A Forrester Consulting
Thought Leadership Paper
Commissioned By Micron

January 2019

Memory And Storage Are Critical To Building Better Artificial Intelligence And Machine Learning Architectures

FORRESTER®

Table Of Contents

Project Director:
Andrew Magarie,
Senior Market Impact Consultant

Contributing Research:
Forrester's Infrastructure &
Operations research group

Executive Summary

While artificial intelligence (AI) and machine learning (ML) continue to make headlines across the tech world, the truth is most organizations are still experimenting and just scratching the surface of the seemingly limitless possibilities for both AI and ML. As companies develop more complex AI/ML models and start working on more advanced use cases, the hardware used to train and run those models will become increasingly more important. Advanced AI/ML use cases require a detailed look at compute, memory, and storage configurations to avoid performance and throughput bottlenecks and drive faster, better results.

The bottom line is: hardware matters. Whether it's at the edge, in the cloud, or on-premises, the right hardware architecture can deliver the performance companies need to transform their business with AI and ML.

Micron commissioned Forrester Consulting to evaluate artificial intelligence and machine learning hardware architecture. Forrester conducted an online survey and three additional interviews with 200 IT and business professionals that manage architecture, systems, or strategy for complex data at large enterprises in the US and China to further explore this topic.

Whether it's at the edge, in the cloud, or on-premises, the right hardware architecture can deliver the performance companies need to transform their business with AI and ML.

KEY FINDINGS

- **AI/ML will continue to exist in public and private clouds.** Early modeling and training on public data is occurring in public clouds, while production at scale and/or on proprietary data will often be in a private cloud or hybrid cloud to control security and costs.
- **Memory and storage are the most common challenge in building AI/ML training hardware.** While the CPU/GPU/custom compute discussion received great attention, memory and storage are turning out to be the most common challenge in real-world deployments and will be the next frontier in AI/ML hardware and software innovation.
- **Memory and storage are critical to AI development.** Whether focusing on GPU or CPU, storage and memory are critical in today's training environments and tomorrow's inference.

As AI/ML Use Cases Expand, Firms Must Put Workloads Where They Belong

We are in the nascent stages of AI and ML, both as an architectural and business discipline. Many organizations have started experimenting with AI/ML on a small scale, with use cases such as customer recommendations, targeted advertising, and predictive analysis. These organizations run on either commodity hardware or a few machines with devoted GPU hardware in the data center; alternatively, third-party cloud providers offer the needed systems. In both cases, while easing setup, this takes much of the control and strategy out of the hands of enterprise architects.

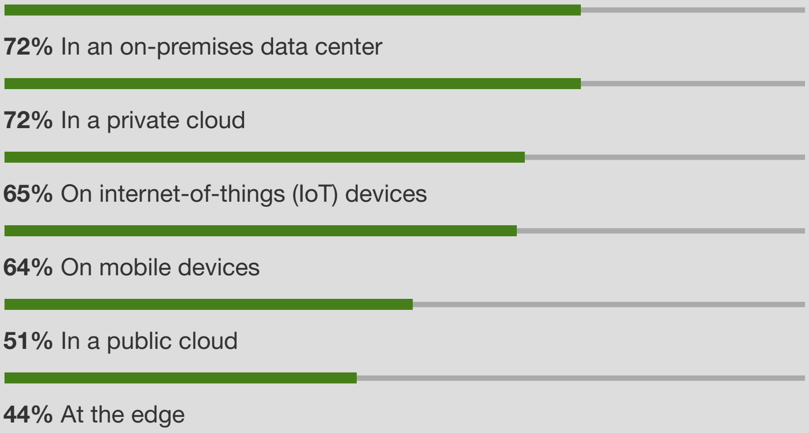
As more advanced use cases such as image recognition, speech recognition, self-automation, and others become widespread, the hardware needed to efficiently and effectively run these use cases — as well as workload placement — must evolve. To find out how architecture will change as use cases advance, we surveyed 200 IT and business professionals that manage architecture or strategy for AI and ML at large enterprises in the US and China. We found that:

› **Advanced analytics are being run across the enterprise today.** Seventy-two percent of firms are running advanced analytics in on-premises data centers today, while 51% analyze in the public cloud and 44% at the edge (see Figure 1). Many organizations are also running mixed workloads. Seventy percent of our survey respondents say they are running mixed complex data sets, and this is only growing.

