

AMD Advancing AI: MI350X and MI400 UALoE72, MI500 UAL256

June 20, 2025

Contents

1 Executive Summary	2
2 MI350X and MI355X Specs	3
3 Competitive Performance per TCO with the HGX B200 NVL8	5
4 NVIDIA is Upsetting Neoclouds with DGX Lepton	6
5 MI355X Is Not a Rack Scale Solution – AMD’s Marketing Spin	7
6 Hyperscale and AI Lab Adoption of new AMD Products	9
7 AMD’s Solving Its Neocloud Rental Market Weakness	10
8 Pushing into High Gear – AMD is Accelerating the Development of the AMD Neocloud Ecosystem	10
9 ROCm Software Improvements	11
10 MI355X PyTorch Continuous Integration (CI) and Testing	12
11 ROCm MLPerf Training Submission	13
12 MIG Partitioning is Wasting Time and Engineering Resources	13
13 MI355X Manufacturing – Updated Chiplet Architecture	14
14 CDNA4 Microarchitecture (UArch)	16
15 AMD Advancing AI Developer Session Track is Disappointing	16
16 RCCL – ROCm Collective Communication Library	17
17 New AMD Initiatives to Pay AI Engineers Market Rate	17
18 MI400 Series Flexible Input Output (I/O)	18

지난 6개월 동안, AMD는 전시 체제를 유지해왔다. AMD는 NVIDIA와 경쟁할 수 있다는 목표를 향해 열심히 그리고 현명하게 작업해왔다. Advancing AI 2025 event에서 AMD는 MI350X/MI355X GPU를 출시했으며, 이는 performance per TCO 기준으로 소규모에서 중간 규모 LLM의 inference에서 NVIDIA의 HGX B200 solution과 경쟁할 수 있다. AMD가 투영하는 reality distortion field에도 불구하고, MI355X는 rack scale product가 아니며, frontier model inference나 training에서 NVIDIA의 GB200 NVL72와 경쟁할 수 없다.

대신, MI400 Series가 진정한 rack scale solution이며, 2026년 하반기에 NVIDIA의 VR200 NVL144 rack scale solution과 잠재적으로 경쟁할 수 있다. 또한 MI400 Series 주변에는 일부 marketing spin이 있는데, AMD가 "IF over Ethernet" protocol을 "UALink Protocol over Ethernet"으로 이름을 바꾸었지만, 이는 진정한 UALink가 아니다.

본 논문에서는 AMD의 새로운 제품들의 상대적 경쟁력을 논의하고 총 소유 비용을 분석할 것이다. 또한 AMD의 새로운 hyperscale 고객인 AWS에 대해 자세히 설명하고, 반대로 기존 고객인 Microsoft로부터의 후속 주문에서 지속적인 실망에 대해서도 다룰 것이다.

최근 NVIDIA는 compute를 상품화하는 것을 목표로 하는 DGX Lepton Marketplace 출시로 상당수의 Neocloud partner들을 화나게 했다. 우리는 이러한 발전이 AMD가 자체 Neocloud ecosystem을 육성할 수 있는 기회의 창을 여는 데도 도움이 되었다고 믿는다. 우리는 AMD가 Neocloud에 더 기꺼이 투자하는 방법, 이러한 Neocloud를 돋기 위해 사용하는 영리한 financial engineering, 그리고 AMD가 자체 내부 개발 R&D cluster에 하는 투자에 대해 설명할 것이다.

1 Executive Summary

1. MI355X는 소규모에서 중간 규모 model inferencing에서 HGX B200과 경쟁력이 있지만 GB200 NVL72와는 경쟁할 수 없다
2. AMD의 marketing RDF에도 불구하고, MI355 128 GPU rack은 "rack scale solution"이 아니다 – 이는 scale up world size가 8개 GPU에 불과한 반면 GB200 NVL72는 world size가 72개 GPU이다. GB200 NVL72는 대규모 frontier reasoning model inference에서 Perf per TCO에서 MI355X를 능가할 것이다
3. MI355X는 HGX B200과 유사한 collective performance를 가질 것이지만 MI355X collective은 GB200 NVL72보다 최소 18배 느리게 실행될 것이며, 더 느릴 수도 있다
4. AMD는 Developer Cloud를 발표했으며, 이는 MI300에 대해 \$1.99/hr/GPU의 on demand pricing을 제공할 것이다. 이는 현재 AMD Neocloud 시장의 \$3.00/hr/GPU와 비교되며, AMD GPU 임대를 NVIDIA GPU 임대와 경쟁력 있게 만들 수 있는 움직임이다
5. NVIDIA의 DGX Lepton Marketplace는 많은 Neocloud를 화나게 했으며, 이는 잠재적으로 AMD가 Neocloud를 설득하여 NVIDIA와 AMD를 모두 지원하도록 하는 기회를 제공한다
6. AMD는 마침내 NVIDIA와 유사한 전략을 채택하고 있으며, 강력한 balance sheet를 사용하여 cloud로부터 GPU의 일부를 다시 임대함으로써 Neocloud와 hyperscale ecosystem이 AMD를 채택하도록 지원하고 있다. 이는 AMD system의 end user 채택을 가속화하는 데 도움이 될 것이다
7. MI400 Series는 2026년 하반기에 NVIDIA의 VR200 NVL144와 잠재적으로 경쟁할 수 있는 rack scale solution이 될 것이다
8. AMD engineering 급여를 시장 수준과 더 경쟁력 있게 만들고 보상을 AMD의 성공과 더 밀접하게 연계시키는 새로운 진행 중인 initiative가 있다. AMD가 이를 AI engineer들에게 언제 발표할지는 아직 알 수 없다
9. MI400 Series rack은 실제로 scale-up networking에 진정한 UALink를 사용하지 않는다. 대신 AMD는 Infinity Fabric Over Ethernet을 "UALink over Ethernet"으로 이름을 바꾸고 이를 scale-up network에 사용한다

10. MI400 Series scale-up network는 Broadcom Ethernet Tomahawk 6 switch를 사용할 것인데, 이는 Marvell과 Astera Labs의 UALink switch가 2026년 말까지 준비되지 않을 것이기 때문이다
11. 이러한 앞선 지적들에도 불구하고, UALink over Ethernet을 사용하는 MI400 Series는 scale up bandwidth 측면에서 VR200 NVL144의 NVLink와 여전히 경쟁력이 있을 것이며, 또한 72개 logical GPU의 scale up world size를 가진다
12. 2027년 말에 AMD는 MI500 UAL256을 출시할 것이며, 이는 VR300 NVL576의 144개 physical/logical chip이 아닌 256개 physical/logical chip을 특징으로 할 것이다

2 MI350X and MI355X Specs

이 series에는 두 가지 버전의 CDNA4 chip이 있다 – 즉 MI350X와 MI355X이다. MI350X는 air cooling되는 1,000W 버전이고 MI355X는 air cooling과 DLC liquid cooling을 모두 지원하는 1,400W 버전이다. MI355X가 1.4배 더 많은 전력을 사용함에도 불구하고, on-paper spec은 TFLOPS throughput 측면에서 MI350X보다 10% 미만 빠르다는 것을 보여준다. 그러나 우리는 published spec이 power 제한으로 인해 종종 달성되지 않기 때문에 MI355X의 realized performance가 10%보다 더 좋을 것으로 예상한다. 이러한 published spec은 peak clock speed가 실제 workload에서 유지될 수 있다고 가정하지만, 이는 AMD와 NVIDIA system 모두에서 단순히 그렇지 않다.

MI350X와 MI355X의 on-paper spec은 모두 BF16/FP8/FP4 data type(dtype)에서 HGX B200과 경쟁력이 있다. 우리는 BF16과 FP8이 training에 사용되고 FP8/FP6/FP4가 inference에 사용될 것으로 예상한다. HGX B200에서 FP6는 FP8과 동일한 physical circuit를 공유하여 동일한 FP8/FP6 on paper FLOP/s를 가진다. MI355X에서 FP6는 FP4와 동일한 physical circuit를 공유하므로 FP6는 FP4와 동일한 peak TFLOP/s speed를 가질 것이다. 이는 MI355X FP6가 B200 FP6보다 2.2배 빠르다는 것을 의미한다. 실제로 MI355X FP6는 AI chip이 항상 power에 의해 제한되기 때문에 MI355X FP4보다 최소 20% 느릴 것이다.

[SemiAnalysis benchmarking](#)이 보여준 바에 따르면 MI300X와 H100이 각각 FP16과 BF16에 대해 동일한 on-paper TFLOP/s를 보여주지만 (즉, NVIDIA의 FP16 TF = BF16 = 989 TFLOP/s, AMD의 FP16 = BF16 = 1307 TFLOP/s) 실제로는 각 card가 FP16 vs BF16을 실행할 때 서로 다른 realized TFLOP를 제공한다. 우리는 가까운 미래에 MI355X FP6 versus FP4에 대한 현실적인 TFLOP/s를 파악하기 위해 microbenchmark를 실행하는 논문을 발표할 것이다.

