Blueshift RESEARCH REPORT

Seth Agulnick, sa@blueshiftideas.com

Nvidia Has an Inside Track on the AI Inferencing Market

Companies: AMD, AMZN, AVGO, FB, GOOG/GOOGL, IBM, INTC, KRX:005930, MSFT, NVDA, TSM, XLNX

February 26, 2020

Report Type: ☐ Initial Coverage ☑ Previously Covered ☑ Full Report ☐ Update Rating: 3.5/5

Research Question:

How big will the market be for AI inferencing chips? How large a share can Nvidia capture?

Summary of Findings

- Nvidia Corp. (NVDA) has carved out an early lead in the emerging market for artificial intelligence (AI) inferencing chips and its extensive software ecosystem puts it in prime position to reap the benefits.
- The opportunity in inferencing chips is huge and likely to dwarf that of training chips in terms of volume, driven by the evolution of natural language processing (NLP). Buyers for now will mainly be hyperscale data center operators like Amazon.com Inc. (AMZN), Alphabet Inc.'s (GOOG/GOOGL) Google, and Microsoft Corp. (MSFT), but their appetite for inferencing chips will be significant.
- The biggest risk to Nvidia's growth in AI inferencing may be hyperscalers using custom chips—both internally and for customer cloud operations. One source also suggested that CPU-based systems-on-a-chip (SoCs) from Intel Corp. (INTC) and others will be good enough for a majority of the inferencing workload. Longer term, Intel and even Advanced Micro Devices Inc. (AMD) could cut into Nvidia's stranglehold on the GPU market.
- Nvidia's CUDA software ecosystem remains a major competitive advantage, especially compared to alternative technologies like field-programmable gate arrays (FPGAs) from the likes of Xilinx Inc. (XLNX), which lack easy connections to deep learning libraries.
- Increased demand for inferencing in edge applications—such as mobile phones and smart appliances—will limit Nvidia's share of the overall market, as such applications require more efficient chips than Nvidia's powerful GPUs. Nvidia's best opportunity in the edge market is with 5G wireless networks, one source said.

	Outlook for Inferencing Market	Nvidia's Outlook in Inferencing	CUDA's Competitive Advantage
Industry Specialists	↑	→	↑
Data Center Executives	↑	→	↑
Software Engineers	↑	→	↑
Hardware Manufacturers	↑	→	↑

Silo Summaries

1) Industry Specialists

All three sources in this silo **expect major growth in sales of AI inferencing chips**. Hyperscalers are already shifting to inferencing from training to handle tasks like natural language processing, and edge applications will drive further growth. The volume of chips needed for inferencing will be much bigger than that for training chips. **Nvidia is in a good position in the early days of inferencing** but faces big challenges as the technology evolves. **Hyperscalers may opt for custom-built chips and could also choose CPU-based SoCs** rather than discrete processors from Nvidia. Intel is already a factor in inferencing and could gain share from its acquisition of Habana. All three sources agreed that **CUDA is a major competitive advantage**.

2) Data Center Executives

Nvidia has a great opportunity to thrive in the inferencing market. **Nvidia has an advantage with the developer ecosystem around its platform**. GPUs are well-suited to AI tasks but hyperscalers could opt for custom chips. Intel could make gains in inferencing.

3) Software Engineers

Demand for AI inferencing is going to rise dramatically. Nvidia faces many challengers in the inferencing market, including AMD and Xilinx in the short term and Intel and Google in the longer term. Its developer ecosystem offers a big advantage.

4) Hardware Manufacturers

Demand for inferencing chips is starting to ramp up and should continue to do so. Ultimately, the number of chips needed for inferencing will far surpass those needed for training. **Hyperscalers will be the main buyers for now, mainly to serve complicated tasks like natural language processing**. At the edge, cheaper and more narrowly focused ASICs will be more popular than Nvidia's powerful GPUs. It will be a while before a winner can emerge in inferencing chip technology because the market is still nascent and **hyperscalers are testing every possibility**. CUDA is a big advantage for Nvidia, but companies like Google and Facebook remain committed to developing alternatives.

Background

Nvidia's data center business got back on track in Q4 after several consecutive disappointing quarters. Rebounding demand for its GPUs from hyperscale data center operators helped Nvidia rocket past Wall Street estimates. Revenue in the segment hit $968 million in the quarter, up 33% from the prior quarter and 43% year over year. Company executives forecast another sequential increase in the current quarter.

Nvidia's optimism around data center growth is partly connected to the emerging market of artificial intelligence inferencing. Nvidia already has a dominant position in chips used for AI training, which involves teaching computer systems to learn without being programmed. Inferencing, on the other hand, centers on using those trained machines to make predictions and recommendations. The inferencing market is still largely untapped but has several powerful catalysts, including the growth of conversational AI. Such human-computer interactions require a chip to process a chain of tasks—from removing noise to speech recognition to language understanding—in a matter of milliseconds. Nvidia's CEO called the inferencing market, "one of the largest computer industry opportunities." In Q4, Nvidia said sales of its T4 GPUs, used for inferencing, had risen four-fold annually.

In the coming weeks, Nvidia is expected to unveil its next-generation GPU architecture, codenamed Ampere. Chips based on Ampere, the successor to Nvidia's Turing architecture, could launch as soon as this spring. Ampere chips, to be built by Taiwan Semiconductor Manufacturing Ltd. (TSM) and Samsung Electronics Co. Ltd. (KRX:005930) using a 7nm process, could offer significant performance and cost benefits—one report suggested as much as a 50% performance improvement at half the cost, compared to Turing's 12nm chip. The launch of Ampere will come as chief rivals like Intel and Xilinx continue to pour money into competing products. Intel, in addition to developing its own AI-focused GPUs, just spent $2 billion to acquire Israeli AI chipmaker Habana Labs. Xilinx's field-programmable gate array circuits, or FPGA chips, are an efficient and flexible alternative in the data center, while startups like Graphcore focus on customizable application-specific integrated circuits (ASICs) for AI uses. Xilinx said that its data center group revenue increased 8% year over year for its fiscal Q3, but was down 16% sequentially from a record quarter.

The software stacks that sit on top of data center chips are becoming increasingly important. Nvidia has nurtured a deep ecosystem of software developers using its CUDA framework, which allows chip buyers to customize their hardware in unique ways. Sources in Blueshift Research's June 28, 2018, report said CUDA [illegible]

Silos

1) Industry Specialists

All three sources in this silo expect major growth in sales of AI inferencing chips. Hyperscalers are already shifting to inferencing from training to handle tasks like natural language processing, and edge applications will drive further growth. The volume of chips needed for inferencing will be much bigger than that for training chips, two sources said, though one said overall revenue in the two markets could be similar because inferencing chips will be less expensive. Two sources were split on how quickly demand will move beyond hyperscalers, with one suggesting large enterprises will be buyers. Nvidia is in a good position in the early days of inferencing but faces big challenges as the technology evolves. Hyperscalers may opt for custom-built chips and could also choose CPU-based SoCs rather than discrete processors from Nvidia, according to one source. Intel is already a factor in inferencing and could gain share from its acquisition of Habana—whose technology was praised by all three sources—and the rollout of its Xe family of GPUs. One source predicted that CPUs will account for 50% of the inferencing market, followed by ASICs at 30%, leaving Nvidia to fight over the remaining 20% for GPUs. One source said Microsoft is a proponent of FPGAs in its Azure cloud operations, but another said FPGA makers like Xilinx will have a hard time expanding to new customers. All three sources agreed that CUDA is a major competitive advantage for Nvidia, one that AMD and Intel are likely to continue to struggle to overcome.

Key Silo Findings

AI Inferencing

- 3 of 3 said the AI inferencing market is growing significantly, driven by hyperscalers and edge applications.
  - 1 pegged the overall AI market for hardware and software at a combined $50 billion globally, with expectations for 40% annual growth.
- 2 said chip volumes for inferencing will far outpace those for training.
  - 1 said every data center machine is going to need an AI accelerator for training.
  - 1 said inferencing chips will have a much lower selling price than training chips, so the overall revenue in the two markets may be similar.
- 1 said large enterprises, not just hyperscalers, will drive demand for inferencing chips but 1 other said traditional enterprise data centers will be slow to add AI accelerators.
- 1 said a key risk for Nvidia is that CPU-based SoCs are good enough for a majority of inferencing workloads, meaning buyers will be less inclined to purchase discrete cards.
- 1 said hyperscalers are likely to use meaningful volumes of custom-built chips for inferencing, partly for technical reasons and partly to keep from becoming dependent on Nvidia.
- 1 said GPUs are great for processing images and sound but not for recommendation workloads.
  - That is why 80% of Facebook's inferencing is done on CPUs.
- 1 said Nvidia is the leader in inferencing—at least a year ahead of Intel and AMD in technology—and has a great growth opportunity.
- 1 said Nvidia has a strong position right now but may not be able to sustain it.
- 1 said inferencing is so new for Nvidia that it is hard to predict how successful the company will be, but its market expertise gives it a good chance to flourish.

