November 2025

SEMIANALYSIS GSA

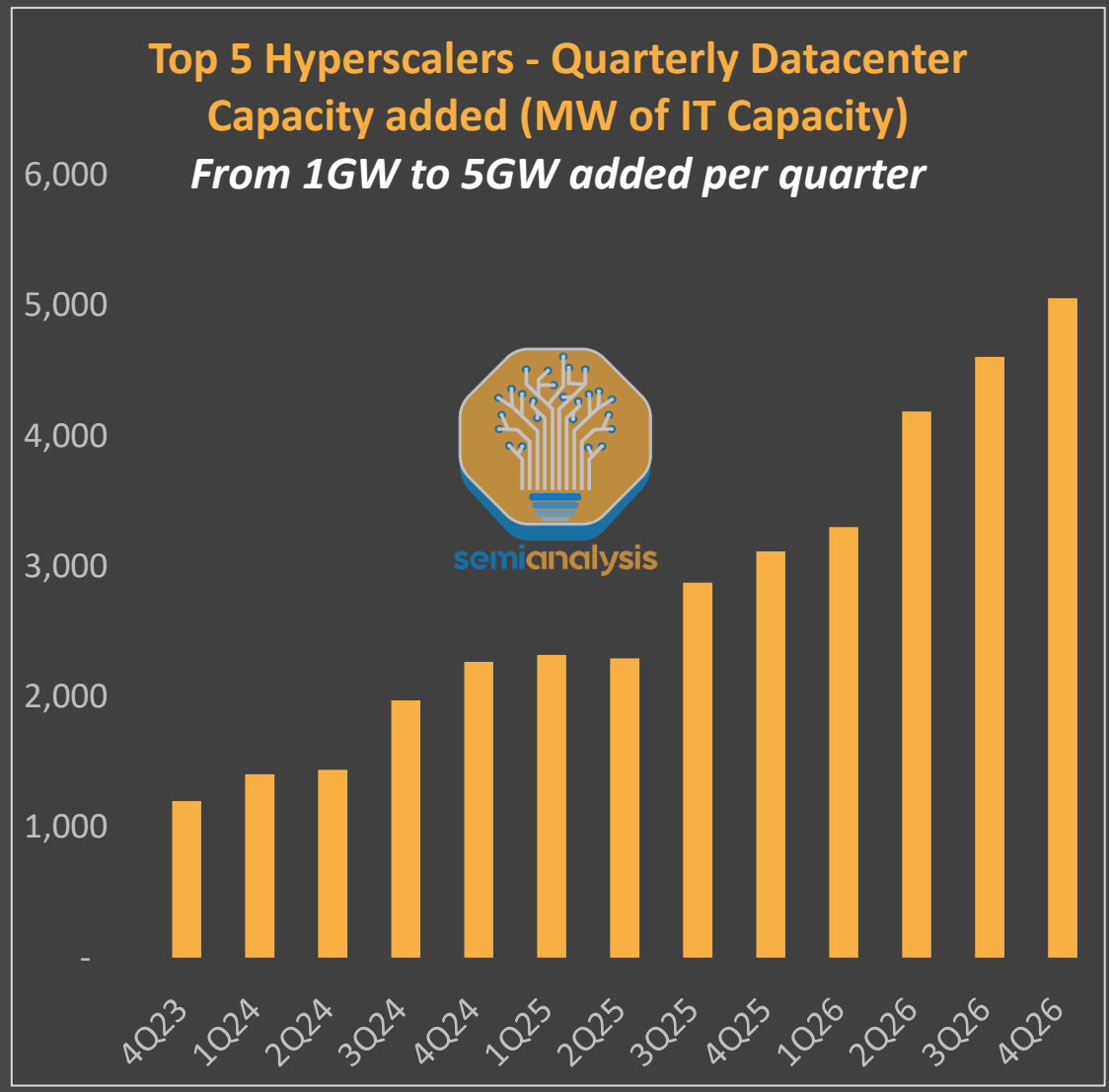
DYLAN PATEL

semianalysis

HYPERSCALER CAPEX – OVER HALF A TRILLION USD BY 2026
DRIVEN BY UNPRECEDENTED DATACENTER CAPACITY GROWTH

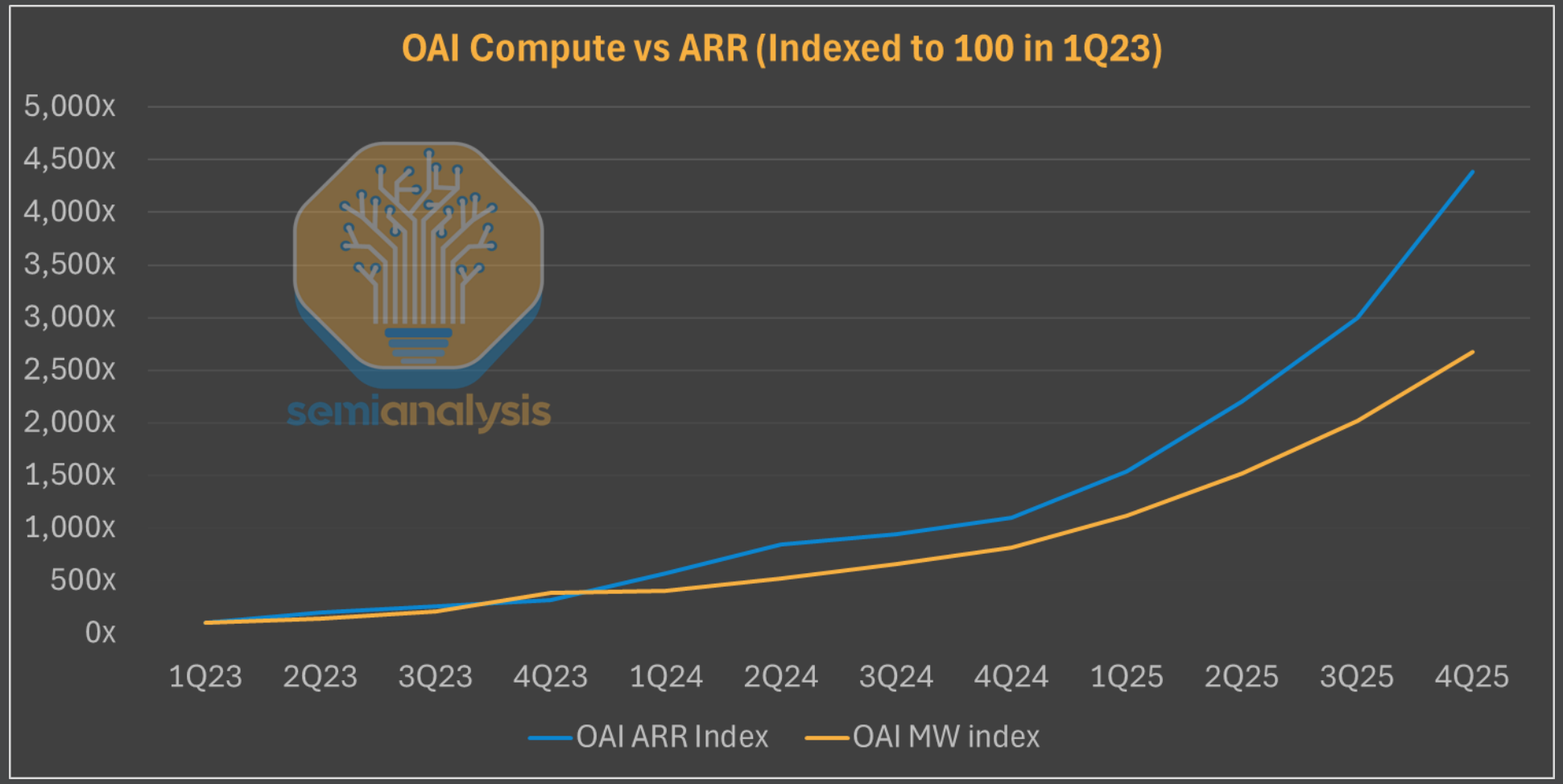
Hyperscaler CapEx (CY) - SemiAnalysis vs. Consensus					
Company	Unit	2023	2024	2025	2026
MSFT*					
SA Current	**$M**	**41,200**	**75,600**	**106,006**	**137,225**
Consensus	$M	41,200	75,600	107,360	125,015
% Difference	%	0%	0%	-1%	10%
META					
SA Current	**$M**	**27,266**	**37,256**	**68,909**	**110,970**
Consensus	$M	27,266	37,256	69,024	98,979
% Difference	%	0%	0%	0%	12%
GOOGL					
SA Current	**$M**	**32,251**	**52,535**	**84,753**	**120,114**
Consensus	$M	32,251	52,535	84,442	95,767
% Difference	%	0%	0%	0%	25%
AMZN					
SA Current	**$M**	**48,133**	**77,658**	**120,390**	**161,575**
Consensus	$M	48,133	77,658	119,057	130,344
% Difference	%	0%	0%	1%	24%
ORCL					
SA Current	**$M**	**6,935**	**10,745**	**28,779**	**43,692**
Consensus	$M	6,935	10,745	31,853	40,135
% Difference	%	0%	0%	-10%	9%

