

Penguin Solutions 2nd Quarter Fiscal 2026 Earnings Call

April 1, 2026

Thank you, operator. Good afternoon and thank you for joining us on today's earnings conference call and webcast to discuss Penguin Solutions' second quarter fiscal 2026 results. On the call today are Kash Shaikh, Chief Executive Officer, and Nate Olmstead, Chief Financial Officer.

You can find the accompanying slide presentation and press release for this call on the Investor Relations section of our website. We encourage you to go to the site throughout the quarter for the most current information on the company. I would also like to remind everyone to read the note on the Use of Forward-Looking Statements that is included in the press release and the earnings call presentation. Please note that during this conference call, the company will make projections and forward-looking statements, including, but not limited to, statements about the market demand, technology shifts, and industry trends; and the company's growth trajectory and financial outlook, business plans and strategy, including investment plans; product development and roadmap; anticipated sales, orders, revenue, and customer growth and diversification; and existing and potential strategic agreements and collaborations.

Forward-looking statements are based on current beliefs and assumptions, are not guarantees of future performance, and are subject to risks and uncertainties, including, without limitation, the risks and uncertainties reflected in the press release and the earnings call presentation filed today as well as in the company's most recent annual and quarterly reports. The forward-looking statements are representative only as of the date they are made and, except as required by applicable law, we assume no responsibility to publicly update or revise any forward-looking statements. We will also discuss both GAAP and non-GAAP financial measures. Non-GAAP measures should not be considered in isolation from, as a substitute for, or superior to our GAAP results. We encourage you to consider all measures when analyzing our performance. A reconciliation of the GAAP to non-GAAP measures is included in today's press release and accompanying slide presentation.

And with that, let me turn the call over to Kash Shaikh, CEO. Kash?

Good afternoon! Thank you for joining our second quarter FY26 earnings call. This is my first earnings call as CEO of Penguin Solutions, and I am excited to step into this role. I want to start by thanking Mark Adams for his leadership and for the strong foundation he built.

Since joining in early February, I've spent significant time with customers, partners, and our teams around the world. I've witnessed the strength of the company, both in our technology and our customer relationships.

What is clear is this: AI is moving from experimentation to production, with workloads increasingly shifting towards real-time inference.

We are already seeing this translate into customer demand beyond hyperscale, across enterprise, neocloud, and sovereign AI markets. We expect this transition to expand our addressable market and drive increased demand for integrated AI infrastructure, where Penguin is already winning. We see this firsthand in the breadth of our deployments – from a sovereign AI Factory, Haein, in South Korea, to enterprise voice AI with Deepgram, to large-scale research systems with Georgia Tech, along with a growing pipeline across all three market segments.

What makes this opportunity so significant is that the architecture of AI is also changing. Model training was largely compute-bound. Inference, powering agentic AI, is memory-bound and latency-sensitive. We believe this is driving a re-architecture of the data center across compute, memory, interconnect, and software.

We also see AI driving memory demand, not only for the high bandwidth memory, or HBM, used with GPUs or other accelerators, but also for general-purpose memory. General-purpose compute wraps around every GPU buildout, and whether it's reinforcement learning pipelines or inference serving, that workload runs on processors backed by significant memory content across the entire system. So, while memory markets are cyclical, we believe AI is adding a more durable layer of demand for memory.

As AI factories scale, I expect customers to increasingly prioritize partners that deliver with speed and precision, along with full-stack AI Factory Platform capabilities—including compute, scalable memory systems, cluster management software, end-to-end services, and a partner ecosystem to deliver a differentiated solution. Time to deployment is now directly tied to time to first token.

Against this backdrop, we are building Penguin into an AI Factory Platform company.

Our AI Factory Platform is built around 6 core elements:

- **First**, Penguin ClusterWare™, our AI Infrastructure Management software;
- **Second**, our new Penguin MemoryAI™ line of systems, designed specifically for AI inference workloads;
- **Third**, Penguin Advanced Computing systems optimized for AI workloads;
- **Fourth**, Penguin OriginAI factory architectures, our reference designs for AI factories; and
- **Fifth and Sixth**, End-to-end Services and our Partner Ecosystem. Production-grade AI factories require full-stack design across compute, memory, storage, networking, and software. We partner with leading AI companies, including NVIDIA and SK Telecom, and partners like Dell. We also offer complete end-to-end services spanning design, build, deploy, and managed services.

We are strategically positioned at the intersection of AI infrastructure and memory, with a long track record in both. Few, if any, companies combine these capabilities at scale.

We believe that together, our AI Infrastructure and memory expertise position us to meet the evolving requirements of AI infrastructure as it shifts towards inference workloads. This supports our ability to develop differentiated solutions.

