Amazon.com, Inc. (NASDAQ:AMZN) Q4 2024 Earnings Conference Call February 6, 2025 5:00 PM ET

Company Participants

Dave Fildes - VP of IR
Andy Jassy - CEO
Brian Olsavsky - CFO

Conference Call Participants

Mark Mahaney - Evercore ISI
Eric Sheridan - Goldman Sachs
Doug Anmuth - JPMorgan
John Blackledge - TD Cowen
Michael Morton - MoffettNathanson

Operator

Thank you for standing by. Good day, everyone, and welcome to the Amazon.com Fourth Quarter 2024 Financial Results Teleconference. At this time, all participants are in a listen-only mode. After the presentation, we will conduct a question-and-answer session. Today's call is being recorded.

And for opening remarks, I will be turning the call over to the Vice President of Investor Relations, Dave Fildes. Thank you, sir. Please go ahead.

Dave Fildes

Hello and welcome to our Q4 2024 financial results conference call. Joining us today to answer your questions is Andy Jassy, our CEO; and Brian Olsavsky, our CFO. As you listen to today's conference call, we encourage you to have our press release in front of you, which includes our financial results as well as metrics and commentary on the quarter. Please note, unless otherwise stated, all comparisons in this call will be against our results for the comparable period of 2023.

Our comments and responses to your questions reflect management's views as of today, February 6, 2025, only and will include forward-looking statements. Actual results may differ materially. Additional information about factors that could potentially impact our financial results is included in today's press release and our filings with the SEC, including our most recent annual report on Form 10-K and subsequent filings.

During this call, we may discuss certain non-GAAP financial measures. In our press release, slides accompanying this webcast and our filings with the SEC, each of which is posted on our IR website, you will find additional disclosures regarding these non-GAAP measures, including reconciliations of these measures with comparable GAAP measures.

Our guidance incorporates the order trends that we've seen to date and what we believe today to be appropriate assumptions. Our results are inherently unpredictable and may be materially affected by many factors, including fluctuations in foreign exchange rates, changes in global economic and geopolitical conditions and customer demand and spending, including the impact of recessionary fears, inflation, interest rates, regional labor market constraints, world events, the rate of growth of the internet, online commerce, cloud services and new and emerging technologies and the various factors detailed in our filings with the SEC.

Our guidance assumes, among other things, that we don't conclude any additional business acquisitions, restructurings, or legal settlements. It's not possible to accurately predict demand for our goods and services and therefore, our actual results could differ materially from our guidance.

And now I'll turn the call over to Andy.

Andy Jassy

Thanks, Dave. Today, we're reporting $187.8 billion in revenue, up 10% year-over-year. Given the way the dollar strengthened throughout the quarter, we saw $700 million more of foreign exchange headwind than we anticipated at guidance. Without that headwind, revenue would have been 11% year-over-year and exceeded the top-end of our guidance.

Operating income was $21.2 billion, up 61% year-over-year and trailing 12-month free cash flow adjusted for equipment finance leases was $36.2 billion, up $700 million year-over-year. We're pleased with the invention, customer experience improvements and results delivered in 2024 and have a lot more planned in 2025.

I'll start by talking about our Stores business. We saw 10% year-over-year revenue growth in our North America segment and 9% year-over-year in our International segment, excluding the impact from foreign exchange rates. Our continued focus on expanding selection, lowering prices and improving convenience drove strong unit growth that even outpaced our revenue growth.

We continue to add to our broad range of selection, giving customers choice across a variety of price points. We welcome notable brands to our store throughout 2024, including Clinique, Estee Lauder, Oura Rings and Armani Beauty. We continue to add to the hundreds of millions of products offered from our selling partners who made up 61% of items that we sold in 2024, our highest annual mix of third-party seller units ever. We also launched Amazon Haul for US customers in Q4, which offers customers an engaging shopping experience that brings ultra-low price products into one convenient destination. It's off to a very strong start.

Customers continue to want Amazon to be the place they rely on for sharp pricing. In the fourth quarter, consumers saved more than $15 billion with our low everyday prices and record-setting events during Prime big deal days in October and Black Friday and Cyber Monday around Thanksgiving. Additionally, for Federal's annual pricing study found that entering the holiday season, Amazon had the lowest online prices for the eighth year in a row, averaging 14% lower prices on average than other leading retailers in the US.

Our speed of delivery continues to accelerate and 2024 was another record-setting year for Prime members. We expanded the number of same-day delivery sites by more than 60% in 2024, which now serve more than 140 metro areas. And overall, we delivered over 9 billion units the same or next day around the world. Our relentless pursuit of better selection, price and delivery speed is driving accelerated growth in Prime membership.

