on 05-Oct-2026

BERNSTEIN
SOCIETE GENERALE GROUP

5 October 2026

US Networking & Communications Equipment

AI Value Chain: How much does a GW of AI ASIC data center capacity actually cost?


Daniel Zhu
+1 917 344 8309
daniel.zhu@bernsteinsg.com


Stacy A. Rasgon, Ph.D.
+1 213 559 5917
stacy.rasgon@bernsteinsg.com


Alex Wang, CFA
+852 2123 2613
alex.wang@bernsteinsg.com


Mark Li
+852 2123 2645
mark.li@bernsteinsg.com


David Dai, CFA
+852 2918 5704
david.dai@bernsteinsg.com


Mark C. Newman
+1 212 845 7822
mark.newman@bernsteinsg.com


Varun Govindaraj
+1 917 344 8543
varun.govindaraj@bernsteinsg.com


Madison Rezaei
+1 917 344 8622
madison.rezaei@bernsteinsg.com


Chad Dillard
+1 917 344 8469
chad.dillard@bernsteinsg.com


Kengo Makita
+81 3 6777 6977
kengo.makita@bernsteinsg.com


Arpad von Nemes
+1 917 344 8461
arpad.vonnemes@bernsteinsg.com


Alrick Shaw
+1 917 344 8454
alrick.shaw@bernsteinsg.com


Ethan Xu
+852 2123 2634
ethan.xu@bernsteinsg.com

Based on a series of conversations with industry experts, supply chain checks, and third party data, we extend our analysis of Nvidia Blackwell and Vera Rubin AI data center pod-level economics to include estimates for AMD and ASIC architectures. This note includes our analysis. Excel backup is available on request.

We revise our Vera Rubin estimates and land at $7.5M per rack vs. our prior estimate of $9.1M. Biggest drivers of the revision are 1) lowered estimates of per-rack NAND content and 2) more conservative expectations for HBM 4 pricing. We also 3) revise our estimates for networking content, as well as 4) our estimates for cooling and power delivery, based on Asia channel checks, and 5) reflect latest DRAM and NAND pricing. Our estimates are now much more in line with supply chain expectations of $7-8M per rack.

We estimate AMD's MI455X architecture to be a $7.3M rack, relatively similar to Vera Rubin. In terms of the high level differences, we think AMD is earning lower margins on GPUs and thus has lower computing content, despite higher CPU value than Nvidia's Vera. Conversely, Networking content appears significantly higher, due to AMD including "up to 3" NICs per GPU.

We expect OpenAI Jalapeno to be $5.0M per pod. We had expected Jalapeno to be a white-label version of the NVDA architecture other than replacing the GPU with a cheaper ASIC. However, on closer scrutiny, the design is radically different - notably, OpenAI intends to expand the scale-up domain and hold data locally in HBM (and DRAM), and reduce dependence on scale-out and NAND. OpenAI's disclosure are less complete than NVDA, AMD. However, we find it reassuring that our estimates are generally fairly consistent with OpenAI's stated design goals.

Our best guess is TPU v7 Ironwood is a $2.7M, however we found disclosures to be lacking, and lack conviction in very granular estimates as result. We're hopeful Google will disclose more details as external sales roll out; we will likely make significant revisions to our estimates as disclosures come in. We also await further details on TPU v8, and hope that as Google's external customer base broaden, they will also disclose more to customers (and thus to the public).

We also update our estimates for out-of-rack cost based on our prior work, and construct a build of hourly cost per GPU. Based on reported power consumption for each XPU pod, we then reach estimates of total capex / GW, ranging from $35B / GW for OpenAI Jalepno to ~$40B / GW for VR NVL72.

While our estimates imply notably higher GW shipments than our US data center model, we believe the disconnect could be explained by the increasing importance of international builds amid US data center delays.

Available compute continues to accelerate, which we see as an unlock for AI adoption.

See the Disclosure Appendix of this report for required disclosures, analyst certifications and other important information. Alternatively, visit our Global Research Disclosure Website.
First Published: 04 Oct 2026 23:16 UTC Completion Date: 04 Oct 2026 23:16 UTC

on 05-Oct-2026

Daniel Zhu +1 917 344 8309 daniel.zhu@bernsteinsg.com

BERNSTEIN TICKER TABLE

Ticker	Rating	Cur	2 Oct 2026 Closing Price	Price Target	TTM Rel. Perf.	Cur	Adjusted EPS 2025A	Adjusted EPS 2026E	Adjusted EPS 2027E	Adjusted P/E (x) 2025A	Adjusted P/E (x) 2026E	Adjusted P/E (x) 2027E
CLS (Celestica)	O	USD	387.32	520.00	39.4%	USD	6.06	11.35	23.79	63.9	34.1	16.3
CSCO (Cisco)	M	USD	112.20	110.00	49.3%	USD	3.73	4.33	5.11	30.1	25.9	22.0
ANET (Arista)	O	USD	207.35	250.00	28.5%	USD	2.96	4.11	5.61	70.0	50.5	36.9
LITE (Lumentum)	O	USD	1,085.42	1,220.00	524.4%	USD	2.05	8.37	22.40	530.6	129.6	48.5
COHR (Coherent)	O	USD	337.04	350.00	183.8%	USD	3.53	5.58	9.51	95.4	60.4	35.4
NVDA (NVIDIA)	O	USD	233.95	400.00	8.9%	USD	4.77	9.20	15.80	49.0	25.4	14.8
EQIX (Equinix)	O	USD	1,025.72	1,270.00	17.8%	USD	14.96	17.79	20.87	68.6	57.7	49.2
AMD (Advanced Micro)	O	USD	633.91	650.00	258.5%	USD	4.17	7.55	15.56	151.9	84.0	40.7
AVGO (Broadcom)	O	USD	355.14	575.00	(10.0)%	USD	6.82	11.64	18.63	52.1	30.5	19.1
3037.TT (Unimicron)	O	TWD	1,305.00	990.00	722.2%	TWD	4.36	20.13	30.18	299.2	64.8	43.2
2308.TT (Delta)	O	TWD	1,885.00	2,620.00	79.3%	TWD	23.09	39.35	57.98	81.6	47.9	32.5
2382.TT (Quanta)	U	TWD	332.00	250.00	(10.2)%	TWD	18.91	21.07	23.08	17.6	15.8	14.4
2360.TT (Chroma ATE)	O	TWD	2,190.00	2,750.00	221.9%	TWD	27.50	41.39	63.61	79.6	52.9	34.4
DLR (Digital Realty)	O	USD	178.55	226.00	(12.2)%	USD	3.87	3.09	3.21	46.1	57.7	55.7
CRWV (CoreWeave)	U	USD	89.62	74.00	(50.1)%	USD	(1.20)	(4.56)	(4.77)	N/M	306	38.8
VRT (Vertiv)	O	USD	252.18	368.00	41.0%	USD	4.20	6.60	9.06	60.1	38.2	27.8
NVT (nVent)	O	USD	169.56	229.00	55.5%	USD	3.34	5.15	6.55	50.7	32.9	25.9
CARR (Carrier)	M	USD	55.12	78.00	(21.4)%	USD	2.57	2.85	3.29	21.4	19.3	16.7
JCI (Johnson Controls)	O	USD	156.24	190.00	28.8%	USD	3.78	5.09	6.35	41.3	30.7	24.6
TT (Trane)	O	USD	463.91	575.00	(5.5)%	USD	13.06	15.28	17.79	35.5	30.4	26.1
APH (Amphenol)	O	USD	86.96	99.00	25.7%	USD	1.67	2.65	3.31	52.1	32.8	26.3
TEL (TE Connectivity)	O	USD	220.36	243.00	(15.3)%	USD	9.26	11.46	12.98	23.8	19.2	17.0
SPX			7,722.72									
ASIAX			1,948.97									