Given the momentum we are seeing in our AI Infrastructure business and the significant market opportunity ahead of us, we are very focused in this area. We plan to invest more in our AI Factory Platform to accelerate AI business growth, specifically in product innovation, go-to-market, and customer engagement.

In March at NVIDIA GTC conference, we announced two AI inference-centric solutions aligned with this strategy.

- **First**, the Penguin MemoryAI™ Server. Building upon our Compute Express Link®, or “CXL”, based memory expansion capabilities, we introduced a new line of scalable memory systems called MemoryAI. CXL is a high-speed interconnect that enables scalable, shared memory across GPUs and CPUs.

We also announced the immediate availability of our new MemoryAI KV Cache Server. Here, “KV”, or key-value cache, stores inference context to accelerate large language model responses.

- **Second**, the expansion of our OriginAI® Factory Architecture portfolio, which now includes blueprints that address the larger workloads and the low-latency demands of AI inference.

We also continue to expand the capabilities of ClusterWare toward a unified control plane for AI Factory infrastructure, integrating with the open ecosystem to deliver repeatable, production-scale deployments.

To accelerate innovation and strengthen our leadership team, we recently appointed Ian Colle as Senior Vice President and Chief Product Officer. Ian brings more than two decades of experience building AI infrastructure platforms and scaling high-performance computing, most recently at Amazon Web Services. He was recently named by HPCwire to its “People to Watch 2026” list, reflecting his reputation in the industry.

Now, let me briefly address our second quarter performance. In Q2, we delivered net sales of \$343 million. Non-GAAP gross margin was 31.2%, and non-GAAP diluted earnings per share were \$0.52. These results reflect strong demand and execution in Memory and continued progress in our AI/HPC business.

Before turning to the segments, I would like to address our updated outlook. As Nate will describe in further detail, following our solid Q2 net sales and EPS performance, we are raising the midpoint of our full year net sales and EPS outlook. We are raising our outlook for our Integrated Memory business, fueled by AI-driven demand, strong execution by our team, and favorable pricing dynamics.

While our second half Advanced Computing net sales outlook is lower than our prior expectations, we are encouraged by strong year-over-year Q2 bookings growth for our non-hyperscaler AI/HPC business, which included 5 new AI/HPC customer wins. That brings our first half total this year to 7 new AI/HPC logos compared to 3 in the first half of last year.

With that context, let me turn to a closer look at each of our segments.

Starting with **Advanced Computing**: net sales for the quarter were \$116M, representing 34% of total company net sales, and declined year-over-year.

Advanced Computing net sales for the second quarter reflect both the timing of large deployments and our transition away from hyperscaler concentration. They also reflect the previously disclosed wind-down of our Penguin Edge business. We believe diversification of net sales and wind-down of Penguin Edge will strengthen the long-term quality of the business.

As I mentioned, we are transitioning our AI Infrastructure business from hyperscaler concentration toward a more diversified customer base across enterprise, neocloud, and sovereign AI. This transition is showing very encouraging progress, but we still have more work to do.

Non-hyperscale AI/HPC net sales grew 50% year-over-year for the first half of the year, representing over 40% of first half segment net sales, supported by strong non-hyperscale year-over-year bookings growth in the quarter including 5 new AI/HPC logos across financial services, biomedical research, and energy. We expect further diversification in the second half of the fiscal year.

Our AI/HPC pipeline continues to strengthen, with opportunities to acquire additional new logos in the second half of the fiscal year across enterprise, neocloud, and sovereign AI customers. As previously discussed, these engagements typically progress over many months from prospecting to design to award, followed by contracting and ultimately system build and deployment. While this sales cycle can be long, often 12-18 months, and can introduce quarterly net sales variability, it also supports deeper customer relationships, repeat business, and more durable long-term growth.

I'm encouraged by the trajectory of the business and the signals we're seeing in the market. Beyond the numbers, we are also seeing increased activity in specific enterprise verticals.

For example, we recently announced our collaboration with Deepgram and Dell to support enterprise voice AI deployments. This win highlights the growing demand for low-latency, production-scale inference infrastructure in real-time applications. In this engagement, Penguin designed and deployed an optimized inference environment built on Dell PowerEdge servers and NVIDIA RTX PRO 6000 Blackwell GPUs. This solution facilitates Deepgram's speech-to-text, text-to-speech, and voice agent functionalities for applications within the healthcare and retail sectors. This case study also demonstrates how our design and integration expertise delivers differentiated value.

As inference workloads scale, we expect these types of deployments to become an increasingly important driver of AI infrastructure demand. Georgia Tech's AI Makerspace, developed in partnership with NVIDIA, is a strong example. Our relationship with Georgia Tech continues to grow and validates Penguin's ability to help organizations move efficiently from concept to production-grade AI infrastructure.