AMD vs NVIDIA Accelerator Specifications								
	B200 HGX NVL8	B300 HGX NVL8	GB200 NVL72	GB300 NVL72	MI355X	MI350X	MI355X vs GB200	MI355X vs B200
Peak TDP	1,000W	1,200W	1,200W	1,400W	1,400W	1,000W	1.2x	1.4x
BF16 Dense TFLOP/s	2,250	2,250	2,500	2,500	2,300	1.0x	1.1x	
FP8 Dense TFLOP/s	4,500	4,500	5,000	5,000	4,600	1.0x	1.1x	
FP6 Dense TFLOP/s	4,500	4,500	5,000	5,000	10,000	9,200	2.0x	2.2x
FP4 Dense TFLOP/s	9,000	13,500	10,000	15,000	10,000	9,200	1.0x	1.1x
Supported FP4 Dtypes	OCP MX4, NVFP4	OCP MX4, NVFP4	OCP MX4, NVFP4	OCP MX4, NVFP4	OCP MX4	OCP MX4		
Memory Bandwidth	8.0 TByte/s	8.0 TByte/s	8.0 TByte/s	8.0 TByte/s	8.0 TByte/s	8.0 TByte/s	1.0x	1.0x
Memory Capacity	180 GB	288 GB	192 GB	288 GB	288 GB	288 GB	1.5x	1.6x
Scale Up World Size	8	8	72	72	8	8	0.1x	1.0x
Scale Up Bandwidth (Uni-di)	900 GByte/s	900 GByte/s	900 GByte/s	900 GByte/s	7x76.8GByte/s	7x76.8GByte/s	0.6x	0.6x
Scale Out Bandwidth (Uni-di)	400 Gbit/s	800 Gbit/s	400 Gbit/s	800 Gbit/s	400 Gbit/s	400 Gbit/s	1.0x	1.0x
Cooling	Air/DLC	Air/DLC	DLC	DLC	Air/DLC	Air		

1. The GB200 NVL72 and the GB300 NVL72 both have 144 compute chiplets across 72 logical GPUs.

Source: SemiAnalysis

4-bit floating point format의 경우, MI355X는 32개 element의 block에 걸쳐 microexponent scale factor가 적용되는 OCP MX4만 지원할 것이다. 대조적으로, NVIDIA의 Blackwell GPU는 OCP MX4와 NVFP4를 모두 지원하지만, NVFP4는 16개 element의 더 작은 block size를 사용하여

QAT/PTQ quantization을 수행할 때 numerical accuracy를 calibrating할 때 더 적은 문제를 야기할 것이다. 우리는 몇몇 vLLM과 open-source inference contributor들과 이야기했으며, 그들은 NVPF4가 MX4보다 information/model quality를 훨씬 더 잘 보존하지만, MX4가 추가적인 runtime quantization software technique으로 잠재적으로 동일한 quality를 달성할 수 있다고 언급했다.

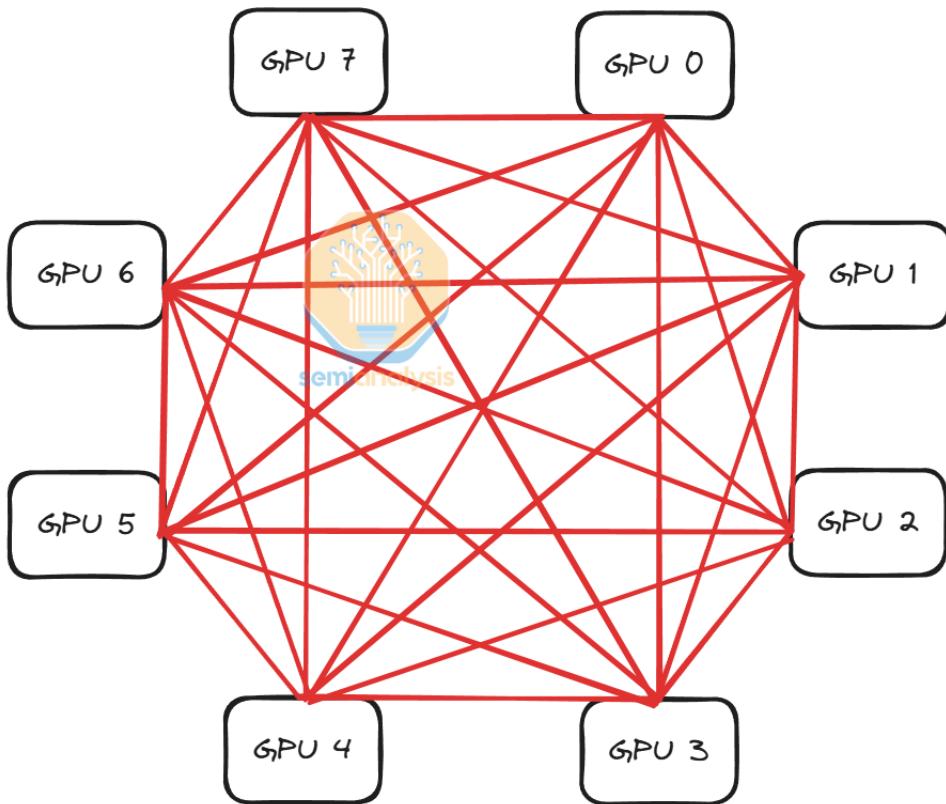
Blackwell Ultra B300 HGX NVL8에서 NVIDIA는 1.4배 더 많은 FP4 tensor core circuit을 위한 공간을 만들기 위해 대부분의 FP64와 int8 tensor core를 제거했다. 이는 B300이 이러한 optimization을 사용하지 않는 MI350과 MI355와 비교할 때 FP4 inference를 지배할 수 있게 한다. 결과적으로 B300의 FP4 TFLOP/s는 200W 적은 전력을 소비하면서 MI355X보다 1.3배 빠르다.

HBM 측면에서 MI350/MI355는 B300과 동일한 memory bandwidth와 capacity를 가지지만, B200의 180GB에 비해 288GB로 훨씬 더 많은 HBM을 가진다. 이는 AMD single node inference에 있어서 중요한 장점이다. 그러나 multi node high rank expert parallelism과 disaggregated prefetch의 시대에서 GPU당 더 많은 HBM을 갖는 것은 여전히 유익하지만 그렇게 중요하지는 않다. Bandwidth가 훨씬 더 중요하며, 이것이 8Hi HBM4가 2개의 서로 다른 high profile ASIC program을 위해 2개의 HBM vendor에 의해 서두르고 있는 이유이다. 자세한 내용은 SemiAnalysis accelerator와 HBM model을 참조하라.

MI350/MI355의 scale-up network의 경우, AMD는 XGMI protocol(PCIe 5.0 PHY Serdes 사용)을 64GByte/s에서 76.8GByte/s로 1.2배 "overclock"할 수 있었다. 이는 32GT/s per link 대신 약 38GT/s per link를 제공하는 PCIe 5.0 PHY extended speed mode를 사용함으로써 이루어진다. 그럼에도 불구하고 비교 가능한 NVIDIA product들은 여전히 MI350/MI355의 scale up network speed를 압도하는데, 이는 HGX B200/B300이 MI350/MI355의 mesh topology 기반 scale-up network보다 1.6배 빠른 switched all to all topology를 사용하기 때문이다. GB200 NVL72/GB300 NVL72의 경우, MI350/MI355의 scale-up solution과 비교하거나 경쟁할 여지가 정말로 없는데, 이는 GB200 NVL72/GB300 NVL72가 단일 scale-up domain 내에서 72개 GPU를 연결하는 진정한 rack scale solution인 반면 MI350/MI355는 scale-up domain에서 8개 GPU만 함께 연결하기 때문이다.

AMD MI355X xGMI Topology

76.8GByte/s Point to Point Mesh Topology



Source: SemiAnalysis

Scale-out domain으로 넘어가면, MI350/MI355는 GPU당 400 Gbit/s의 속도를 지원한다 – 이는 B200과 GB200 NVL72와 동일하지만, GPU당 800 Gbit/s networking을 제공하는 B300 HGX NVL8과 GB300 NVL72에 의해 곧 능가될 것이다. AMD 전체적으로는 scale-out networking에서 뒤처질 것인데, NVIDIA가 올해 말 800GbE ConnectX-8 NIC의 대량 배포를 시작할 것인 반면 AMD의 800GbE "Vulcano" NIC은 2026년 하반기까지 대량 배포를 시작하지 않을 것이기 때문이다.

3 Competitive Performance per TCO with the HGX B200 NVL8

우리는 MI355X가 소규모에서 중간 규모 LLM production inference workload에 대해 HGX B200과 경쟁할 수 있다고 생각한다. 이는 MI355X의 total cost of ownership이 자체 소유 cluster에 대해 HGX B200보다 33% 낮으면서도, 훨씬 더 많은 HBM memory capacity를 제공하고, 약간 더 많은 FP8 및 FP4 TFLOP/s와 두 배의 FP6 TFLOP/s를 제공하기 때문이다. AMD의 AI Software King인 Anush의 리더십 하에 AMD software의 급속한 개선은 또한 MI355X의 상대적인 performance per TCO 우위를 더욱 높일 것이다.

MI355X의 경쟁력에 대한 AMD의 주장은 direct to chip liquid cooling (DLC)이 필요하지 않다는 사실을 중심으로 한다. 어느 정도까지는 분명히 장점이 있지만, AMD가 여전히 차세대 MI355X를 이미 시장에 출시된 지 오래된 NVIDIA의 "economy-class" HGX 제품들의 경쟁자로 홍보하고 있다는 사실에는 어느 정도 아이러니가 있다. AMD의 MI355X는 앞서 언급한 더 작은 scale-up

world size로 인해 frontier reasoning inference에서 NVIDIA의 flagship GB200 NVL72와 정면으로 경쟁할 수 없으므로, 대신 air-cooled HGX B200 NVL8 및 air-cooled HGX B300 NVL8과 경쟁하도록 위치하고 있다. 그렇긴 하지만, 이 제품 segment는 MI355X의 software 품질과 AMD가 기꺼이 판매하려는 가격에 따라 의미 있는 volume을 출하할 것이다. 우리는 대규모 scale-up world size의 이점을 얻지 못하는 소규모에서 중간 규모 model 사용자들 사이에서 가장 많은 견인력을 얻을 수 있을 것으로 예상한다. 하지만 대규모 disaggregated deployment의 이점을 얻거나 대규모 scale up network를 활용할 수 있는 mixture of experts를 사용하는 reasoning model 및 frontier inference deployment에 관해서는, GB200 NVL72가 특히 inference에서 performance 및 perf per TCO에서 여전히 지배적일 것이다.