O - Outperform, M - Market-Perform, U - Underperform, NR - Not Rated, CS - Coverage Suspended

CLS, CSCO, ANET, LITE, COHR, 3037.TT, 2308.TT, 2382.TT, 2360.TT estimate is Reported EPS; CLS, CSCO, ANET, LITE, COHR, 3037.TT, 2308.TT, 2382.TT, 2360.TT valuation is Reported P/E (x); CRWV valuation is EV/EBIT (x); NVDA base year is 2026;

Source: Bloomberg, Bernstein estimates and analysis.

INVESTMENT IMPLICATIONS

CLS (OP, US$520.00): We rate Celestica Outperform, PT=US$520.00)

CSCO (OP, US$110.00): We rate Cisco Outperform, PT=US$110.00)

ANET (OP, US$250.00): We rate Arista Outperform, PT=US$250.00)

LITE (OP, US$1,220.00): We rate Lumentum Outperform, PT=US$1,220.00)

COHR (OP, US$350.00): We rate Coherent Outperform, PT=US$350.00)

NVDA (OP, US$400.00): We rate Nvidia Outperform, PT=US$400.00.

AMD (OP, US$650.00): We rate Advanced Micro Devices Outperform, PT=US$650.00

AVGO (OP, US$575.00): We rate Broadcom Outperform, PT=US575.00

Unimicron (OP, NT$990.00): We rate Unimicron Outperform, PT=NT$990.00.

Delta (OP, NT$2,620.00): We rate Delta Outperform, PT=NT$2,620.00.

Quanta (UP, NT$250.00): We rate Quanta Underperform, PT=NT$250.00.

人在草木间

on 05-Oct-2026

Chroma (OP, NT$2,750.00): We rate Chroma Outperform, PT=NT$2,750.00.

Digital Realty (OP, $226): We rate DLR Outperform, PT=$226.

Equinix (OP, $1,270): We rate Equinix Outperform, PT=$1,270.

Coreweave (UP, $74): We rate Coreweave Underperform, PT=$74.

We rate VRT Outperform, TP $ 368

We rate NVT Outperform, TP $ 229

We rate CARR Market-Perform, TP $78

We rate JCI Outperform, TP $190

We rate TT Outperform, TP $ 575

We rate APH Outperform, TP $99

We rate TEL Outperform, TP $243

on 05-Oct-2026

DETAILS

Based on a series of conversations with industry experts, supply chain checks, and third party data on out-of-rack costs of land and physical infrastructure, we extend our analysis of Nvidia Blackwell and Vera Rubin AI data center pod-level economics to include estimates for AMD and ASIC architectures. This note includes our analysis and key takeaways. See AI Value Chain: How much does a GW of data center capacity actually cost, and what goes into it? for our original view on the Blackwell cycle, and our Black Book for our broader perspective on AI infrastructure*:* Artificial Intelligence: The AI Infrastructure Value Chain

For a summary of our estimated pod BOM breakdown, see Exhibit 1, which is based on the disclosed and estimated specs in Exhibit 2. Details on our estimates can be found at the end of this note.

EXHIBIT 1: **Datacenter XPU Pod Capex Breakdown by Architecture**

	Nvidia			AMD	Google	OpenAI
	H100 DGX x8	GB200 NVL 72	VR NVL 72	MI455X Helios	TPU v7 Ironwood	Jalapeno
		4	225			
$K / Rack						
CPU ex. DRAM	34	92	180	268	32	320
DRAM	263	280	888	1,212	126	395
CPU	**297**	**372**	**1,068**	**1,481**	**158**	**715**
XPU ex. HBM	182	2,112	3,960	2,520	1,408	2,034
HBM	59	300	697	952	275	846
XPU	**240**	**2,412**	**4,657**	**3,472**	**1,683**	**2,880**
Computing sub-total	**537**	**2,784**	**5,725**	**4,952**	**1,841**	**3,595**
Front-end		144	80	70		32
Scale-up Switches		180	277	200		300
OCS					154	77
Cabling		123	158	140	43	194
Backplanes & Other		102	154	250	102	102
Scale-up		**405**	**589**	**590**	**299**	**673**
SuperNICS	112	144	198	324		0
Scale-out Switching		108	108	237	113	0
Transceivers		127	194	245	128	109
Scale-out	22	**380**	**500**	**806**	**241**	**0**
Networking sub-total	**134**	**929**	**1,170**	**1,466**	**539**	**814**
Direct-attached Storage	106	222	222	532	92	214
Cooling	16	51	102	102	40	70
Chassis	31	100	100	100	48	96
Power Delivery	14	44	102	80	30	144
Others			100	100	100	100
Pod Total	**838**	**4,130**	**7,520**	**7,333**	**2,691**	**5,032**

Source: Expert Conversations, Company Disclosures, LIterature Search, Channel Checks, Bernstein estimates and analysis

We also update our estimates for out-of-rack cost based on our prior work. Based on reported power consumption for each XPU pod, we then reach estimates of total capex / GW below in Exhibit 3. We rely on published specs and industry feedback to estimate provisioned power for each accelerator pod, and maintain our estimate that the rack is 78% of the data center power load[1]. We also update our estimates for out-of-rack spend (for more details on our updated estimates for out-of-rack cooling and power, see: Data Center Cooling and Power: What's the $M per MW for different types of equipment), and land closer to the lower end of our previous $12-15B range. On net, we estimate spend per W as ranging from $34.6/W for Jalapeno to $39.5/W for Vera Rubin.