> “[Where we run the workloads] really depends on the usage, the scenario use case, and the stage of R&D. For research using proprietary data, we use on-premises, but we also have photo recognition projects running in the public cloud, the reason being that it’s actually very easy to maintain.”
>
> *Former director of AI, global technology company, current AI startup founder*

Figure 1

“Where are you currently analyzing complex data sets today?”
(Select all that apply)



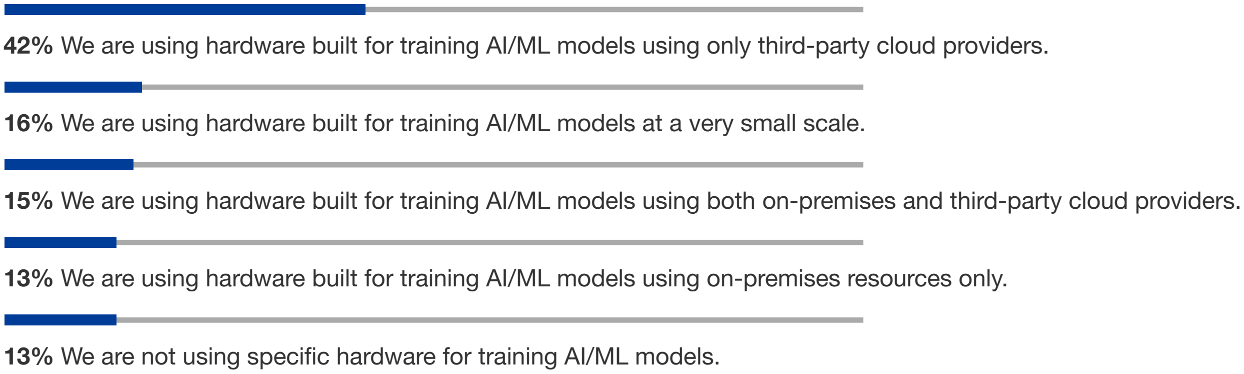
Base: 200 IT and business professionals that manage architecture or strategy for complex data sets at large enterprises in the US and China
Source: A commissioned study conducted by Forrester Consulting on behalf of Micron, August 2018

70% of survey respondents say they are running mixed complex data set workloads today.

- **Far fewer organizations run dedicated hardware for AI training and inference.** Only 28% of surveyed organizations are using hardware for training AI/ML models on-premises or mixed with third-party cloud providers (see Figure 2). Far more common, 42% exclusively use third-party cloud providers for their current AI/ML model training, due to ease of use, tool kits, reduced maintenance, etc. A further 29% are not using specialized hardware or using it at a very small scale.
- **Those who leverage specific hardware solutions are finding that some assembly is required.** Components of dedicated hardware for AI require customization that off-the-shelf offerings can't yet provide. Close to half of respondents say they source, or plan to source, AL/ML hardware components by buying and customizing or integrating them to address this.

Figure 2

"Please describe the scope of your current hardware capabilities when it comes to training AI/ML models?"



Base: 200 IT and business professionals that manage architecture or strategy for complex data sets at large enterprises in the US and China
Source: A commissioned study conducted by Forrester Consulting on behalf of Micron, August 2018

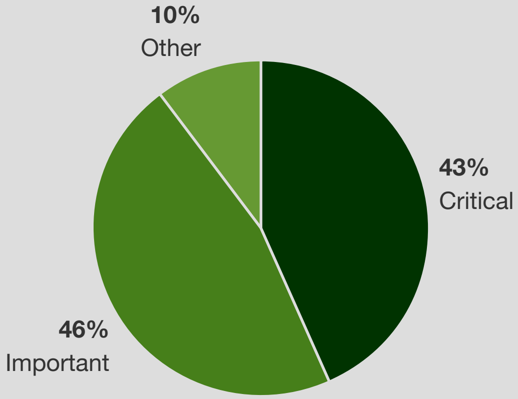
PUT THE WORKLOAD WHERE IT BELONGS: THE IMPORTANCE OF AI/ML LOCALITY