For just $14.99 a month, Prime members get unlimited free shipping on 300 million items often the same-day or one-day delivery, exclusive shopping events like Prime Day, access to a vast collection of premium programming and live sports on Prime Video, ad-free listening of 100 million songs and podcasts with Amazon Music, access to unlimited generic prescriptions for only $5 a month, unlimited grocery delivery and orders over $35 from Whole Foods Market and Amazon Fresh for $9.99 a month, a free Grubhub+ membership with free unlimited delivery and our latest benefit of a $0.10 per gallon fuel discount at BP, Amoco and AMPM stations.

When you think about this as a whole and also compared to many other membership services that are comparably and more expensively priced and offer just one benefit like video, Prime is a streaming deal and we have more coming for our Prime members in 2025. We also remain squarely focused on cost to serve in our fulfillment network, which has been a meaningful driver of our increased operating income.

We've talked about the regionalization of our US network. We've also recently rolled out our redesigned US inbound network. While still in its early stages, our inbound efforts have improved our placement of inventory so that even more items are closer to end customers. Ahead of Black Friday in November, we'd improved the percentage of ordered units available in the ideal building by over 40% year-over-year.

We've also spent considerable time optimizing the number of items we send customers in the same package, which reduces packaging is more convenient for customers and less expensive for us to fulfill. And our per unit transportation costs continue to decline as we build out and optimize our last mile network. Overall, we've reduced our global cost to serve on a per unit basis for the second year in a row, while at the same time, increasing speed, improving safety and adding selection.

As we look to 2025 and beyond, we see opportunity to reduce costs again as we further refine inventory placement, grow our same-day delivery network and accelerate robotics and automation throughout the network.

In advertising, we remain pleased with the strong growth on a very large base, generating $17.3 billion of revenue in the quarter and growing 18% year-over-year. That's a $69 billion annual revenue run rate, more than double what it was just four years ago at $29 billion.

Sponsored products, the largest portion of ad revenue are doing well and we see runway for even more growth. We also have a number of newer streaming offerings that are starting to become significant new revenue sources. On the streaming video side, we wrapped up our first year of prime video ads and we're quite pleased with the early progress and head into this year with momentum.

We've made it easier to do full-funnel advertising with us. Full-funnel is from the top of the funnel with broad-reach advertising that drives brand awareness to mid-funnel, where sponsored brands let companies specify certain keywords and audiences to attract people to their detail pages or brand store on Amazon. To bottom of the funnel, where sponsored products help advertisers service relevant product ads to customers at the point of purchase. We make this easy for brands to sign up for and deploy across our growing advertising. We also have differentiated audience features that leverage billions of customer signals across our stores and media destinations.

From Amazon Marketing Cloud's secure clean rooms, providing advertisers the ability to analyze data, produce core marketing metrics and understand how their marketing performs across various channels to our new multi-touch attribution model that helps advertisers understand how their marketing is working. If an advertiser uses streaming TV, display, sponsored products and other ad types in their campaign, multi-touch attribution will show the relative contribution of each to their sales.

Moving on to AWS. In Q4, AWS grew 19% year-over-year and now has a $115 billion annualized revenue run rate. AWS is a reasonably large business by most folks' standards. And though we expect growth will be lumpy over the next few years as enterprise adoption cycles, capacity considerations and technology advancements impact timing, it's hard to overstate how optimistic we are about what lies ahead for AWS's customers and business.

I spent a fair bit of time thinking several years out. And while it may be hard for some to fathom a world where virtually every app has generative AI infusing it with inference being a core building block just like compute, storage and database, and most companies having their own agents that accomplish various tasks and interact with one another. This is the world we're thinking about all the time and we continue to believe that this world will mostly be built on top of the cloud with the largest portion of it on AWS.

To best help customers realize this future, you need powerful capabilities in all three layers of the stack. At the bottom layer for those building models, unique compelling chips. Chips are the key ingredient in the compute that drives training and inference. Most AI compute has been driven by NVIDIA chips and we obviously have a deep partnership with NVIDIA and will for as long as we can see into the future.

However, there aren't that many generative AI applications at large scale yet. And when you get there as we have with apps like Alexa and Rufus, cost can get steep quickly. Customers want better price performance and it's why we built our own custom AI silicon. Trainium 2 just launched at our AWS Reinvent Conference in December and EC2 instances with these chips are typically 30% to 40% more price performant than other current GPU-powered instances available. That's very compelling at scale. Several technically capable companies like Adobe, Databricks, poolside, and Qualcomm have seen impressive results in early testing of Trainium 2. It's also why you're seeing Anthropic build their future frontier models on Trainium 2.