Now, turning to **Integrated Memory**: net sales for the quarter were \$172M, representing 50% of total company net sales, and grew 63% year-over-year. AI-driven demand remains strong across networking, telecommunications, and computing market segments. Pricing dynamics were favorable, and although supply remained tight, we continued to manage constraints effectively through our supplier relationships and disciplined procurement.

Stepping back, our AI/HPC and Memory segments, taken together, enable us to integrate compute and memory architecture in ways that meet the requirements of production AI environments. Memory architecture is becoming increasingly central to AI performance, particularly as inference workloads scale.

Our early investments in CXL position us well as customers evaluate more dynamic memory architectures. Furthermore, we are beginning to see this demand translate into customer deployments, including a recent substantial order for CXL cards from a generative AI company building solutions for inference workloads. This reinforces our strategic position at the intersection of Memory and AI infrastructure to capitalize on the next phase of AI, focused on inference powering agentic AI workloads.

These solutions are sold to enterprise AI infrastructure buyers, the same customers we serve in our AI/HPC business. For example, we sold our CXL-powered KV Cache servers to a tier-one financial institution for their on-premise AI Factory.

In parallel, we continue to advance development of our Photonic Memory Appliance, or PMA (formerly referred to as OMA), which is designed to extend memory capacity and bandwidth for large-scale AI environments. We were an early investor in a photonic memory company, Celestial AI, reflecting our long-standing focus on memory architecture innovation and our early conviction in the importance of optical interconnects for next-generation AI systems. Celestial AI was recently acquired by Marvell in a multi-billion dollar deal. Beyond the portion of the proceeds we received from the acquisition as an investor, we are positioning ourselves for future growth in this market. As inference workloads expand, technologies like PMA can help address key memory scaling challenges in the next-generation AI systems.

Last but not least, **LED**: net sales for the quarter were \$56M, representing 16% of total company net sales, and were down 7% year-over-year. The business continues to operate with focused leadership and dedicated operational discipline. While market conditions remain mixed, we are maintaining a disciplined approach to investment and capital allocation. We are focused on optimizing portfolio value while concentrating resources on areas where we see the strongest long-term returns.

In close, the demand for data center AI infrastructure and Memory is expanding rapidly. AI factories are becoming infrastructure that powers artificial intelligence across a range of industries. As AI shifts toward inference and agentic systems, and scales across large enterprise, neocloud, and sovereign AI environments, we expect demand to accelerate.

At the same time, memory is becoming a defining constraint, and a defining opportunity. Penguin sits at the intersection of AI infrastructure and memory innovation, and we believe that is a powerful position to be in.

Our focus is clear: we are prioritizing four areas:

- **First**, to invest in product innovation across our AI Factory Platform, particularly at the intersection of AI infrastructure and memory, to drive profitable growth;
- **Second**, to execute with speed and precision;
- **Third**, to deepen customer engagement and our ecosystem to support long-term growth; and
- **Fourth**, to continue diversifying our customer base while building toward more consistent and predictable growth.

We believe this focus positions us well to execute in a rapidly evolving market while continuing to build a durable and scalable business.

With that, I'll turn it over to Nate.

Thanks, Kash.

I will focus my remarks on our non-GAAP results, which are reconciled to GAAP in our earnings release tables and in the investor materials available on our website. With that, let me now turn to our second quarter results.

In the quarter, total Penguin Solutions net sales were \$343 million, down 6% year-over-year. Non-GAAP gross margin came in at 31.2%, which was up 0.4 percentage points versus Q2 last year. Non-GAAP operating margin was 13.2%, down 0.2 percentage points versus last year and non-GAAP diluted earnings per share were \$0.52, flat year-over-year.

In the second quarter of fiscal 2026, our overall services net sales totaled \$64 million, up 1% versus the prior year. Product net sales were \$279 million in the quarter, down 8% versus the prior year.

Net sales by business segment were as follows:

- In Advanced Computing: Q2 net sales were \$116 million, which was 34% of total company net sales and down 42% year-over-year. This sales decline reflects both the ongoing wind down of our Penguin Edge business and hyperscale hardware sales in Q2 last year which did not recur in Q2 this year.

Drilling down deeper in our Advanced Computing results, our non-hyperscale, AI/HPC net sales were down 35% year-over-year in the quarter but up 50% for the first half of the year. Given the project nature of the business, where sales can be lumpy from one quarter to the next, we believe looking at the multi-quarter trend is a helpful way to evaluate the growth in this portion of our business.

In addition to solid first half growth in our non-hyperscale AI/HPC business, we continue to make good progress on diversifying our net sales to new customer segments. For the first half of the year, the non-hyperscale AI/HPC business represented more than 40% of total Advanced Computing net sales versus approximately 20% in the first half of last year. We expect to see our mix of net sales from enterprises, neoclouds and sovereign AI customers increase further in the second half of this fiscal year.