Competition

- 1 said that by 2025 CPUs will account for 50% of inferencing demand, with ASICs garnering 30%, and GPUs and other architectures accounting for the remaining 20%.
- 1 said Microsoft is a big fan of FPGAs for inferencing.
- 1 said Intel is a threat to Nvidia on multiple levels: with its CPUs, the acquisition of Habana, and its upcoming Xe family of GPUs.
  - Intel already has a large share of the inferencing market.
- 3 praised Habana's chip technology.
  - 1 said it demonstrated performance advantages over other architectures.
- 1 dismissed Wave Computing Inc. as a key player because of financial, execution, and management turnover issues.
- 1 said Xilinx can grow with existing customers but it is hard to see where it can grab new ones.
- 1 said Graphcore has good technology but it is not clear where its market lies.
- 1 praised AMD's chips but said it lags behind Nvidia in how much silicon it has dedicated to machine learning.

- 1 said startups that are building processors specifically for inferencing may show performance advantages over Intel, Nvidia, and others, but will have trouble displacing the bigger players' relationships with big customers.
- 1 said GPUs are not cost-effective solutions for edge applications.

CUDA and Other Key Issues

- 3 said CUDA is a big, structural advantage for Nvidia.
  - 1 said AMD's failure to shift developers away from CUDA is a key reason that it will struggle to gain share. Intel may have the same problem with its competing platform, oneAPI.
- 1 said Ampere will provide good performance benefits, though not as big as some headline numbers might suggest.

1) Software and systems engineer with expertise in accelerators and high-performance computing; repeat source

The inferencing market will be much bigger than the training market and hyperscalers are already shifting to inferencing. Nvidia is going to have multiple challenges capturing the inferencing market. For one, GPUs are not the best solution for tasks like recommendation workloads. For another, hyperscalers are going to deploy a significant volume of custom-made accelerators rather than rely on Nvidia. Further, competitors like Intel are going to offer CPU-based SoCs that will be good enough to keep some hyperscalers from investing in discrete cards like Nvidia's. Traditional enterprise data centers will not be as aggressive as hyperscalers in deploying inferencing accelerators. CUDA remains a key advantage for Nvidia on the software side.

AI Inferencing

- "The inference market is much bigger relative to the training market. Every machine is going to have the inference accelerator. … The inference workload is way, way bigger [than training]."
- "The growth of the AI inference market is already happening. At the hyperscalers, there is already much more inference going on. Training is a big budget item that they like to talk about. There's more noise being made about the heroic training efforts, but if you look at what's actually getting deployed in terms of volume, it is shifting more toward inference."
- "The question is, who [among the hardware makers] is going to win? [Will inferencing accelerators be] integrated in the processor or is that an extra processor that's added to the machine? That's not entirely clear just yet."
- "For now, we're tending to see that some of the models only run well on the CPU. At Facebook, 80% of their workload is inference on the CPU. The custom accelerators can give you much better performance, but you have to figure out if it is worth the extra cost to have a discrete one or if there is something you can integrate in a future processor that's good enough, even if it is not the best you can possibly do."
- "We see Intel adding instructions specifically to accelerate inference on the CPU side, we see SoC vendors trying to put a core processor on the same chip, and then we see people with discrete accelerators. It's the discrete market that has the biggest questions about long term. I think that will tend to get absorbed for inference."

> **The growth of the AI inference market is already happening. … if you look at what's actually getting deployed in terms of volume, it is shifting more toward inference.**
>
> *Software and systems engineer with expertise in accelerators and high-performance computing*

- "The flexibility and the programmability of [Nvidia's] GPUs will help. They'll continue to be popular. They're not going away any time soon for training and inference. But I think that, while they'll sell many thousands of units, and even tens of thousands, their main competition are their own customers. [CEO] Jensen [Huang] should realize that their big competitors right now are the ones they're actually selling the chip to. They're building them as well."
- "If you think about the case of every single machine getting an inference accelerator, the hyperscalers are going to do it themselves. We see Amazon who's doing their own CPUs. I think what's going to happen is a two-horse race between the hyperscalers doing it themselves and their microprocessor vendor—Intel or AMD or one of the ARM folks—managing to integrate enough of an accelerator into the SoC."
- "It's hard to predict who will win. I suspect that we'll end up split between Intel with their integrated stuff, the custom SoCs being done by the hyperscalers with integrated stuff, and then a mix of discrete cards as well. But the discrete cards will be lower volume. They'll only deploy them for workloads that you can't handle on the SoC."

- "In the not too distant future, every SoC will have inference acceleration on it. While there'll still be a need and market for discrete inference accelerators, it remains to be seen how much of the market it'll be. The danger for folks like Nvidia is that the CPU-based SoCs are good enough for large parts, or even the majority."
- "It'll be most interesting to watch the evolution of Amazon's strategy. They have Inferentia as their ML [machine learning] chip. It's cleverly designed so that you can gang up several to reach very high performance standards, and be connected to [Intel-based] x86 chips. They also have their new high-end ARM-based server processors. I'm very curious to see if they integrate the Inferentia IP within their future ARM processors, or keep it discrete only."
- "The question, in a way, is how much will the hyperscalers shift to their own silicon? I don't think the large buyers will want to have Nvidia have a monopoly on this. They deeply wish for a competitor to Intel in CPUs. AMD is strong at the moment, but historically hasn't been a continued strong presence, so Intel tends to have an effective monopoly at times. [Hyperscalers] desperately do not want a company like Nvidia to do the same."
- "I think we'll see [hyperscalers] put in a lot of effort to making sure they are not dependent on Nvidia. That makes life hard for Nvidia."
- "On the public cloud side, we'll continue to see companies like Google buy enough GPUs because people want to use them in the public cloud. The question is, at what point do [public cloud operators] manage to shift the demand from the public cloud users away from Nvidia and towards their own solutions?"
- "We're seeing Google trying to push its TPU. People aren't buying into it entirely on the public cloud side because they don't want Google's hegemony and to be wedded to the TPU. But I think that's a transient thing. We'll see the large services deploy interesting enough software stacks that are specialized for video or images or text processing. They'll manage to draw a lot of the traffic away from Nvidia for inferencing on the public cloud. Nvidia is strong right now, but it's not obvious that they have sustainable position."

> **I think we'll see [hyperscalers] put in a lot of effort to making sure they are not dependent on Nvidia. That makes life hard for Nvidia.**
>
> *Software and systems engineer with expertise in accelerators and high-performance computing*

- "The open question is the traditional enterprise data center. I think some of that load will be picked up by the CPUs, because people like Intel will push their own accelerators—for instance, the Xe program—and, of course, Nvidia continues to sell in there."
- "But I don't expect the enterprise side to be as aggressive in rolling out inference accelerators everywhere. The main reason for that is not technical on the machine learning side or on the hardware side, but it's a case of it being less obvious what the killer app is in the enterprise."
- "Given that all those enterprises have the cloud at their disposal, if they have a bursty workload, and for experimental things, I can see them using a lot of the ML services that Microsoft, Google, and Amazon will provide. If you have images to process, you may as well use one of those services. So the question is, what is the killer vertical app for the enterprise that demands that they add accelerators to all their servers? And I don't know what that is yet."
- "On the data center side, at a very coarse grain, we can break [workloads] into two pieces. First are the perception-based workloads—processing image and sound. These are typically … well-optimized by GPUs for inference."
- "Second are the recommendation workloads, e.g., what stories to show on Facebook, what movies, etc. These are typically not well optimized by GPUs at all. So, depending on the customer, usefulness of GPU for inference varies wildly."
- "A really big example would be Facebook. Around 80% of their AI inference runs on CPU, not GPU. That's because they're dominated by recommendation needs. Of course, they do some work on images, too, but it's not nearly as big."
- "I think Nvidia has a story for inference in perception, and we can see the T4 GPU selling well. But it's not as widely applicable as Jensen might hope."

Competition

- "Are there threats to Nvidia on the perception side? Yes, that's where a lot of AI accelerators are targeting. While there are too many chip startups, and we're seeing the first zombies and consolidation going in, there are also emerging successes."
- "Still early days, but there's the obvious example of Habana—a large acquisition [by Intel] partly driven by [Habana's] success with customers and partly by [Intel's] Nervana failure, Graphcore, and so on. I think we'll also see Intel compete better on the GPU side with their Xe family."

- "[Nvidia] can do OK, but we'll see people deploy custom inferencing accelerators. Certainly, people deploying at hyperscales are motivated to build their own inference processors—whether that's Microsoft or Amazon or Google. Google led the way with the initial TPU."
- "Microsoft loves FPGAs. They're still buying them. They tend to do soft accelerators; their equivalent of the TPU is done inside another vendor's FPGA. We'll continue to see that. We'll also continue to see automotive companies continue to use FPGAs because of hard, real-time constraints, where you have to be very predictable. When you have to make functional safety people happy, FPGAs tend to be useful. We'll see inference workloads there in the FPGAs."
- "I think Intel will carve out a fairly large percentage of market share in inferencing. For many people's workloads, they're better suited, mainly because of memory capacity. For recommendation types of problems, the CPU is the best possible thing to use just for memory capacity problems. The data access patterns are very unkind to a GPU."
- "I think Intel is already a surprisingly large percentage [of inferencing], even if it's not visible, and they'll continue to be. I think what we'll see publicly play out is [Intel's] Xe products pull some of the Nvidia market. I think [Intel] is going to start competing in the next couple of years for some of the stuff that maps well to the GPU with their new GPUs. That will be straight GPU vs. GPU. Historically, they haven't done well with GPUs, but there are reasons to think that this time they'll have better technology at least. We'll have to wait and see what happens in the market. That will be a source of pain for Nvidia."
- "Xe is an IT program built out independently of Habana. It's going to be interesting because they're going to have two very credible parts. It's tricky to know right now what [Intel's] strategy is. Habana is a really good technology. Nervana was a disappointment. Habana is actually credible and getting customer support. What remains to be seen is if they try to converge the Habana and Xe lines or whether they keep them separate."