*Includes Financial Leases
MSFT and ORCL are quarterly calendarized

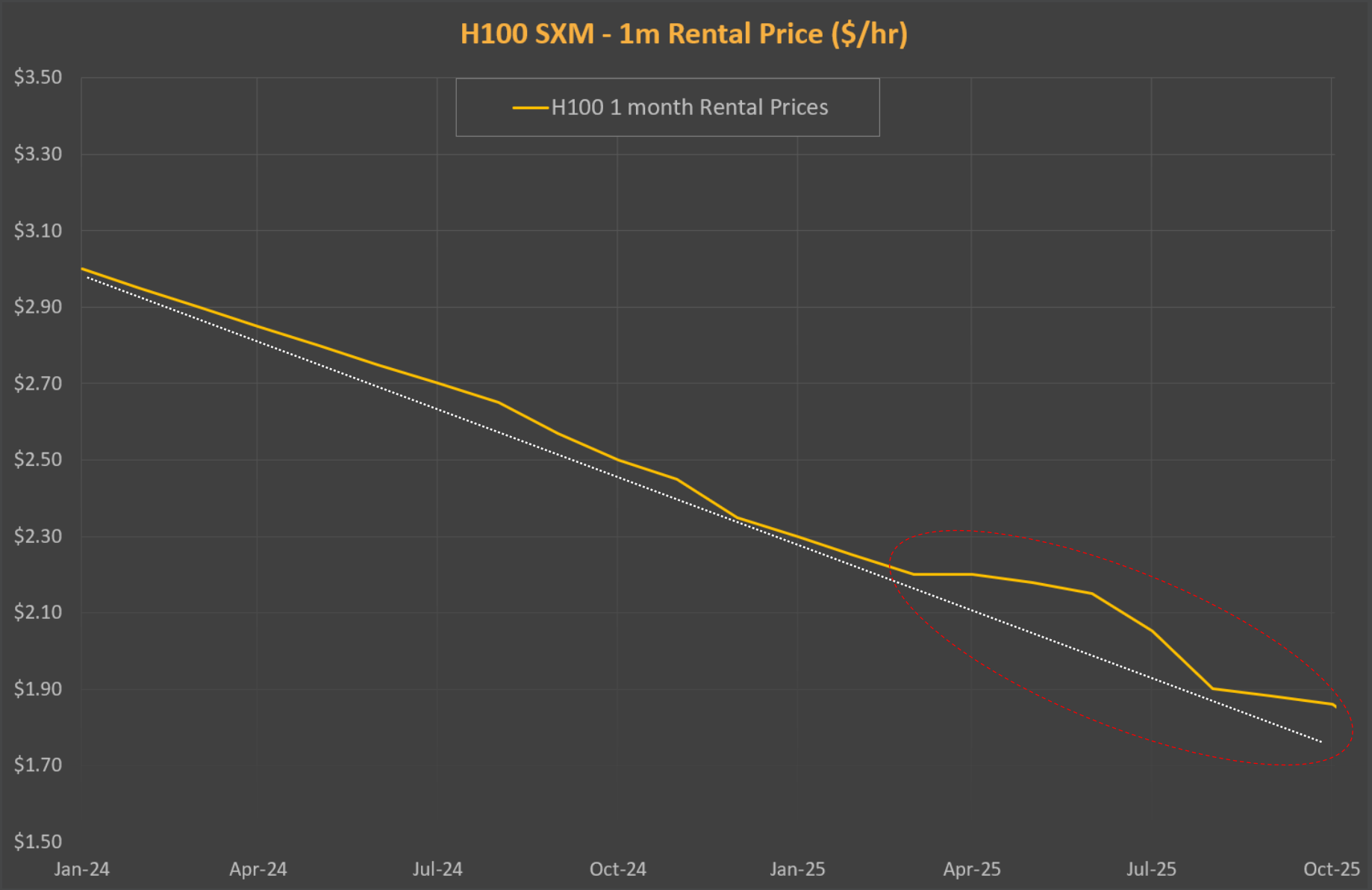


AI REVENUE GROWTH OUTSTRIPPING AI POWER GROWTH

- Annual Recurring Revenue (ARR) trends serve as a proxy for AI demand, while critical IT power (MW) reflects compute supply.
- Revenue is accelerating faster than GPU and power capacity, highlighting demand growth exceeding infrastructure deployment.



GPU BUYING MINDSET – GPU HOARDING?



- **At the start of 2025, H100 on-demand prices surged** to nearly $3 per GPU-hour as buyers scrambled for capacity, attracting speculative H100 buying.
- **H100 rental prices fell faster than expected, curtailing speculative builds.**
- **By 3Q25, H100s pricing stabilized as demand from inference workloads accelerated.** Thus, H100 rental prices flatlined and even increased in some cases!

GPU MEGA DEALS – LONG TERM CONTRACTS WITH OFFTAKES

Major AI Cloud Contracts												
	Units	IREN-Microsoft	Lambda-Microsoft	AWS-OpenAI	Coreweave-Meta	OCI-OpenAI	Nebius-Microsoft[1]	Coreweave-OpenAI	Coreweave-OpenAI	Coreweave-OpenAI	Nscale-Microsoft	Nscale-Microsoft[2]
Announcement Date	Date	3-Nov-25	3-Nov-25	3-Nov-25	30-Sept-25	10-Sept-25	8-Sept-25	10-Mar-25	15-May-25	25-Sept-25	16-Sept-25	15-Oct-25
Contract Value	USD	$9.7B	Muti-Billion	$38.0B	$14.2B	$300.0B	$17.4B	$11.9B	$4.0B	$6.5B	$6.2B	$14.0B
Term Duration (Years)	Years	5	?	7	6	5	5	5	5	5	5	5
Implied Annual Revenue	USD	$1.9B	?	$5.4B	$2.4B	$60.0B	$3.5B	$2.4B	$.8B	$1.3B	$1.2B	$2.8B
Chip Type	Type	GB300	?	GB200s/GB300s	GB300	GB300	GB300	GB200	GB200	GB300	GB300	GB300
Critical IT Power Contracted[3]	MW	170	?	?	230	4500	300	195	65	105	116	260
GPU Capex Disclosed	USD	$5.8B										