- In Integrated Memory: Q2 net sales were \$172 million, which was 50% of total company net sales and up strongly with 63% growth year-over-year; and
- In Optimized LED: Q2 net sales were \$56 million, which was 16% of total company net sales and down 7% versus the same quarter last year.

Non-GAAP gross margin for Penguin Solutions in the second quarter was 31.2 percent, up 0.4 percentage points year-over-year, and up 1.2 percentage points sequentially, with strong margin performance in each business driven primarily by product mix in Advanced Computing, favorable pricing in Memory and tariff recovery in LED. We currently project lower gross margins in the second half, driven by a higher mix of lower margin AI hardware and memory sales, rising memory costs in our AI factory solutions, and less tariff cost recovery in LED.

Non-GAAP operating expenses for the second quarter were \$62 million, down 3% year-over-year and relatively flat sequentially. We expect a modest sequential increase in operating expenses in the second half, reflecting normal seasonality and increased investments in R&D, including for our ClusterWare software and MemoryAI solutions.

Q2 non-GAAP operating income was \$45 million, down 8% year-over-year and up 9% versus last quarter. Operating margins were down 0.2 percentage points versus the prior year but up 1.1 points sequentially, driven by higher sequential gross margins in both Memory and Advanced Computing.

Non-GAAP diluted earnings per share for the second quarter were \$0.52, flat versus Q2 last year and up 7% versus the prior quarter. Adjusted EBITDA for the second quarter was \$50 million, down 6% year-over-year, and up 11% versus the prior quarter.

Turning to the balance sheet; for working capital, our net accounts receivable totaled \$371 million, compared to \$330 million a year ago with the increase driven by higher memory sales volumes and variations in sales linearity across the quarters. Days sales outstanding were healthy at 50 days, consistent with the prior year and down 1 day versus last quarter.

Inventory totaled \$322 million at the end of the second quarter, up from \$200 million a year ago, reflecting increased memory costs, growth in our memory business, and strategic purchases to fulfill memory and AI demand in the second half of the year. Days of inventory was 51 days, up from 37 days a year ago and 38 days last quarter, primarily due to our strategic memory purchases and the timing of receipts and shipments.

Accounts Payable were \$401 million at the end of the quarter, up from \$238 million a year ago due primarily to higher memory costs, growth in our memory business and the timing of purchases and payments. Days payable

outstanding was 63 days compared to 44 days last year and 55 days last quarter. The year-over-year and quarter over quarter movements were due to the timing of purchases and payments.

Our cash conversion cycle was 38 days, an improvement of 5 days compared to Q2 last year and up 3 days versus last quarter due to timing of purchases and payments.

Consistent with past practice, days sales outstanding, days payables outstanding, and inventory days are calculated on a gross sales and gross cost of goods sold basis, which were \$672 million and \$578 million, respectively, in the second quarter.

As a reminder, the difference between gross and net sales is primarily related to our memory business's logistics services, which are accounted for on an agent basis, meaning that we only recognize the net profit on logistics services as net sales.

Cash, cash equivalents and short-term investments totaled \$489 million at the end of the second quarter, down \$158 million from Q2 last year and up \$28 million sequentially. The year-over-year fluctuation was primarily due to proceeds from the issuance of preferred shares in Q2 of last year offset by debt repayments for our term loan in Q4 of last year. Sequentially, the cash increase was due to cash generated from operating activities as well as approximately \$32 million received from proceeds from the disposition of our investment in Celestial AI in connection with its sale to Marvell Technology. These sources of cash were partially offset by our share repurchase activity in the quarter. We ended the quarter with \$450 million of debt, down \$20 million versus last quarter due to the retirement of our 2026 convertible notes. In total, we closed the quarter in a net cash position and based on our current debt maturity schedule have no further scheduled debt payments due until 2029.

Second quarter cash flows provided by operating activities totaled \$55 million, compared to \$73 million provided by operating activities in the prior year quarter. The decrease in cash flow in the quarter versus last year was due primarily to investments in net working capital to support growth for the second half of this fiscal year.

For those of you tracking capital expenditures and depreciation, capital expenditures were \$2 million in the second quarter, and depreciation was \$5 million for the quarter.

Wrapping up our cash flow activities, we spent \$32 million dollars to repurchase approximately 1.7 million shares in the second quarter under our stock repurchase program. As of February 27, 2026, an aggregate of \$64.5 million remained available for the repurchase of our common stock under the Current Authorizations.

