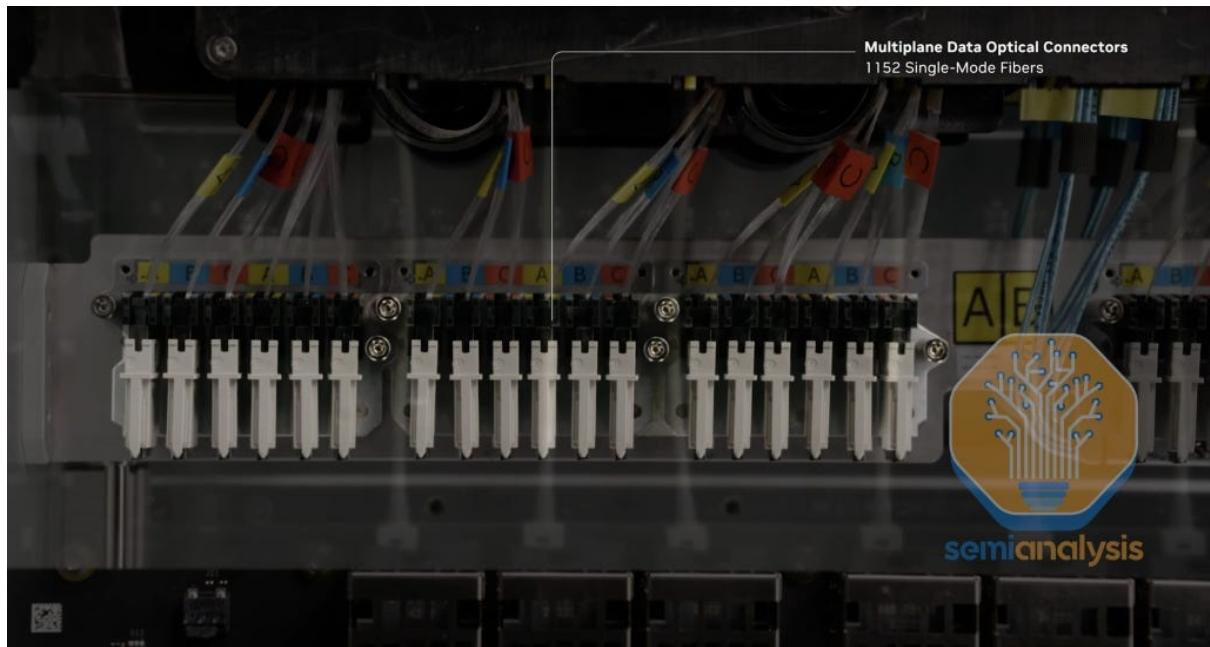


Figure 100: Source: NVIDIA

5.5 MPO Connectors

shuffle box의 outskirt에는 MPO connector가 있으며, 이는 shuffle box 내에 수용된 fiber를 external port에 연결한다. Fiber Optic MPO cable은 그런 다음 이 port에 plug하여 switch를 다른 distant switch 또는 distant NIC에 연결할 수 있다. MPO connector의 manufacturing process는 주로 unit 내부의 MT ferrule의 injection molding, vacuum adhesive filling 및 fiber threading을 포함한다. X-800 Q3450 CPO switch의 경우 144개의 MPO connector가 필요하다.

MPO connector를 생산할 수 있는 여러 player가 있으며, US Conec (U.S.), T&S Communications (China), Senko (Japan), Broadex (China) 및 Optec (China)가 포함된다. 그들은 또한 전체 network design에 component를 공급하기 위해 fiber network contract manufacturer (Corning과 같은)와 협력해야 한다.