All clusters assume a 3-Layer InfiniBand Network

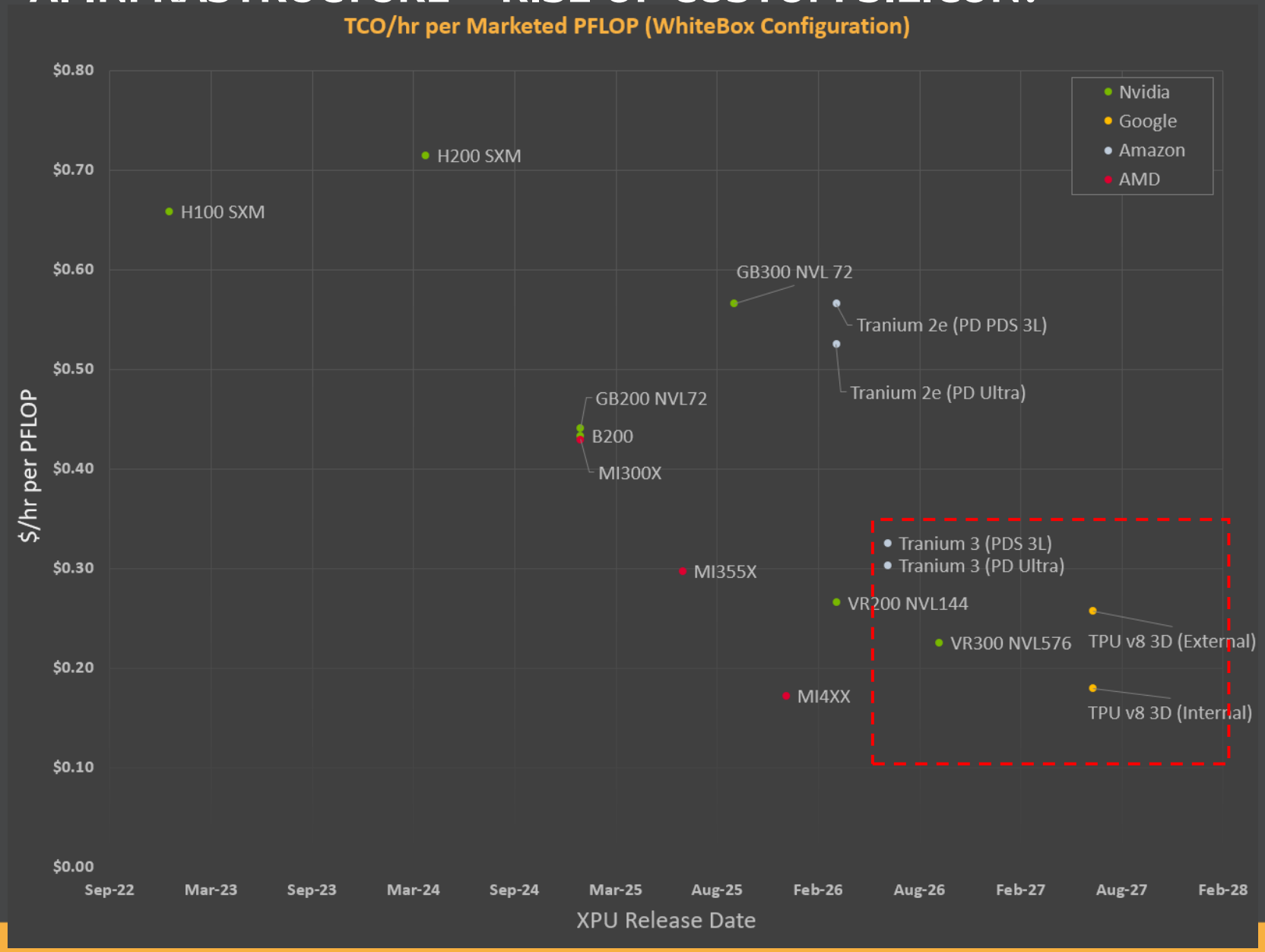
1. Option to increase contract value to $19.4B

2. Option to add further 700MW in late 2027

3. IREN disclosed 200MW of Crit IT power, with likely 170MW used for GPUs and remaining 30MW likely reserved for other non-AI compute that supports this cluster in some form

- **We think these deals are well thought out and mark a departure from H100 speculative buying** as the deals are either backstopped or are done by Hyperscalers with strong free cash flow and credit profiles.
- **In contrast to the speculative H100 purchases,** these long-term commitments are done with a through IRR analysis to ascertain the Perf/TCO for chips. We think this is healthy and suggests demand for AI infrastructure is structural.
- **We still expect more deals to come,** with Microsoft, Meta, OpenAI, Anthropic and xAI as key players. We have yet to see signs of Google and AWS using Neocloud capacity at a significant scale, but we don't rule it out.
- **Advanced AI labs and hyperscalers continue to dominate demand for high-end chips,** using custom kernels, FP4/FP8, and NVLink to extract maximum performance.

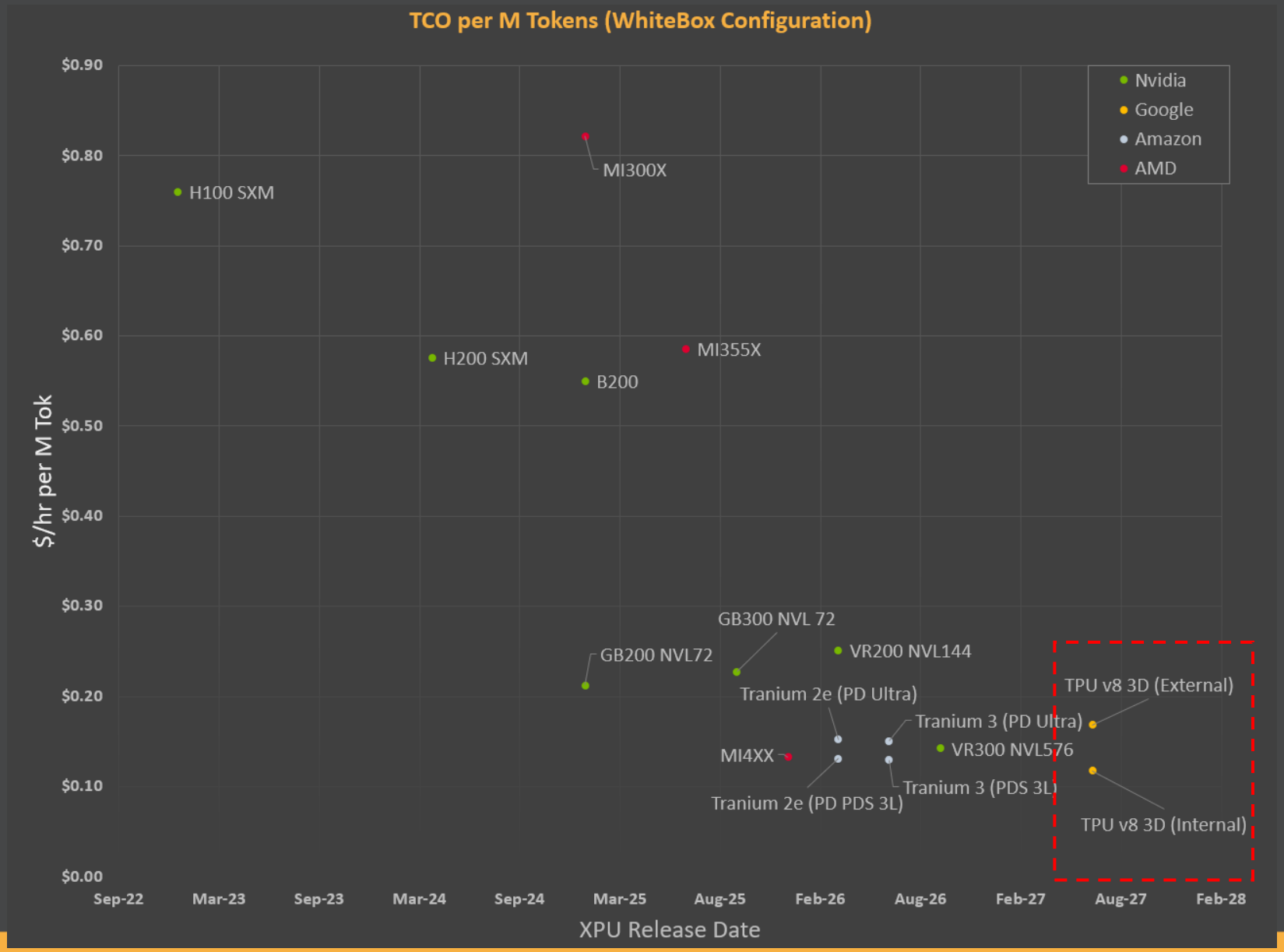
AI INFRASTRUCTURE – RISE OF CUSTOM SILICON?