And now turning to our outlook. Given our solid H1 performance and an improved H2 outlook for our memory business, we are raising our full-company net sales and non-GAAP diluted EPS outlook for the year, which at the midpoint now calls for 12% net sales growth and \$2.15 of non-GAAP diluted EPS, up from our previous outlook of 6% net sales growth and \$2.00 of non-GAAP diluted EPS. As a reminder, our full year outlook assumes that we will continue to diversify our customer sales mix and does not include any Advanced Computing AI hardware sales to hyperscale customers. And also consistent with our assumptions from last quarter, our FY26 financial

outlook reflects the ongoing wind down of our high margin Penguin Edge business. We expect sales from this business to essentially cease by the end of fiscal 2026.

The combined effect of these two assumptions in our FY26 outlook remains approximately a 14 percentage point unfavorable year-over-year impact to our total company net sales growth and approximately a 30 percentage point unfavorable impact to Advanced Computing.

With that said, our full year net sales outlook reflects the following full year growth ranges by segment:

- for **Advanced Computing**, we now expect full year net sales to change between minus 25 and minus 15% year-over-year. While our Advanced Computing net sales outlook for this fiscal year is lower than our previous forecast, we are encouraged by our AI/HPC bookings, including several new logos, and pipeline growth. As it has previously, this outlook reflects the Penguin Edge and hyperscale hardware sales impacts mentioned earlier.
- for **Memory**, we now expect net sales to grow between 65 and 75% year-over-year driven by strong demand and a favorable pricing environment.
- and for **LED**, we continue to expect net sales to decline between minus 15 and minus 5% year-over-year.

Our non-GAAP gross margin outlook for the full year is now 28.0% plus or minus 0.5 percentage points. We adjusted our gross margin outlook down by 1 percentage point to account for a higher mix of memory sales, which have a lower gross margin than our company average, and higher memory costs in our AI hardware business.

Our full year expectation for total non-GAAP operating expenses remains \$250 million, and we have narrowed that range to plus or minus \$5 million.

For FY26, we now expect a non-GAAP diluted share count of approximately 53 million shares, down from our prior outlook, primarily reflecting the impact of our recent share repurchases.

Our non-GAAP full year diluted earnings per share is now expected to be approximately \$2.15, plus or minus 15 cents.

Our forecasted FY26 non-GAAP tax rate remains at 22%.

While we expect to use this normalized non-GAAP tax rate throughout FY26 and beyond, the long-term non-GAAP tax rate may be subject to changes for a variety of reasons, including the rapidly evolving global and US tax environment, significant changes in our geographic earnings mix or changes to our strategy or business operations.

Our outlook for fiscal year 2026 is based on the current environment, which contemplates, among other things, the global macroeconomic environment and ongoing supply chain constraints, especially as they relate to our Advanced Computing and Integrated Memory businesses. This includes extended lead-times for certain components that are incorporated into our overall solutions, impacting how quickly we can ramp existing and

new customer projects and fulfill customer orders. Our outlook also contemplates the industry-wide higher costs for memory, which may slow customer demand for our products and solutions and may lower our gross margins in our Advanced Computing and Memory businesses.

Overall, we believe our focused execution, disciplined expense management, and balance sheet strength provide a strong foundation for sustained, profitable growth. We expect these qualities to support our continued progress as we pursue opportunities to enhance long-term shareholder value.

Please refer to the non-GAAP financial information section and the reconciliation of GAAP to non-GAAP measures tables in our earnings release and the investor materials on our website for further details.

With that Operator, we are ready for Q&A.

QUESTIONS AND ANSWERS

Operator

(Operator Instructions) Katherine Murphy from Goldman Sachs.

Katherine Murphy - Goldman Sachs - Analyst

I'll ask about the raised memory segment outlook for 65% to 75% growth, how much of this is from increased favorable pricing versus demand for new product categories?

And as a follow-up, how should we think about the impacts to the operating margin outlook for this segment and the investments that need to be made into new technologies like CXL and photonic memory appliances?

Nate Olmstead - Penguin Solutions, Inc. - Senior Vice President and Chief Financial Officer

Kat, it's Nate. So, on the memory outlook, listen, we're really pleased with the demand that we're seeing as well as the favorability that we see in the pricing environment.

I would say for the increase that we're seeing in the second half, that's majority pricing, but demand is also very strong across telco, networking, AI-driven demand is just very strong. In fact, to get to the high end of that outlook really just refers to our ability to secure materials, which is really the only inhibitor we see right now to raising that outlook here in the second half.

So, we are chasing the materials. We're using the balance sheet to strategically purchase ahead where we can, but the demand is very strong in memory. In terms of the investments, we've reflected it in the outlook.

So, I kept the OpEx for the year at \$250 million, plus or minus \$5 million. We're balancing the portfolio as we always do, to look for opportunities to accelerate our investments in innovation in AI or in the memory solutions that we've been talking about. But that's all included in the outlook.

I expect the operating margins for memory to remain pretty healthy in the back half of the year. I do expect some pressure on gross margins in AI as we see a higher mix of new hardware shipments in the second half as well as factoring in some of the higher memory input costs that we have in that business.