Total Cost of Ownership and Power Efficiency by GPU						
	Unit	B200 HGX NVL8	B300 1200W	GB200 NVL72	GB300 NVL 36/72	MI355X
Total Upfront Cluster Capex, per Logical GPU	USD	\$44,954	\$53,742	\$54,123	\$61,644	\$24,711
Total Upfront Cluster Capex, per GPU per Hour	USD/hr/GPU	\$1.54	\$1.85	\$1.86	\$2.12	\$0.85
TDP per GPU	W	1,000	1,200	1,200	1,400	1,400
Dense Flops - FP4	TFLOPS	9,000	13,500	10,000	15,000	10,000
Dense Flops - FP6	TFLOPS	4,500	4,500	5,000	5,000	10,000
Dense Flops - FP8	TFLOPS	4,500	4,500	5,000	5,000	5,000
TCO per PFLOPS of FP4	USD/hr/PFLOPS	\$0.17	\$0.14	\$0.19	\$0.14	\$0.08
TCO per PFLOPS of FP6	USD/hr/PFLOPS	\$0.34	\$0.41	\$0.37	\$0.42	\$0.08
TCO per PFLOPS of FP8	USD/hr/PFLOPS	\$0.34	\$0.41	\$0.37	\$0.42	\$0.17
TFLOPS of FP4 per Watt	TFLOPS/W	9.0	11.3	8.3	10.7	7.1
TFLOPS of FP6 per Watt	TFLOPS/W	4.5	3.8	4.2	3.6	7.1
TFLOPS of FP8 per Watt	TFLOPS/W	4.5	3.8	4.2	3.6	3.6

Source: SemiAnalysis

4 NVIDIA is Upsetting Neoclouds with DGX Lepton

이번 주 GTC Paris에서 Jensen Huang은 DGX Lepton과 그 비즈니스 전략에 대해 더 자세히 논의 했으며, 이는 전 세계적인 규모에서 AI 컴퓨팅의 상품화로 이어질 수 있다고 했다. 이는 고객들이 이론적으로 동일한 소프트웨어 사용자 인터페이스와 경험을 유지하면서 추론 워크로드를 다양한 클라우드 간에 자동으로 원활하게 이동할 수 있다는 것을 의미한다. 이는 주로 추론과 소규모 훈련 워크로드에 집중하는 사람들에게 특히 매력적이다. 대규모 추론 배포나 대규모 훈련에서는 DGX Lepton을 사용할 것으로 예상하지 않기 때문이다. DGX Lepton이 성공한다면, 모든 Neocloud에서 정확히 동일한 feature set, value proposition 및 performance를 가진 표준 사용자 경험을 만들어 내어 모든 Neocloud들을 가격 경쟁의 바닥으로 몰아넣을 것이다. 이들은 효과적으로 Neocloud margin을 초저가 commodity 수준의 margin으로 전환시킬 것이다.

Uber/Lyft가 고객과 driver를 연결하는 platform인 것과 같은 방식으로, DGX Lepton은 GPU compute를 위한 그러한 platform이 되려고 하는 것으로 보인다. 유명하게도, Uber/Lyft는 그들의 platform에 종속된 저margin gig economy worker들의 전체 군대를 만들어냈다. DGX Lepton은 Neocloud들에게 동일한 효과를 가질 수 있다. 반면에, Uber/Lyft와 마찬가지로 DGX Lepton은 소비자들에게는 훌륭할 것이다. middleman margin을 낮춤으로써, NVIDIA는 NVIDIA 자체의 엄청난 margin에는 아무런 영향 없이 end user들을 위한 TCO당 performance를 효과적으로 증가시켰다. Compute는 더 저렴해질 것이고, 경험은 표준화될 것이다.

다양한 Neocloud들과의 논의에서, 많은 곳들이 앞서 언급한 이유들로 DGX Lepton Market-place에 대해 만족하지 않고 있다. DGX Lepton에 대해 만족하지 않음에도 불구하고, 많은 곳들이 여전히 NVIDIA와의 좋은 관계를 유지하기 위해 참여해야 한다고 느끼고 있다. [The Information](#)이 최근 게시한 기사는 또한 Neocloud들의 매우 복잡한 감정과 DGX Lepton에 대한 전반적인 불만을 자세히 설명했다. [NVIDIA Lepton team의 일부인 일부 engineer들도 Neocloud들과의 업무 관계가 어떻게 발전할지에 대해 불안해하고 있다고 전해진다.](#)

Jensen^{o]} DGX Lepton에서 취할 수 있는 한 가지 대안적 접근법은 Lepton의 놀라운 software platform을 완전히 open-source로 만들고, 참여하는 Neocloud들이 DGX Lepton marketplace에 참여할 때뿐만 아니라 Lepton의 software를 self-host할 때도 Lepton의 software를 무료로 배포할 수 있도록 허용하는 것이다.

이는 Neocloud들이 NVIDIA의 marketplace와 독립적인 여러 sales channel을 가질 수 있게 하면서도 여전히 소비자들에게 강력한 performance와 더 나은 경험을 제공하여 전체 ecosystem의 기준을 높일 것이다.

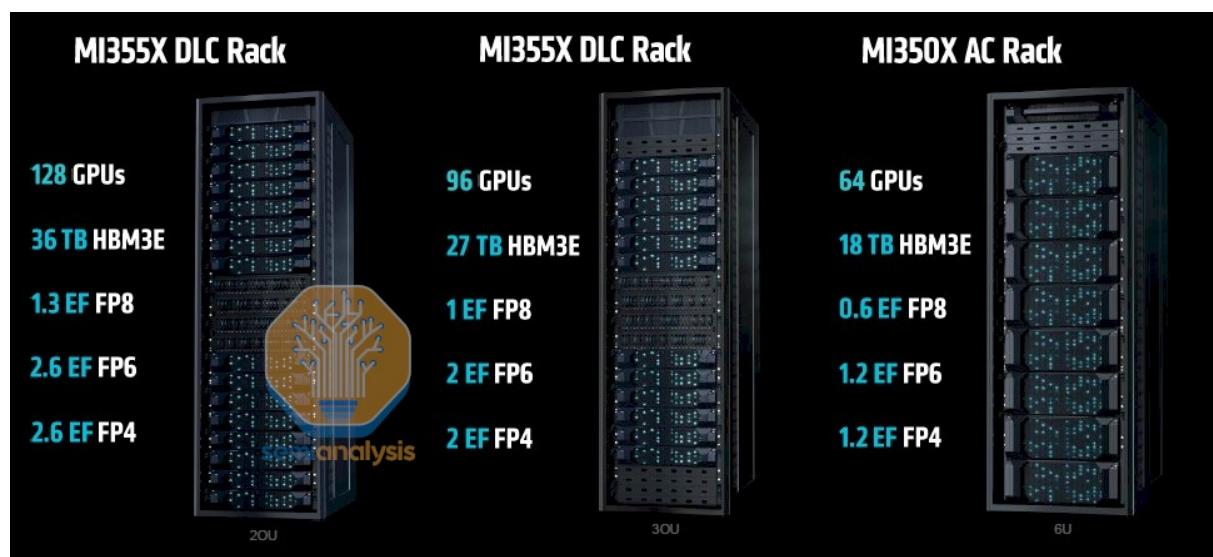
진행 중인 DGX Lepton drama의 한 가지 결과는 Neocloud들이 단일 vendor에 완전히 의존하는 아이디어를 재검토하기 시작했다는 것이다. 많은 곳들이 궁극적으로 이 위험을 완화하기 위한 대안을 찾을 수 있다는 것이다. 이러한 발전은 AMD가 Neocloud engagement를 빠르게 확대하고 AMD GPU를 hosting하는 Neocloud의 수를 빠르게 확장할 수 있는 완벽한 기회를 만들어냈다.

5 MI355X Is Not a Rack Scale Solution – AMD's Marketing Spin

AMD는 MI355X가 어떤 정의로도 rack scale solution^{o]} 아님에도 불구하고 MI355X를 "rack scale solution"으로 마케팅해왔다. MI355X "128 GPU Rack"은 전체 rack에 걸친 coherent scale up domain 없이 동일한 rack에 배치된 16개의 MI355X UBB8 server일 뿐이다.

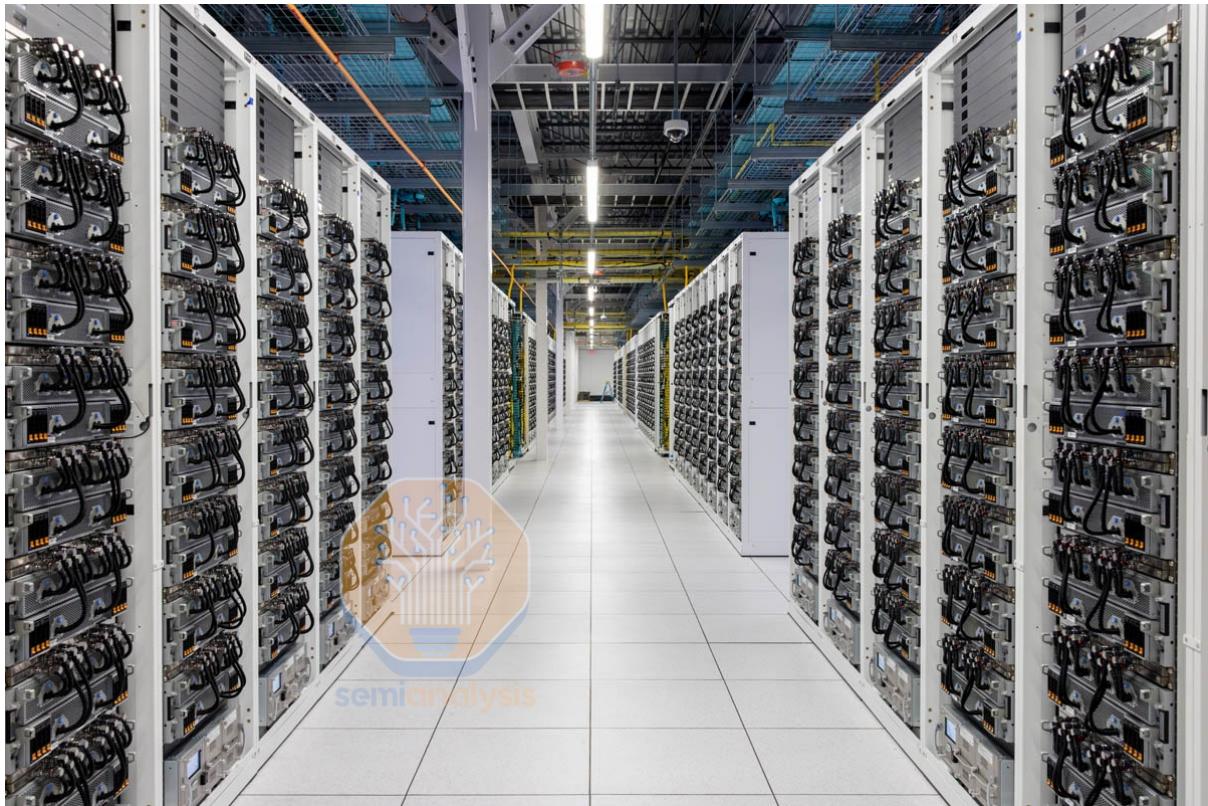
MI355 "128 GPU Rack"은 temu dot com의 rack scale이다. MI355 DLC Rack을 "rack-scale solution"이라고 부르는 것은 다가오는 Hollywood blockbuster에서 Matt Damon 대신 Jesse Plemons를 고용하자고 producer에게 제안하려는 것과 같다.