When architecting hardware specifically for AI/ML, survey respondents say that the location of compute and memory is a crucial component of performance and success. Eighty-nine percent of respondents say that it is important that compute and memory are architecturally close together (see Figure 3). This is even more essential for organizations that analyze data outside of their own data centers or the cloud — 51% of companies analyzing data sets at the edge say locality is critical, compared to just 37% who are not. As these more advanced use cases continue to grow in popularlity, this idea will play a greater role in how hardware solutions are architected.

89% of respondents say that it is important that compute and memory are architecturally close together.

Figure 3

"How important is the locality of compute and memory to your AI/ML workloads (i.e., is it important that they are architecturally close together)?"



Base: 200 IT and business professionals that manage architecture or strategy for complex data sets at large enterprises in the US and China
Source: A commissioned study conducted by Forrester Consulting on behalf of Micron, August 2018

51% of companies analyzing complex data sets on the edge think locality of compute and memory is critical (compared to **37%** not analyzing data at the edge)

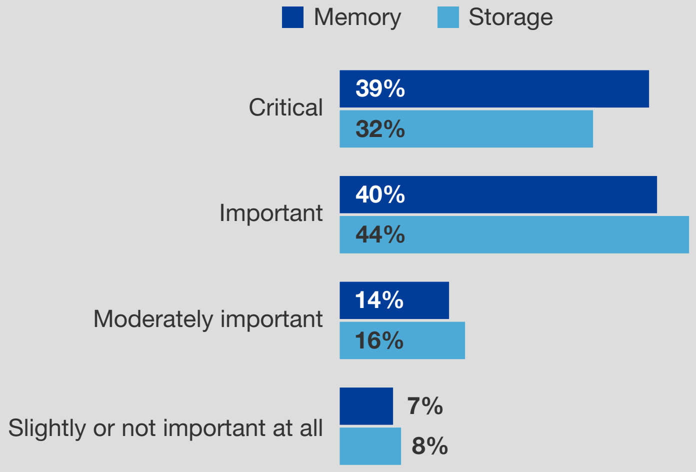
Memory And Storage Bottlenecks Limit AI/ML Analytics

Throughput and performance are two of the most important aspects of memory and storage when it comes to both AI/ML training and inference. While most organizations are satisfied with the kind of analytics their current hardware can support today, there are some architectural bottlenecks that limit throughput and performance and as a result, bottleneck AI/ML analytics. Our survey shows that:

- **Enterprises recognize a need to upgrade or rearchitect their memory and storage.** For AI/ML training, 79% say upgrading or rearchitecting memory is important or critical, while 76% say it is important or critical to do the same with storage (see Figure 4).
- **Storage and memory limit both training and inference performance and throughput.** Available memory and storage are the top two hardware-related challenges for AI/ML training and inference today, with close to two-thirds of respondents indicating that they are bottlenecks to performance and throughput (see Figure 5).
- **These challenges escalate as AI/ML moves more to the edge.** Data centers are not always the ideal location for all AI/ML workloads. Respondents understand this and predict that they will move workloads away from data centers to the public cloud and edge in the next three years (see Figure 6). When this happens, the location of workloads' compute and memory resources becomes even more important and firms must pay greater attention to how their hardware is architected.

Figure 4

"How important is it that you upgrade or rearchitect your memory and storage in order to meet your goals for AI/ML training in the future?"