Operator

Brian Chin from Stifel.

Brian Chin - Stifel Nicolaus & Company, Inc. - Analyst

Thank you, I'd like to ask a few questions, and maybe first question, I guess in advanced computing, what changed that caused you to lower the midpoint of your prior guidance to the new range you've communicated? And can you describe how booked you are to that midpoint of that new range?

Kash Shaikh - Penguin Solutions, Inc. - President and Chief Executive Officer

So, one of the main factors is the lag between our bookings and the revenue. Our revenue lags about 3 to 6 months from the time of the bookings. And this is primarily driven by the timing of the deployments. In some cases, the material availability and so on and so forth.

And given where we are in terms of our fiscal year, we have five months remaining. So going forward, most of the bookings that we are expecting may not materialize into the revenue for the second half of this fiscal, but we believe that it will have a positive impact, obviously, going into the first half of the next fiscal.

So that's one of the reasons that we are lowering the guidance for advanced computing driven by the deployments, but we are seeing strong momentum in our pipeline as well as bookings.

Bookings grew very significantly in Q2 for non-hyperscale AI HPC business, which is very strategic for us, and we are encouraged to see the progress.

We closed five new logos with the AI/HPC in Q2 and first half, that figure total to seven new logos as compared to three new logos last year. So, we are very confident in our ability to execute. The main issue at this point is timing.

Brian Chin - Stifel Nicolaus & Company, Inc. - Analyst

Okay. Yes. I appreciate that, Kash. And it sounds like you're pretty well booked into the fiscal second half, lowered outlook and that some of these new bookings are more kind of beyond a six-month window.

Also thinking about growth in the business, obviously, there's that sort of headwind that you helped to clarify in terms of reduction in hardware revenue to the new hyperscaler, the wind down of Penguin Edge.

And so, 30-percentage-point impact, if we kind of net that against the guidance, maybe 10% growth for this year, net of that in that segment.

So, moving forward, as you survey the business and you haven't been in the role that long, can you think about what that sort of apples-to-apples growth rate was or is tracking to for this fiscal year? How are you thinking about sort of target growth rates for the advanced computing business moving forward?

Kash Shaikh - Penguin Solutions, Inc. - President and Chief Executive Officer

So overall, let me give you a data point. So, the first half of this fiscal, our net sales grew about 50% year-over-year for non-hyperscale AI/HPC business, representing 40% of the overall mix of advanced computing, which is almost 2x of what we closed last fiscal.

So, the growth is substantial in terms of the bookings as well as the revenue that we see. And we expect that to continue.

And as we continue to close the booking, converting the pipeline, we see strong pipeline across all three segments that I mentioned between enterprises on-premises AI deployments significant activity with sovereign AI customers as well as neocloud customers.

Operator

Matthew Calitri from Needham & Company.

Matthew Calitri - Needham & Company LLC - Equity Analyst

Matt Calitri here from Needham. Do the new memory launches mark a shift in strategy on that front? Just curious because in the past, the company has talked kind of more about the niche parts of the integrated memory business and noted it's early on things like CXL front.

But now it sounds like memory is expected to be a larger driver as part of this AI factory platform. So just wondering if anything has changed there and what gives you confidence there's durable demand here?

Kash Shaikh - Penguin Solutions, Inc. - President and Chief Executive Officer

Yes. So, it is a part of our strategy. The MemoryAI appliances that we launched about a month ago, starting with GPC is a part of us investing more in our AI factory platform strategy.

So, there are six elements to this strategy and MemoryAI is one of the strategic elements where it is very timely if you look at how AI is transitioning from model training to inference.

And in the workloads where you are focused on inference, memory becomes an increased requirement because of lower latency as well as larger context size for inference, powering the agent. So, this is very strategic for our business.

In fact, we are leading the market in this area, taking advantage of our unique position at the intersection of memory and AI infrastructure and combining the deep understanding and architecture, we introduced this MemoryAI, KV cache server as one of the products in the line of memory, we are working on other products, and we will continue to invest and in fact, invest more in this area to take advantage of the market opportunity because the timing is perfect and our leadership in the MemoryAI line of products.

To give you a proof point, one of the new logos we acquired, Tier 1 financial institution, not only we are deploying the AI infrastructure AI factory deployment for them, they also purchased our CXL-based KV cache server, which is a proof point of as customers are transitioning from training and bringing AI on-premises in their factories, deploying on-premises, focusing on inference, and powering agentic AI, it is very strategic for us and the timing is ripe. So, we expect to see this demand, and we plan to continue to invest in this area.

Matthew Calitri - Needham & Company LLC - Equity Analyst

Awesome. That's great to hear. And then met with the new CEO in the seat and some moving pieces around sales cycles and supply chains, did you change the guidance philosophy at all or embed any additional conservatism? Any color on the puts and takes there would be helpful.