> **[Intel] is going to start competing in the next couple of years for some of the stuff that maps well to the GPU. … That will be a source of pain for Nvidia.**
>
> *Software and systems engineer with expertise in accelerators and high-performance computing*

- "I don't think Wave is important. They've got execution problems and financial problems. They have high executive turnover. They're not likely to be a factor long term."
- "Xilinx, I think, is going to have some good winds. However … it's not clear where their growth is coming from in terms of breaking out towards new customers [beyond Microsoft]. I think we'll see organic growth within their existing customers, and they can do a better job on that."
- "Graphcore's technology is good. It's on par with Habana, so the question becomes where their market is. They've done a deal with Microsoft to start deploying a little in the public cloud. But the consolidation has started with Intel picking up Habana. Given the amount of money Graphcore has raised, there's a limited number of people who can acquire them, so it's about trying to figure out who's actually going to be the customer and who's going to be the acquirer."
- "The technology is actually good, but they're running slightly late and they didn't execute as well as Habana. They're shipping hardware, but their software stack is a little behind. They could be stuck trying to raise money, but they've raised quite a bit already and that narrows their options."
- "AMD is executing super well on the CPU side. On the GPU side, they've been doing quite well with things like Navi and onwards. If you look at the raw performance of the hardware, it's actually pretty good relative to Nvidia. I think Nvidia will be a little bit ahead."
- "The other big problem with AMD for GPUs is they haven't put in tensor cores yet. They haven't dedicated as much silicon as Nvidia is willing to dedicate to machine learning."

CUDA and Other Key Issues

- "[AMD's] biggest problem of all is their software stack. They've been historically weak on this and they're continuing to be weak. They've failed to really shift people who are using CUDA. They had various failed efforts to try to get people off CUDA. That never worked."
- "I think Intel is going to have the same kind of problem with oneAPI, which is Intel's attempt to compete with CUDA. If you look at the machine learning stacks, like TensorFlow and PyTorch, AMD doesn't seem to be putting enough effort in to optimize those and support them well."
- "I'm actually a fan of AMD for a couple of reasons on a technical level, but I wouldn't use AMD right now because of the software reasons. I think they'll execute fine in hardware, but the software side will let them down and they won't get to penetration."

- "Nvidia's Ampere is basically the next generation from the V100. It's their heroic training chip. It should basically double, in terms of performance, compared to V100. Relative to V100, A100 will be twice the capacity. The problem is there's a very big gap between the theoretical performance of the V100 and the delivered performance in machine learning."
- "Nvidia markets the big headline number, but the practical number is actually much, much lower. I think the same will be true on Ampere. There's no doubt they will have fine-tuned some things and they will have good benchmarks but, in practice, people will discover the delivered performance and practice numbers being much lower. That's true for a lot of hardware. It's not just an Nvidia problem, but it's particularly noticeable on these heroic training chips that there's a large gap."
- "I don't think Ampere will be a factor in inferencing. They'll have this as a replacement for T4 and the T4 is a pretty popular inference accelerator. If you're using Nvidia cards, T4 is a good one to buy for inferencing."
- "I think there will be a successor to T4 that has better performance, but it will still have the same structural problem. If you're the SVP of infrastructure at Google, you're going to look at it and say you'll [buy] some small number, maybe something like 10,000, but if they want to think about their million-parts order, they'll think twice before deciding to give it to Nvidia or not."

Interview summary from June 28, 2018, report: New workloads for machine learning are driving growth for Nvidia's GPUs. However, although AI training tasks work well on GPUs, inferencing does not. Even Intel's Xeon CPU has better performance for certain inference workloads than Nvidia's GPUs. Large cloud providers like Google and Amazon are going to drive demand for GPUs, while enterprises are more likely to use cloud services for machine learning rather than buying the hardware themselves. Another major challenge for Nvidia is that the major cloud operators all have active projects in the field, and the chips are not that complicated. It will be relatively easy to compete as a vendor, whether a large chip vendor like Intel or AMD or a smaller company such as Graphcore and Cerebras Systems. CUDA is still a significant advantage for Nvidia but is rapidly eroding in the machine learning era.

2) Industry veteran and senior AI executive

The market for AI inferencing chips will far exceed the training market in terms of volume but is likely to be on par in revenue because of lower selling prices. Nvidia is still fairly new in the AI inferencing market but should not be discounted as a strong competitor because of its expertise on the training side. The CUDA ecosystem gives Nvidia a huge advantage in both training and inferencing. Many competitors are still in the proof-of-concept stage. Chip performance is not the only predictor of success; existing penetration with large enterprises is also important. Intel's Habana acquisition stands out as a major development because Habana is showing advantages over other architectures.

AI Inferencing

- "It's a combination of data and the proliferation of IoT [Internet of Things] and edge devices, as well as the technology, [that is driving demand for inferencing]. The AI inferencing market has really taken off as IoT and edge-connected devices have become more prevalent. This is mainly fueled by the amount of data out there available at the edge."
- "The technology is also a driver. The technology is becoming much more powerful within the same cost and power-consumption envelope. You can do more in compute at the edge and that's become an enabler."
- "Enterprises can do inferencing in the data center but a lot of it is done at the edge."
- "Devices like [Amazon's] Alexa are already fairly sophisticated. With the advent of increasing performance capabilities in those devices, we are going to be able to do much more. Conversational AI uses a different set of networks compared to computer vision AI. These are recurrent neural networks, LSTMs [long short-term memory], which look at the context of a sentence rather than a specific word. It's a bit more complex in terms of computational abilities."
- "We are getting more capabilities. In inferencing chips, we are going to be able to do more. We are somewhat meeting the standard requirements

> **The AI inferencing market has really taken off as IoT and edge-connected devices have become more prevalent. This is mainly fueled by the amount of data out there available at the edge.**
>
> *Industry veteran and senior AI executive*

today. As we all experience with Alexa, Alexa sometimes gets it wrong. But I expect it will continually get better."

- "Google's BERT is a big workload model for neural networks to process conversational AI. A lot of companies are looking at that model today so that when they build hardware, it will be able to address and support that model. It's the talk of the town. It's hard to tell, though, if it's a watershed development because so many things in AI are watershed developments."
- "There are multiple components in AI that can be watershed developments, like neural networks, hardware architecture, and the compiler technology. It would be dangerous to say one thing is more watershed than the other."
- "In terms of dollar value, I expect the inferencing market to be on par with the training market. Even though we're experiencing growth in inferencing chips in terms of volume, the ASP of training chips is higher because training chips have to do bigger performance calculations and deal with very complex algorithms. They tend to be very complex chips."
- "Over time, inferencing may not be as big as the training market in terms of dollars, but in terms of the number of chips, it will far exceed the training market. I think inferencing chips will be on par and even overtake training chips in the next two or three years."
- "I don't think demand will be just from hyperscalers. Large enterprises are developing their own inferencing solutions. Companies that are in the verticals are doing deployments. It's hard to tell how much of the demand will come from hyperscalers and how much from enterprises because the market is still evolving. Very often, the large enterprises are developing their own and not using hyperscalers."
- "Hyperscalers tend to use a mix of chips. In general, hyperscalers buy standard chips like CPUs and GPUs on the market, but they also create some of their own [custom chips] because of their high-performance demand. They also have the volume economics to justify [custom chips]. The drivers are performance plus cost. They will do customization for specific workloads."
- "The hyperscalers' appetite for inferencing chips is big. We are seeing public information that the hyperscalers are interested in inferencing."
- "Historically, Nvidia has been involved with training [and] it's been Intel CPUs that have done inferencing. But the inferencing market now has a lot of novel architectures and some of those have started to come to market."
- "Nvidia is new in the inferencing market. They introduced Tesla, one of their newer architectures that supports inferencing, but it's new for them, so it's hard to tell how successful they'll be in it. At the same time, it's also hard to discount them because they obviously know their market well."