We're collaborating with Anthropic to build Project Rainier, a cluster of Trainium 2 ultra servers containing hundreds of thousands of Trainium 2 chips. This cluster is going to be 5 times the number of exaflops as the cluster that Anthropic used to train their current leading set of cloud models. We're already hard at work on Trainium 3, which we expect to preview late in 2025 and defining Trainium 4 thereafter. Building outstanding performing chips that deliver leading price performance has become a core strength of AWS', starting with our Nitro and Graviton chips in our core business and now extending to Trainium in AI and something unique to AWS relative to other competing cloud providers.

The other key component for model builders is services that make it easier to construct their models. I won't spend a lot of time in these comments on Amazon's SageMaker AI, which has become the go-to service for AI model builders to manage their AI data, build models, experiment, and deploy these models, except to say, the SageMaker's HyperPod capability, which automatically splits trading workloads across many AI accelerators, prevents interruptions by periodically saving checkpoints and automatically repairing faulty instances from their last saved checkpoint, and saving training time by up to 40%.

It continues to be a differentiator, received several new compelling capabilities at Reinvent, including the ability to manage costs at a cluster level and prioritize which workloads should receive capacity when budgets are reached, and is increasingly being adopted by model builders. At the middle layer for those wanting to leverage frontier models to build GenAI apps, Amazon Bedrock is our fully managed service that offers the broadest choice of high-performing foundation models with the most compelling set of features that make it easy to build a high-quality Generative AI application. We continue to iterate quickly on Bedrock, announcing Luma AI, poolside, and over 100 other popular emerging models to Bedrock and Reinvent.

In short order, we also just added DeepSeek's R1models to Bedrock and SageMaker. And additionally, we delivered several compelling new Bedrock features at Reinvent, including prompt caching, intelligent prompt routing, and model distillation, all of which help customers achieve lower cost and latency in their inference.

Like SageMaker AI, Bedrock is growing quickly and resonating strongly with customers. Related, we also just launched Amazon's own family of frontier models in Bedrock called Nova. These models compare favorably in intelligence against the leading models in the world, but offer lower latency, lower price, about 75% lower than other models in Bedrock, and are integrated with key Bedrock features like fine tuning, model distillation, knowledge bases of RAG and agentic capabilities. Thousands of AWS customers are already taking advantage of the capabilities and price performance of Amazon Nova models, including Palantir, Deloitte, SAP, Dentsu, Fortinet, Trellix, and Robinhood, and we've just gotten started.

At the top layer of the stack, Amazon Q is the most capable Generative AI-powered assistant for software development and to leverage your own data. You may remember that on the last call, I shared the very practical use case where Q Transform helped save Amazon's teams $260 million and 4,500 developer years in migrating over 30,000 applications to new versions of the Java JDK. This is real value and companies asked for more, which we obliged with our recent deliveries of Q transformations that enable moves from Windows.NET applications to Linux, VMware to EC2, and accelerates mainframe migrations.

Early customer testing indicates that Q can turn what was going to be a multiyear effort to do a mainframe migration into a multi-quarter effort, cutting by more than 50% the time to migrate mainframes. This is a big deal and these transformations are good examples of practical AI. While AI continues to be a compelling new driver in the business, we haven't lost our focus on core modernization of companies' technology infrastructure from on-premises to the cloud.

We signed new AWS agreements with companies, including Intuit, PayPal, Norwegian Cruise Line Holdings, Northrop Grumman, the Guardian Life Insurance Company of America, Reddit, Japan Airlines, Baker Hughes, the Hertz Corporation, Redfin, Chime Financial, Asana, and many others. Consistent customer feedback from our recent AWS Reinvent gathering was appreciation that we're still inventing rapidly in non-AI key infrastructure areas like storage, compute, database and analytics.

Our functionality leadership continues to expand, and there were several key launches customers were abuzz about, including Amazon Aurora DSQL, our new serverless distributed SQL database that enables applications with the highest availability, strong consistency, PostgreS compatibility, and 4 times faster reads and writes compared to other popular distributed SQL databases. Amazon S3 tables, which make S3 the first cloud object store with fully managed support for Apache Iceberg for faster analytics.

Amazon S3 Metadata, which automatically generates queryable metadata, simplifying data discovery, business analytics, and real-time inference to help customers unlock the value of their data in S3. And the next generation of Amazon SageMaker, which brings together all of the data, analytics services, and AI services into one interface to do analytics and AI more easily at scale.

As 2024 comes to an end, I want to thank our teammates and partners for their meaningful impact throughout the year. It was a very successful year across almost any dimension you pick. We're far from done and look forward to delivering for customers in 2025.