우리가 더 자세히 설명하겠지만, 이는 MI355X "rack-scale solution"^{o]} GB200 NVL72와 비교하여 18배 더 나쁜 collective performance를 가진다는 것을 의미한다. MI355X의 경우, UBB8 server A의 GPU는 동일한 rack의 UBB8 server B의 다른 GPU와 Ethernet을 통해 400Gbit/s로만 통신할 수 있는 반면, GB200NVL72의 경우 다른 compute tray의 GPU들은 900GByte/s로 통신한다.



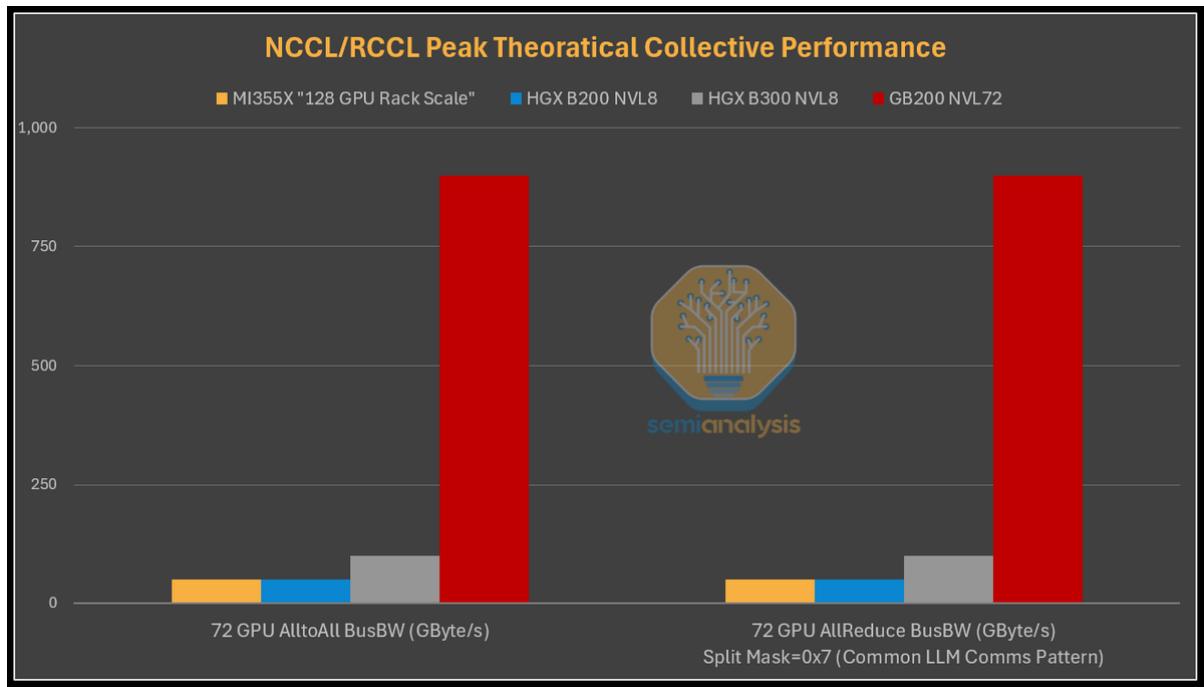
Source: AMD

만약 MI355 128 GPU rack^{o]} "rack scale solution"으로 간주된다면, 많은 H100 rack들도 "rack scale solution"이라고 부르지 않을 이유가 있을까? 분명히, MI355가 "rack scale solution"으로 라벨링된다면, H100도 "rack scale solution"으로 간주되어야 한다. 이는 아무도 rack 당 64개 GPU를 가진 xAI의 H100 배포를 "rack scale solution"이라고 부르지 않기 때문에 터무니없는 제안이다. MI355와 마찬가지로, 이 H100 배포는 모든 64개 GPU에 걸친 coherent scale up domain을 가지지 않으며, rack에 있는 8개의 HGX H100 NVL8 server일 뿐이다.



XAI의 "rack scale solution". Source:[ServeTheHome](#)

mixture of experts model의 inference 및 training에 관해서는, 가장 중요하고 communication 집약적인 collective는 token을 올바른 expert로 routing하는 all to all operation이다. all to all communication의 경우, MI355X는 GB200 NVL72보다 18배 느리고 HGX B300 NVL8보다 2배 느리다. 2D+ parallelism을 사용하여 model을 training하는 경우, 일반적인 LLM pattern은 0x7의 split mask를 가진 all reduce를 사용하는 것이며, 이 operation에서 MI355X는 GB200 NVL72와 비교하여 역시 18배 느리다. 이 예시는 MI355X가 명백히 rack scale이 아니며 GB200 NVL72와 같은 league에 있지 않다는 것을 보여준다.



Source: SemiAnalysis

6 Hyperscale and AI Lab Adoption of new AMD Products

MI355 rack이 어떻게 마케팅되는지에 대한 어리석음에도 불구하고, 우리가 total cost of ownership과 강력한 잠재적 perf per TCO에 대해 제기하는 요점들은 Hyperscaler들과 대규모 AI Lab 고객들에게 명확히 공감을 얻었으며, 우리는 이러한 고객들과의 강력한 engagement와 좋은 주문 momentum을 보고 있다.

AWS는 AMD의 Advancing AI event의 title sponsor였으며, 이제 대규모로 임대를 위한 AMD GPU 구매 및 배포에 대한 첫 번째 진지한 추진에 나설 것이다.

AMD에 관해서는 보통 inference use case에 집중하던 Meta가 이제 AMD에서 training도 시작하고 있다. 그들은 72 GPU rack의 핵심 추진력이며 MI355X와 MI400에 참여할 것이다. Meta의 PyTorch engineer들은 이제 AMD의 engineer들만이 AMD torch에서 작업하는 대신 AMD Torch에서도 작업하고 있다.

OpenAI의 경우, Sam Altman이 AMD event 무대에 올랐다. OpenAI는 우리의 [AMD와 NVIDIA를 benchmarking한 첫 번째 기사](#) 이후 AMD가 얼마나 빠르게 움직이고 있는지를 좋아한다.

x.AI는 production inference를 위해 이러한 다가오는 AMD system들을 사용할 예정이며, AMD의 존재감을 확장하고 있다. 과거에는 protection inference의 작은 비율만이 AMD를 사용했고 대부분의 workload는 NVIDIA system에서 실행되었다.

GCP는 AMD와 협상 중이지만, 그들은 꽤 오랫동안 논의를 해왔다. 우리는 AMD가 몇 개의 핵심 Neocloud들에게 제공하고 있는 것과 동일한 거래를 GCP에게 제공해야 한다고 생각한다 - 즉, AMD의 내부 연구 개발 요구를 위해 compute를 다시 임대하겠다고 제안함으로써 AMD rental 제품을 bootstrap하는 것이다.

Neocloud capacity의 빠른 배포 측면에서 명확한 선구자인 Oracle도 30,000개의 MI355X를 배포할 계획이다.

Microsoft는 MI355의 적은 양만 주문하며 방관하고 있는 유일한 hyperscaler이지만, MI400 배포에 대해서는 긍정적으로 기울고 있다.

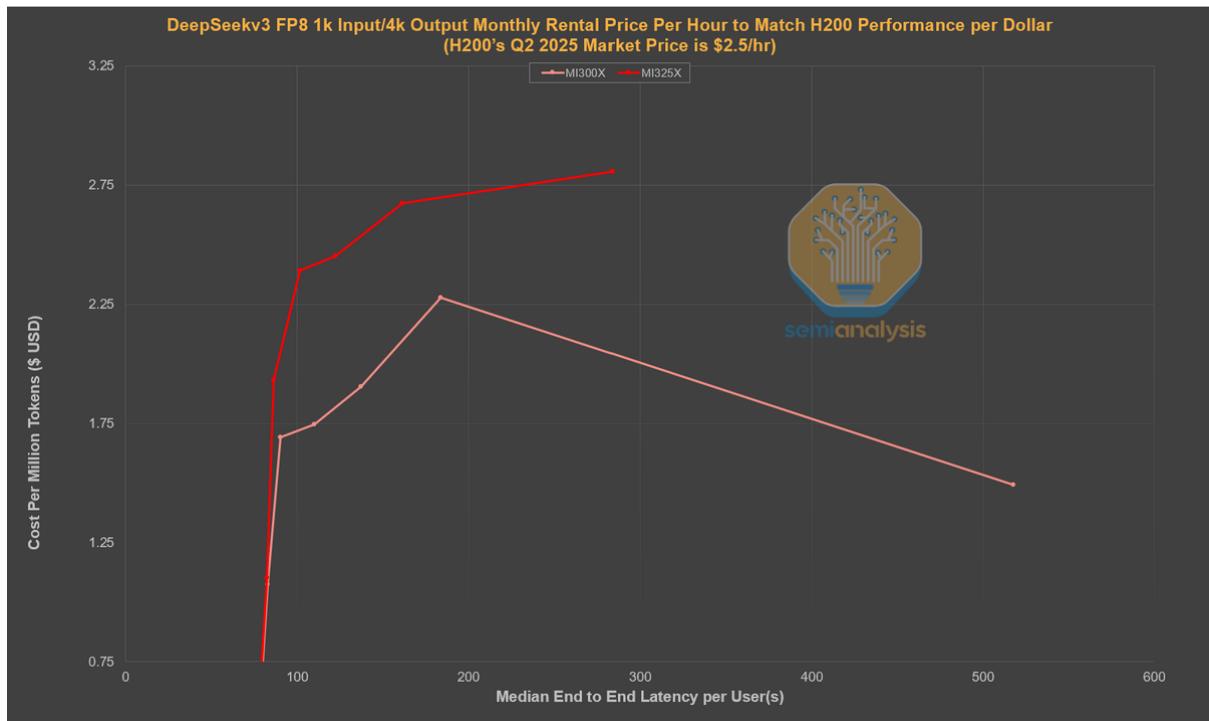
이러한 hyperscaler들 중 많은 곳들이 legacy datacenter 설계 architecture로 인해 air-cooled data center가 풍부하며, 매력적인 perf/TCO proposition을 고려할 때 air cooled MI355X를 채택하는 것을 매우 기뻐한다. 전반적으로, 우리는 이러한 모든 hyperscaler들이 MI355를 배포할 것으로 예상하며 많은 곳들이 MI400 true rack scale solution도 배포할 것이다.