Competition

- "In general, purpose-built [custom] architectures will always outperform. GPUs and even CPUs were never meant to be AI-accelerated processors. For example, GPUs used to be graphic processors and now they are adapted for inferencing, while CPUs are more general-purpose processors."
- "We're going to see companies—startups and others—that are building specific processors geared for inferencing that will probably do better in terms of performance. But bear in mind that it's not just performance that needs to be important but also their business model, their footprint with large enterprises in that market."
- "In other words, companies that already have an entrenched footprint with large customers will continue to do better even if they don't have the most optimized architecture. It would take a long time for newcomers to acquire those customers, even if they have a great and superior architecture."
- "In terms of competitors, there is a lot of great and innovative technology. It would be hard to make a bet at how they will pan out."
- "All those [competitor] companies have promises in terms of performance targets. A few of them have demonstrated them, like Habana. They can show the advantages over other existing architecture, but a lot of the [competitor products] are still in the early proof-of-concept stage in the inferencing market. We don't know how the performance will pan out in production."
- "Intel obviously acquired Habana for a good reason. Still, I would be very cautious even in providing a good qualitative comparison. It would be not be fair. It's still all new and uncertain."

We're going to see companies—startups and others—that are building specific processors geared for inferencing that will probably do better in terms of performance. But bear in mind that it's not just performance that needs to be important but also their business model, their footprint with large enterprises in that market.

Industry veteran and senior AI executive

CUDA and Other Key Issues

- "CUDA is a huge advantage for Nvidia because of the established ecosystem and they have an opportunity to continue it and expand on it, whether it's training or inferencing."

3) Nabeel Mahmood, technologist, futurist, board member, and advisor to CEOs and CIOs

Nvidia is the leader in GPUs for AI inferencing and should see double-digit revenue growth over the next five years. However, Intel and AMD are likely to cut into its lead on the GPU side. Further, GPUs are likely to account for less than 20% of the inferencing market by 2025, as they are not well suited for mobile, IoT, and other edge applications. CPUs and custom chips will be used for the lion's share of inferencing. CUDA is a big advantage for Nvidia and that gap could widen.

AI Inferencing

- "The demand [for inferencing chips] will primarily be from hyperscalers and edge. I would say hyperscalers and edge have an equal appetite for inferencing chips. Of course, it's all interconnected and entwined. The demand from both is comparable."
- "I believe Nvidia is the leader in AI inferencing GPUs, whereas Intel and AMD are a couple of years behind. Nvidia has had the opportunity to corner the market and capture initial success and we can expect them to gain double-digit revenue growth in AI inferencing chips over the next five years."
- "Of course, there's always brand loyalty and preference that we're going to see from the hyperscalers and edge guys. That will help Intel and AMD. In order for Nvidia to stay ahead of the game, they're going to have to be even better."
- "AI training is a huge growth driver for the data center platform. This involves training—teaching an artificial neural network how to make inferences from data like humans do—and applications—machines applying their training to new data."
- "Also helping to power growth [of inferencing] are high-performance computing and virtualized computing."
- "Autonomous vehicles, intelligent machines, supply chain modernization, intelligent services, and smart cities are also growth drivers."
- "Another driver is professional visualization platforms where you use AI to fill in the dots. You leverage AI to make a decision to generate the design."
- "Conversational AI converges three separate technologies: artificial intelligence, messaging apps—for example, Nextiva—and speech recognition, natural language processing. It is extremely difficult for chatbots and intelligent personal assistants to operate with human-level comprehension because of the inability to deploy extremely large AI models in real time. With using GPUs, we'll see two times the latency reduction and five times throughput improvement during inference."
- "Google's BERT is moving conversational AI forward a little. Conversational AI is still basically NLP. There is an impact from BERT, but there isn't enough data and the infrastructure is not ready for it yet. The impact is slow. NLP is being used in chatbots and messaging that a lot of companies use today. But they're unable to mine and map the data."
- "The AI market for both hardware and software is currently approximately $50 billion globally. I would expect it to grow by 40% CAGR [compound annual growth rate] over the next five years."
- "In the last decades, we've been trying to customize everything for everybody, hence the increased cost of performance and a lower ROI. We have this challenge in the whole infrastructure environment. If you customize everything for every single deployment and we do not leverage standards, it's going to take forever. I believe the players need to achieve a level of standardization to reduce the cost of manufacturing, the cost of deployment, and increase ROI. That's what's going to keep the shareholders happy."
- "We will need to have significant reduction in custom chip play and there needs to be a level of standardization. In general computing, a top-of-the-line computer that a decade ago cost $3,000, today costs about $300. We were able to achieve that not only through mass production but through standardization that resulted in a price drop."

It is extremely difficult for chatbots and intelligent personal assistants to operate with human-level comprehension because of the inability to deploy extremely large AI models in real time. With using GPUs, we'll see two times the latency reduction and five times throughput improvement during inference.

Nabeel Mahmood, technologist, futurist, board member, and advisor to CEOs and CIOs

- "For this market, too, there has to be a level of standardization and a reduction in cost of manufacturing, where it can be streamlined for the mass market. Otherwise we're going to come to a screeching halt."

Competition

- "CPUs will account for 50% of AI inferencing demand by 2025, with ASICs—custom chips designed for specific activities—at 30%. GPUs and other architectures will pick up the rest."
- "GPUs are not as cost-effective for automating inferencing within mobile, IoT, and other edge computing uses. Technologies including CPUs, ASICs, FPGAs, and various neural network processing units have performance, cost, and power-efficiency advantages over GPUs in many edge-based inferencing scenarios, such as autonomous vehicles and robotics."
- "If Nvidia can capture more than half of the GPU AI inferencing market share, Intel can probably be second, followed by AMD and others."
- "It's still premature to talk about Habana. Habana will get Intel to that second place. They're very secretive, even if they have a lot of initiatives going on. That's probably part of their play. But, other than Habana, we haven't seen much coming out of Intel."
- "The largest three have the market cornered in a lot of ways but we're still early in the process. They have capital, and good R&D, and they have the level of relationships where spending several billions of dollars is not that big a thing."
- "When you look at Wave or Graphcore, they are not the front runners. They're facing big players and they're going to run into capital issues. They could become targets of acquisition down the road. If market players come up with great technology, that's what happens. I believe the little players are going to be acquired and the larger players are going to leverage their innovation. Maybe Intel or AMD or Nvidia will make that play down the road. But I don't think anything major is going to happen for the next 18 to 24 months."

CUDA and Other Key Issues

- "CUDA is a significant advantage for Nvidia. If Intel and AMD don't bring new innovative platforms to the front in the next 24 months, that gap is going to get bigger and bigger. CUDA will become an even more significant advantage for Nvidia."
- "I believe Nvidia's Ampere can have an impact in inferencing."

2) Data Center Executives

Nvidia has a great opportunity to continue to thrive in the inferencing market, where growth will be driven by demand from hyperscalers and edge applications. The path to success for Nvidia is not an easy one, however, as competition will be fierce. Nvidia has an advantage with the developer ecosystem around its platform but will need to maintain that edge in inferencing. One source said the inferencing and training markets should grow proportionally, while another forecast a similar adoption curve with inferencing as we saw with the cloud. GPUs are well-suited to AI tasks while FPGAs lack the developer ecosystem. Intel could make gains in inferencing, though not necessarily because of Habana. It is not clear yet whether Ampere will deliver on rumored performance advantages or drive a buying cycle in the data center.

Key Silo Findings

AI Inferencing

- 1 of 3 said hyperscalers and edge applications will drive demand for inferencing chips.
- 2 said Nvidia is well-positioned to thrive in the inferencing market, though 1 of those said it will need to appeal to developers to maintain an edge over competitors.
- 1 said the training and inferencing markets will grow in tandem.
- 1 said companies with significant customer support operations will need natural language processing capabilities and GPUs are the best option for such workloads.
- 1 said the growth of the inferencing market will follow a similar curve to that of cloud adoption, where there was a very slow ramp followed by an enormous burst.
    - It could take as long as 10 years for the inferencing market to reach maturity.
- 1 said inferencing will be the domain of the biggest of the big, like Facebook and Google, for quite a while before it starts to trickle down.
- 1 said self-driving cars are likely to be the biggest driver of demand for inferencing.

Competition
- 1 said GPUs are a good fit for AI workloads while technologies like FPGAs lack the developer ecosystem that have sprouted around GPUs.
- 1 said AMD and others will have a hard time challenging Nvidia because of the code libraries around Nvidia's technology and the number of developers trained in it.
- 1 said CPUs could evolve as good fits for AI computation.
- 1 said Intel has an opportunity in AI inferencing, though its acquisition by itself of Habana may not be a big development.
- 1 said makers of ASICs and FPGAs will have a hard time growing sales outside of the big cloud operators.

CUDA and Other Key Issues
- 1 said it is possible Ampere could deliver major performance gains, but it is hard to know yet and, thus, hard to predict whether it will drive a buying cycle.

1) Cloud AI executive at a hyperscaler

The training and inferencing markets will grow proportionally, as models will need to get continually retrained. Demand for inferencing chips will be driven by the major cloud operators and by edge applications. Nvidia has a strong position in the inferencing market but its ability to fend off competition will come down to execution and success with the developer community. The latter is an area where GPUs have an advantage over other chip technologies like FPGAs.