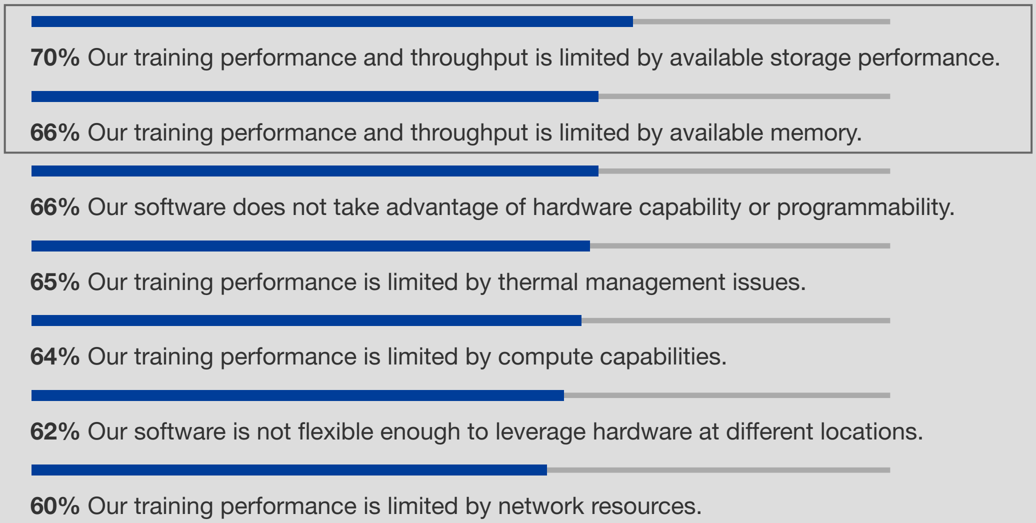


Base: 200 IT and business professionals that manage architecture or strategy for complex data sets at large enterprises in the US and China

Source: A commissioned study conducted by Forrester Consulting on behalf of Micron, August 2018

Figure 5

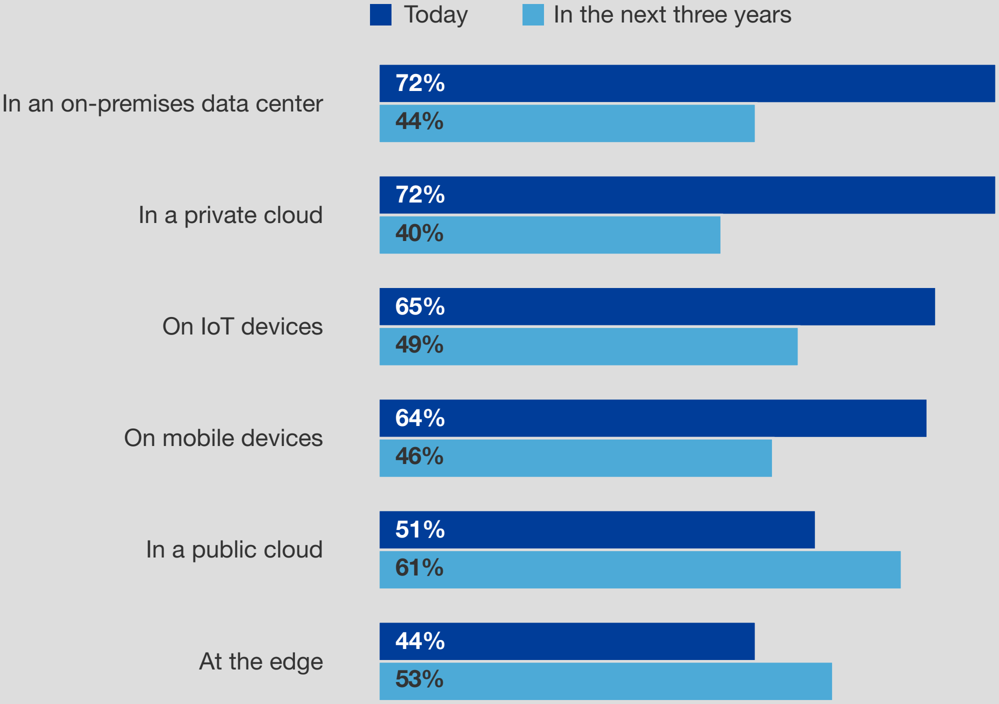
"What hardware-related challenges are you having with training AI/ML models today?" (Rank up to your top five)



Base: 200 IT and business professionals that manage architecture or strategy for complex data sets at large enterprises in the US and China
Source: A commissioned study conducted by Forrester Consulting on behalf of Micron, August 2018

Figure 6

"Where are you currently analyzing complex data sets today? Where do you expect to be analyzing complex data sets in the next three years?"