Nate Olmstead - Penguin Solutions, Inc. - Senior Vice President and Chief Financial Officer

Yes. Matt, no, no change in the philosophy. We Kash and I very quickly aligned, I think, on how we think about tracking the business and looking at things.

And in fact, I think with our new CRO, which, who came in a couple of quarters ago, he's done a nice job of adding some more rigor to the planning process in our AI business and just improving the visibility there a little bit.

But it's a challenging environment from a supply chain standpoint, and we of course, got a lot of experience managing supply chain in our memory business, and I think that that's an advantage for us in an environment like this.

Operator

Samik Chatterjee from JPMorgan.

Samik Chatterjee - JPMorgan Chase & Co - Analyst

This is MP on behalf of Samik Chatterjee. So, my first question is I just wanted to double click on your advanced computing guidance. You mentioned that a lag of three to six months for the revenue which we will book in your second half.

But was there a change observed for the bookings which you did in first quarter? Or any change in relative to what were you expecting to do in 2Q? And I have a follow-up as well.

Nate Olmstead - Penguin Solutions, Inc. - Senior Vice President and Chief Financial Officer

Yes. MP, I think bookings were strong in Q2, really good growth sequentially and year-over-year. I do think that the deployment cycle has lengthened a little bit with some of the supply constraints. In particular, on memory, things have gotten a little bit longer.

But we're really pleased with the five new logos. And I think demand is good. We're seeing good strength in the pipeline, and it's also diversifying nicely across the non-hyperscale segments such as enterprise and neocloud and sovereign.

So, I think we feel really good about the demand. I think this is just an issue of a little bit of timing as we can convert bookings into revenue.

Samik Chatterjee - JPMorgan Chase & Co - Analyst

Okay. And my second question would be also on advanced computing and your AI factory related business. Like does NVIDIA coming up with their own reference designs for factory level solutions.

Like how does that play relative to you? Like, is that a tailwind for you? Or is that a headwind for you? Like can you please help us understand?

Kash Shaikh - Penguin Solutions, Inc. - President and Chief Executive Officer

Yes. We believe this is an advantage for us. So, we work very, very closely with NVIDIA and some of the wins that I mentioned, for example, the financial institution recently along with our MemoryAI product in this transaction.

NVIDIA worked very closely with us, and we are working with NVIDIA leveraging their reference design, combining that with our AI factory platform and complementing NVI (NVIDIA's Link Fusion), as an example, to provide full stack to our customers.

So, their blueprints are more complementary to our AI factory platform and the components that make up for it. So, we are actually quite excited about those blueprints and working very closely with NVIDIA to capture the opportunities, especially as NVIDIA is increasingly focused on enterprise, it aligns with our strategy and go to market.

Operator

Ananda Baruah from Loop Capital.

Ananda Baruah - Loop Capital Markets LLC - Analyst

A couple, if I could. Kash and maybe Nate as well. Earlier remarks were that you're seeing increased momentum across neo cloud, sovereign and enterprise. And you mentioned one or two of the new wins. Do you have, and I think Kash you had mentioned you've made some specific or at least general inferencing remarks, including around agentic.

Do you have any specific context you can give us around what your customers are telling you their thrust in inferencing is right now? And maybe the degree to which agentic is showing up there.

Like we just want to get a sense of what the customer's activity tone is like behaviorally, say, over the last 90 to 180 days. Do you have anything there you can share with us to make it a little bit more experiential for us? And then I have a quick follow-up.

Kash Shaikh - Penguin Solutions, Inc. - President and Chief Executive Officer

Sure. We believe we are early in the adoption of inference with these customers, but it is increasingly deployed as in customers as they move towards Agentic. Inference provides the opportunity for powering the agentic.

And when you think about inference, I'll give you an example of why the architecture is changing and why memory is becoming increasingly critical in inference as compared to the model training.

So, for example, let's say, if you are writing a book and if you have to write a new sentence without having the memory as a supporting component for you, you will have to reread the entire book before writing the next sentence. So, in the inference, you're doing inference on a lot of data you already have.

And if you have a component where the book you have written so far is stored. So before writing the new sentence, you don't have to reread the book that kind of how it is changing for the enterprises and other segments. And we see customers already deploying it.

And the architecture is changing, which is why not only we have the opportunity and advantage to provide them our AI infrastructure as well as the services increasingly, we are seeing the demand for our MemoryAI portfolio where they are deploying AI infrastructure and increasingly inference, they need products like that to be able to provide that memory component for the inference, so the responses of LLM can be much faster than they would be otherwise.

Ananda Baruah - Loop Capital Markets LLC - Analyst

I got it. That's helpful. And just one last one quick follow-up, I'm mindful of the time here in case if anybody behind me. The CXL product, it sounds like, to the earlier question, it sounds like you guys are a little bit more enthusiastic about the CXL server today than you were maybe 90 days ago--you have the new products out of GTC.