AI Inferencing
- "[Nvidia] has a very strong position [in inferencing] and has a lot of opportunities to grow. There's strong competition, though, as well. It will depend on the execution, ability to adjust to the market, and success with the developer community."
- "[The emerging drivers for the development of new inferencing technology] are inference efficiency on devices [with] low power [and] no-accuracy-loss computations; the high cost of inferencing at scale in internet services and app backends; [the need to] bring the models where data is, making the models available to data owners; and the ability to control the lifecycle of models, easy to reprogram, monitor, and operate at scale."
- "I think these markets [training and inferencing] are proportional. Static models degrade over time and they get obsolete as data changes. True AI makes the apps adapt as environment changes and models need to get retrained. More AI usage will drive both training and inferencing."
- "The hyperscalers make the inferencing chips available to others via cloud services. They will drive a lot of demand as hubs for other services and apps. Another demand generator is IoT/phone/edge devices."
- "I expect [hyperscalers] will [deploy custom chips for inferencing]. There are huge benefits for creating those for internal workloads, and then those get offered publicly. The hyperscale cloud providers have big leverage with the scale and diversity of workloads, as well as amortized operational cost. They will compete on price and performance with GPU/CPU vendors, while running GPU/CPU as well."

> **[Nvidia] has a very strong position [in inferencing] and has a lot of opportunities to grow. There's strong competition, though, as well. It will depend on the execution, ability to adjust to the market, and success with the developer community.**
>
> *Cloud AI executive at a hyperscaler*

Competition
- "[GPUs will] not necessarily [be the chip of choice for inferencing]. The jury is still out. GPUs have the architecture to fit AI workloads well and a large developer ecosystem and tools. FPGAs lack the latter. CPUs can evolve to have specialized compartments for AI computation. To a large degree, it will be determined by the developer tools and ease of use."
- "The opportunity [for Intel] is still there. But I doubt a single startup acquisition [like Habana] can make a big difference. It is not the first in this kind of acquisition by Intel. It will require a strong technical and marketing strategy to leverage Intel's CPU market share to drive AI as well."
- "[Other players like Xilinx, Graphcore, Wave, and AMD] remain relatively niche."

- "The tools and ease of use, and adoption to users' workloads and ML frameworks, is what will determine the leaders. I am guessing the ASIC and FPGA makers outside of hyperscale clouds might have a challenge in scaling sales."
- "SoCs are not any bigger threat to GPUs as they are for CPUs. SoCs versus CPUs were around for many decades, and have their customers. The more general applicability of GPUs will make the tools more widely known and adopted."

CUDA and Other Key Issues

- "It's hard to tell [whether Ampere might drive a buying cycle] without seeing the benchmark numbers. It is possible [that Ampere can deliver 50% better performance at half the cost], but we still need to see the representative benchmarks."

2) VP of engineering for an online database firm

GPUs are the best choice for inferencing in data centers. Nvidia will be difficult to beat, not just because of its hardware but because of the training and software ecosystem around its GPUs. Demand for GPUs among enterprises will grow—slowly in the next year or two as large companies test smaller systems and then very quickly after that. Large companies require thousands of GPUs each for natural language processing. Amazon's cloud will be a big driver of GPU demand.

AI Inferencing

- "I believe most tasks in most companies will be solvable by generic GPUs, like Nvidia's. I don't see much demand for custom chips."
- "I feel there will be a big demand for GPUs for natural language processing over the next two to three years, and I don't think other companies will be able to compete with Nvidia."
- "It's proven that GPUs give significant improvement for running very large neural networks, which are used everywhere now for natural language processing. This is not yet widely adopted in industry, but it will be more and more adopted and will create demand for GPUs."
- "Typically, the best approach [to NLP] is to solve this problem with very large neural networks like BERT, which has proven very effective. It takes a long time to train [a system], but companies can do it now, and it has proven to be much more effective than many traditional technologies, like Conditional Random Fields."
- "[BERT] provides a very significant improvement for inference time and functionality on GPUs, where efficiency is very important. Nvidia very recently improved their CUDA stack and the inference times are significantly improved."
- "I think it will take three to five years to get to the point where 30% to 50% of companies will have really big NLP stacks. A big chunk of companies will have these stacks for customer support. And I think the majority of them will move from traditional NLP to running large neural networks, and I think this will [create demand for] GPUs."
- "I think, over a five-year period, I believe 50% of companies will have an NLP stack—say 50% of them will run BERT and a large percentage of those will run on GPUs. So maybe like 20% of companies will need to have GPUs to run their stack. That would be a good estimate."
- "I don't think [the need for NLP] is going to create demand for Nvidia's most recent GPUs [like Ampere]. Many companies that don't have a GPU stack now will probably just go for [Nvidia's] T4 chip."
- "For this type of processing, many companies will move from their own metal stack to Amazon. It will be very important as to what chips are actually available from AWS. That will be an important driver for what chips will be used and what tech will be used."
- "Many companies are moving over to AWS, and the decision about what chips will be in demand are what chips will be available [at AWS]. I know that Amazon makes Nvidia available, and I think Amazon gives them priority. I believe there are more Nvidia chips on Amazon than any other."
- "Assuming it's a big company in retail or an airline, a Fortune 500 with customer support applications from 100,000 to a million or several million minutes of calls—thousands of customer calls simultaneously—to process this you will need several thousand GPUs [per company]."
- "I think, over five years, a big chunk of Fortune 500 companies will require 1,000 or thousands of GPUs to enable these kinds of applications."

> **It's going to be hard for anybody to compete with Nvidia—not just because their GPU is better, but because they've managed to create a whole infrastructure.**
>
> *VP of engineering for an online database firm*

- "Typically, you start by building certain prototypes for a smaller set of clients, and this might take more like 100 GPUs and about one or two years to build and check if it works."

Competition

- "AMD has been trying to get into this market, but I think it will be very hard to fight against Nvidia, because Nvidia created a very strong support [system] with the number of people trained on how to use Nvidia GPUs."
- "It's going to be hard for anybody to compete with Nvidia—not just because their GPU is better, but because they've managed to create a whole infrastructure. If you want to build inference [technology], or NLP for your applications, currently most of the libraries that are available have versions running on Nvidia GPUs; you have people who know how to run it on Nvidia GPUs; you have increasing demand because almost every big company is building one or another version of a conversational AI/NLP stack. Most libraries will get significant improvement in performance if you use Nvidia GPU chips."
- "Even if [SoCs] will be [a threat to Nvidia], it will take a long time to get support such as libraries and developers to be widely used. So they are not a threat in near future."

CUDA and Other Key Issues

- Did not discuss.

3) Data center infrastructure executive for a foreign currency trading firm

Growth of the AI inferencing market will follow a similar pattern to that of cloud adoption—a slow build followed by a huge burst over about a decade. Autonomous vehicles are the application most likely to be the big driver of inferencing adoption.

AI Inferencing

- "If you look at the trend in cloud [adoption], it was a very slow adoption rate at first and then there was a huge ramp about four years ago. It's still climbing a little, but not much."
- "As far as adoption [of AI inferencing] is concerned, you're going to see that same adoption schedule as you saw in the cloud. It'll be a 10-year ramp-up, because you're developing skill sets around it. It's going to be the top-tier companies—Facebook, Google, and those types of companies that have lots of money that they can throw at it. And then it will start to trickle down."
- "You've got to figure out what you're going to use [AI inferencing] for in the first place. That's always the biggest struggle because otherwise it's just a data analytics platform. And people mistake it for that all the time. They're not doing AI, but just very intelligently analyzing trends in your data. That's huge and a big deal to be able to do that, but to manipulate code as data is being transferred, that's the inferencing side. And being able to really do that, like Facebook can—these guys are the ones that are really using it, and it's pretty amazing what they can do."
- "AI has been talked about the past two years [a lot more]. People don't really know what it is, but it's a buzzword that CIOs like to throw around, just like the cloud was, and then all of a sudden everybody's going to want to use it. So there will be this very steep curve and people will start dedicating dollars and making investments."
- "I just feel that this technology is so far ahead of time that it's going to be the top-tier engineers that are working for the 'bigs' [Facebook, Google, etc.] and then that talent will start to trickle down as companies decide to invest in the tech."

> **Inferencing is a watershed technology and, when it hits the right market, it's going to hit it huge. Teaching machines how to learn is one thing, but allowing them to learn and then execute and predict, that's a whole other side of the technology.**
>
> *Data center infrastructure executive for a foreign currency trading firm*

- "Once the appetite gets whet, you're going to see a pretty steep curve and then it will taper off, just like the cloud has, because a lot of people are taking their apps back. That's another trend that's emerging in cloud."
- "Inferencing is a watershed technology and, when it hits the right market, it's going to hit it huge. Teaching machines how to learn is one thing, but allowing them to learn and then execute and predict, that's a whole other side of the technology."
- "I would say the self-driving automobile would be the big driver [of AI inferencing]. That requires miniaturization, which all drives back to the hardware. I think that's where it's going to be the most prevalent—in satellite technology,

automated cars. But that's a daunting task, and I think the adoption of that is very niche. It's going to be slow going. I think that will be more of a 10- to 15-year type of [ramp]."

Competition

- Did not discuss.

CUDA and Other Key Issues

- Did not discuss.

3) Software Engineer

The AI inferencing market is at least as important as the training market, according to the one source in this silo. Demand for AI inferencing is going to rise dramatically, driven by the use of public clouds. Nvidia faces many challengers in the inferencing market, including AMD and Xilinx in the short term and Intel and Google in the longer term. Chinese hardware manufacturers are also a threat because they supply China's huge demand for AI at a time when China has been forced, because of trade issues, to become more self-sufficient with its technology.