7 AMD's Solving Its Neocloud Rental Market Weakness

AMD 채택 증가의 주요 과제 중 하나는 현재 100개가 넘는 NVIDIA 중심 Neocloud와 비교하여 AMD 중심 Neocloud가 매우 적다는 것이다. 이러한 공급 부족과 rental market에서의 offering 다양성 부족은 AMD GPU 임대에 대한 인위적으로 높은 가격으로 이어져 AMD GPU의 전반적인 cost-competitiveness를 침식한다.

2025년 Q2 현재까지, H200에 대한 현재 1개월 term contract market rental 가격은 약 \$2.50/hr/GPU 이다, 낮은 품질의 cloud에 대해서는 넓은 변동성과 더 낮은 가격을 보인다. MI325X 임대를 위한 1개월 계약은 존재하지 않는다. MI300X 임대를 위한 1개월 계약은 \$2.50/hr로 가격이 책정되어 있어, MI300X를 H200과 비교하여 임대하는 데 경쟁력이 없게 만든다. 아래에서 우리는 MI300X 와 MI325X가 NVIDIA H200 임대와 경쟁력을 갖기 위해 필요한 대략적인 MI300 및 MI325X 1 개월 임대 가격을 보여준다. 이 분석은 [우리의 실제 inference benchmark](#)에 크게 기반했다.

reasoning inference task (1k input, 4k output)의 경우, MI300X는 H200과 경쟁력 있는 dollar 당 performance를 갖기 위해 1개월 계약에서 \$2.10-2.40/hr 미만으로 가격이 책정되어야 한다. MI325X는 상호작용성에 따라 경쟁력을 갖기 위해 \$2.75/hr/GPU에서 \$3.00/hr/GPU 사이로 가격이 책정되어야 한다. 이는 광범위한 협상 없이는 어떤 AMD Neocloud도 제공하지 않는 가격 범위이며, 이는 이러한 시장 비효율성의 결과로 부분적으로 NVIDIA가 현재 임대에 대한 dollar당 performance에서 승리하고 있다는 것을 의미한다.



Source: SemiAnalysis

8 Pushing into High Gear – AMD is Accelerating the Development of the AMD Neocloud Ecosystem

몇 달 전까지만 해도 AMD는 Neocloud ecosystem 내에서 자사 제품의 강력한 성장을 추진하는 데 집중하지 않았으며, GPU cloud들이 AMD GPU를 hosting하고 이를 임대하지 못할 위험을 감수할 만한 충분한 incentive를 제공하지 않았다. 지난 몇 달 동안 AMD leadership은 건강한 Neocloud ecosystem을 구축하는 것이 중요하다는 것을 인식했으며, 이는 developer 채택을 촉진하고 부풀려진 AMD GPU 임대 가격을 낮추는 데 도움이 된다. 최종 결과는 end user에게 더 높은 dollar당

performance와 AMD에 익숙하고 더 넓은 AMD ecosystem에 기여할 수 있는 더 많은 developer 들이다.

이를 위해 AMD는 AWS, OCI, Digital Ocean, Vultr, Tensorwave, Crusoe 및 기타 Neocloud 들에게 이러한 Hyperscaler들과 Neocloud의 AMD 채택을 지원하고 business case의 위험을 줄일 수 있는 놀라운 incentive를 제공했다. AMD가 체결한 거래는 고객들이 더 많은 AMD GPU를 구매하려는 의지에 대한 대가로, AMD가 내부 AMD software 개발 목적을 위한 장기 계약 형태로 이 capacity의 상당 부분을 다시 임대한다는 것이다. 이는 NVIDIA가 이미 GCP, OCI, AWS, Azure, CoreWeave로부터 NVIDIA의 대규모 내부 compute 요구를 위해 GPU의 대규모 cluster들을 다시 임대하는 방식과 유사하다. 일부 Neocloud의 경우, AMD는 Neocloud가 capacity를 완전히 판매 할 수 없는 경우 AMD 자체가 backstop으로 이를 임대할 수 있도록 투자 case의 위험을 완전히 제거하는 incentive를 제공하고 있다. 우리는 현재 유사한 incentive 구조를 제공받으며 AMD와의 잠재적 partnership을 탐색하고 있는 많은 Neocloud를 알고 있다.

이러한 incentive가 마련된 상황에서, 이러한 Neocloud들이 단기적으로만 NVIDIA cluster를 임대하고 상당한 가격 및 점유율 위험을 감수하는 경쟁사들과 비교하여 AMD와 협력함으로써 덜 위험한 business case를 구축할 수 있다고 주장할 수 있다.

AMD의 developer cloud 출시는 또한 AMD의 compute를 경쟁력 있는 가격으로 보편적으로 이용 가능하게 만드는 핵심 전략이다. 이 출시의 일환으로, AMD는 MI300X GPU 임대 가격을 대폭 낮춰 더 넓은 developer 계층에 대한 접근을 민주화했다. 안타깝게도 우리가 테스트했을 때 기본 quota는 0개 GPU로 설정되어 있었고 GPU quota 증가를 얻는 것이 어려웠다. 우리는 AMD가 새로운 사용자의 기본 quota를 최소 16개 MI300X GPU로 설정하여 developer들을 ecosystem에 더 효과적으로 도입할 것을 권장한다. AMD developer cloud on demand 가격이 훨씬 더 합리적인 1.99\$/hr/GPU 가격으로 설정되어 있기 때문에, on demand MI300을 제공하는 AMD Neocloud 들은 오늘날의 높은 수준인 \$3/hr/GPU에서 \$2/hr/GPU로 가격을 낮춰 이에 맞춰야 할 것으로 예상한다.

Create GPU Droplet

Choose a GPU Plan

GPU Plans

AMD MI300X
1 GPU - 192 GB VRAM - 20 vCPU - 240 GB RAM
Boot disk: 720 GB NVMe- Scratch disk: 5 TB NVMe

AMD MI300Xx8
8 GPU - 1.5 TB VRAM - 160 vCPU - 1920 GB RAM
Boot disk: 2 TB NVMe- Scratch disk: 40 TB NVMe

Summary

GPU \$15.92/hr

Type: AMD MI300Xx8
GPU: 8
VRAM: 1.5 TB
vCPU: 160
RAM: 1920 GB
Boot Disk: 2 TB NVMe SSD
Scratch Disk: 40 TB NVMe SSD

Total cost \$15.92/hour

Create GPU Droplet

Your current limit does not allow creating any GPU Droplet. Request an Increase ↗

Source: SemiAnalysis, AMD

9 ROCm Software Improvements

AMD는 inference capability와 performance에 중점을 둔 ROCm 7을 발표했다. inference throughput performance의 경우, AMD는 ROCm 6 대비 ROCm 7의 평균 3.5배 개선과 DeepSeek R1을 serving할 때 NVIDIA B200 대비 ROCm7의 1.3배 개선을 자랑했다. 우리는 이러한 주장들을 검증하기를 기대한다.

AMD는 또한 distributed inference에서 open ecosystem과 협력하는 데 전념하고 있다. inference framework인 vLLM과 SGLang을 지원하는 것 외에도, AMD는 distributed inference 기법인 PD disaggregation을 가능하게 하는 NVIDIA Dynamo의 대안인 orchestration framework llm-d를 지원한다. llm-d stack은 여전히 NVIDIA Dynamo KVCache manager와 동일한 기능을 제공하는 상당한 기능들이 부족하다. KVCache manager는 많은 inference workload에 대해 throughput에서 여러 배의 개선을 unlock할 수 있는 inference workload에 대한 대규모 TCO 이익을 제공할 수 있기 때문에 매우 중요하다.

kernel 작성 library인 Triton에 대한 ROCm의 지원도 지난 몇 버전에서 크게 개선되었다. ROCm은 작년에 Triton에 대한 기능적 지원을 달성했으며, ROCm 7은 performance 개선에 중점을 둔다. 우리는 AMD가 노력을 계속하고 FlexAttention과 같은 고급 기능에 대한 지원을 확장하기를 바란다.

최근 ByteDance Seed는 compute와 GPU communication overlap을 가능하게 하는 Triton 기반 library인 Triton Distributed를 만들었다. AMD는 Triton Distributed에 큰 관심을 보였으며 이에 대한 더 큰 지원에 대해 이야기했다. 그러나 OpenAI(Triton의 maintainer)가 ByteDance의 Triton Distributed 기능들의 기여를 원래 Triton library로 다시 받아들일지는 불분명하다. OpenAI가 Triton을 위한 distributed compute-comms kernel 구현에 대한 자체 경로를 추구하고 있을 가능성성이 있다.