With that, I'll turn it over to Brian for a financial update.

Brian Olsavsky

Thanks, Andy. Starting with our top-line financial results, Worldwide revenue was $187.8 billion, an 11% increase year-over-year, excluding the impact of foreign exchange. This equates to an approximate $900 million headwind from FX in the quarter, which is about $700 million higher than what we'd anticipated in our Q4 guidance range. Excluding that additional FX headwind, we would have exceeded the top end of our revenue guidance range. Worldwide operating income was $21.2 billion, our largest operating income quarter ever, and was $1.2 billion above the high end of our guidance range. Across all segments, we continued to innovate for customers while operating more efficiently at the same time.

In the North America segment, fourth-quarter revenue was $115.6 billion, an increase of 10% year-over-year. International segment revenue was $43.4 billion, an increase of 9% year-over-year, excluding the impact of foreign exchange. Worldwide paid units grew 11% year-over-year as our focus on low prices, broad selection, and fast shipping continues to resonate with customers.

Shifting to profitability. North America segment operating income was $9.3 billion, an increase of $2.8 billion year-over-year. Operating margin was 8%, up 190 basis points year-over-year. In the International segment, operating income was $1.3 billion, an improvement of $1.7 billion year-over-year. Operating margin was 3%, up 400 basis points year-over-year. This marks the eighth consecutive quarter where we've seen year-over-year margin improvement in both the North America and International segments.

2024 also marks the second year in a row where we've lowered our global cost to serve on a per-unit basis. In the fourth quarter, we saw strong productivity in our transportation network from improved inventory placement, higher units per package, and reduced travel distances. We also saw improved productivity in our fulfillment centers.

Overall, our teams executed extremely well throughout the quarter and particularly during our peak seasons. I want to thank them for all they do to deliver for our customers. Looking ahead, we have several opportunities to keep lowering our costs through even better inventory placement, which also allows us to deliver items to customers faster.

In the US, we're tuning our inbound network and continuing to expand our same-day delivery network. Globally, we're adding automation and robotics throughout our network. While these efforts will take time to implement and progress may not be linear, we have a good plan to continue to drive improvements in our cost structure.

Advertising remains an important contributor to profitability in the North America and International segments. This quarter, we saw strong advertising revenue growth on an increasingly large base. We will also continue to invest in experiences that have potential to be important to customers in Amazon long term in areas like Alexa, healthcare, and grocery, as well as Kuiper, including the planned launches of our production satellites in the coming months.

As a reminder, we currently expense the majority of the costs associated with the development of our satellite network. We will capitalize certain costs once the service achieves commercial viability, including sales to customers.

Moving next to our AWS segment. Revenue was $28.8 billion, an increase of 19% year-over-year. AWS now has an annualized revenue run rate of $115 billion. During the fourth quarter, we continue to see growth in both generative AI and non-generative AI offerings as companies turn their attention to newer initiatives, bring more workloads to the cloud, restart or accelerate existing migrations from on-premise to the cloud and tap into the power of generative AI. Customers recognize to get the full benefit of generative AI, they have to move to the cloud.

AWS reported operating income of $10.6 billion, an increase of $3.5 billion year-over-year. This is a result of strong growth, innovation in our software and infrastructure to drive efficiencies and continued focus on cost-control across the business. As we've said in the past, we expect AWS operating margins to fluctuate over time, driven in part by the level of investments we're making. Additionally, we increased the estimated useful life of our servers starting in 2024, which contributed approximately 200 basis points to the AWS margin increase year-over-year in Q4.

Now turning to our capital investments. As a reminder, we define these as a combination of cash CapEx plus equipment finance leases. Capital investments were $26.3 billion in the fourth quarter, and we think that run rate will be reasonably representative of our 2025 capital investment rate. Similar to 2024, the majority of the spend will be to support the growing need for technology infrastructure. This primarily relates to AWS, including to support demand for our AI services as well as tech infrastructure to support our North America and International segments.

Additionally, we're continuing to invest in capacity for our fulfillment and transportation network to support future growth. We're also investing in same-day delivery facilities and our inbound network as well as robotics and automation to improve delivery speeds and to lower our cost to serve. These capital investments will support growth for many years to come.

Turning to our revenue guidance for Q1. Net sales are expected to be between $151 billion and $155.5 billion. I'd like to highlight two items impacting our Q1 revenue guidance. First, we estimate the year-over-year impact of changes in foreign exchange rates based on current rates, which we expect to be a headwind of approximately $2.1 billion in Q1 year-over-year or 150 basis points. As a reminder, global currencies can fluctuate during the quarter and just as we saw in Q4 with the strengthening of the dollar versus most other currencies.