Custom Silicon vs Merchant Silicon Training Performance/TCO Comparison :

- Merchant Silicon like VR300 and MI4XX have the better TCO/hr per marketed PFLOP compared to custom silicon, despite the additional chip provider margin
- However, one caveat is that these comparisons are done across marketed PFLOP, which can differ from real world training outcomes based on Model Flops Utilization %

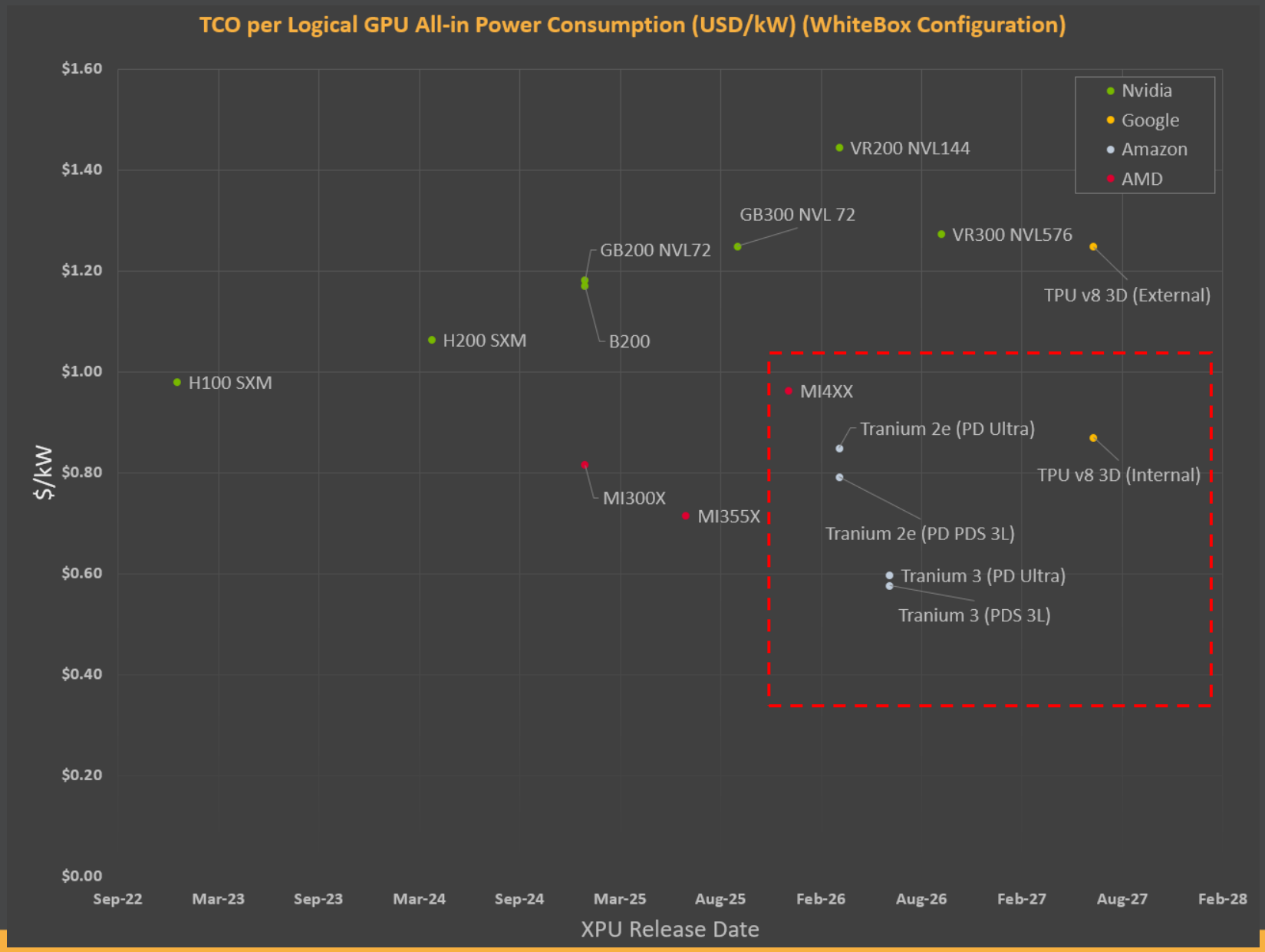
AI INFRASTRUCTURE – RISE OF CUSTOM SILICON?



Custom Silicon vs Merchant Silicon Inference Performance/TCO Comparison :

- Custom Silicon like TPU v8 3D (Internal) and Trainium3-Teton3 PDS have a better TCO/M Tok compared to merchant silicon – driven by a much lower TCO
- While M Tok assumptions are based on chip specs, real world inference throughput can differ with model architecture, token serving quality and assumptions networking efficiency

AI INFRASTRUCTURE – RISE OF CUSTOM SILICON?



Custom Silicon vs Merchant Silicon $/W Comparison:

- As custom silicon like Trainium and TPUs have started external sales, they have become an alternative to merchant silicon. This is especially so for deployments looking to balance cost vs performance (Etc. Anthropic, Fluidstack)
- We expect custom silicon chips deployed in the same generation as merchant silicon to have a lower TCO/kW – a key benefit given GW scale deployments

AI INFRASTRUCTURE – RISE OF CUSTOM SILICON?