We also construct a breakdown of hourly cost per GPU, given our estimates for utilization, differences in depreciation time frames for in-rack and out of rack costs, and power costs. Even at an elevated cost of $0.15/kWH, it costs ~$1.3B in electricity to run a GW of data center capacity for a year. Personnel costs are also negligible, with even the largest data centers reportedly operating with 8-10 people costing $30-80k per year each. By contrast, even on a 6 year depreciation cycle, capex costs ~$6.6B in annual depreciation for Rubin. Given the relatively high cost of capex relative to ongoing operating costs, and

[1] Based on assuming 1.15 PUE and server rack as 90% of total IT equipment power

on 05-Oct-2026

the fact that hardware has a shorter depreciation lifespan than physical infrastructure (**Exhibit 4**), we net out at an estimate of hourly economics which is actually be even more heavily weighted towards servers, storage and networking compared to what cash expenditure would imply - **Exhibit 5**.

Layering our estimates on per-rack power consumption with our estimates for shipments for each accelerator implies 36 GW shipped in 2026, increasing to 43 GB in 2027. While this is notably higher than our estimate of 18 GW in US capacity adds for 2026E and 26 GW in 2027E, we believe the disconnect stems from timing differences and increasing importance of international builds amid US data center delays. We combine our estimate of per-rack power consumption with our estimates for shipments from each accelerator (which you can find in detail in the Asia semiconductor team's CoWoS/HBM Model). Assuming that power consumption is proportional to wafer consumption, and given that rack architectures covered by our analysis account for ~40% of wafer shipments, this implies 36 GW in 2026E, ramping to 43 GW in 2027E (**Exhibit 6**). This is notably higher than our estimate of 18 GW in US capacity adds in 2026E and 26 GW in 2027E (for more details, see: U.S. Data Center Capacity: Your seat is skewing your numbers. Five ways to forecast DC capacity (and our definitive view). However, we believe the disconnect stems from timing differences as well as the increasing importance of international builds - amid increasingly widespread operational, power availability, and policy delays in the US, AI labs appear to increasingly be turning to non-US data center partners, such as OpenAI's deal with Firmus for AI data center capacity across Australia, Signapore, Indonesia and Malaysia. Accordingly, we belive the % of data center capacity deployed in the US may decrease compared to historical norms.

Available compute continuing to accelerate. End user adoption appears to have accelerated in 2026, driving dramatic growth among AI labs. With the Rubin cycle as well as more widespread adoption of AI ASICs driving a further step-up in available compute, we believe that this could help to unlock further AI adoption.

EXHIBIT 2: **Published and Estimated Specs by Architecture**

	Nvidia			AMD	Google	OpenAI
	H100 DGX x8	GB200 NVL 72	VR NVL 72	MI455X Helios	TPU v7 Ironwood	Jalapeno
Specs						
XPU Count	64	72	72	72	64	128
CPU Count	8	36	36	18	16	32
XPU HBM Memory (TB)	5.1	13.4	20.7	31.1	12.3	27.6
CPU DRAM Memory (TB)	16	17	54	74	7.68	24
E1.S NAND Storage (TB)	276	576	576	1,382	240	556
Scale-up Switch count		9	9	6		10
Scale-out Ports per GPU		3.4	3.4	8	4	1
NICS per GPU		1	1	3		
Copper DPs per rack		5,184	5,184	5,184	1,800	8,192
Excluded: ICMS NAND Storage (TB)			1,217			

Published figures in blue, Bernstein estimates in red
Source: Company websites, Bernstein estimates and analysis

on 05-Oct-2026

EXHIBIT 3: **Datacenter Infrastructure Capex Breakdown**

	GB200 NVL 72	VR NVL 72	MI455X Helios	TPU v7 Ironwood	Jalapeno
$K / Pod	**4,130**	**7,520**	**7,333**	**2,691**	**5,032**
Pod Power (kW)	132	220	250	80	180
Pod as % of total power	78%	78%	78%	78%	78%
Total Power per Rack (kW)	**169**	**281**	**319**	**102**	**230**
				- 0	
Pod per GW	5,929	3,557	3,130	9,783	4,348
Pod Cost per GW ($B)	**24.5**	**26.8**	**23.0**	**26.3**	**21.9**
Cooling	1.8	1.8	1.8	1.8	1.8
Power	2.5	3.2	3.2	3.2	3.2
Others	3.2	3.2	3.2	3.2	3.2
Mechanical and Electrical Total	**7.5**	**8.2**	**8.2**	**8.2**	**8.2**
Land & Building	4.5	4.5	4.5	4.5	4.5
Total Out-of-Rack Cost	**12.0**	**12.7**	**12.7**	**12.7**	**12.7**
Total Datacenter Capex / GW	**36.5**	**39.5**	**35.7**	**39.0**	**34.6**

Source: Omdia, Expert Conversations, Channel Checks, Company Disclosures, Literature Search, Bernstein estimates and analysis

EXHIBIT 4: **Hyperscalar Disclosures on Depreciation Lifespans for Relevant Categories**