Key Silo Findings

AI Inferencing

- 1 of 1 said public cloud providers like Amazon and Microsoft are going to have a huge appetite for inferencing chips.

Competition

- 1 said Nvidia faces competition in the inferencing market from a host of different fronts and chip makers, including Intel's CPUs, Xilinx's FPGAs, and Google's TPUs.
- 1 said Google's TPUs are especially promising because of the TensorFlow library and Google's reputation for good documentation.
- 1 said Microsoft uses a mix of GPUs, CPUs, and FPGAs in its Azure cloud.
- 1 said AI software development on FPGAs is difficult because the most popular deep learning libraries cannot be used directly.

CUDA and Other Key Issues

- 1 said CUDA provides Nvidia with a significant competitive advantage—a gap that could get even wider in the inferencing market.
- 1 said the release of Ampere is not likely to spur an immediate buying cycle, despite its performance benefits, largely because it is likely to be expensive.

1) AI software engineer in the telecom industry

AI Inferencing

- "The inferencing market is as important as the training market. I believe the demand for inferencing chips in data centers is going to rise dramatically. The growth of AWS is one of the drivers, as companies use it more and more for production and even for deploying their models. Another driver is Microsoft's Azure."

Competition

- "In terms of share, Nvidia dominated the [AI] market, but now it has a lot of competitors and there are many promising technologies. For one, we see Intel getting into graphics and AMD is also getting stronger. The FPGA community is also developing graphics, and there are TPUs. Nvidia now has much more competition than before."
- "The competitors include Google with the TPU; Intel, with a sudden inclination towards GPUs; AMD, which is getting better; Xilinx, with FPGAs; and local Chinese companies. China is the largest investor into AI."
- "Also, with the trade war, a lot of Chinese companies and startups will be more inclined to use their own technology—Chinese competitors—in order to be more self-sufficient. There are a lot of local companies making chips, decreasing Nvidia's market in China."

> **Nvidia dominated the [AI] market, but now it has a lot of competitors and there are many promising technologies. … Nvidia now has much more competition than before.**
>
> *AI software engineer in the telecom industry*

- "The market for AI hardware in China is very strong. China also has the most AI consumers, such as police, malls, public places, and road surveillance. China had the first AI reporter on national television."
- "Microsoft's Azure uses FPGAs, or at least some hybrid form. Take a look at Project Catapult. Longs Peak is one of the internal names of the FPGAs. They are mostly believed to be Xilinx FPGAs. Amazon uses GPUs. CPUs don't have that much computational power."
- "FPGAs are tricky because there aren't a lot of toolsets available. You have to create everything from scratch."
- "Toolsets are available, but it is difficult to inference a deep learning and AI solution on FPGAs. Famous libraries like TensorFlow, Keras, and PyTorch cannot be directly used to deploy them. On the contrary, these libraries provide strong GPU support, internally utilizing CUDA and cuDNN, making the life of developers easier. CuDNN is the Deep Neural Net library used by TensorFlow, etc."
- "The basic workflow is as follows: An idea is born, then a developer starts designing the AI product using TensorFlow or Keras or PyTorch. These are libraries helping developers to make deep neural networks. In layman terms, instead of writing the entire paragraph of code, developers can just write a few words and get things done using them, because, in the background, the paragraph is already written by these libraries. These libraries internally access cuDNN, the library provided by Nvidia, to better access the GPU. Basically, cuDNN can interact with the GPU. So it acts as a bridge between the hardware and the software. CuDNN makes code to CUDA and deploys on GPUs."
- "The developer doesn't have to interact with the hardware at all. They can focus only on development. Designing and deploying stuff on FPGAs needs specialized knowledge and there aren't many open source projects. Microsoft has manpower and resources to pursue it as an inferencing option, but it is still not completely FPGA internally. Azure is a mix of FPGA, GPU, and CPU. Microsoft hasn't made anything publicly available as to how they are inferencing on FPGA. Nothing open to the community, either."
- "There's a lot of community support, on the other hand, for GPUs. This gives GPUs an advantage."
- "Amazon still uses GPUs, based on my knowledge. Google has started focusing on TPUs."
- "A lot of people speculate that Google's TPUs might be the way. They look very promising. Since TPU is associated with a company like Google, the community can expect good documentation and fantastic work. Google's TensorFlow is one of the most widely used deep learning libraries. GPUs and TPUs might be the way forward, as FPGAs are still lagging."
- "Intel is also releasing their first graphic processing unit. However, graphics are not Intel's strength, while Nvidia is very strong in graphics and they have been working on graphics for a very long time."
- "AMD was always in the shadows of Nvidia. They will be releasing a new GPU that is supposed to be more powerful than Nvidia's. It might be a game changer for Nvidia, but AMD is considered more at the low-end of GPUs."
- "I would say that the immediate threats [to Nvidia AI inferencing chips] are AMD and Xilinx. The long-term threats are Xilinx, AMD, Intel, Google, local Chinese companies. Finally, companies with resources can make their own or use alternatives for inferencing."

CUDA and Other Key Issues

- "CUDA does give Nvidia a big advantage. CUDA makes development with Nvidia products faster."
- "The CUDA advantage will increase for the inferencing market because of the size of the community that can support it. The CUDA community will be able to use the new inferencing products easily because of the availability of documentation. The community support and documentation have a big impact on companies' decisions to go with a product. People want to launch their products as soon as possible. Time is crucial, so they look for how easy something is to develop."
- "Ampere is a graphic card/GPU. A new card is usually always more powerful than its predecessor. It will be used in new high-end PCs, especially ones like Razer."
- "Data centers have quite strong and expensive hardware. Replacing that whenever a new GPU is released is unfeasible. It depends on the company when they decide to upgrade their hardware. Ampere is the first GPU to be made with 7nm process. That means it is more power efficient and powerful. But, to be honest, I won't be expecting companies to upgrade their data centers instantly. Also, AMD is also releasing a similar 7nm Navi GPU."

> **CUDA does give Nvidia a big advantage. … The community support and documentation have a big impact on companies' decisions to go with a product. People want to launch their products as soon as possible. Time is crucial, so they look for how easy something is to develop.**
>
> *AI software engineer in the telecom industry*

- "The Ampere series will surely slowly be consumed into data centers due to their low power and performance benefits, but it's too soon to comment. I haven't found any benchmarks and people are just speculating. Once its officially released and benchmarks are available, more things can be estimated."
- "Ampere is suspected to be priced high. Data centers might need a lot of them. If the upgrade cost overpowers the benefits and cost reductions in terms of power, it might not be absorbed in data centers."
- "Ampere will definitely decrease training time because it is more powerful. It will definitely help developers. For inferencing, it will have some impact, but data centers already have powerful machines and it will be very costly to implement Ampere. I don't know how often they replace the machines and replace the graphic cards. That's an expensive proposition."

4) Hardware Manufacturers

Both sources in this silo said demand for inferencing chips is starting to ramp up and should continue to do so. Ultimately, the number of chips needed for inferencing will far surpass those needed for training. Hyperscalers will be the main buyers for now, mainly to serve complicated tasks like natural language processing. Eventually, inferencing in edge devices will become more common and drive volumes of inferencing chips. At the edge, cheaper and more narrowly focused ASICs will be more popular than Nvidia's powerful GPUs, though Nvidia can make inroads with edge applications like wireless communication networks. It will be a while before a winner can emerge in inferencing chip technology because the market is still nascent and hyperscalers are testing every possibility. CUDA is a big advantage for Nvidia, one source said, because so much development has already been done around it, but companies like Google and Facebook remain committed to developing alternatives.

Key Silo Findings

AI Inferencing

- 2 of 2 said demand for inferencing chips is in its early stages and has significant room to grow.
  - 1 said demand will be orders of magnitude larger than for training chips.
- 1 said the development of natural language processing is fueling the growth of inferencing.
- 1 said hyperscalers will be the main buyers of inferencing chips.
- 1 said inferencing in edge devices will grow as a percentage of total inferencing as it becomes more feasible in more devices.

Competition

- 1 said ASICs are becoming more popular for inferencing, especially at the edge, because they can be more narrowly focused and less expensive than GPUs like Nvidia's.
- 1 said Nvidia's best opportunity in edge applications is with wireless telecommunications networks.
- 1 said Nvidia's strength in data center training applications—including its software and loyal customer base—will not necessarily translate to inferencing as small size, low cost, and energy efficiency are paramount in edge applications.
- 1 said hyperscalers are testing all possible chip solutions because the inferencing market is in its infancy and key issues around algorithms, instruction sets, and compression technology are still evolving.
  - The biggest data center operators are using chips from Graphcore, Xilinx, Intel, Broadcom Inc. (AVGO), AMD, and Nvidia.

CUDA and Other Key Issues

- 1 said CUDA is a key advantage because Nvidia developed it early, thus many algorithms and much technology is based on it.
  - Others like Facebook and Google, however, are not conceding the field to CUDA and will present a challenge to Nvidia's software dominance.