AI Cloud Operating Cost of Ownership Summary

Chip	Unit	MI300X	MI355X	MI4XX	H100 SXM	H200 SXM	B200	GB200 NVL72	GB300 NVL 72	VR200 NVL144	VR300 NVL576	TPU v8 3D (Internal)	TPU v8 3D (External)	Tranium 2e (PD Ultra)	Tranium 2e (PD PDS 3L)	Tranium 3 (PD Ultra)	Tranium 3 (PDS 3L)
Customer Profile	Unit	Hyperscaler	Hyperscaler	Hyperscaler	Hyperscaler	Hyperscaler	Hyperscaler	Hyperscaler	Hyperscaler	Hyperscaler	Hyperscaler	Hyperscaler	Neocloud Giants	Hyperscaler	Hyperscaler	Hyperscaler	Hyperscaler
Total Cost per Unit per Hour	**USD/hr/GPU**	**$1.12**	**$1.49**	**$3.44**	**$1.30**	**$1.41**	**$1.95**	**$2.20**	**$2.83**	**$4.43**	**$7.53**	**$1.58**	**$2.27**	**$0.68**	**$0.74**	**$0.81**	**$0.87**
Capital Cost as % of Total Ownership Cost	%	67.4%	63.8%	73.8%	74.0%	75.0%	77.6%	78.0%	79.4%	82.4%	80.4%	70.0%	76.9%	66.9%	65.2%	55.4%	54.2%
Marketed TFLOPS (FP8)	TFLOPS	2,615	5,000	20,000	1,979	1,979	4,500	5,000	5,000	16,667	33,333	8,800	8,800	1,299	1,299	2,668	2,668
Effective training TFLOPS (FP8)	TFLOPS	650	1,449	6,622	750	800	1,800	2,500	2,500	7,500	15,000	2,640	2,640	390	390	800	800
Inference Throughput[1]	Tok/s/GPU	380	705	7,213	477	683	985	2,885	3,462	4,905	14,715	3,741	3,741	1,247	1,559	1,496	1,870
Memory Bandwidth per Logical GPU	TB/s	5	8	20	3	5	8	8	8	13	32	13	13	4	4	4	4
Marketed TFLOPS (FP8) / Memory Bandwidth	TFLOPS/TB/s	493	625	1,020	590	412	563	625	625	1,282	1,042	677	677	371	371	667	667
TCO per PFLOP	$/hr per PFLOP	$0.43	$0.30	$0.17	$0.66	$0.71	$0.43	$0.44	$0.57	$0.27	$0.23	$0.18	$0.26	$0.53	$0.57	$0.30	$0.33
TCO per effective training PFLOP	$/hr per PFLOP	$1.73	$1.03	$0.52	$1.74	$1.77	$1.08	$0.88	$1.13	$0.59	$0.50	$0.60	$0.86	$1.75	$1.89	$1.01	$1.09
TCO per M Tokens	$/M tokens	$0.82	$0.59	$0.13	$0.76	$0.58	$0.55	$0.21	$0.23	$0.25	$0.14	$0.12	$0.17	$0.15	$0.13	$0.15	$0.13
TCO per Memory Bandwidth	$/hr per TB/s	$0.21	$0.19	$0.18	$0.39	$0.29	$0.24	$0.28	$0.35	$0.34	$0.24	$0.12	$0.17	$0.19	$0.21	$0.20	$0.22
TCO per GPU All-in Power Consumption	USD/kW	$0.82	$0.72	$0.96	$0.98	$1.06	$1.17	$1.18	$1.25	$1.45	$1.27	$0.87	$1.25	$0.85	$0.79	$0.60	$0.58

Networking uses WhiteBox Ethernet Networking Cluster Configuration

1. DeepSeek R1 FP8. Uses 8k input, 1k output and 1k input, 1k output tokens 50/50 mix, 30 interactivity.

Custom Silicon vs Merchant Silicon Performance/TCO Comparison :

- Custom silicon seems to be more efficient for inference workloads with the caveat that it depends on model architecture, familiarity with software stack, assumptions around networking and token serving quality. This is further support by custom silicon's superior TCO per Memory Bandwidth
- For training workloads, merchant silicon still remains the better option. Despite the much higher TCO, merchant silicon's much higher marketed PFLOP per GPU outweighs the increase in TCO

AI INFRASTRUCTURE – RISE OF CUSTOM SILICON?

Inference Tokenomics - OAI Inference Economics						
		VR200	TPU v8 (Int)	Tr3 PDS		
Measure	Unit	2030	2030	2030	The Internet	Analysis Method
Global (Ex China) Population	M	6,906	6,906	6,906	6,906	SemiAnalysis Estimates
WAU Penetration Rate	%	44.3%	44.3%	44.3%	66.0%	
Year End Weekly Active Users (WAU)	**M**	**3,061**	**3,061**	**3,061**	**4,558**	
MAU to WAU Ratio	%	80.0%	80.0%	80.0%	80.0%	
Monthly Active Users (MAU)	**M**	**3,826**	**3,826**	**3,826**	**5,698**	
ChatGPT M Tokens per MAU	MTok/MAU/Month	3.74	3.74	3.74		AI Tokenomics Model
ChatGPT Tokens Per Day	**Trillion Tok/Day**	**477**	**477**	**477**		
API Tokens Per Day	**Trillion Tok/Day**	**482**	**482**	**482**		
Inference Throughput/GPU[1]	Tokens/s/GPU	4,905	3,741	1,870		InferenceMAX™ Benchmarking
Effective Utilization	%	40.0%	40.0%	40.0%		
GPUs Required	GPUs	5,660,764	7,422,771	14,845,542		
TCO/Hr	$/Hr/GPU	$4.43	$1.58	$0.87		AI Datacenter Model
Annual TCO	**$Mn/Year**	**$219,673**	**$102,684**	**$113,082**		

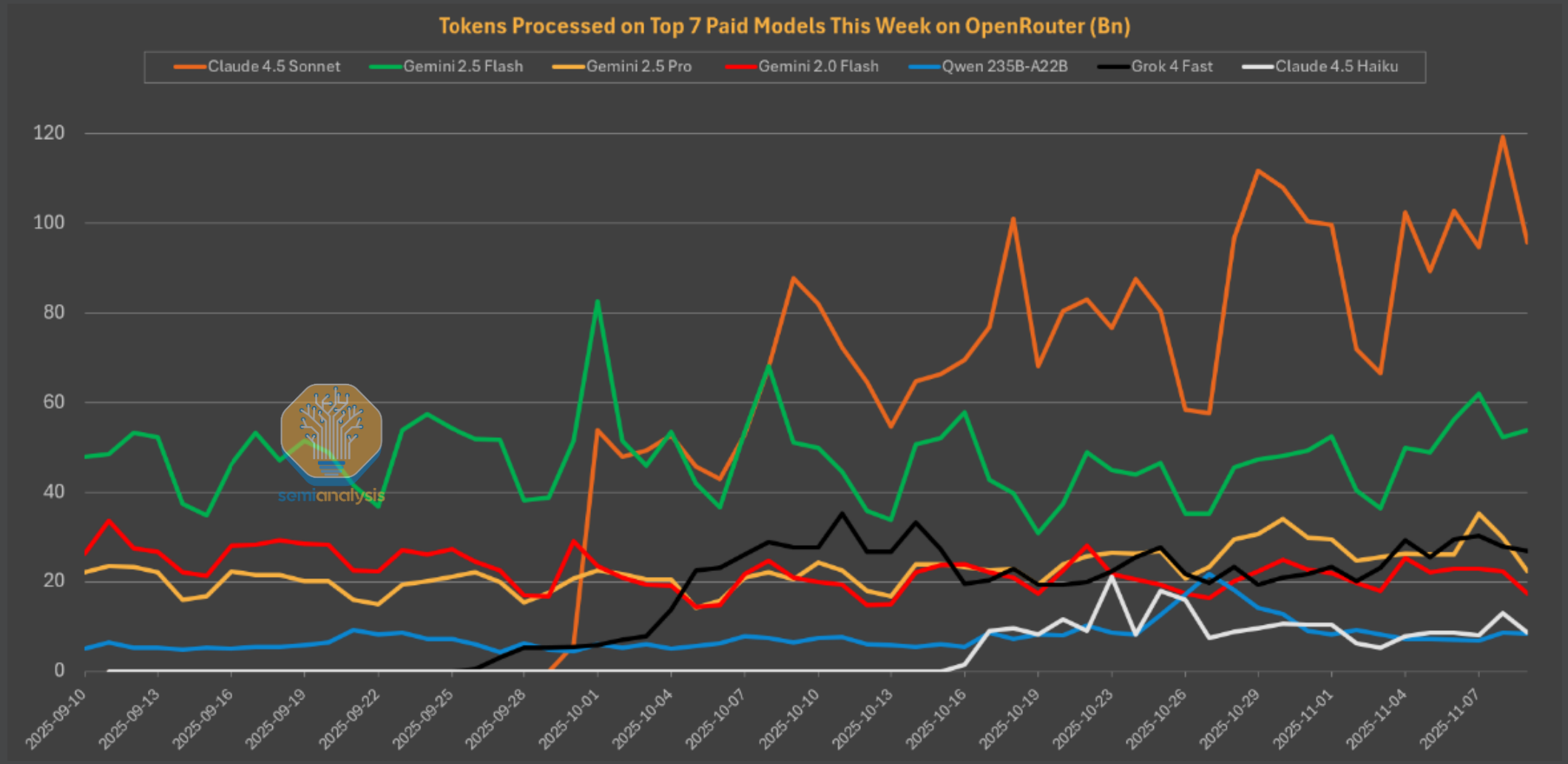
1. DeepSeek R1 FP8 on H200 using SGLang. Uses 8k input, 1k output tokens, 43 interactivity. For 2025 only, future periods are for different models and GPU assumptions.