5.6 MT Ferrules

MT ferrule은 FAU, shuffle box 및 MPO connector에서 사용되는 핵심 요소이다. 그들은 여러 fiber를 parallel 방식으로 정렬하는 데 사용된다. US Conec (U.S.), T&S (China), Senko (Japan), Fukushima (Japan), FOCI (Taiwan), Sumitomo (Japan) 및 TFC (China)를 포함한 많은 기업이 MT ferrule을 제조할 수 있다. MT ferrule은 특히 제조하기 어렵지 않지만 필요한 precision과 robustness를 달성하는 데는 여전히 많은 engineering이 필요하다. 기업은 fiber connection 동안 insertion loss를 최소화하는 ferrule을 생산하기 위한 molding capability에서 경쟁한다.

이러한 기업 중 US Conec은 또한 기술을 개발한 30년 이상의 경험을 가지고 있으며 NVIDIA의 Q3450 CPO system의 주요 공급업체 중 하나가 될 것으로 예상된다. Fukushima도 강력한 mold design 및 manufacturing capability를 가지고 있어 경쟁력 있는 가격으로 high-quality MT ferrule을 생산할 수 있다. 한편 FOCI, TFC 및 T&S는 주로 자체 in-house FAU 및 shuffle box용 MT ferrule을 생산한다. Rationale은 quality control 및 cost efficiency를 향상시키기 위해 가능한 한 많은 process를 vertically하게 통합하는 것이다.

115.2T Nvidia X800-Q3450 CPO Switch Cost and Power (Current Pricing)						
Item	Unit Cost	Quantity	Extended Price	BOM (%)	Power (W)	Extended Power (W)
Switch Chip		4			400.0	1,600
Optical Engines		72			7.0	504
FAUs (included in OE)		72			0.0	0
External Laser Source (ELS)		18			8.0	144
MPO connectors and jumper cables		144			0.0	0
Shuffle box		1			0.0	0
Fibers (per meter)		2000			0.0	0
Others						1,300
Total			\$70,640	100%		3,548
Gross Margin			\$134,216	60%		
Selling Price			\$176,600			
3y Service and Warranty			\$28,256			
Selling Price with Service			\$204,856			

Bill of Material estimates are based on current scale of production and could improve as production scales up

Figure 101: Source: [SemiAnalysis AI Networking Model](#)

5.7 Manufacturing, Assembling, and the Testing Processes

OE fabrication: 위에서 논의한 바와 같이 TSMC는 PIC, EIC를 fabrication하고 통합하며 CPO system의 역할을 하는 데 중요한 역할을 할 것이다. 그들의 COUPE platform은 차세대 CPO endpoint를 위한 선택 솔루션으로 보인다. Global Foundries와 Tower는 강력한 SiPho capability를 가진 foundry이지만, leading edge CMOS 및 advanced packaging의 부족은 미래의 더 높은 bandwidth OE를 제공할 수 있는 능력을 제약한다.

Advanced packaging: Outsourced Semiconductor Assembly and Test (OSAT) provider는 back-end process에 집중할 것이다. OE packaging, OE testing 및 system packaging (laser 및 coupler integration & testing)을 포함한다.

ASE/SPIL (3711.TW), Amkor (\$AMKR) 및 Shunsin (6451.TW)은 그러한 솔루션의 주요 provider이다. 그 중 ASE는 미래 Rubin-rack CPO system에 대한 involvement를 포함하여 NVIDIA의 supply chain에서 핵심 공급업체로 두드러지는 반면, Shunsin은 Broadcom과 긴밀한 관계를 유지한다.

다른 주목할 만한 이름은 Fabrinet (\$FN), TFC Optical (300394.SH) 및 Foxconn (2354.TW)이다. Fabrinet은 오랫동안 NVIDIA의 in-house optical module unit을 위한 module assembler로 봉사해 왔으며 이제 OE packaging, testing 및 full system assembly에서 capability를 적극적으로 구축하고 있다. Fabrinet은 또한 Micas 및 Foxconn과 함께 Broadcom의 CPO system assembly 작업의 잠재적 후보 중 하나로 간주된다.