1) Former AI product executive for an ASIC developer

The inferencing market is heading toward an inflection point, driven by the development of natural language processing. The number of chips needed for inferencing will be orders of magnitude larger than for training. The Super 7 hyperscalers will be the largest consumers of inferencing technology. They have not yet decided which will be the chip of choice because the technology is still evolving and innovations are changing the landscape constantly. Data center operators

will not want to tie themselves to any one technology or company. CUDA is a huge advantage for Nvidia but protecting the lead it has developed will be challenging.

AI Inferencing

- "NLP is the key technology segment driver for AI inferencing. That can be as simple, and as profound, as Alexa. There are a lot of holes with Alexa, like having to speak quietly or not having too many people around. Alexa still does not have the meta-vocabulary that natural language processing would have."
- "But it can get more sophisticated, where I speak in gibberish or with a different accent, and you'll hear English. These are contextual things for the next generation of complexity."
- "Video and imaging have been very hot; autonomous driving has also been a part of it. This includes being able to recognize a dog or a shadow versus a child running after a ball when the sun is directly hitting the camera or the car. These, too, are contextual. Imaging is evolving, too, and there's more work to be done. There have been a lot more sophisticated algorithms for imaging and video."
- "You're supposed to be able to train and then inference any language, any context of that language, and any discussion and conversation. You're supposed to be able to understand the real question in there. For example, if I use a twist in my voice and make it sound like a question, I can expect a different type of answer compared to when I make a statement."
- "These things take a lot of computational power. That translates directly into both power consumption of equipment that's sitting in a cloud or an edge device, or in energy consumption—how you cool things, how you manage that power, how you get power to it, and how you manage the energy that comes out of it. All that is part of the complexities that big data centers have to think about."
- "The winners at the end of the day get access to trillions of dollars of software and hardware as their prize and billions of dollars in cost savings or in cost avoidance. The more energy conservative or more computationally conservative the innovations are—the chip technology and the software technology that sits on top of it—the more cost savings they add to Google or Amazon and the higher probability of them winning a bigger percentage of the market."
- "Natural language processing and any other inferencing technology has many applications. It's not just Alexa at home. It could be a doctor speaking into a microphone and medical emergency people doing on-the-scene work. You need to be able to communicate, translate, discuss, and debate things. There are many different types of verticals that can take advantage of the technology."
- "As part of that, the product appeal and effectiveness of what Google, Amazon, and Microsoft sell will be key drivers for creating AI inferencing chips."
- "The semiconductor technology [for conversational AI] does exist from a foundational perspective. However, the software and the algorithms are not quite there yet in terms of translating into some sort of an architecture, going to the transistors, and then having the transistors connected. The logic is built based on the algorithm and the architecture. The distance between what the data scientists need to come up with and how the transistors will flop is quite wide. The range is very big between a sophisticated natural language processing algorithm and the chip technology that exists today. The line hasn't quite been drawn yet."
- "But early versions do exist because there are simpler algorithms that will do limited language processing. Alexa and Siri are examples of these. Those technologies exist in a phone or on a box connected to a power outlet at home. You may need something more powerful than that as far as energy consumption is concerned."
- "For example, if I spoke with an accent and somebody else showed up at my house and wants to use the same speaker or same technology, I might need two of the same chips inside of the console. Because the device hasn't been designed to take on the power consumption of two devices, it might be too much for it. For that, we need the next generation semiconductor technology to integrate those two devices together and yet still bring the power use down. When it comes to semiconductors, every time you improve on the technology curve, you gain more processing power and improve on the power consumption."

> **Once you train the model and have a model, then you can have many instances of that model being inferenced. You train once, but every time you speak into the microphone, you infer it. … Therefore, the inference market is orders of magnitude larger than the training market.**
>
> *Former AI product executive for an ASIC developer*

- "You can define the [inferencing] market by the number of components or the value. Once you train the model and have a model, then you can have many instances of that model being inferenced. You train once, but every time you speak into the microphone, you infer it. That's why the number of inference chips is going to be magnitudes larger than training."
- "You train to get better and better, but you don't need great numbers of [training chips] running in parallel. At some point, you've trained your model and you've got what you need. Out in the marketplace, though, any time anybody talks into a Google box or with Alexa or with Siri, people are doing real-time inferencing. And for that real-time inference to happen, it needs to happen on a chip. Therefore, the inference market is orders of magnitude larger than the training market."
- "In terms of the timeframe, it's already happening now with not-so-sophisticated algorithms. You and I can already ask Google what the temperature outside is and it's going to answer. Alexa can turn on the TV or turn the lights off. There's a good amount of inferencing that's being done now. All the videos being uploaded to YouTube, all the pictures being uploaded to Facebook, a lot of the surveillance cameras that are doing facial recognition, all those are already doing inferencing on images. The inference market is absolutely in the middle of its inflection point but it still has a lot of room to grow."
- "The Super 7—Google, AWS, Microsoft, Facebook, Tencent [Holdings Ltd./TCEHY], Baidu [Inc./BIDU], and Alibaba [Group Holding Ltd./BABA]—are the seven largest consumers of servers and AI technology for their cloud. Some people even call them the Super 7+, with the plus including Apple, IBM [Corp./IBM], Tesla [Inc./TSLA], and lots of others that are jumping in the market and creating their own back end. Companies like Netflix [Inc./NFLX], Twitter [Inc./TWTR], and [Snap Inc.'s/SNAP] Snapchat are taking advantage of existing clouds. They might become customers of Azure or AWS or Google Cloud. The largest consumption of any inferencing technology will come most likely from the Super 7."

Competition

- "There is no one, simple, cohesive answer [as to which inferencing chips will win out]. A person working for Intel might say CPUs can do it. They just bought Habana and Habana can do it. Or, if not Habana, a combination of Intel CPUs and Habana can do it. With their $2 billion acquisition of Habana, they are definitely hoping it will provide a large return on investment."
- "[Intel] can also argue they bought Altera for $17 billion plus the Habana acquisition and the $400 million they spent on Nervana and other expenses and the Intel processors—they are hoping that $50 billion to $70 billion of investment will return at least $150 billion to $200 billion over time. They believe in this. Every expert will give a different answer to the question about the chip of choice."
- "Not even the Super 7 can say yet what the specific answer is. This is because a lot of algorithms haven't been discovered yet. How those algorithms will be executed on an instruction set, on a technology, what computation engine they will have, and what compression technology will come into play—all of that will change significantly over time. No one can say where it will go or that Nvidia will win over AMD or Intel, etc."
- "The Super 7 are, therefore, doing it all. They are using Graphcore, they are using Xilinx, they are using Intel, they are using Broadcom ASICs, they are using AMD GPUs, and they're using Nvidia's custom ASICs. They're all running them in different applications in different modes, and sometimes in the same mode in applications to see what the differences are."
- "[Testing multiple options] is not just about [avoiding] being stuck in a one-sided relationship with Nvidia or Intel. Let's say Facebook picked Intel and Google used Nvidia. If Nvidia turned out five years from now to be better for the algorithm that was created, Facebook would be toast. That's why everybody is keeping all their options open."
- "There is a lot of innovation still ahead. This is still in its infancy. You don't take one over the other for a cost advantage. You pick both to make sure the innovation in each platform is something you can take advantage of, should the need be there."
- "Habana doesn't even exist yet. Graphcore was there but their software didn't come out. Cerebras [Systems] was there but the chip didn't come out yet. There are five, 10, or 100 more technologies, companies, and chip architectures that will come out over the next three to five years that will be better than the systems that there are

> **Not even the Super 7 can say yet what the specific answer is. This is because a lot of algorithms haven't been discovered yet. How those algorithms will be executed on an instruction set, on a technology, what computation engine they will have, and what compression technology will come into play.**
>
> *Former AI product executive for an ASIC developer*

today. No one has created a lock on the market and its direction, owing to the fact that the vertical is completely in its infancy. It's too early to call a winner."

- "Habana, of course, has an advantage in a certain area, because they came later. Was Nervana too early? But it had some good things going for it. Did it give Intel some serious advantages because they were the first guys and now Intel can take some of those algorithms and create some instructions for the X86? Absolutely, yes. [Nervana] became obsolete as quickly as it came out because five other companies showed up."
- "Whether the brand is Nvidia, Intel, or AMD, it's going to be very dependent on who captures what algorithm and how quickly they can deliver a chip and its efficiency in terms of power dissipation, computational power, and it being more consistent with the algorithms coming out."

CUDA and Other Key Issues

- "Nvidia did a ton of things early, when nobody else was doing it, by creating the libraries in the environment that are very CUDA-driven. From mid-2014 or 2015 when Alexa came out and all the way to today, an entire ecosystem, every single computer science department of every single university that was interested in doing artificial intelligence had no other choice but to use Nvidia. Lots of algorithms and technology have been written and driven off CUDA."
- "In the last two to four years, every one of those Super 7 companies—Facebook, Google, AWS, and even others trying to compete in that marketplace and that can't directly use CUDA because CUDA is a technology that only works directly on technology from Nvidia—had to come up with their own alternatives."
- "There are two ways of looking at it. If there is no future for replacing CUDA, then the valuation for Nvidia should be even greater. The others can just go home if they can't discover a way around it. That's clearly not the case."
- "Does CUDA have a lot of momentum, has it captured a lot of imagination, and written a lot of algorithms based on the technology? Yes. And is that going to be harder to beat than if nobody was there? Absolutely. But to protect that area, it's not as easy as it might seem."