Category	Company	Ticker	Label	Disclsoure
IT Hardware	Google	GOOGL	Servers & Network Equipment	Six Years
	Amazon	AMZN	Servers and Networking Equipment	Five Years
	Meta	META	Servers and network assets	Four to Five years*
	Microsoft	MSFT	Computer Equipment	Two to Six Years
	Oracle	ORCL	Computer, network, machinery and equipment	1-6 years
	Nebius	NBIS	Server and network equipement	4.0 years
	Coreweave	CRWV	Technology Equipement	6 years
Mechanical & Electrical	Meta	META	Equipment and other	One to 25 years
	Microsoft	MSFT	Furniture and equipment	One to 10 years
	Nebius	NBIS	Infrastructure systems and equipment	3.0-10.0 years
	Coreweave	CRWV	Data center equipment	8-12 years
Land & Building	Google	GOOGL	Data Center and Office Buildings	7 to 40 years
	Meta	META	Buildings	25 to 30 years
	Microsoft	MSFT	Buildings and Improvements	Five to 15 years
	Oracle	ORCL	Buildings and Improvements	1-40 years
	Nebius	NBIS	Buildings	20.0 years

Source: Company Filings, Bernstein analysis

on 05-Oct-2026

EXHIBIT 5: **Hourly accelerator cost by architecture**

	H100 DGX x8	GB200 NVL 72	VR NVL 72	MI455X Helios	TPU v7 Ironwood	Jalapeno
Annual Cost per GPU						
GPU / GW	779,221	426,877	256,126	225,391	626,087	556,522
Pod capex ($/ GPU)	13,090	57,355	104,450	101,848	42,051	39,316
Out of pod capex per GPU ($ / GPU)	13,741	28,111	49,585	56,346	20,285	22,820
Assume: GPU lifespan	5	5	5	5	5	5
Total Annual depreciation ($/GPU)	3,992	14,282	25,848	26,004	10,439	10,145
Power price ($/kWh)	0.15	0.15	0.15	0.15	0.15	0.15
Hours in a year	8,766	8,766	8,766	8,766	8,766	8,766
Annual power cost per kW	1,315	1,315	1,315	1,315	1,315	1,315
Total power per rack (kW)	82	169	281	319	102	230
Total power per GPU (kW)	1.28	2.34	3.90	4.44	1.60	1.80
Annual power cost ($/GPU)	**1,687**	**3,080**	**5,134**	**5,834**	**2,100**	**2,363**
Other operating cost ($/GW)	2,000,000	2,000,000	2,000,000	2,000,000	2,000,000	2,000,000
Other operating cost ($/GPU)	2.57	4.69	7.81	8.87	3.19	3.59
Total operating cost ($/GPU)	5,682	17,367	30,990	31,847	12,542	12,512
Cost breakdown						
Power cost	1,687	3,080	5,134	5,834	2,100	2,363
Other operating cost	3	5	8	9	3	4
Cash costs	**1,690**	**3,085**	**5,142**	**5,843**	**2,103**	**2,366**
Depreciation	3,992	14,282	25,848	26,004	10,439	10,145
Non-cash costs	**3,992**	**14,282**	**25,848**	**26,004**	**10,439**	**10,145**
Total costs	**5,682**	**17,367**	**30,990**	**31,847**	**12,542**	**12,512**
Utilization-adjusted total cost						
Utilization	40%	40%	40%	40%	40%	40%
Utilization-adjusted cash costs	676	1,234	2,057	2,337	841	947
Non-cash costs	3,992	14,282	25,848	26,004	10,439	10,145
Total utilization-adjusted cost	4,668	15,516	27,905	28,341	11,280	11,092
Utilization-adjusted hourly cost ($)	0.53	1.77	3.18	3.23	1.29	1.27

Source: Bernstein estiamtes and analysis

on 05-Oct-2026

EXHIBIT 6: **Estimated GW Shipments by Rack Architecture**

	2023	2024	2025	2026	2027
AI accelerator shipments (K)					
H100	809	1,420	0	0	0
B200	0	706	4,207	2,311	80
R200	0	0	0	2,124	3,095
MI450X	0	0	0	0	94
TPU v7	0	0	952	576	0
Racks (K)					
H100	12.6	22.2	-	-	-
B200	-	9.8	58.4	32.1	1.1
R200	-	-	-	29.5	43.0
MI450X	-	-	-	-	1.3
TPU v7	-	-	14.9	9.0	-
% of Wafers					
H100	23%	16%	0%	0%	0%
B200	0%	14%	34%	14%	0%
R200	0%	0%	0%	23%	28%
MI450X	0%	0%	0%	0%	1%
TPU v7	0%	0%	8%	4%	0%
Others	77%	71%	58%	59%	71%
kW per rack					
H100	82	82	82	82	82
B200	169	169	169	169	169
R200	281	281	281	281	281
MI450X	319	319	319	319	319
TPU v7	102	102	102	102	102
GW shipped					
H100	1.0	1.8	-	-	-
B200	-	1.7	9.9	5.4	0.2
R200	-	-	-	8.3	12.1
MI450X	-	-	-	-	0.4
TPU v7	-	-	1.5	0.9	-
Others	3.5	8.4	15.6	20.8	30.6
Overall	**4.5**	**11.9**	**27.0**	**35.5**	**43.3**

Source: Bernstein estiamtes and analysis

Our detailed estimation methodology follows below, and the excel backup is available on request:

We revise our Vera Rubin estimates, and land at $7.5M per rack, vs. our prior estimate of $9.1M. Biggest drivers of the revision. are 1) lowered estimates of per-rack NAND content, in part because we are re-evaluating how we account for out-of-rack NAND (more on this later) and 2) more conservative expectations for HBM 4 pricing. Our estimates are now much more in line with supply chain rumors of $7-8M per rack. We also 3) revise our estimates for networking content, largely based on expert conversations, as well as 4) our estimates for cooling and power delivery, based on Asia channel checks, and 5) reflect latest DRAM and NAND pricing. For a details, see

on 05-Oct-2026

- **Computing.** Our views on GPU and CPU (ex. DRAM and NAND) are largely unchanged. Our estimates are consistent with widespread reporting that the Rubin GPUs are selling for $55k per GPU[2]. At 72 GPUs per rack, this implies $3.96M in GPU cost alone, nearly half of the rack cost. Similarly, the Vera CPU is reportedly selling for $5k per CPU, pointing to $180k in total cost for 36 CPUs per rack.

- **Memory and storage.** We revise our estimates and arrive at $1.8M per rack in memory storage content. We continue to remind investors that, given continued volatility in DRAM and NAND prices, our own estimates will likely become stale relatively quickly, and that investors following the space should frequently reflect fluctuations in DRAM and NAND prices in order to maintain accurate forecasts.