Is that accurate statement that you're expecting maybe because of these new products, a little bit more and certainly some of the NVIDIA announcements of CES as well? But are you expecting a little bit more revenue a little bit sooner than maybe you were at CXL 90 days ago?

And a quick second part to that. Do you need photonics to work before you really get CXL amplification like you need CPO or photonics to work before you can really amplify CXL for scale up? That's it for me.

Kash Shaikh - Penguin Solutions, Inc. - President and Chief Executive Officer

Yes. So let me address your CXL question first. I think CXL adoption is timely given the transition to inference because, as I mentioned with inference, you need increased memory, faster LLM responses.

And what CXL provides Compute Express Link is you can share the memory between GPUs and CPUs. So, what it allows is new memory pooling, which is an advantage in inference workloads.

So, while CXL was obviously available for the last few quarters, it is driving that, that inference adoption is driving the adoption for CXL and this transaction that I mentioned where we received an order, it's actually an enterprise generative AI company working on inference workloads.

So, you can imagine CXL cards make sense for them because those workloads need increased memory and the memory pooling capabilities provided by CXL between GPUs and CPUs are an advantage for those kinds of customers.

And then in terms of photonic memory appliance that we are working on in our partnership with Celestial AI, which is now obviously Marvell, that provide increased capability because, obviously, when you have photonic connectivity, then you have increased capacity to share the memory. So, it takes it to the next level.

However, CXL in itself is an advantage, we can take it to the next level with the photonic appliance. There is another element which is KV cache that I mentioned, MemoryAI KV cache server, which is essentially providing much more responsiveness for larger context workloads, again, used in inference.

So various requirements, you can think of it as inference as various requirements related to memory and the type of workloads it has and some of it is latency.

So, these components between CXL or the CXL-based KV cache, which provides increased responses and larger memory side, largest context sizes and then taking it to the next level, photonic memory make up various use cases for inference or inference gets mainstream. We will have an advantage of this portfolio helping with various use cases of inference.

Operator

Kevin Cassidy from Rosenblatt.

Kevin Cassidy - Rosenblatt Securities, Inc. - Analyst

Yes. Just the gross margin for the memory, your gross margin was up in the quarter and memory revenue was up strong, and I just want to understand that what the dynamics are there.

Nate Olmstead - Penguin Solutions, Inc. - Senior Vice President and Chief Financial Officer

Yes, sure, Kevin. We saw a little favorability in memory margins. Some of that is mix, a little bit stronger demand in flash actually, which is a little bit higher margin product for us within the portfolio.

And then also some of the pricing increases, we were able to capture a little bit of margin upside on that just based on the timing of our inventory purchases relative to the timing of shipments and sales to customers.

Kevin Cassidy - Rosenblatt Securities, Inc. - Analyst

Okay. So, you kind of, as you look out to the second half of the year, you see that catching up price increases compared to.

Nate Olmstead - Penguin Solutions, Inc. - Senior Vice President and Chief Financial Officer

Yes. So, the price increase is slow, right? If that's an assumption that you use, if price increases are going to slow, then we would see, we would expect to see less margin favorability from that because there'd be less of a timing difference between or less of a price variation between the timing between purchasing inventory and selling.

But we have been using the balance sheet to try to secure inventory where we can. It's a tight market, so it's not unlimited supply. But where we can, we're using the balance sheet to try to gain a little bit of an advantage.

Kevin Cassidy - Rosenblatt Securities, Inc. - Analyst

Okay. And maybe just as we're talking about memory as you get to the CXL systems, would you expect that's going to be a higher margin than the module business?

Nate Olmstead - Penguin Solutions, Inc. - Senior Vice President and Chief Financial Officer

Yes, we do. It's really a solution. It's got software aspects to it, some good differentiation on the hardware as well. So, I see that as a nice margin opportunity for us down the road.

Operator

At this time, there are no further questions. I will now hand the call over to Kash Shaikh, CEO, for closing remarks.

Kash Shaikh - Penguin Solutions, Inc. - President and Chief Executive Officer

Thank you, operator. We see AI shifting towards inference with demand expanding beyond hyperscaler to enterprise, neocloud and sovereign AI customers. We are still in the early shift in this transition, but the combination of our customer demand, product innovation and booking momentum gives us the confidence in the path ahead.

We believe we are well positioned at the intersection of AI compute infrastructure and memory, and we are making good progress diversifying our customer base.

My focus is on strong execution across product innovation, customer engagement and diversification, disciplined capital allocation and investment in our AI/HPC business to support the long-term growth. We look forward to updating you on our progress.

Operator

This concludes today's call. Thank you for attending. You may now disconnect.