Second, a reminder that we are comping the impact of last year's leap year. The extra date contributed approximately $1.5 billion of additional net sales across our businesses in Q1 2024 or about 120 basis points to the year-over-year growth rate, which impacted all segments. Q1 operating income is expected to be between $14 billion and $18 billion. This guidance includes the estimated impact of certain updates to the useful life of our fixed assets. I'll provide a bit more detail in a moment, but on an aggregate basis, we estimate this will decrease full-year 2025 operating income by approximately $400 million for the assets on our balance sheet as of December 31, 2024.

First, in Q4, we completed a useful life study for our servers and networking equipment and observed an increased pace of technology development, particularly in the area of artificial intelligence and machine learning. As a result, we're decreasing the useful life for a subset of our servers and networking equipment from six years to five years, beginning in January 2025. We anticipate this will decrease full-year 2025 operating income by approximately $700 million.

In addition, we also early retired a subset of our servers and network equipment, we recorded a Q4 2024 expense of approximately $920 million from accelerated depreciation and related charges, and expect this will also decrease full-year 2025 operating income by approximately $600 million. Both of these server and network equipment useful life changes primarily impact our AWS segment.

Lastly, we also completed a useful life study for certain types of heavy equipment used in our fulfillment centers and are increasing the useful life from 10 years to 13 years beginning in January 2025. We anticipate this will increase full-year 2025 operating income by approximately $900 million.

As we turn the page to 2025, we're energized by the great work our teams have delivered. We remain focused on driving an even better customer experience and we believe putting customers first is the only reliable way to create lasting value for our shareholders.

With that, let's move on to your questions.

Question-and-Answer Session

Operator

At this time, we will now open the call up for questions. [Operator Instructions] Thank you. Our first question comes from the line of Mark Mahaney with Evercore ISI. Please proceed with your question.

Mark Mahaney

Thanks. Two quick questions. So, Brian, that's $100 billion CapEx we should think about in 2025. And then, Andy, were there any -- would you describe that AWS growth is being currently moderated down by supply constraints? Do you see those across the industry or do you see those materially impacting AWS today? Thank you very much.

Andy Jassy

So I'll take both of those. This is Andy. On the CapEx side, as Brian mentioned earlier, we spent $26.3 billion in CapEx in Q4. And I think that is reasonably representative of what you could expect in the annualized CapEx rate in 2025. The vast majority of that CapEx spend is on AI for AWS. It's -- the way the AWS business works and the way the cash cycle works is that the faster we grow, the more CapEx we end up spending because we have to procure data center and hardware and chips and networking gear ahead of when we're able to monetize it.

We don't procure it unless we see significant signals of demand. And so, when AWS is expanding its CapEx, particularly in what we think is one of these once-in-a-lifetime type of business opportunities like AI represents, I think it's actually quite a good sign medium-to-long term for the AWS business. And I actually think that spending this capital to pursue this opportunity, which from our perspective, we think virtually every application that we know of today is going to be reinvented with AI inside of it and with inference being a core building block just like compute and storage and database.

If you believe that, plus that altogether new experiences that we've only dreamed about are going to actually be available to us with AI. AI represents for sure the biggest opportunity since cloud and probably the biggest technology shift and opportunity in business since the Internet. And so, I think that both our business, our customers and shareholders will be happy medium-to-long term that we're pursuing the capital opportunity and the business opportunity in AI.

We also have CapEx that we're spending this year in our Stores business really with an aim towards trying to continue to improve the delivery speed and our cost to serve. And so you'll see us expanding the number of same-day facilities from where we are right now. You'll also see us expand the number of delivery stations that we have in rural areas, so we can get items to people who live in rural areas much more quickly. And then a pretty significant investment as well on robotics and automation, so we can take our cost to serve down and continue to improve our productivity.

So that's the CapEx piece. I think the second question you asked, Mark, is really around AWS growth and whether this is being moderated down at all by supply chain constraints. It is hard to complain when you have a multi-billion dollar annualized revenue run rate business in AI like we do and it's growing triple digit percentage year-over-year. It's hard to complain. However, it is true that we could be growing faster if not for some of the constraints on capacity. And they come in the form of -- I would say, chips from our third-party partners come in a little bit slower than before with a lot of midstream changes to take a little bit of time to get the hardware actually yielding the percentage healthy and high-quality servers we expect. It comes with our own big new launch of our own hardware and our own ships in Trainium2, which we just went to general availability at re:Invent but the majority of the volume is coming in really over the next couple of quarters, the next few months.