더욱이, 중국에 대한 상당한 chip 수출 제한을 고려할 때, ByteDance가 서구 GPU를 위한 open-source library에 기여하는 것에서 물러날 가능성성이 있다. 그렇긴 하지만, ByteDance는 AMD에 크게 투자하고 있으며 우리는 그들이 의미 있는 AMD 기반 rental GPU capacity를 차지할 것으로 예상한다. 그러나 ByteDance는 여전히 주로 NVIDIA 진영에 남아있을 것인데, 이는 그들의 compute capacity 확장의 대부분이 NVIDIA 기반 capacity 임대에서 나올 것이기 때문이다. ByteDance의 compute 대부분은 Cloud 임대 또는 중국 외부에 위치한 대규모 전용 bare metal cluster에서 나오며, 그들의 Neocloud 및 Cloud provider 대부분은 여전히 주로 NVIDIA compute capacity에 의존한다.

더 낮은 수준에서, AMD는 인기 있는 data transfer interface인 Mooncake Transfer Engine과 expert parallel communication library인 DeepEP를 통합하고 있다고 주장했다. 그러나 이 글을 쓰는 시점에서 우리는 아직 DeepEP나 Mooncake가 포함된 open source ROCm repo를 보지 못했다.

마지막으로, AMD는 Developer Cloud와 Developer Credits program을 발표했다. compute 접근을 신청하기 위한 간단한 interface 외에도, AMD는 developer들이 ROCm PyTorch, HipBLAS와 같은 ROCm library들, 그리고 이러한 ROCm library들을 위한 개발 도구들을 쉽게 설치할 수 있도록 Python package "rocm"을 만들었다. 모든 code는 GitHub repo [ROCM/TheRock](#)에서 open source로 제공된다.

10 MI355X PyTorch Continuous Integration (CI) and Testing

AMD는 MI355 chip을 위한 CI와 자동화된 testing을 Pytorch에 추가하는 작업을 시작했다. MI355X PR 중 어느 것도 아직 merge되지 않았지만, AMD가 첫날부터 open source PyTorch MI355X CI에 대해 생각하고 있는 것을 보는 것은 훌륭하다. NVIDIA의 경우, Blackwell의 대량 배송 이후 6개월이 지났지만 open source PyTorch를 위한 CI를 시작하지 않았으며 내부 Blackwell CI에만 집중해왔다. 실제로 Meta는 PyTorch CI 비용의 대부분을 지불하여 월 100만 달러 이상을 지출하는 반면, AMD는 AMD에서 open source PyTorch CI를 위해 자체적으로 지불한다. NVIDIA는 지금 까지 open source PyTorch CI에 의미 있는 자금이나 compute를 기부하지 않았지만, DGX Cloud에서 많은 compute credit을 기부하고 다양한 Neocloud provider로부터 임대한 GPU capacity를 Meta open source PyTorch에 기부함으로써 기여할 계획을 세우고 있다.

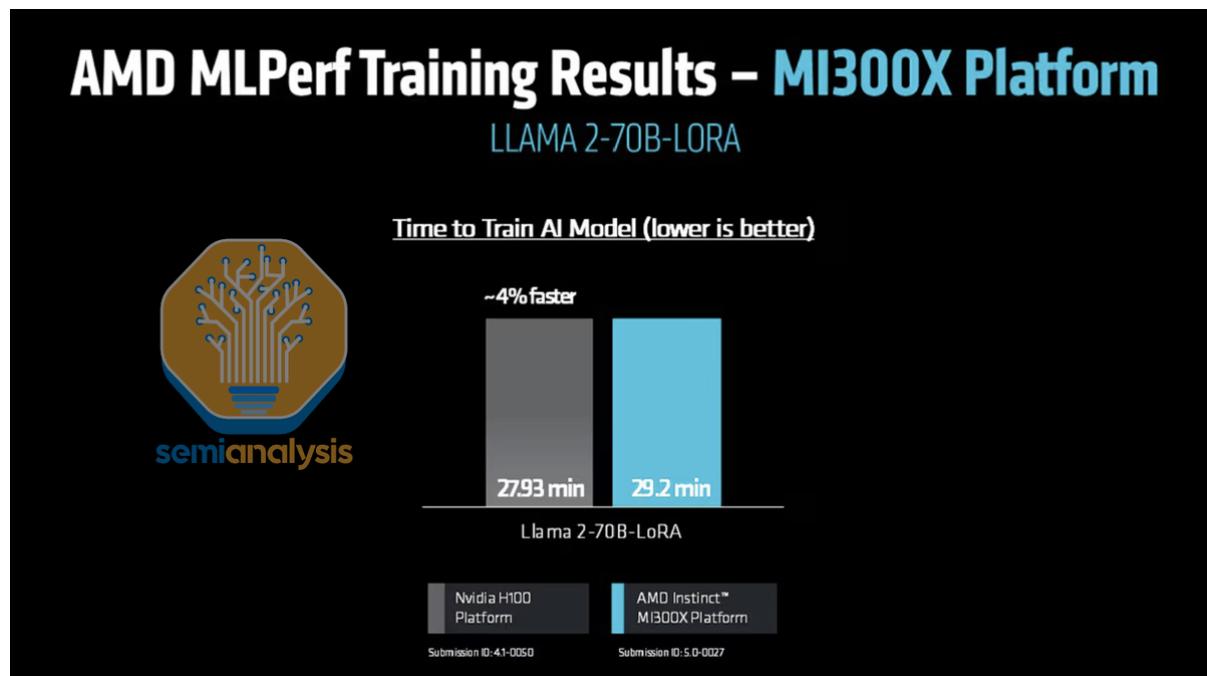
NVIDIA는 open source B200 PyTorch CI 추가 작업을 적극적으로 진행하고 있으며 PyTorch CI 목적으로 PyTorch Foundation에 48개의 B200을 기부하기로 약속했다. 모든 사람이 0일차부터 CI를 갖는 것을 선호하겠지만, 6개월 후에 Blackwell open source CI를 PyTorch에 추가하는 것은 늦었지만 안 하는 것보다는 낫다. 우리가 AMD의 CI 부족에 대해 조명한 것이 그들이 여기서 상당한 진전을 이루도록 nudge했을 가능성이 있다. NVIDIA는 Blackwell을 위한 PyTorch CI에

훨씬 더 많이 투자해야 한다. 또한, consumer AI가 안정적이 되도록 하기 위해 그들의 consumer GPU들이 PyTorch와 인기 있는 inference library들을 위한 CI에 추가되어야 한다. 현재 NVIDIA consumer GPU들은 CI resource 부족으로 인해 특정 framework를 사용할 때 일부 불안정성을 경험한다.

11 ROCm MLPerf Training Submission

지난달, AMD는 single node Llama2 70B LoRA finetuning과 BERT training을 위한 첫 번째 MLPerf Training run을 제출했다. 이는 training이 single AMD node에서 작동할 수 있음을 보여주는 매우 중요한 발전이다. 다음 단계로, AMD는 MLPerf Llama 405B multi-node training benchmark와 같은 더 많은 실제 training benchmark에 참여해야 한다. 우리는 그들이 이 테스트에서 경쟁력 있는 결과를 보여줄 수 있다고 생각한다.

benchmarking에 관해서는, AMD가 자신들의 solution이 잘 작동할 때 MLPerf run에 대한 따라하기 쉬운 재현 가능한 지침을 제시함으로써 명확하게 보여주는 방식이 마음에 듈다. 이는 재현하기 매우 어려운 NVIDIA의 MLPerf submission과 대조적이다.



Source: AMD

12 MIG Partitioning is Wasting Time and Engineering Resources

AMD는 현재 GPU partitioning을 지원하는 것을 목표로 하는 그들의 pet project에 많은 engineering resource와 돈을 낭비하고 있다. 이 project는 사용자가 하나의 GPU를 8개의 더 작은 GPU로 나눌 수 있게 해준다. 어떤 고객도 이것을 요구하지 않는다. Meta, OpenAI, x.AI 모두 이것을 요구하지 않는데, 이는 모든 online inferencing workload가 최소한 하나의 GPU를 필요로 하기 때문이다. 우리는 AMD hardware engineer들이 GPU당 대량의 HBM을 가진 가장 진보된 chip 중 하나를 개발하기 위해 열심히 일했는데 이 GPU를 8개 부분으로 나누고 싶어한다는 것이 비논리적이라고 생각한다.

실제로, Meta, OpenAI, x.AI 모두 이것의 반대를 원하며 DeepEP 및 disaggregated prefill과 같은 기법을 사용하여 최소 16개 GPU를 사용하는 multi-node inferencing에 대한 더 나은 지원을 AMD가 갖기를 원한다.

AMD Instinct™ MI350 Series GPU Partitioning
Flexible Partitioning for Bare Metal and SR-IOV

Maximize GPU Utilization with up to 8 spatial partitions

HBM can be one of two NUMA per socket modes⁺

- Instinct MI300X enabled NPS1 & NPS4
- Instinct MI350 Series GPUs now support NPS1 & NPS2

Instinct MI350 Series GPUs can support

- Up to 520B parameter AI Model a SPX+NPS1
- Up to 8x Instances of Llama 3.1 70B in CPX+NPS2