Base: 200 IT and business professionals that manage architecture or strategy for complex data sets at large enterprises in the US and China
Source: A commissioned study conducted by Forrester Consulting on behalf of Micron, August 2018

HARDWARE CHALLENGES ARE NOT THE ONLY BARRIER TO AI/ ML SUCCESS

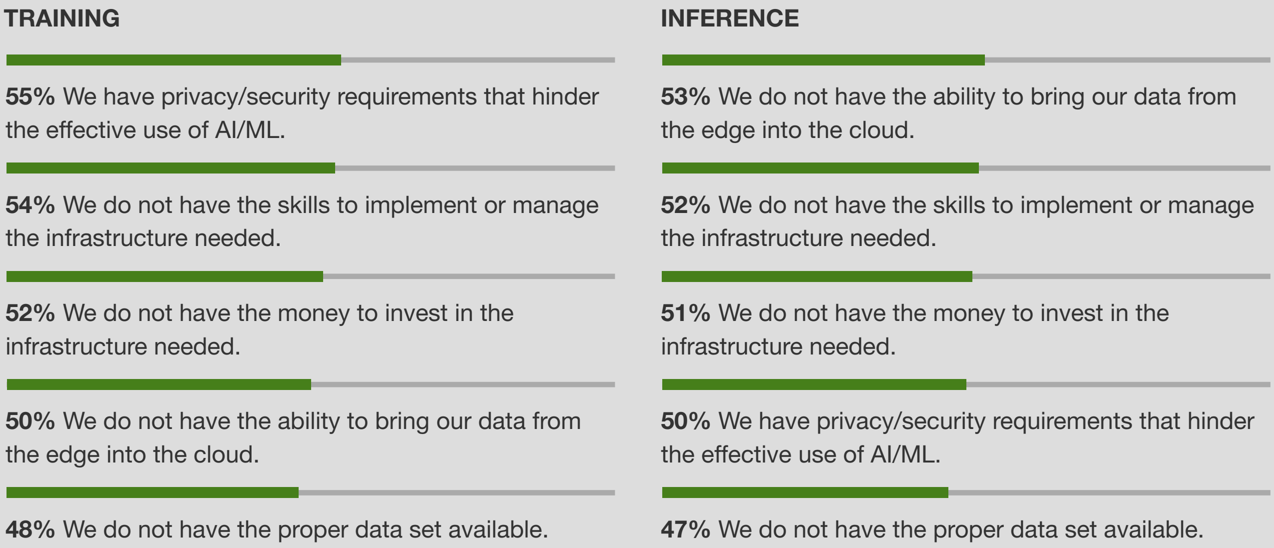
Architects are running into issues managing infrastructure, as well as working with data privacy and governance challenges (see Figure 7).

- **Most lack skills to implement and manage infrastructure.** As firms tackle more advanced use cases for AI and ML, the standard toolkits and base models provided by cloud vendors and third parties are not sufficient. Firms need to implement custom infrastructure to support these use cases, and the skills to do that are in low supply. Over 50% of firms say that they do not have the skills to implement or manage the hardware for both AI/ML training and inference.
- **Data privacy and governance challenges grow.** Across the globe, there is growing concern about how data, especially personal data, is used by companies. Regulations like GDPR highlight this and foreshadows additional regulations to come. Many current AI/ ML practices are based on potentially sensitive data and further regulations may hinder those use cases. Concerns over privacy and security requirements hindering the effective use of AI/ML is the number one concern when it comes to training AI/ML. Organizations need to review their data governance practices for potential data privacy challenges and make changes that will allow for the effective use of AI/ML.