  - **High-Bandwidth memory (HBM).** HBM remains hotly debated. Nvidia's published spec (link) lists 20.7 TB of HBM 4 (Exhibit 2). Based on our latest conversations with management, we expect Nvidia to give a more conservative HBM spec for Rubin Ultra. However, we also VR NVL 72 HBM 4 content to remain unchanged[3] (for more details, see U.S. Semiconductors and Semicap Equipment: Notes from the road - takeaways from management meetings in Silicon Valley). In terms of pricing, we revise our price estimate from $48/GB to $30.6 GB, given weaker signals from the channel and lower conviction that HBM margins will equalize wiith traditional DRAM. However, we continue to believe that Nvidia can price on price increases to end customers and collect a mark-up of 10%; with $30.6/GB * 110% * 20.7 TB, we land at $697K in HBM content per rack. For more details, see: Global Memory: Cycle is still elevated, but beware of the long-term threat from China

  - **Direct-attached storage content.** We remind investors that direct-attached storage is somewhat less transparent. Storage is excluded from Nvidia's specs, and there is likely room for configuration variance. Our previous estimate assumed that it does not make sense to add external storage arrays until in-rack NAND content is maximized, given the lower performance of external storage. Since external storage appears to be significant, we assumed that we could safely assume the maximum spec for in-rack NAND, which would imply 2.2 PB per rack (144x E1S drives with 15.36 TB each and 18 M.2 drives with 2 TB each). We also included 1.2 PB of out-of-rack ICMS storage. However, our most recent belief is that in-rack NAND content is closer to 8 TB per GPU, or 576 TB per rack. We have also decided to deal with out-of-rack storage seperately, given that it is outside the scale-up domain, and accordingly are now excluding ICMS (and CMX, which is still upcoming) from the Vera Rubin NVL 72 pod bill-of-materials. It remains somewhat unclear to us why AI data center vendors are buying external storage arrays when there is still capacity to increase in-rack NAND content, which would improve performance and save on facilitate floor area.

  - **NAND and CPU DRAM price.** We continue to model 54 TB in LPDDR5X CPU DRAM content, in line with Nvidia's published spec. However, we update our DRAM price assumption from $14.85/GB to $16.45/GB, to reflect market price increases. Similarly, we update our estimate for TLC NAND from $0.37/GB to $0.38/GB to reflect contract price increases. 54 TB in DRAM at $16.45 / GB gives $888k per rack, and similarly $0.38 / GB for 576 TB of NAND points to $222k in storage content.

  - **Memory and storage prices remain volatile; investors (and industry participants!) will need to track prices closely.** At the risk of stating the obvious, overall memory prices remaining more volatile than history. Moreover, we continue to believe that Nvidia likely has dynamic pricing mechanisms to protect its own margins, and price fluctuations are being passed on to end customers. As a result, investors (and industry participants!) that are leveraging this analysis will need to frequently reflect memory / storage price fluctuations in order for estimates to remain accurate. Our latest estimates for DRAM and NAND prices can be found in our monthly memory price tracker: MEMORY TRACKER (Aug): Conventional DRAM & NAND price to rise nearly 20% in 3Q26

- **Networking.** We model $1.2M in networking content, split between ~$600k in scale-up, ~$500k in scale-out, and a ~$80k front-end. For more details on networking, see our recent initiation: US Networking: Scaling the bandwidth wall. Initiating with a positive view

  - **Front-end.** We believe Vera Rubin includes 18 Bluefield 4s per rack, translating to ~$8k per rack. Notably, this is lower than for Blackwell; while Bluefield 3 is ~10% cheaper, the count appears to have halved from 36 to 18.

[2] Technically Rubin is 2 GPU dies in a single package, but virtually everyone in the industry and Nvidia's own specifications are counting Rubin GPU units on a per-package basis.

[3] We will address Rubin Ultra when more details become available, likely at GTC

on 05-Oct-2026

- **Scale-up switches.** For industry feedback is that NVLink list price is $25k-40k for 9 switches, or $225k-$360k. However, we also received the feedback that the upper end of the range looks too high, and NVLink is frequently discounted. We estimate slightly below the midpoint of the range, and estimate $280k in scale-up.

- **Cabling + backplane.** Industry participant feedback points to ~$160k per rack in cabling content. Consistent with this, we observe that number of DPs per rack is unchanged from Blackwell, and it would be surprising if cabling cost per DP was increasing meaningfully.. For backplanes, we take our estimate down to 1.5x our estimate for Blackwell, or ~$150k. For more on copper connectors, see our multi-industrials team's initiation: US Multi Industrials: Can we (inter)connect? Initiating coverage with Outperform on Amphenol and TE Connectivity

- **Scale-out switches and SuperNICs.** In theory, the number of switch ports per XPU should be fixed by the network architecture. In theory, a non-blocking scale-out network should have 3 scale-out switch ports for 2-tiers and 5 switch ports for 3-tiers (1 for GPU to leaf switch, 2 for leaf to spine connection, 2 for spine to core connection, if present). Comparing Dell'Oro data to our XPU shipment estimate, we land at an empirical estimate of 3.4 scale-out ports per GPU, consistent with the general feedback that most networks are 2-tier, with some early adoption of the third core tier. 3.4 ports / GPU * 72 GPUs * $0.44 / port[4] in AI back-end switches translates to $108k per rack in scale-out switching. Feedback from industry participants is that the NICs are another ~$200k in content per rack.

- **Transceivers.** For Rubin, the number of transceivers per GPU should be one more than the number of scale-out switch ports (each scale-out switch port + one on the GPU-side NICs). That translates to 4.4 transceivers per GPU. Our Asia industry contacts point to $610 per transceiver or $194k in total transceiver content.

- **Power, cooling, chassis, and others.** We model ~$400k across power, cooling, chassis, and other smaller components.

  - **Power.** We previously estimated $150k, which is the price of a standalone power rack. However, we now believe that only 20% of Vera Rubins will adopt the sidecare, and the remaining Rubin racks should have power content closer to $80-100k. We start from the midpoint at $90k, and then take a blended average of 20% of racks at $150k and 80% at $90k to arrive at an average of $102k in power content per rack.

  - **Cooling.** Supplier feedback is that cooling content is roughly double that of GB200, with the cold plate alone being a 50% increase. We revise our estimate to $102k, double our Blackwell estimate of $51k.