It comes in the form of power constraints where I think the world is still constrained on power from where I think we all believe we could serve customers if we were unconstrained. There are some components in the supply chain like motherboards too that are a little bit short in supply for various types of servers. So I think the team has done a really good job scrapping and providing capacity for our customers so they can grow. We're still growing at a pretty reasonable clip, as I mentioned earlier, but I do think we could be growing faster if we were unconstrained. I predict those constraints really start to relax in the second half of 2025. And as I said, I think we could be growing faster even though we're growing a pretty good clip today.

Operator

And our next question is from Eric Sheridan with Goldman Sachs. Please proceed.

Eric Sheridan

Thanks so much for taking the question. I'll just ask one that's building on Mark's questions there. Andy, when you think about the news that came out of China over the last couple of weeks and think longer-term about bending the cost curve lower with AI. I understood the commentary around CapEx for 2025. But when you look at where you sit in the industry, the move towards open source elements of custom silicon, how do you think about bending the cost curve and either speeding up or amplifying time deployment to market or possibly higher returns on capital for AI? Thanks so much.

Andy Jassy

Yeah. Well, I'd say a few things because there are a few questions built into that. First of all, I think like many others, we were impressed with what DeepSeek has done. And I think in part, impressed with some of the training techniques, primarily in flipping the sequencing of reinforcement training, reinforcement learning being earlier and without the human in the loop. We thought that was interesting ahead of the supervised fine-tuning.

We also thought some of the inference optimizations they did were also quite interesting. For those of us who are building frontier models, we're all working on the same types of things and we're all learning from one another. I think you have seen and we'll continue to see a lot of leapfrogging between us. There is a lot of innovation to come.

And I think if you run a business like AWS and you have a core belief like we do that virtually all the big generative AI apps are going to use multiple model types and different customers are going to use different models for different types of workloads. You're going to provide as many leading frontier models as possible for customers to choose from. And that's what we've done with services like Amazon Bedrock. And it's why we moved so quickly to make sure that DeepSeek was available both in Bedrock and in SageMaker, faster than you saw from others and we already have customers starting to experiment with that.

I think what's -- one of the interesting things over the last couple of weeks is sometimes people make the assumptions that if you're able to decrease the cost of any type of technology component, in this case, we're really talking about inference that somehow it's going to lead to less total spend in technology. And we just -- we have never seen that to be the case. We did the same thing in the cloud where we launched AWS in 2006, where we offered S3 object storage for $0.15 a gigabyte and compute for $0.10 an hour, which of course is much lower now many years later.

People thought that people would spend a lot less money on infrastructure technology. What happens is companies will spend a lot less per unit of infrastructure and that is very, very useful for their businesses, but then they get excited about what else they could build that they always thought was cost-prohibitive before and they usually end up spending a lot more in total on technology once you make the per unit cost less.

And I think that is very much what's going to happen here in AI, which is the cost of inference will substantially come down. What you heard the last couple of weeks that a DeepSeek is a piece of it, but everybody is working on this. I believe the cost of inference will meaningfully come down. I think it will make it much easier for companies to be able to infuse all their applications with inference and with generative AI. And I think it's going to -- if you run a business like we do where we want to make it as easy as possible for customers to be successful building customer experiences on top of our various infrastructure services, the cost of inference coming down is going to be very positive for customers and for our business.

Operator

And the next question is from Doug Anmuth with J.P. Morgan. Please proceed.

Doug Anmuth

Thanks for taking the questions. I'll stick with AWS to start. Just, Brian, maybe you can talk a little bit more about margins there just given that they've kind of moved between the mid-20s to high-30s over the past two years. How should we think about that more normalized, especially as you're investing that much more in generative AI? And then just on the Store side, can you talk about the impact of less volume going through your shipping partner UPS going forward? And are you able to manage that incremental shipping that's required? Thanks.

Brian Olsavsky

Yeah, sure, Doug. Thanks for your question. First on AWS, yeah, we have seen a lot of fluctuation in operating margin AWS. And we've said historically that they will be lumpy, as you say, over time. And the stage we're in right now, AI is still early stage. It does come originally with lower margins and a heavy investment load as we've talked about. And in the short-term, over time, that should have a -- be a headwind on margins. But over the long-term, we feel the margins will be comparable in non-AI business as well. So we're very pleased with the strong growth, focus on driving efficiencies in all of our data centers, saving power, reusing power in new generative AI applications and just generally reducing costs. So very pleased with the performance of the AWS team and look forward to strong 2025.