Custom Silicon vs Merchant Silicon Inference TCO Comparison :

- For a given workload, custom silicon can meaningfully reduce annual TCO for inference despite having lower specs and predicted token throughput

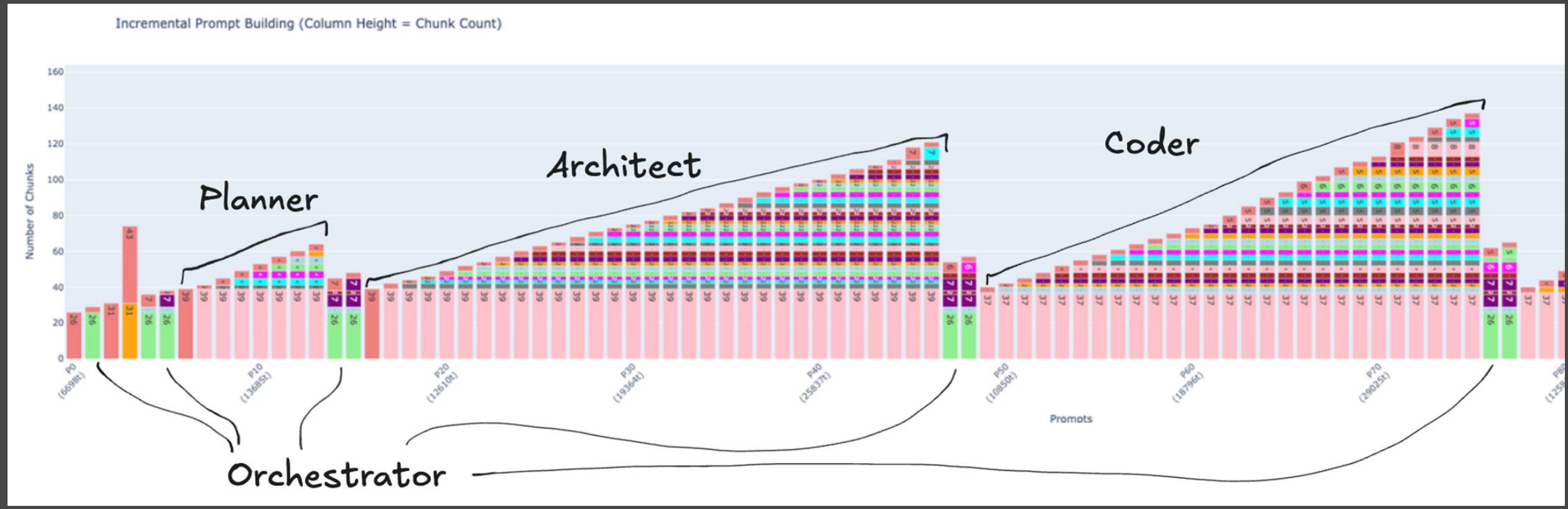
OPEN MODEL ADOPTION AT SCALE

- Adoption of open-source models at scale trails closed models according to OpenRouter, an API platform