  - **Chassis.** We estimate ~$100k per rack for the chassis.

- For more details on our previous views, see: AI Value Chain: How much does a GW of Vera Rubin data center capacity actually cost?

We estimate AMD's MI455X architecture to be a $7.3M rack - relatively similar to Vera Rubin. In terms of the high level differences, we AMD is earning lower margins on GPUs and thus has lower computing content, despite higher CPU value than Nvidia's Vera. Conversely, Networking content appears significantly higher, due to AMD including "up to 3" NICs per GPU. Without firm anchors for storage content, we revert back to our previous max-spec methodology for estimating, although we acknowledge that this may be aggressive.

- **Computing.** Feedback from industry participants is that the AMD MI455X GPU is a $35k GPU, putting its price roughly comparable to a GB200. For a 72 GPU rack, that translates to $2.5M in GPU content, excluding HBM. The CPU is also rumored at $8-15k; based on our conversations, we take the high end, and land at $268k for 18 trays.

- **Memory and storage.** In total, we estimate $2.7M in memory/storage content per rack - signficantly higher than for Nvidia. However, this is in part due to using a relatively aggressive estimation methodology. If we benchmark to 8 GB of NAND per GPU, similar to Nvidia, we would land at $2.4M per rack, which would still be higher but also closer.

  - **HBM.** For HBM, we follow AMD's published spec of 31.1 TB per rack. In addition, while we assume Nvidia is marking up HBM 4 price by ~10%, we assume no such mark-up for AMD. $30.6 / GB * 31.1 TB / rack = ~$950k in HBM per rack.

[4] Per-port prices are also estimated from Dell'Oro data

on 05-Oct-2026

  - **Direct-attached storage.** Lacking clear feedback from industry about AMD GPU content, we revert to our previous methodology of estimating the maximum spec. The Helios rack reportedly has 5x E1.S slots for each of the 18 compute trays; assuming 15.36 GB per E1.S drive implies 1.4 PB of NAND capacity. We have not heard feedback on the size of the boot drives, but we assume it is immaterial - if it is a 2 GB M.2 drive (similar to Nvidia), this would be 0.036 PB in additional NAND. At $0.38/GB, that implies ~$530k in direct-attach NAND.

  - **CPU DRAM.** Somewhat surprisingly, AMD does not seem to include CPU DRAM in its published spec. However, each CPU reportedly has 16 channels, and based on Micron's spec, each channel should be able to support 256 GB. 256 GB / channel * 16 channels / CPU * 18 CPUs per rack = 74 TB in DRAM. At $16.45 / GB, that implies $1.2M in DRAM content per rack.

- **Networking.** We weren't able to get as many inputs on AMD Helios networking as for Rubin, which in turn means we are less confident in very granular estimates. However, AMD includes notably more scale-out than Nvidia, due to the designing including "up to 3" NICs per GPU. Our conversations do suggest that scale-up content for Helios is roughly similar to VR NVL 72, but with a different mix - using ethernet based UALink scale-up switches from Celestica instead of Nvidia's captive solution lowers the cost of scale up. Conversely, we do believe that the cost of backplans is higher. We net out at $70k in front end, $200k for scale up, $140k for cabling, and $250k for backplanes. For scale-out, we heard $1.8 per NIC - assuming a blended average of 2.5 NICs per GPu implies $324k in total content. We assume constant switch pricing (based on Dell'Oro data) to get to $237k for scale-out switches. We believe Helios is using transceivers which are more in line with Blackwell than Rubin, and assume ~$400 per transciever, or $245k in trasnceiver content for the rack.

- **Power, cooling, chassis, and others.**

  - **Cooling.** Based on conversations with suppliers, we believe AMD's cooling content is similar to Vera Rubin. We model $102k/rack, exactly in line with our estimate for Rubin.

  - **Power.** For AMD, power content is likely to be lower than Rubin, given less talk in the AMD ecosystem about HVDC as well as seemingless less non-GPU chips in the rack that consume power. We model $80k per GPU, in line with the low end of the range shared by supply chain checks for Rubin.

  - **Chassis.** For the chassis, we benchmark to our own estimates for Rubin, and estimate $100k in chassis and another $100k in miscellaneous smaller components.

We expect OpenAI Jalapeno to be $4.9M per pod. We had expected Jalapeno to be something of a white-label version of the Nvidia architecture other than replacing the GPU with a cheaper ASIC. However, on closer scrutiny, the design is in fact radically different - notably, OpenAI intends to expand the scale-up domain and hold data locally in HBM (and DRAM), and reduce dependence on scale-out and NAND. OpenAI's disclosure are much less complete than Nvidia or AMD (which makes sense - without external sales, there is less need to share detail with customers. However, we find it reassuring that our estimates are generally fairly consistent with OpenAI's stated design goals.

- **The archictecture.** OpenAI's design is radically different than the Nvidia reference architecture, which may be more familiar to investors at this point. The OpenAI Jalapeno pod is technically a two-rack system, optimized for inferencing. The first rack consists of a host rack which has 16 "Katsu" trays that each contain two Turin-class CPUs. The second is an ASIC rack with 16 "Vindaloo" ASIC trays each containing 8 ASICs and 8 "Chana" switch trays, connected by PCIe DAC.

- **Computing.** Our conversations suggest that Jalapeno is a $22.5k ASIC including HBM. The ASIC also reportedly includes 216 GB of HBM per compute die, or 27.6 TB in total content across 128 ASICs. Assuming HBM 4 at $30.6 / GB, that implies ~$850M in HBM content, which in turn backs into an estimate of $2M in ASIC content and $16k per ASIC, which we see as reasonable. We also assume that the Turin CPU is $10k each, or $320k in total.

- **Memory and storage.** In total, we estimate $1.45M in memory and storage content across NAND, DRAM, and HBM. However we acknowledge that we are making some assumptions which are potentially aggressive, especially on NAND, and believe there may be downside to implied NAND demand estimates.

  - **HBM.** As mentioned, we estimate $850M in HBM content (216 GB).

on 05-Oct-2026

  - **DRAM.** Jalapeno reportedly includes 1.5 TB of DRAM for each of the 16 CPU trays, implying 24 TB in total content. At $16.45/GB, that implies ~$400K in DRAM.