Andy Jassy

I'll take the UPS one, which is really the -- UPS has been a partner of ours for many years and we expect that we'll continue to be partners with UPS for many years. As you know, increasingly over the last several years, particularly accelerated by the pandemic. We have shipped a much larger percentage of our shipments through our own logistics network, our own last mile transportation network. And yes, I think that's in part because we needed to scale up so fast in the pandemic with everything being shut down and needing to serve more of the total market segment share of retail units during that time and needing to do it at a low-cost structure because, of course, our customers expect low prices and that's the nature of the business.

I think UPS has decided that serving Amazon is a lower margin for them. And so I think they've walked away from some of the volume that they otherwise could have had in the partnership. We're able to handle it with our own logistics capability, and we'll see how it continues to evolve.

Operator

And the next question comes from the line of Brian Nowak with Morgan Stanley. Please proceed.

Brian Nowak

Thanks for taking my questions. Andy, maybe to drill a little bit into the robotics acceleration that you talked about. Any new data points that you can share on learnings from Shreveport and how do we think about sort of the scalability of savings or the timing that we could see a real impact on profitability from robotics?

And then maybe just a bigger picture, Gen AI, GPU-enabled changes, you've -- other examples of how you see the Amazon retail shopping experience changing throughout 2025 as you're better using Gen AI or GPU-enabled machine learning and things?

Andy Jassy

Yes. Okay. Well, on the robotics piece, what I would tell you is, since we've been pretty substantially integrating robotics into our fulfillment network over the last many years, we have seen cost-savings and we've seen productivity improvements and we've seen safety improvements. And so, we have already gotten a significant amount of value out of our robotics innovations. What we've seen recently, and I think maybe part of what you're referencing in Shreveport, is that the next tranche of robotics initiatives have started hitting production. And we've put them all together for the first time as part of an experience in our Shreveport facility.

And we are very, very encouraged by what we're seeing there, by both the speed improvements that we're seeing, the productivity improvements, the cost to serve improvements. It's still relatively early days, and these all being put together only in Shreveport at this point, but we have plans now to start to expand that and roll that out to a number of other facilities in the network, some of which will be our new facilities and others of which will retrofit existing facilities to be able to use those same robotics innovations.

I'll also tell you that this group of, call it, a half dozen or so new initiatives is not close to the end of what we think is possible with respect to being able to use robotics to improve the productivity, cost to serve, and safety in our fulfillment network. And we have kind of the next wave that we're starting to work on now. But I think this will be a many-year effort as we continue to tune different parts of our fulfillment network where we can use robotics. And we actually don't think there are that many things that we can't improve the experience with robotics.

On your other question, which is, I think, really about how we might use AI in other areas of the business, in AWS, maybe more. And I think you asked about our retail business. The way I would think about it is that there's kind of two macro buckets of how we see people, both ourselves inside Amazon, as well as other companies using AWS, how we see them getting value out of AI today.

The first macro bucket, I would say, is really around productivity and cost savings. And in many ways, this is the lowest-hanging fruit in AI, and you see that all over the place in our retail business. For instance, if you look at customer service and you look at the chatbot that we've built, we completely rearchitected it with Generative AI. It's delivering -- it already had pretty high satisfaction. It's delivering 500 basis points better satisfaction from customers with the new Generative AI-infused chatbot.

If you look at our millions of third-party selling partners, one of their biggest pain points is because we put a high premium on really organizing our marketplace so that it's easy to find things, there's a bunch of different fields you have to fill out when you're creating a new product detail page, but we've built a Generative AI application for them where they can either fill in just a couple of lines of text or take a picture of an image or point to a URL, and the Generative AI app will fill in most of the rest of the information they have to fill out, which speeds are getting selection on the website and easier for sellers.

If you look at how we do inventory management and trying to understand what inventory we need and what facility at what time, the Generative AI applications we've built there have led to 10% better forecasting on our part and 20% better regional predictions. In our robotics, we were just talking about the brains in a lot of those robotics, or Generative AI-infused to do things like tell the robotic claw, what's in a bin, what it should pick up, how it should move in, where it should place it in the other bin that it's filling next to. So it's really in the brains of most of our robotics.

So we have a number of very significant, I'll call it, productivity and cost savings efforts in our retail business. They're using Generative AI. And again, it's just a fraction of what we have going. I'd say the other big macro bucket are really altogether new experiences. And again, you see lots of those in our Retail Business, ranging from Rufus, which is our AI-infused shopping assistant, which continues to grow very significantly, to things like Amazon Lens, where you can take a picture of a product that's in front of you, check it out in the app, you can find it in the little box at the top.

You take a picture of an item in front of you, and it uses computer vision and Generative AI to pull up the exact item in a search result; to things like sizing, where we basically have taken the catalogs of all these different clothing manufacturers and then compare them against once another -- one another. So we know which brands tend to run big or small relative to each other. So when you come to buy a pair of shoes, for instance, it can recommend what size you need.