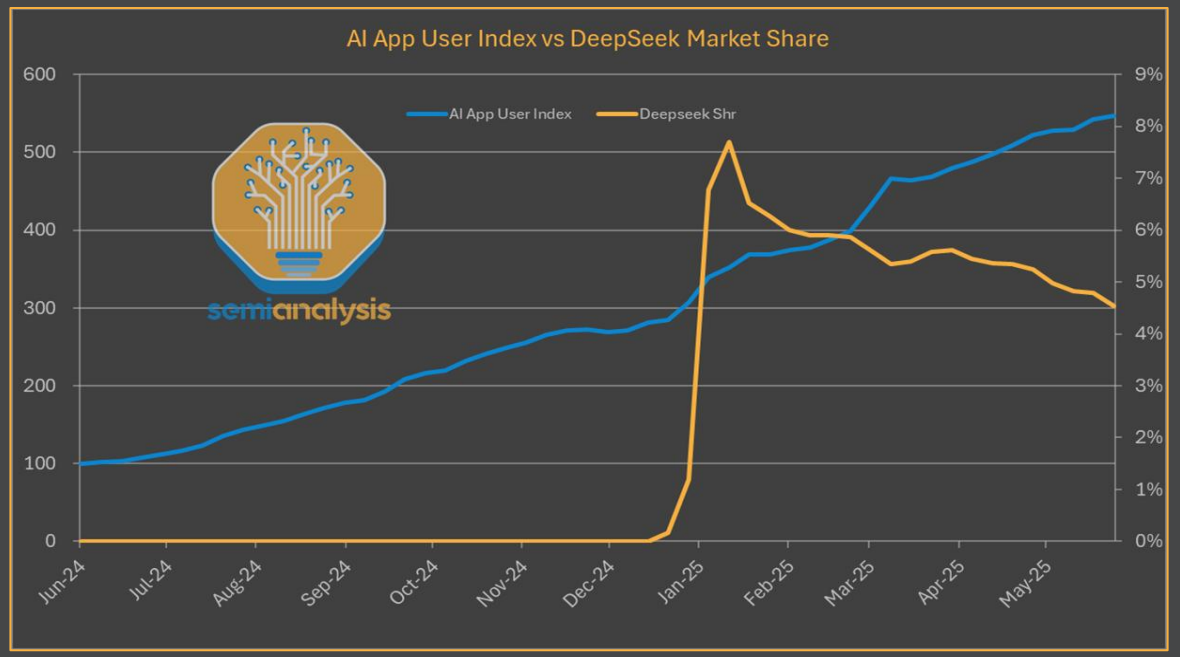


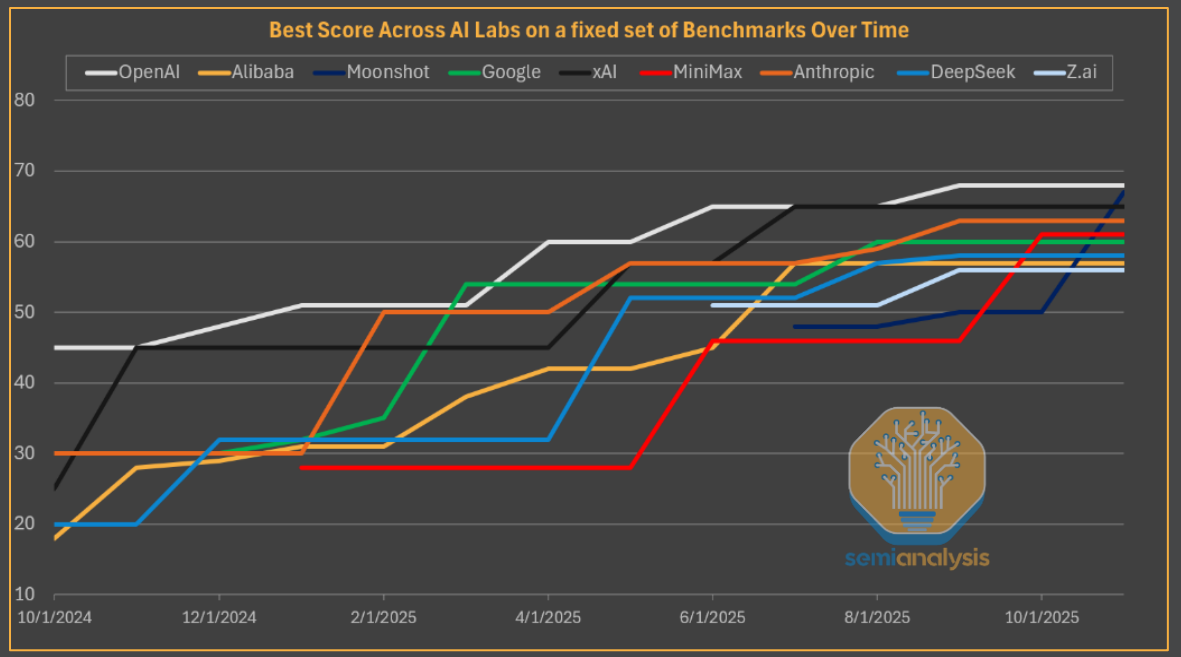
LARGE FRONTIER MODELS VS SMALLER OPEN MODELS

- The launch of Cursor's Composer model and Cognition's SWE 1.5 model is evidence that agentic AI workflows may look to offload some of the easier tasks to smaller, faster open source models over time



OPEN SOURCE VS CLOSED MODELS



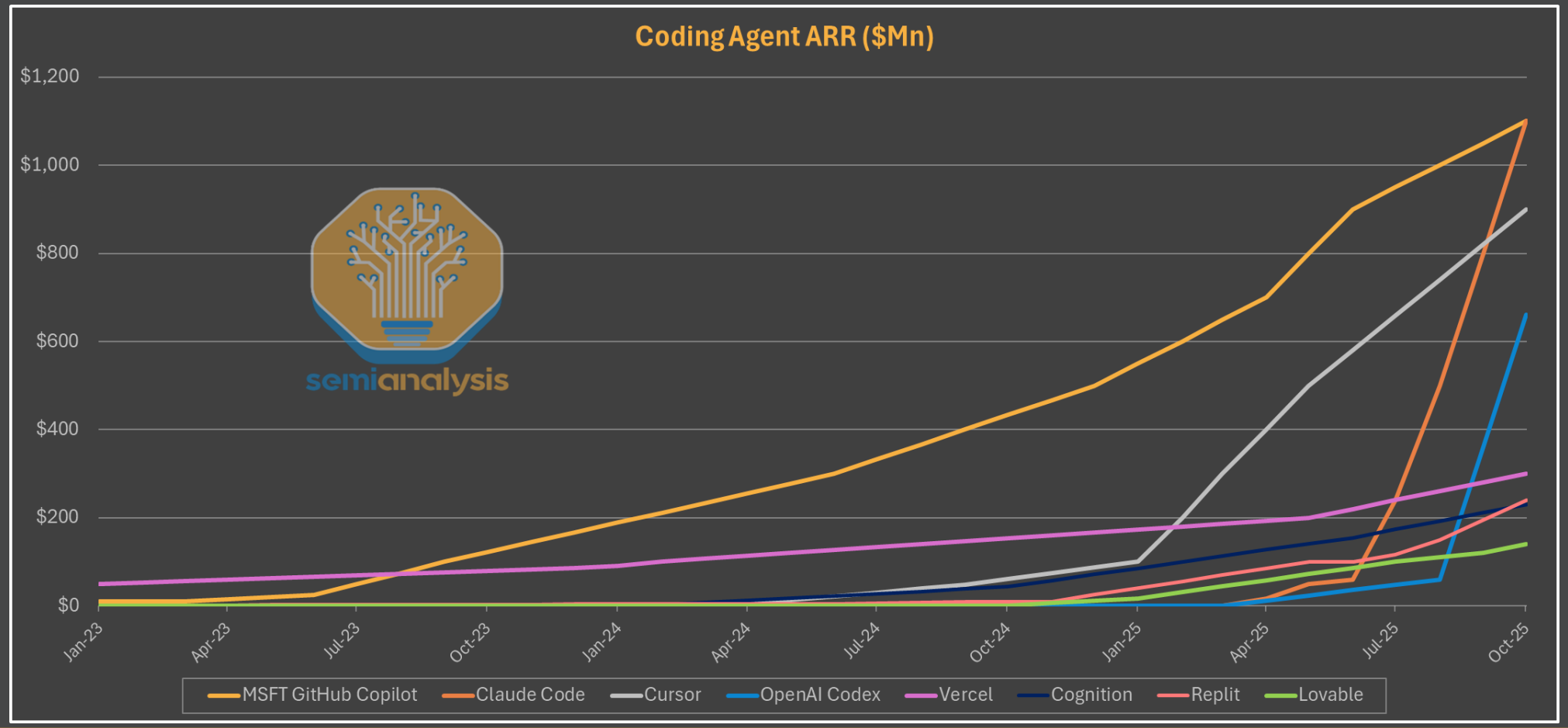


Open models lag slightly in capability, face difficulty in retaining market share:

- Open Models, in terms of benchmarks, are catch up to closed models after a small lag
- Frontier labs and their hyperscaler distribution partners still have an advantage in inference service quality, making AI model serving a game of infrastructure AND model quality
- Startups using closed-weight model APIs may find their sofware product is more capable at the expense of lower margins and commoditization