  - **NAND.** Jalapeno also includes 2 E1.S slots and 2 M.2 slots per CPU tray. Assuming 15.36 GB per E1.S and 2 GB per M.2, that would imply a maximum configuration of 556 GB per rack. However, our latest estimate is that GB200 is closer to 4 GB per E1.S slot, which would instead imply 192 GB per rack. We continue to anchor to the top of the range, but acknowledge significant uncertainty, and hope to get more clarity as OpenAI scales deployments.

- **Networking.** OpenAI's network design is where it differs most from more established designes. One of Jalapeno's stated goals is to seek to largely eliminate the scale-out layer in favor of maximizing the scale-up domain. The design is not quite that radical - Jalapeno also includes an OCS layer which in some ways replicates the scale-out function, but it doesn't appear to include normal scale-out switches. Conversely, the greater reach and demand on the scale-up network leads to a meaningful increase in scale-up content.

  - **Scale-up switches.** Our conversations suggest that the Jalapeno scale-up switch is $15-20k of Broadcom switching silicon, or perhaps ~$30k per switch after factoring CLS ODM content. The switch trays are in turn divided into 6 local switch trays that each contain one Tomahawk 6 switch ASIC, and 2 global domains with 2 switches each, for a total of 10 ASICs, or perhaps ~$300k in switch content.

  - **Front-end.** The design also includes 2x 200G front-ends for each of the 16 CPU trays. Assumiing that these are $1k each, that points to $32k in front-end networking.

  - **Cabling.** We assume copper cabling is about $0.02 per DP, based on benchmarking to GB200. Semianalysis reports that Jalapeno includes 8,192 DPs in total, which then points to $194k in copper cabling.

  - **Backplanes & Others.** Lacking clear data points on backplanes for Jalapeno, we estimate $102k based on benchmarking to GB200.

  - **Optical Circuit Switching (OCS).** The Lumentum MEMs OCS switch used by Google is rumored to cost $600 per port. Assuming OpenAI Jalapeno is using a similar solution, with 128 ports (one per ASIC), that would imply $77k in OCS content.

  - **Transceivers.** We believe Jalapeno is using higher speed transceivers, and raise price estimates to $850 per transceiver. However, Jalapeno likely only needs one transceiver per ASIC, as the optics only need to connect to the OCS layer, and moreover there is no need for OCS-side transceivers (since OCS keeps the signal as optical instead of needing to receive an optical and convert to electricity). As a result, despite the higher ASIC count, we believe transceiver volumes on Jalapeno are much lower, and we net out in $109k in trasnceivers.

- **Power, cooling, chassis, and others.** Overall, we have relatively few data points for Jalapeno. We assume cooling content is proportional to power draw, and accordingly reach $70k per rack by benchmarking to GB200. We assume $3k per tray for the chassis and land at $96k. We also assume power delivery costs $800 per kW, and that there is $100k in other content.

Somewhat surprisingly, despite the relative maturity of the program and early external sales, we found disclosures around Google's TPU v7 Ironwood architecture to be somewhat lacking. Our best guess is that this is a $2.7M rack, but we are hopeful that Google will disclose more details as external sales roll out, and we will likely make significant revisions to our estimates as disclosures come in. We also await further details on TPU v8, and hope that as Google's external customer base broaden, they will also disclosure more to customers (and thus to the public).

- **Computing.** For the ASIC, we assume based on performance and die size that the silicon cost is comparable to a GB200. However, assuming that Broadcom is earning 55% margins vs. Nvidia at 70% implies that this is a $22k ASIC, vs. GB200 being closer to $30k. While we don't believe the value of the Axion CPU has been disclosed (and may not be priced as a standalone product), benchmarking to comparable products from Arm points to a ~$2k CPU, or $32k in content for 16 CPUs. In total, that implies ~$1.8M in computing (including DRAM and HBM).

- **Memory and storage.** We estimate that TPU includes ~$500k in memory and storage content per TPU, including DRAM and HBM which would be within the compute package.

  - **HBM.** TPU's spec includes 192 GB in HBM 3E per ASIC. At $22.4 / GB, that implies $275k in HBM content.