+ Memory partitioning can only be changed via GPU reset

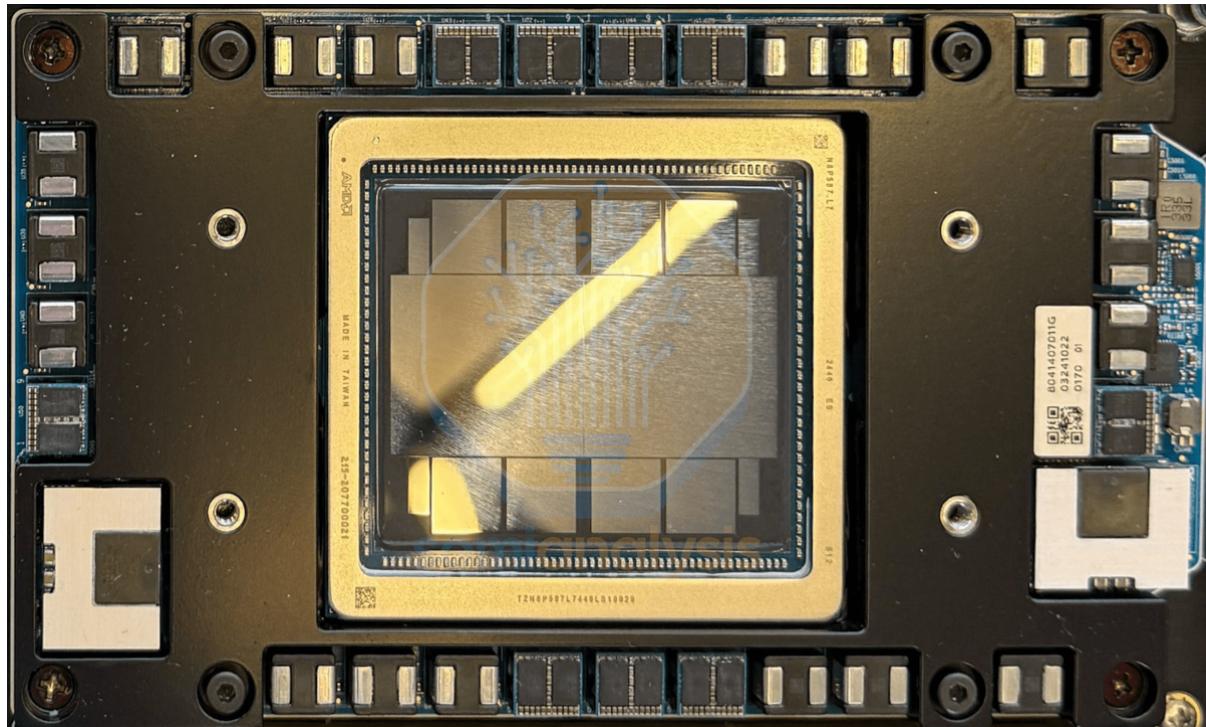
See endnote MI350-012, MI350-037

6 Advancing AI 2025

AMD together we advance...

Source: AMD

13 MI355X Manufacturing – Updated Chiplet Architecture



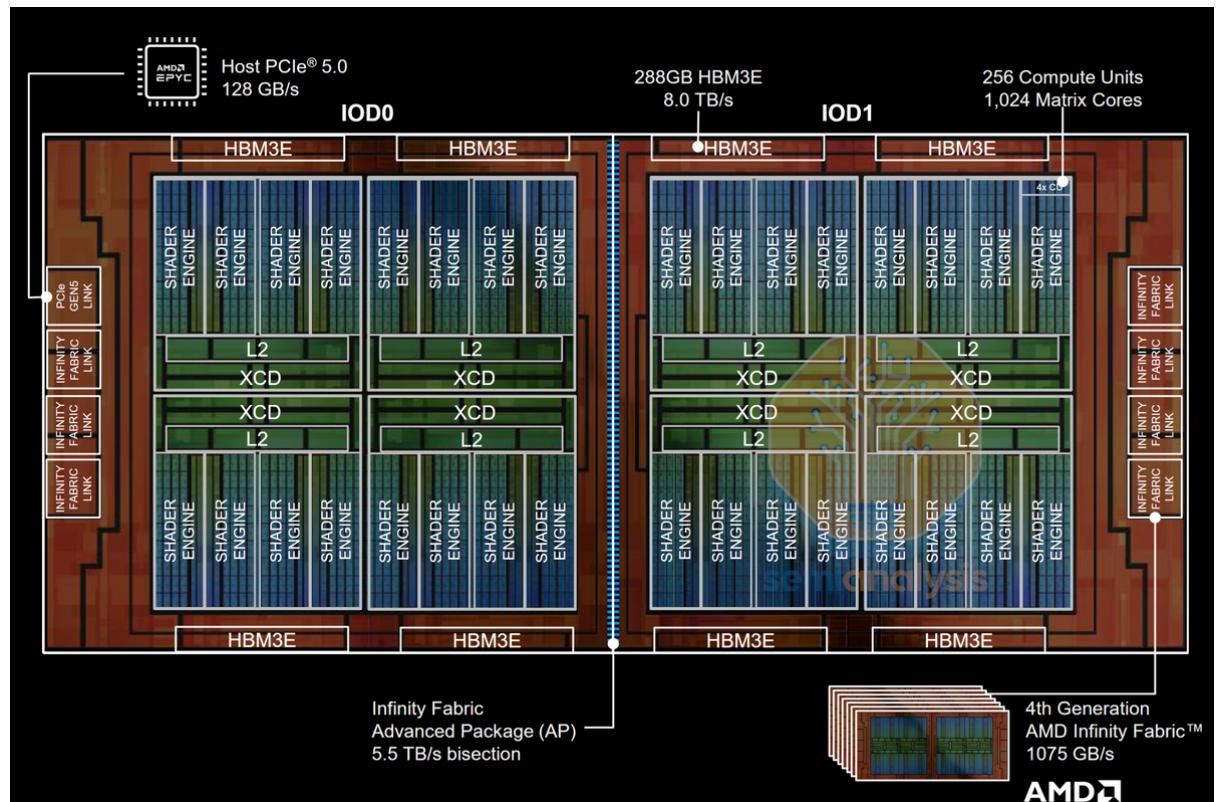
Source: SemiAnalysis

AMD는 MI300 출시 이후 2년 동안 그들의 chiplet architecture를 개선하는 데 사용했다. 위 이미지의 silicon에서 볼 수 있듯이, chip layout이 약간 조정되었으며, base Active Interposer Dies (AID) 가 4개 quadrant에서 2개의 reticle 크기 절반으로 병합되었다. HBM 위치의 minor adjustment는

structural support silicon die를 HBM site 사이에서 모서리로 이동시켰다.

cross-chiplet communication에 대한 이점은 명확하며, 2.5D Infinity Fabric Advanced Package link의 전체 축을 제거하여 더 적은 chip boundary crossing을 요구함으로써 power와 area를 절약한다. 또한 MI300의 반대편 모서리 quadrant가 서로 통신하기 위해 die를 가로질러 두 번 점프해야 했던 two-hop scenario를 제거한다.

그러나 이러한 배치는 또한 3D stacking yield에 추가적인 중요성을 부여한다. AMD는 TSMC의 SoIC hybrid bonding process를 계속 사용하는데, 이제 각 base die에 두 배 많은 Accelerator Complex Dies (XCD)를 부착해야 하므로 문제가 있을 경우 잠재적으로 yield loss와 추가적인 silicon 낭비를 복합시킬 수 있다. AMD가 이 경로를 선택한 것은 TSMC의 SoIC flow의 성숙도와 SoIC의 lead customer로서 5년에 걸친 AMD의 Foundry Technology & Operations team과의 깊은 partnership을 말해준다.



Source: AMD Advancing AI

여전히 TSMC N6에 있지만, base die는 여러 speed upgrade를 받았다. 남은 die-to-die link는 4.8TB/s bisection equivalent에서 MI350의 5.5TB/s bisection equivalent로 upgrade되었다. scale-up을 위한 Infinity Fabric의 속도는 20% 향상되었다. 더 중요하게는, memory controller가 이제 더 빠른 HBM3E를 처리할 수 있다. AMD는 AID 및 HBM attach를 위해 검증된 CoWoS-S를 고수했으며, footprint가 MI300과 동일하게 유지되었다고 언급했다.

compute die의 경우, XCD는 N5에서 TSMC의 N3P node로 이동했으며, 아래에 자세히 설명된 업데이트된 CDNA4 architecture를 가진다. 이번에 AMD는 MI300의 40개 중 38개와 비교하여 die에 인쇄된 36개 CU 중 32개만 활성화했다. 흥미롭게도, AID에서 XCD의 orientation이 변경되어 data bond pad가 AID의 중앙 영역에 위치하게 되었다. 그런 다음 data는 256MB의 Memory Attached Last Level (MALL) cache를 통해 밖으로 진행한 후 HBM에 도달한다.

전체적으로, 새로운 chip은 1,850억 개의 transistor를 포함하며, 이는 MI300 대비 21% 증가이다. 우리는 각 AID에 약 230억 개의 transistor가 들어가고, 각 XCD에 174억 개의 transistor가 들어간다고 추정한다. 이는 N5에서 N3P로 이동하면서 30%의 transistor budget 증가를 의미한다.

14 CDNA4 Microarchitecture (UArch)

AMD의 architecture 설계는 전통적인 HPC 중심에서 AI workload에 최적화된 것으로 점진적으로 이동해왔다. CDNA 4에서 우리는 AMD가 architecture에 관해서는 AI로 더욱 pivot하면서 legacy HPC의 잔여 영향이 계속 사라지는 것을 보지만, CDNA4는 여전히 FP64 matrix core에 많은 floor area를 낭비한다.

CDNA 4는 256개의 compute unit (CU), 160 KB의 local data share (LDS – SMEM equivalent), 그리고 FP16에 대해 CU당 cycle당 4,096 FLOP으로 실행되는 matrix core를 제공한다. CDNA 3와 비교하여, 이는 CU 수의 16% 감소, LDS capacity의 1.5배 증가, 그리고 matrix core throughput의 2배 증가이다. 이러한 변화들은 모두 더 큰 array size를 가진 AI workload로 수렴하는 architecture의 신호이다. HPC workload는 일반적으로 많은 수의 CU로부터 이익을 얻는 반면, AI workload는 각 CU가 큰 matrix를 계산하는 것으로부터 이익을 얻으며, 이 두 요구사항은 power 및 area budget에서 경쟁한다. LDS capacity의 증가는 matrix core가 너무 빨라서 AMD가 core에 data를 충분히 빠르게 공급하기 위해 secondary buffer size를 증가시켜야 한다는 것을 보여준다. AMD가 일반적인 staging buffer VGPR (RMEM equivalent) 대신 LDS를 증가시켰다는 점을 고려할 때, 우리는 차세대 matrix core가 matrix core performance를 계속 scaling하기 위해 큰 architectural 변화를 요구할 것이라고 의심한다.