TFC는 지난 3-4년 동안 CPO design에 대해 NVIDIA와 긴밀하게 협력해 왔으며 FAU의 주요 공급업체가 될 것이다. 회사는 Suzhou, China의 advanced packaging facility에 투자해 왔으며, 이는 CPO supply chain에서 더 큰 역할을 확보하려는 야망을 signal한다.

Electro-optical (E/O) testing equipment: Testing process 동안 위에서 언급한 service vendor는 system reliability를 보장하기 위해 electro-optical (E/O) testing tool을 사용한다. 그러나 업계는 여전히 개발 중이며 photonic engine을 위한 표준화된 testing methodology에 완전히 수

럼하지 못했다. 각 vendor는 이 emerging space에서 발판을 얻기 위해 다양한 솔루션을 개발하고 있다.

CPO supply chain의 다른 주요 equipment provider에는 Keysight, Ficontec, Teradyne, Advantest, FormFactor, Chroma, Anritsu 및 Multilane이 포함된다. 그 중 Keysight는 field의 large player이다. 그들은 premium quality (그리고 premium 가격!)를 가진 high-speed testing equipment를 제공하는 것으로 알려져 있다. 예를 들어 그들은 [최근 1.6T optical transceiver testing을 위한 2개의 새로운 oscilloscope를 발표했다](#). 시장에서의 position으로 인해 그들은 CPO 추세로부터 잘 benefit을 받을 수 있다.

Ficontec은 photonics testing에 강력한 기반을 가지고 있으며 이제 electrical testing capability도 적극적으로 확장하고 있다. 주요 강점 중 하나는 wafer-level photonic testing이다. 이는 전통적으로 inefficient한 photonic testing process의 efficiency를 향상시킨다. 예를 들어 회사는 [최근 기존 semi ATE architecture와 호환되는 PIC용 새로운 high-throughput wafer-level testing tool을 도입했다](#). Machine은 업계 최초의 double-sided wafer testing machine으로 주장된다. Testing equipment 외에도 Ficontec은 FAU assembly 및 coupling machine도 제공하며, 이는 다음 섹션에서 더 자세히 살펴볼 것이다.

Teradyne도 field의 large player로 electrical testing에서 강력한 historical presence를 가지고 있지만, 회사는 photonic testing에 진입하는 것에 대해 **very serious**해 왔다. 예를 들어 회사는 최근 packaged optical testing을 전문으로 하는 startup을 인수했으며, 이는 CPO를 위한 capability를 구축하기 위한 strategic push를 반영하는 move이다.

Chroma는 3D sensing 및 optical communication에 사용되는 laser diode를 위한 photonic testing equipment를 공급해 왔으며, 이 expertise를 활용하여 CPO space에 진입할 잠재력을 가지고 있다. 그러나 회사의 innovation pace는 현재 일부 경쟁자보다 뒤처진다.

Nvidia CPO Supply Chain Market Map					
Laser Source	ELS Module	FAU	FAU Align Tools	FAU Assembly	Shuffle box
LITE	TFC	TFC	finconTEC	Fabrinet	T&S
COHR	O-Net	Senko	All Ring	FOCI	Browave
AVGO	Innolight	FOCI		Foxconn	Molex
Furukawa	Eoptolink	Sumitomo		ASE	
Yuanjie					
MPO Connector	MT Ferrule	Fibers	Foundries	OSAT/ Assembly	E/O Testing
US Conec	T&S	Corning	TSMC	Amkor	ficonTEC
Senko	Senko	Sumimoto	Tower	ASE/ SPIL	Keysight
T&S	US Conec	Nittobo	GF	Fabrinet	Teradyne
Browave	Fukushima			TFC	Formfactor
	FOCI				Chroma
	Sumitomo				Multilane

Vendors highlighted green are key players in the sectors & confirmed to be part of Nvidia's CPO supply chain
For more information on market shares and competitive dynamics, contact sales@semianalysis.com.

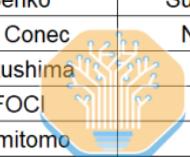


Figure 102: Source: [SemiAnalysis AI Networking Model](#)