> "Regulation is going to be pretty interesting in the next few years to come . . . I think we are in the wild west days of AI where a company can collect any kind of data and do anything with it forever. But companies will need to adapt, especially so in consumer markets where you're collecting individuals' data."
>
> *Sr. product manager, American multinational technology company*

Figure 7

"What business-related challenges are you having with AI/ML models today?" (Rank up to your top three)



Base: 200 IT and business professionals that manage architecture or strategy for complex data sets at large enterprises in the US and China
Source: A commissioned study conducted by Forrester Consulting on behalf of Micron, August 2018

Rearchitect Memory And Storage To Reap Rewards Of Better AI/ML Architecture

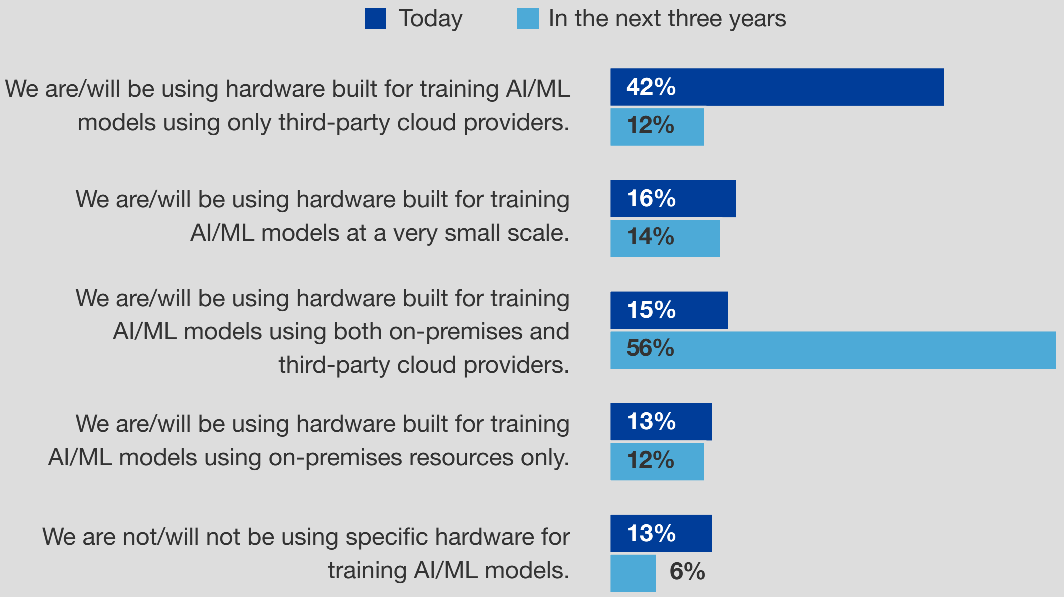
Despite the current challenges, the future is bright for AI/ML analytics. New use cases and models are being built every day, and firms are just scratching at the surface of neural network potential. While many are in the experimentation phases with AI/ML today, almost all plan to expand their footprint. In the next three years, 89% of survey respondents plan to run mixed complex data set workloads — a sign of growing maturity in handling diverse data sets. Rearchitecting memory and storage for AI/ML is seen as a core component of advancing AI/ML practices and will lead to success for the future. Our survey shows:

- **Organizations will adopt a balance between on-premises and cloud architecture for AI/ML.** While today, close to half (42%) of organizations use only third-party cloud for their analytics, most plan to move some of those workloads on-premises, with hardware designed for AI/ML (see Figure 8). Fifty-six percent of organizations see themselves using hardware built for AI/ML both on-premises and in the cloud in the next three years. This follows the emerging trend of "putting the workloads where they belong" — placing workloads in the ideal location depending on the data used, the use case, and other determining factors. This also means that many organizations will be building hardware designed for AI/ML for their data centers in the near future.
- **Organizations understand the importance of moving memory and computing closer.** Moving memory and compute closer together for AI/ML workloads is seen as essential to success by our respondents. Ninety percent of firms plan to move computing and memory closer together to improve AI/ML workloads in the future (see Figure 9). This is especially true of organizations that are experimenting with AI/ML outside of the cloud or data center. Of respondents who are currently running advanced analytics at the edge, this rises to 95%.