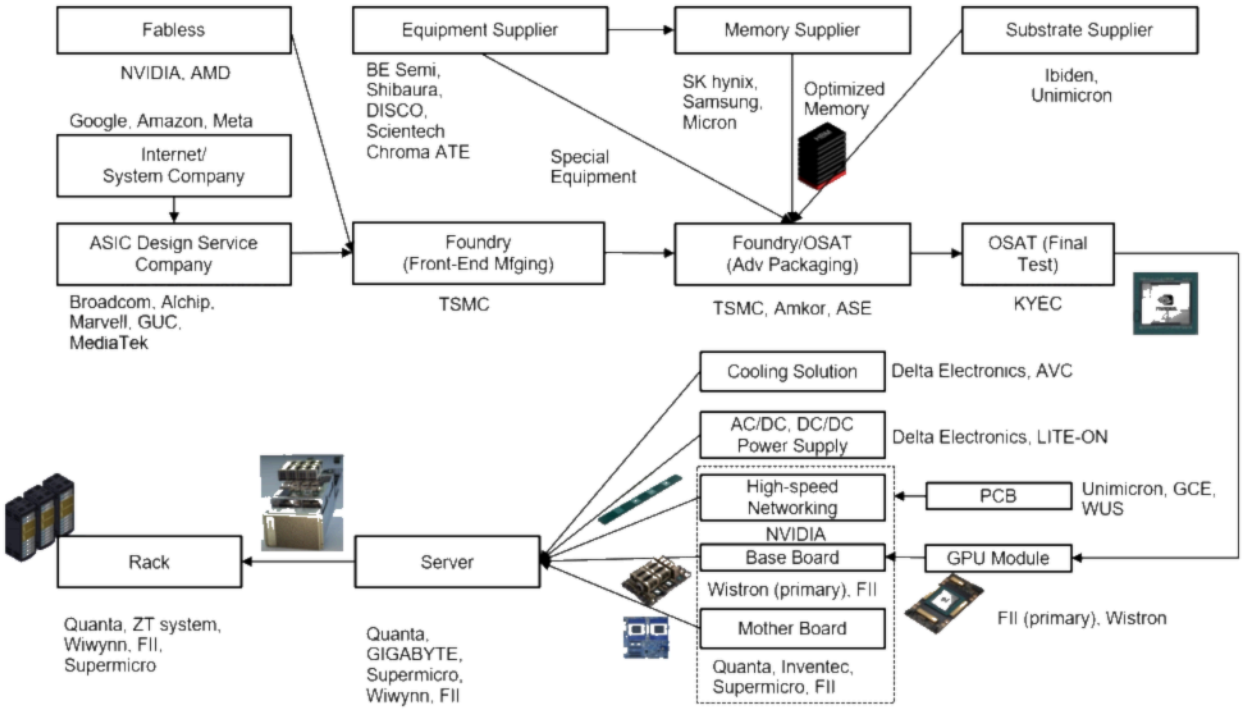
on 05-Oct-2026

  - **DRAM.** The TPU v7 Ironwood's spec (link) includes 960 GB of DRAM for each of the 4-ASIC CPU hosts, in turn translating to 7.68 TB in total DRAM content. At $16.45 / GB, that implies $126k / rack in DRAM content.
  - **NAND.** While Google's spec does not include NAND content, and none of our contacts seemed very confident in Google's NAND content in practice. However, we did hear a rumor that it includes 15 TB of NAND per CPU host, or 240 TB in total NAND. At $0.38 per GB, that points to $92k in NAND content.
- **Networking.** We estimate another $500k in networking conttent per TPU rack.
  - **Scale-up OCS.** Where Jalapeno attempts to replace scale-out with OCS, Google in some ways is doing the same for scale-up. At 64 ASICs per pod and 4 ports per ASIC, as well as $600/port for MEMs OCS, that would imply ~$154k in OCS content per rack.
  - **Cabling.** We believe TPU includes 1.8k copper DPs. Benchmarking again to $0.024 in value per DPU, from Blackwell, that points to $43k in copper content.
  - **Backplanes & Other.** Lacking clear data points, we anchor to $102k from benchmarking vs. GB200.
  - **Scale-out.** Anchoring to industry-wide switch prices of ~$400 per port, 4 ports per ASIC, and 64 ASICs leads us to land at $113k in scale-out content.
  - **Transceivers.** We believe TPU v7 is using $400 transceivers; with one more transceiver per GPU than scale-out switches, this points to ~$130k in transceivers.
- **Power, cooling, chassis, and others.** We believe the remaining content is perhaps $200k per TPU rack.
  - **Cooling.** Based on conversations with suppliers, we believe TPU cooling content is slightly lower than Blackwell. We model $40k, vs. Blackwell at $51k.
  - **Power.** We expect ~$30k in TPU power content, consistent with industry feedback.
  - **Chassis.** We assume the chassis is proportional to rack power, and land at $48k per rack, vs. $100k for GB200.
  - **Others.** We also assume that TPU has ~$100k in miscellaneous smaller components.

For reference: Exhibit 7 includes a map of AI server supply chain. **Exhibit 8** includes companies in the AI server hardware supply chain. Detailed discussions of competition landscape in each sub-segment can be found in our Networking Initiation: US Networking: Scaling the bandwidth wall. Initiating with a positive view, as well as the following notes: server ODM/OEMs, power components, liquid cooling, copper connectors and optical transceivers.

on 05-Oct-2026

EXHIBIT 7: **AI Server Supply Chain**



Note: Bernstein covers Quanta, Delta Electronics, Unimicron, Chroma ATE, TSMC, Samsung, SK hynix, Micron, NVIDIA, AMD, Broadcom, DISCO, Ibiden, BE Semi, but not the rest.
Source: Nvidia, SK hynix, Bernstein analysis

人在草木间

on 05-Oct-2026

EXHIBIT 8: **Summary of AI supply chain companies**

Position in Supply Chain	Ticker	Company	Mkt Cap (US$M)	Forward P/E	Revenue (US$M)		GPM %		EPS
					2025	25-27 CAGR	2025	2027E	25-27 CAGR
OEM	DELL US Equity	Dell	274,150	23.0x	113,538	30%	20.4%	17.5%	41%
	SMCI US Equity	Super Micro	28,207	14.5x	21,972	52%	11.2%	8.3%	26%
	HPE US Equity	HPE	71,096	14.1x	34,296	21%	31.5%	35.9%	44%
	2317 TT Equity	Hon Hai	130,390	14.8x	260,716	24%	6.2%	5.9%	25%
	601138 CH Equity	FII	230,115	21.9x	125,673	52%	7.0%	6.9%	53%
ODM	2382 TT Equity	Quanta	49,592	15.1x	68,329	43%	7.0%	5.0%	21%
	6669 TT Equity	Wiwynn	32,896	14.4x	30,587	43%	8.3%	7.0%	28%
	3231 TT Equity	Wistron	17,789	11.0x	70,351	45%	6.1%	5.3%	39%
Copper connectivity	APH US Equity	Amphenol	180,562	28.3x	23,095	29%	36.9%	38.4%	30%
	002475 CH Equity	Luxshare	80,230	20.9x	46,259	23%	11.9%	12.5%	32%
Optical transceiver	300308 CH Equity	Innolight	210,677	40.3x	5,323	92%	42.0%	44.5%	102%
	300502 CH EQUITY	Eoptolink	114,080	35.6x	3,458	78%	47.8%	48.1%	68%
	300394 CH Equity	TFC Opitcal	55,264	87.4x	719	60%	54.0%	52.0%	59%
	COHR US Equity	Coherent	82,540	52.1x	5,810	28%	37.9%	40.5%	53%
Power / thermal component	2308 TT Equity	Delta	200,179	45.1x	17,853	38%	34.3%	37.7%	66%
	VRT US Equity	Vertiv	124,421	43.9x	10,230	32%	36.3%	39.4%	43%
	FLEX US Equity	Flex	58,441	32.7x	27,914	25%	9.2%	10.5%	44%
	3017 TT Equity	AVC	33,775	23.8x	4,493	45%	25.8%	30.6%	63%
	2301 TT EQUITY	Lite-On	18,291	23.7x	5,344	21%	23%	25%	30%

Bernstein covers Dell, SuperMicro, HPE, Quanta, Luxshare, Delta, but not the rest
Source: Bloomberg data, Bernstein analysis