To even what we're doing in Thursday Night Football, where we're using Generative AI for really inventive features like defensive alerts, where we predict which players is going to blitz the quarterback or defensive vulnerabilities where we were able to show viewers what area of the field is vulnerable. So we're using it really all over our retail business and all the businesses in which we're in. We've got about 1,000 different Generative AI applications we've either built or in the process of building right now.

Operator

And the next question comes from the line of John Blackledge with TD Cowen. Please proceed.

John Blackledge

Great. Thanks. Could you talk about the current speed of delivery? Maybe how much more room to go there and how is it driving the everyday essentials business? And then somewhat relatedly, any further color on inbound network efficiencies you would expect to see this year as you guys try to continue to lower the cost to serve? Thank you.

Andy Jassy

Yes. I would say on speed of delivery that we measure this very carefully and we measure both, what the conversion rate is of somebody who views a product detail page with a faster delivery promise versus those that are slower, as well as what we see downstream from customers once they bought with a fast promise and what they end up buying throughout the year.

And we have not yet seen diminishing returns and being able to continue to improve the speed of delivery. It doesn't mean that there won't be instances in which people are happy to take products later. We have a program where if people want to pick a day during the week where they want to combine a bunch of their shipments and have it delivered, then to be more sustainable and more environmentally friendly, they can. And we have plenty of customers who choose that, but we time in and time out, see that people choose to buy from us more frequently, and we're able to deliver to their homes or wherever they are much more quickly. And it leads to actually using us for more of their everyday purchases when we can deliver more quickly.

And I think that if you look at what we're doing with Prime Air, the promise there is for a number of items that we'll be able to deliver items to customers inside an hour. And I think when you're ordering everyday essentials where you need something more quickly, it's a big deal. And you see it, it's had a big impact on our everyday essentials. It's had a big impact on our pharmacy business, where people are able to get items same day now in lots of cities throughout the US And they're just using us much more frequently than they had before.

On the inbound network efficiencies, what I would tell you is we've made a pretty significant architectural change in our inbound network, that we've been working on for the better part of the year that we rolled out just a few months ago. Again, it's -- what we find when we make big architectural changes like this is that you tend to get some low-hanging fruit efficiencies early, but then there's all sorts of tuning and refinement you have to do once you actually see it working live across the really vast network. And we have all sorts of ways here that, where I think it's early, and I think we're going to get additional efficiencies throughout the year. But I expect that we'll have opportunities to keep taking our cost to serve down this year and that will be a big part of it.

Operator

And our final question comes from the line of Michael Morton with MoffettNathanson. Please proceed.

Michael Morton

Hi, thank you so much for the question. I wanted to follow up on Andy's remarks about how you're using AI within the Amazon e-commerce experience. But I wanted to talk about the other side of the coin, and that's really the e-commerce discovery process that leads people to Amazon. There's a lot of companies rolling out agents and assistants, and I would love to hear how Amazon is planning for potential disruption in this funnel and what the plans are. You spoke about Rufus, maybe make that more prominent. But what do you think the changes coming to the e-commerce funnel over the next several years would be great?

Andy Jassy

Well, I would say that, I think retailers ourselves and probably lots of other retailers are all going to have their own say on how they want to interact with agents. I think that it's an emerging space. And if you think about the investments that we've made in our fulfillment network, in our website, in all the selection that we've built, how we organize it, most retailers are going to have kind of terms in which they're going to interact with agents and will be no different that way.

And I think that -- I do think that Rufus, if you look at what it -- how it impacts the customer experience and if you actually use it month-to-month, it continues to get better and better. It's already -- if you're buying something and you're on a product detail page, our product detail pages provide so much information that sometimes it's hard if you're trying to find something quickly to scroll through and figure and find that little piece of information. And so we have so many customers now who just use Rufus to help them find a quick fact about a product. They also use Rufus to figure out how to summarize customer reviews, so they have to read 100 customer reviews to get a sense of what people think about that product.

If you look at the personalization, really most prominently today, your ability to go into Rufus and ask what's happened to an order or what did I just order or can you pull up for me this item that I ordered two months ago. The personalization keeps getting much better. And so, we expect throughout 2025 that the number of occasions where you're not sure what you want to buy and you want help from Rufus are going to continue to increase and be more and more helpful to customers.

Dave Fildes

Thank you for joining us on the call today for your questions. A replay will be available on the Investor Relations website for at least three months. We appreciate your interest in Amazon and look forward to talking with you again next quarter.

Operator

And ladies and gentlemen, that does conclude today's teleconference. You may disconnect your lines at this time. Thank you for your participation.