90% of firms plan to move computing and memory closer together to improve AI/ML workloads in the future.

Figure 8

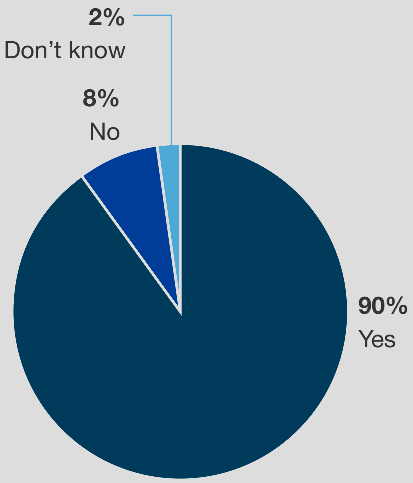
“Please describe the scope of your current hardware capabilities when it comes to training AI/ML models? Where do you see the scope in three years?”



Base: 200 IT and business professionals that manage architecture or strategy for complex data sets at large enterprises in the US and China
Source: A commissioned study conducted by Forrester Consulting on behalf of Micron, August 2018

Figure 9

“Are you planning on moving computing and memory closer together to improve AI\ML workloads?”



97% of companies analyzing complex data sets on the edge plan to do this (compared to **85%** not)

Base: 200 IT and business professionals that manage architecture or strategy for complex data sets at large enterprises in the US and China.
Source: A commissioned study conducted by Forrester Consulting on behalf of Micron, August 2018

BENEFITS OF BETTER AI/ML ARCHITECTURE

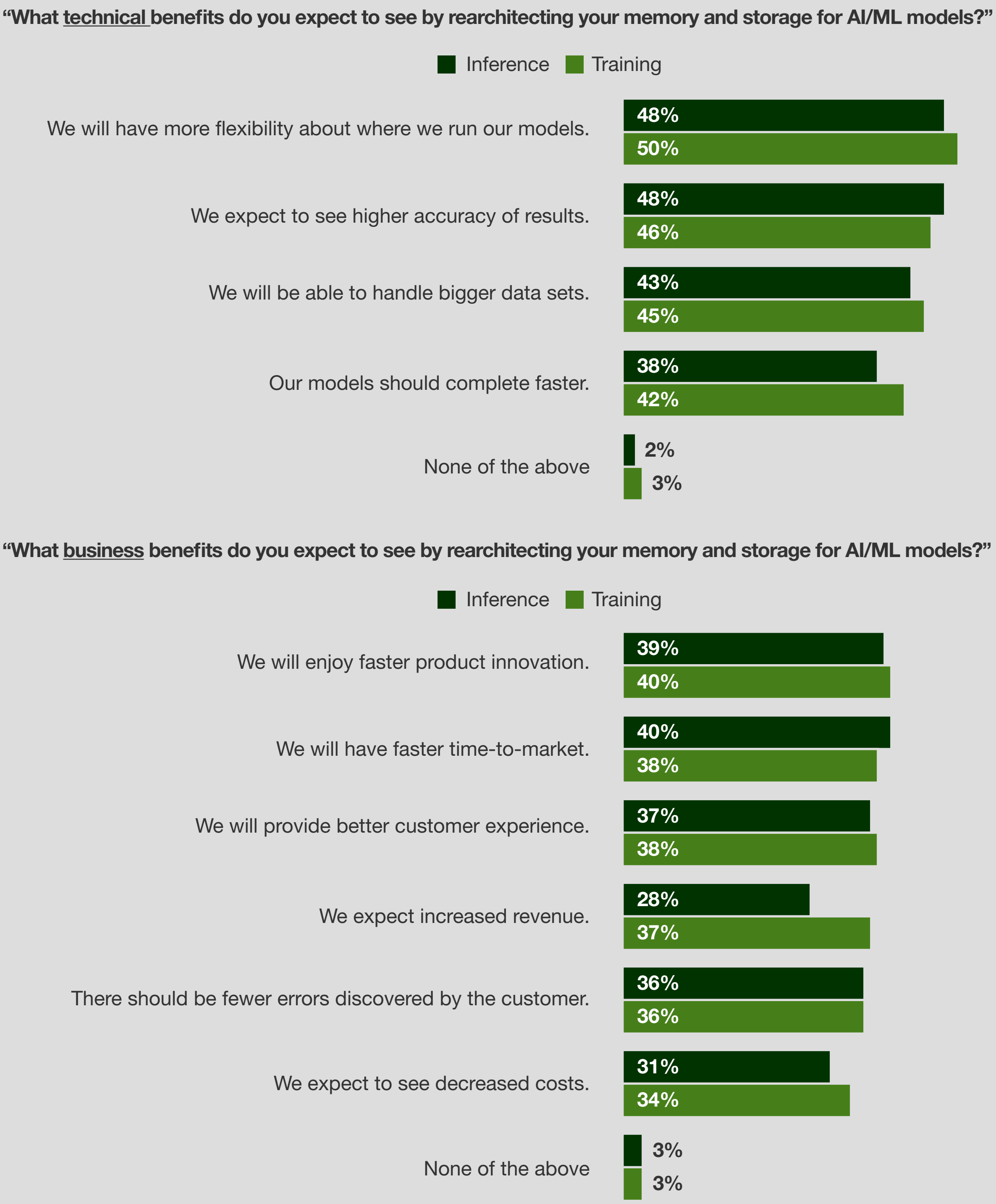
As AI/ML moves beyond experimentation, stakeholders need to show its value and move to more advanced use cases to gain buy-in for custom hardware architectures. To this end, it is important to show the benefits of custom AI/ML architectures, not only related to the use case itself but tied to key business drivers and KPIs. Our survey shows that respondents expect that rearchitecting the memory and storage components of AI/ML architecture will yield the following technical benefits (see Figure 10):