STARTUPS: RISE OF THE TOKEN CONSUMERS

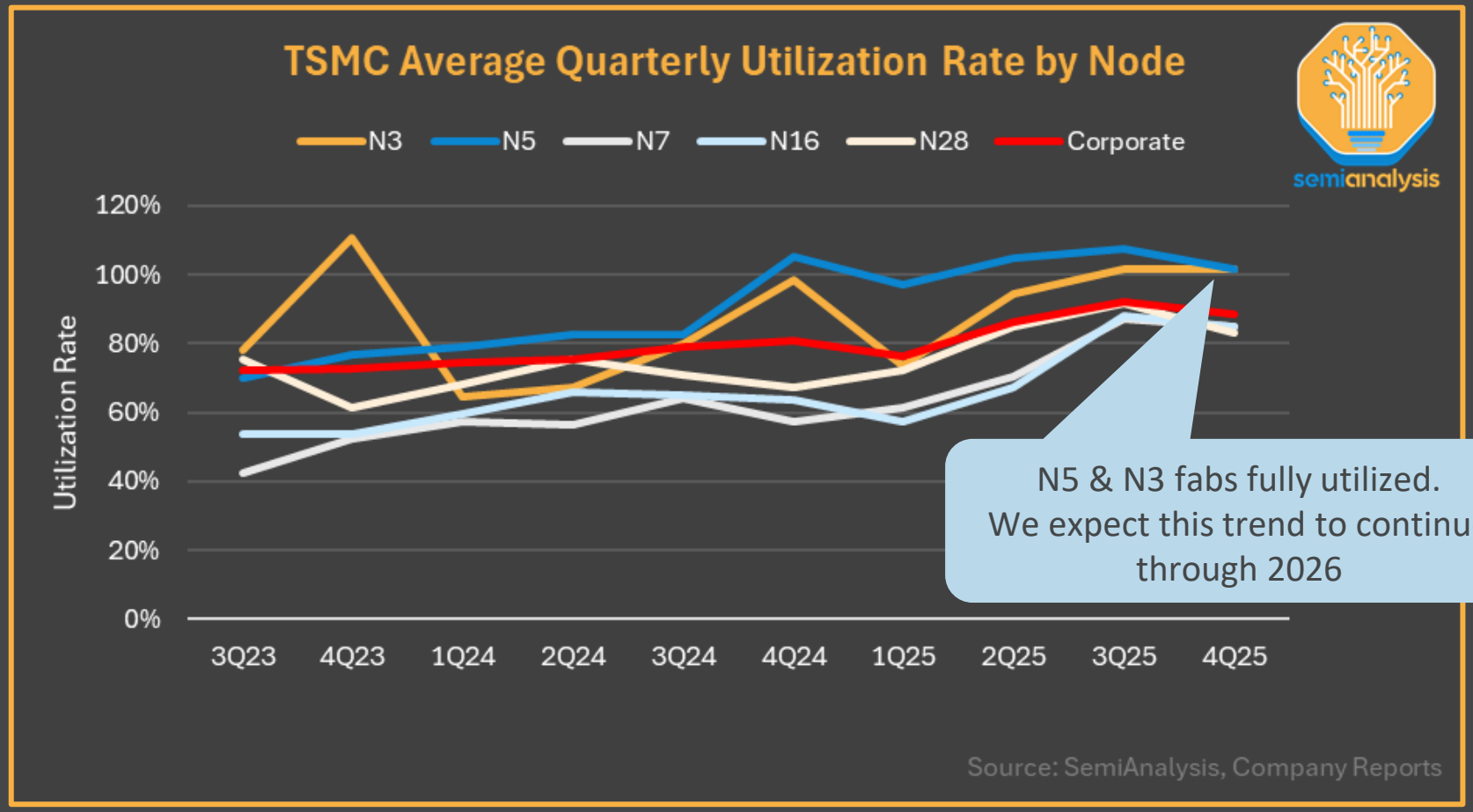
- As foundation labs and hyperscalers drive down the cost per unit of intelligence, more capital is needed to fund the monetization and product-led innovation required to make use of these tokens
- Coding agents lead the way but more software startups will find product-market-fit with token bundling and monetization



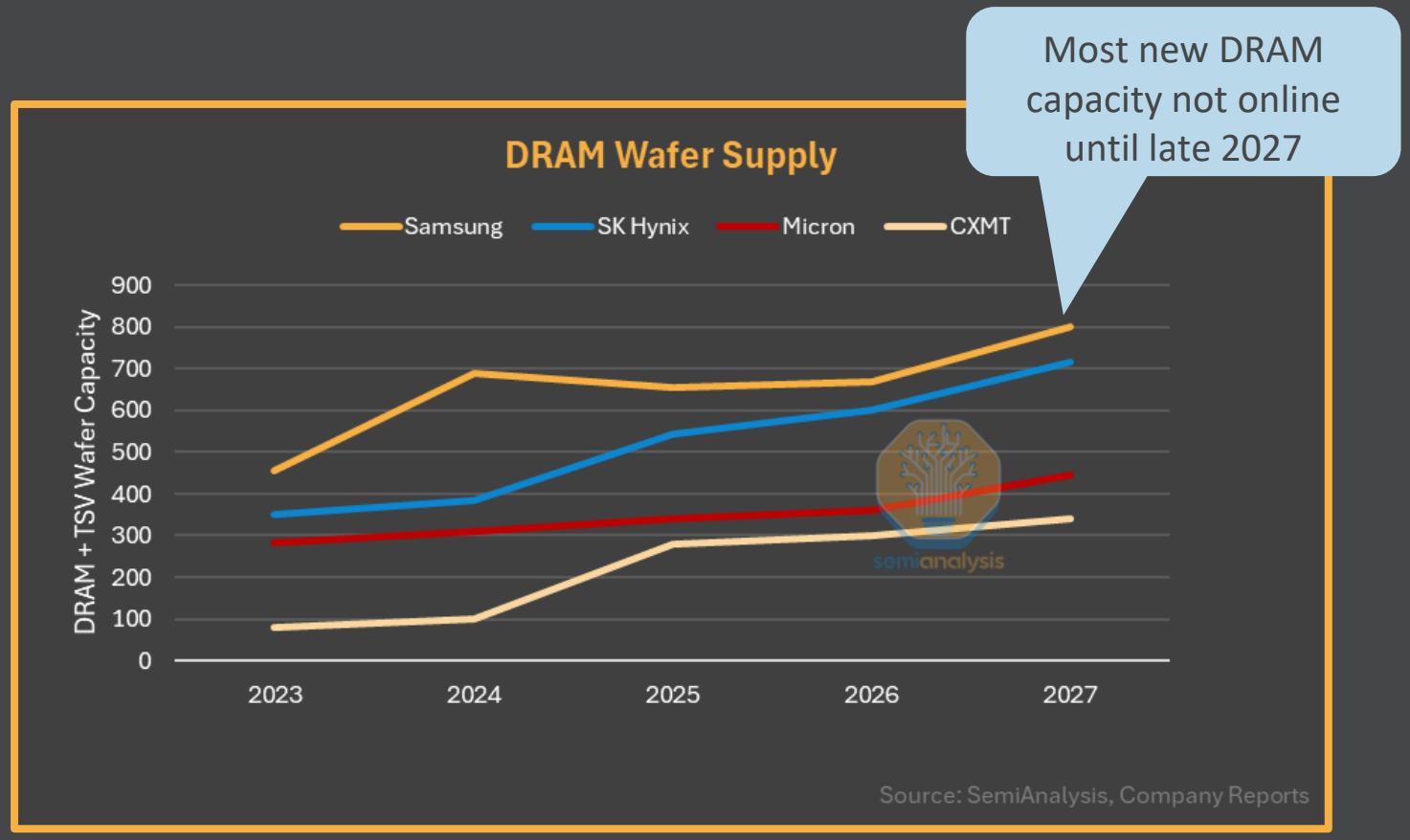
LOGIC OVERSUPPLY?

Advanced Logic / Foundry will be supply constrained :

- TSMC advanced nodes are fully booked in the short-term, and will continue to be supply constrained because:
  - 1) TSMC continues to be conservative in fab expansion capex
  - 2) Demand is accelerating
- AI wafer demand is increasing at an accelerating rate
  - Once negligible, it will now drive new foundry capacity adds
  - Requirements in 2027+ require large increase in foundry capex (or cause acute capacity shortage)



MEMORY OVERSUPPLY?



Memory will be Supply Constrained :

- Memory, particularly DRAM, will face a severe supply shortage for the next ~18 months. Drivers are
  - 1) Lack of available fab shells mean capacity cannot be added. Multiple shells are in progress but most will not reach high volume production until H2 2027
  - 2) Demand for memory rapidly increasing, particularly in AI datacenter storage due to image and video models along with recovery in server and increase in mobile DRAM content per unit
  - 3) Capacity conversions from traditional DRAM to TSV / HBM production mean net DRAM supply is declining in the short-term

Q&A

semianalysis