CDNA 4는 FP8에 대해 FP16 대비 2배의 throughput을, FP4에 대해 4배의 throughput을 제공한다. 흥미롭게도, CDNA 4의 FP6 throughput은 FP6와 FP4가 data path를 공유하기 때문에 이론적으로 FP4 throughput과 동일하다. 그러나 FP6 throughput은 실제 설정에서 power 제한으로 인해 여전히 FP4 throughput보다 약간 낮을 것이다. 이는 FP6 throughput이 FP8과 동일하게 라벨링되는 NVIDIA Blackwell과 다르다.

그러나 NVIDIA의 Blackwell 설계와 비교하여, CDNA 4는 asynchronous feature, data transfer acceleration hardware (sm90/sm100 TMA 등), TMA multicasting 또는 specialized memory (sm100 TMEM)가 없다. 이는 NVIDIA의 SM100 대비 CDNA4에서 intelligence 단위당 더 나쁜 picoJoule을 초래한다. 이 글을 쓰는 시점에서, 우리는 여전히 WGMMA equivalent가 있는지 확인하기 위해 MFMA operation의 변화를 보기 위한 ISA에 대한 세부사항을 기다리고 있다. 그렇긴 하지만, CDNA 4는 또한 이러한 feature들이 performance를 더욱 scaling하는 데 필요하다는 것을 보여주므로, 우리는 CDNA-NEXT에서 급격한 architectural 변화를 볼 것으로 예상한다.

15 AMD Advancing AI Developer Session Track is Disappointing

AMD는 올해 [ROCM blog](#)의 developer content에 큰 개선을 이루었다. 우리는 AMD가 stack 전반에 걸쳐 많은 developer session을 host할 것이라는 희망을 가지고 AMD Advancing AI에 왔지만, talk와 session의 set은 우리를 실망시켰다. RCCL에서 Composable Kernel, rocSHMEM, aiter 등에 이르기까지 대부분의 AMD library에 대한 talk가 없었다. 우리는 AMD가 올해 후반에 더 집중된 conference에서 developer들이 그들의 관심 영역에 더 집중할 수 있도록 talk와 seminar의 set을 확대하기를 바란다.

Developer Sessions		
Talks	NVIDIA GTC 2025	AMD Advancing AI 2025
Template Kernel Libraries	CUTLASS [S72720]	No Composable Kernel Session
PyTorch	CUDA PyTorch [S71946]	No ROCm Pytorch Talk
PyTorch Optimization	PyTorch Crushing White Space [s73733]	No PyTorch Optimization Talk
JAX	CUDA JAX [S73266]	No ROCm JAX Talk
vLLM	vLLM Talk [s72114]	vLLM Talk From Simon
SGLang	No SGLang Talk	SGLang Talk from xAI & Imsys
High Level Communication Libraries	NCCL [S72583]	No RCCL Talk
Low Level Communication Libraries	NVSHMEM [s72578]	ByteDance Triton Distributed Talk But No ROCSHMEM Talk
System Level Profiling	Nsight Analysis [dlit74509]	No rocProfiler or rocprofiler-systems Talk
Low Precision Numerics	MX8/MX6/MX4/NVFP4 Numerics [s72458]	No MX8/MX6/MX4 Talk
AI Kernel Libraries	Blackwell cuDNN [s73071]	No AITER/AOTriton/MIOpen Talk
GEMM Libraries	cuBLAS [s72434]	no rocBLAS/hipBLASLt Talk
FP8 Training	Transformer Engine [s72778]	No Transformer Engine or TorchAO Talk
Current Gen Triton Talk	Blackwell Triton [s72876]	No CDNA4 Triton Talk
Kernel Beginner Tutorial	No	Triton Workshop Tutorial
Disaggregated Optimization Inference	Dynamo [s73042]	No lilm-d or disaggregated prefill talk
Python DSL	1000 Ways to Program CUDA in Python[s74639]	Modular Mojo Talk

Source: SemiAnalysis, NVIDIA, AMD

16 RCCL – ROCm Collective Communication Library

AMD는 그들의 새로운 400G NIC이 Ultra Ethernet (UEC) ready가 될 것이며 기존 RoCEv2 protocol뿐만 아니라 새로운 Ultra Ethernet transport (UEC) protocol도 지원할 것이라고 발표했다. UEC mode에서, 이 NIC은 Bluefield-3와 달리 NIC reordering buffer를 사용하지 않고 GPU memory로의 out of order direct placement와 함께 packet spraying을 지원할 수 있을 것이다. AMD의 새로운 in house 400G NIC은 NVIDIA의 CX-7 NIC이나 Broadcom의 Thor-2 NIC에 의존하는 대신 software를 더 쉽게 수직 통합하고 out of the box experience를 개선할 수 있게 해줄 것이다. Oracle뿐만 아니라 Tensorwave와 같은 AMD Neocloud들이 AMD의 NIC 채택을 약속했지만, Meta는 초기 testing이 아직 AMD NIC 채택을 편안하게 만들지 못했기 때문에 보류하고 있으며, 대신 그들의 MI355X cluster에 ConnectX-7 NIC을 사용할 것이다. Broadcom의 Thor 2와 Thor 3 NIC은 AMD와 NVIDIA의 NIC을 그들의 solution에 수직 통합하는 전략으로 인해 시장 채택에서 어려움을 겪었다. 그러나 우리는 다양한 ASIC program에서 Broadcom의 NIC을 위한 자리가 있다고 생각한다.

AMD의 in house 400GbE NIC은 또한 RING 및 PAT와 같은 algorithm을 위한 all-gather collective를 offload하는 능력과 같은 여러 흥미로운 feature를 지원한다. AMD는 CPU proxy thread도 NIC으로 offload될 것이라고 주장하지만, 이것이 그들이 IBGDA를 사용하는 것인지 아니면 다른 것을 하는 것인지 확실하지 않다.

ROCm 7.0의 RCCL communication library도 출시되었지만, 안타깝게도 이것은 다시 한 번 NVIDIA의 NCCL의 단순한 carbon copy fork로 보이며, 따라서 AMD의 multi node capability를 저해하는 핵심 bottleneck으로 남아있다. 우리의 AMD 2.0 기사에서 권장한 바와 같이, 우리는 여전히 AMD가 NVIDIA의 software를 fork하는 것에 의존하는 대신 그들의 communication library를 처음부터 완전히 다시 작성해야 한다고 생각한다.

17 New AMD Initiatives to Pay AI Engineers Market Rate

대부분의 AMD AI engineer들이 시장 수준보다 다소 낮은 보상을 받아왔다는 것은 업계 내에서 잘 알려져 있다. 유일한 예외는 최근 몇 달 동안의 몇몇 신규 채용과 인수를 통해 들어온 engineer들에게 제한되는 것으로 보인다. 예를 들어, 2년 전 인수인 NodAI에서 들어온 대부분의 AI engineer들은 경험과 skillset을 동일하게 맞춰도 기존 AMD engineer들보다 상당히 높은 보상을 받고 있다. 흥미롭게도, AMD의 HR 부서는 이미 몇 분기 전에 이 문제를 제기하고 이러한 급여 격차를

인식하여 내부적으로 flag pole을 올렸지만, AMD management는 아직 이것을 낮은 우선순위 문제 이상으로 끌어올리지 못했다. 이 시점에서, [AMD가 그들의 AI engineer들에게 시장 수준보다 훨씬 낮은 급여를 지급한다고 설명하는 우리의 공개 기사 이후](#), AMD의 Head of HR은 즉시 이 문제를 최우선 순위로 끌어올렸으며 이러한 대규모 급여 격차를 해결하는 process를 적극적으로 우선시하고 있다 – 하지만 구현은 여전히 진행 중이다. AMD가 수십억 달러의 현금을 보유하고 있다는 점을 고려할 때, 우리는 AMD가 올바른 일을 하고 AMD의 성공과도 연계된 경쟁력 있는 total compensation을 그들의 top individual contributor들에게 지급하기를 희망한다.

18 MI400 Series Flexible Input Output (I/O)

AMD는 NVLink보다 훨씬 나쁜 Infinity Fabric을 배포한 MI300X에서의 실수로부터 배웠다. 그들은 또한 NVSwitch equivalent를 실행할 hardware talent가 없다는 것을 인식했다. 더 나아가, 그들은 또한 너무 많이 수직화함으로써 industry ecosystem을 침범하고 싶지 않다. 따라서 그들은 모든 것을 지원하는 shotgun approach를 택했다.

flexible I/O lane이 등장한다. PCIe 및 Scale Up과 같은 각각의 다른 I/O type에 대해 별도의 SerDes 및 I/O path를 사용하는 대신, AMD는 많은 다른 standard를 지원할 수 있는 144개의 I/O lane을 제공한다. 이러한 I/O lane은 PCIe 6.0, 64G에서의 Infinity Fabric, 128G에서의 UALink, 128G에서의 xGMI 4 (UALink의 어느 정도 superset), 그리고 212G에서의 Infinity Fabric over Ethernet을 지원할 수 있다. 이 approach는 AMD silicon team이 다양한 use case에 대해 최대한의 유연성을 갖도록 한다.

Flexible I/O를 통해, AMD는 scale up UALink 또는 UALink over Ethernet을 배포할 수 있다. 그들은 GPU에 직접 연결된 SSD를 지원할 수 있다. 그들은 UALink를 통해 NIC을 연결할 수 있다. 가능성은 거의 무한하다. 이는 system에 대한 엄청나게 큰 permutation array이며 많은 변화와 진화를 허용한다.

그러나 이러한 다른 형태의 I/O를 허용하는 silicon engineering을 실행하는 것은 쉽지 않다. AMD는 이 모든 다른 permutation과 작동하는 SerDes와 data path를 만들어야 한다. 이는 engineering 위험으로 가득한 엄청나게 어려운 engineering path이다.

다음 section에서, 우리는 MI400 true rack-scale solution에 대해 훨씬 더 깊이 들어가서 핵심 scale-up architecture 선택을 논의하고, elevation diagram과 board 설계 illustration의 도움으로 전체 rack 설계를 설명할 것이다. 우리는 또한 자세한 bill of materials breakdown과 total cost of ownership 및 performance per TCO 분석을 제공할 것이다.