2) Marketing executive for a developer of edge AI chips

Nvidia's strength in the training market will not necessarily translate to inferencing, especially as the inferencing market shifts toward highly efficient, inexpensive chips for edge devices. There is also a trend toward application-specific chips, rather than GPUs, for AI tasks.

AI Inferencing

- "I would expect tremendous growth in inferencing in both [data centers and at the edge] for a long period of time. For the foreseeable future, inferencing in the data centers should continue to have spectacular growth as more and more services adopt AI and require AI to be processed in a cloud in a data center. So that's not going to be shrinking any time soon."
- "There are huge opportunities for inferencing in the data center."
- "That being said, inferencing at the edge is going to be experiencing growth as well. [It will be used in] appliances, smart vacuum cleaners, smart doorbells. Factory robotics is very popular."
- "What you'll see is the percentage of total inferencing start to be eaten into by edge AI because, as more and more chip solutions become feasible in more and more applications, you're going to see deployments. Initially those will be small deployments, people dipping their toes in the water."
- "Whether it's a mobile phone or it's a security camera or it's a smart robot vacuum cleaner for your floor or a doorbell or an industrial manufacturing solution to detect defects on an assembly line, those are all examples of edge AI. As more companies learn how to do it, and as more specialized chip solutions come on the market that can meet their requirements—which is cost, size, performance, energy efficiency—you're going to see the edge AI starting to eat into the data center as a percentage of the total."
- "[Edge AI] is also an application for enterprise. A lot of enterprises—especially SMBs—would be interested to have the power of AI but not have to rely on sending their data somewhere else. Very specific applications in the medical field and health care, where privacy and security of data are concerns, [will be key drivers of inferencing at the edge]. Inferencing there allows the results to be determined by a

What you'll see is the percentage of total inferencing start to be eaten into by edge AI because, as more and more chip solutions become feasible in more and more applications, you're going to see deployments.

Marketing executive for a developer of edge AI chips

properly implemented solution, and the data doesn't have the possibility of going to a cloud that can get hacked or a database where it can be stored and misused."

- "A lot of the solutions that inference in the data center, they still rely on networks being used to send the data, where the algorithm can be processed in a data center that could be quite remote. There are big data centers now in Salt Lake City [for example] that are processing data for people in New York City, in California, or even internationally, to do their inferencing."
- "All along the distribution path from the point where that data's collected—the camera or the audio sensor or the thermometer or whatever sensor is being used to collect that data—that needs to go hundreds of thousands of miles, over many nodes in a network, to get processed through the data center and then sent back. The result has to get sent back to the edge, so you're introducing seconds and sometimes minutes of time, and AI—if it's not convenient—is frustrating. If Alexa takes a couple of minutes to get back to you, you may as well pull out your phone and look it up."

Competition

- "Nvidia is coming at [AI inferencing] by repurposing the GPU that was originally designed for graphics, and they found that they could do AI as well, because AI uses a lot of graphical information. Most AI is using either images or audio, which can be processed as if it was an image. Audio can be converted into a thermal 2D image and then processed by a GPU the same way."
- "What we're seeing is that AI for edge devices is application-specific design. And that's why ASICs as a category for architecture for chips are the new trend in the industry for AI, because AI implementations tend to be very, very specific. Are you using this chip to do object detection, or is it audio? Are you doing it to understand gestures or movement? Or is it more about classifying? What am I seeing? Is this an employee or not?"
- "That's the power of an ASIC. You can come up with a chip that is very low cost, because it's not designed to do a million different things—it's designed to do a very narrow range of things related to processing images, processing sounds."
- "The power of a GPU is that it's flexible, but it carries the overhead of being able to do other things besides AI. That makes it less energy efficient when used only for AI."
- "Nvidia has products designed for the edge, but the edge is a range of applications. Where I see Nvidia now is focusing on the wireless networks. Wireless network providers are trying to put AI in some parts of their infrastructure, so they could do more local processing of some AI, instead of it going all the way to a data center in Salt Lake City or somewhere remote. So they have edge inferencing solutions that are powerful—but they're large, they're expensive, and they use a lot of energy. Maybe if it's going into a base station in a suburban area, maybe that's suitable. Perhaps that's a place where they can start."

The power of a GPU is that it's flexible, but it carries the overhead of being able to do other things besides AI. That makes it less energy efficient when used only for AI.

Marketing executive for a developer of edge AI chips

- "I know that they're trying to become more cost effective and more energy efficient and trying to get to more basic inferencing applications."
- "Nvidia is positioned very well, especially with certain types of applications in the data center, with training. [They are strong in] certain computing applications. They have a lot of software and a lot of expertise. They have a big customer base that is loyal. But being strong in those particular areas is not a guarantee [of continued success]."
- "In order for them to stay strong, they have to continue to address the areas that are exposed by competition. I think the edge is where they're probably most vulnerable. The edge on mobile devices, the edge on low-cost IoT devices—smart home, smart city—the enterprise. All those places that really care about small size, low cost, energy efficiency, because those people are running a business or they're paying their electric bills or having to carry their devices around. Those are the vulnerabilities that Nvidia will have to address in inferencing."
- "In training, they're super strong. The only one that will perhaps challenge them is Google, which has a huge investment in their TPUs and they've got the TensorFlow tool chain and all kinds of free, or almost-free, solutions that a developer can get from them. A company like Google could be a threat to Nvidia's dominance in training."

CUDA and Other Key Issues

- Did not discuss.

Secondary Source

This secondary source focused on rumored specs for Nvidia's upcoming Ampere GPUs. A Twitter user suggested Ampere will have a massive die size and double the tensor core performance. Some claims in the unverified leak seem rather far-fetched.

Feb. 23 Hot Hardware article

- "A series of Twitter posts are drawing attention because they purportedly detail NVIDIA's next-generation Ampere GPU, with some interesting (and potentially far-fetched) information. We're highly skeptical, for reasons we will discuss in a moment. However, it is fun to speculate as we await the next big GPU from NVIDIA, in what figures to be a busy year overall in the graphics space."
- "Twitter user KittyCorgi (@CorgiKitty) has drawn some attention over a handful of tweets regarding Ampere. The latest one, posted on Friday, claims Ampere will feature a die size of around 826mm^2. To put that into perspective, NVIDIA's current generation Turing GPU (12nm) tops out at 754mm^2 (TU012), with 18,600 million transistors, AMD's 7nm Navi GPU tops out at 251mm^2 (Navi 10), 14nm Vega (Vega 10) is 495mm2, and 7nm Vega (Vega 20) is around 331mm2."
- "What makes KittyCorgi's claim interesting is not just the raw size, but the notion of NVIDIA unpacking a 7nm GPU that big. If Ampere does end up measuring 826mm^2, it seems more likely it would leverage a 12nm manufacturing process instead, which goes against previous rumors. But hey, anything is possible, right?"
- "That said, the supposed leaker makes the following claims about Ampere:
    - INT32 Unit remains unchanged
    - Double the FP32 Unit for shader proportion
    - Double the Tensor Core performance
    - Enhanced L1 data cache for more comprehensive functions"
- "According to KittyCorgi, the higher end GA103 GPU will feature 60 streaming multiprocessors (SMs), 10GB or 20GB of memory, and a 320-bit memory bus. That would give the GPU 3,840 CUDA cores, which is less than the GeForce RTX 2080 Ti, though that's assuming NVIDIA sticks with 64 CUDA cores per SM. Being a much denser GPU, according to the information posted, we could be looking at 128 SMs per core, which would give the GA103 GPU 7,680 CUDA cores."
- "Nothing is official until NVIDIA says so, it is simple as that. But if you are looking for reasons why these specifications could be legitimate, it really boils down to Navi. How so? We know AMD is coming out with an upgraded version of Navi later this year, one that will bring performance to the fold, along with support for real-time ray tracing. A custom version of this so called 'Big Navi' will be featured in both Sony's PlayStation 5 and Microsoft's Xbox Series X game consoles, and AMD confirmed we will see Big Navi in the PC space this year as well."
- "Past rumors suggest Ampere will deliver big performance gains, both in rasterized rendering performance and ray-traced workloads. The supposedly leaked specifications would certainly keep NVIDIA in the lead, and probably by a big margin."
- "While it is fun to consider a 7nm GPU with an 826mm^2 die size and 20GB of video memory piping data through a 320-bit bus, it all seems…optimistic. That's a tough ask of TSMC (or any chip fab), and we see problems with yields at that density. All of the resulting defective dies means we would be looking at an astronomical price tag for working parts, and quite frankly, we don't see it happening. As in, go ahead and bet the farm against these leaked specs playing out."
- "If this does somehow come to pass, we are almost certainly looking at a server GPU, with consumer variants being less ambitious with better yields and better pricing. Still, our BS meter is ringing with this one."

Additional research by Eva Cahen and Emily Carr.