- **Greater flexibility on where models can be run.** Half of survey respondents believe rearchitecting AI/ML memory and storage will give them greater flexibility to run AI/ML training in different locations, which is extremely important if firms follow the creed of putting workloads where they belong.
- **Increased accuracy of results.** Nearly half of survey respondents say rearchitecting AI/ML memory and storage will lead to better and more accurate models.
- **Ability to handle larger data sets.** Forty-five percent of respondents expect that rearchitecting AI/ML memory and storage will allow them to handle larger data sets, a key to opening up some of the more advanced use cases that are underpinned by complex neural networks.

Respondents also believe these technical benefits will translate to greater success with business goals.

- **Faster delivery of products and services.** Forty percent of firms say that improved AI/ML architecture will lead to faster product innovation, and 38% say they will be able to deliver faster time-to-market.
- **Better customer experience.** Thirty-eight percent expect their customer experience to improve with enhanced AI/ML memory and storage, and 36% believe that the higher accuracy of models will lead to fewer errors discovered by customers.
- **Greater profits.** Better AI/ML hardware architecture impacts the bottom line. Thirty-six percent of firms say improvement of memory and storage will lead to increased revenues, and 34% expect to see cost savings as a result.

Figure 10


“What technical benefits do you expect to see by rearchitecting your memory and storage for AI/ML models?”

“What business benefits do you expect to see by rearchitecting your memory and storage for AI/ML models?”

> “Our business drivers are primarily around customer experience/customer lifetime value. **It’s really how do we drive better customer engagement and experience. That should eventually feed into other stuff**, like revenue generation, cost savings, and the like.”
>
> *Lead scientist, data and machine learning, Online printing and customization company*

Base: 200 IT and business professionals that manage architecture or strategy for complex data sets at large enterprises in the US and China
Source: A commissioned study conducted by Forrester Consulting on behalf of Micron, August 2018

Key Recommendations

Forrester's in-depth survey of IT and business professionals that manage architecture or strategy for complex data sets about AI/ML architecture yielded several important recommendations:

Put workloads where they belong, in the cloud, data center, or at the edge. While organizations are experimenting with AI and ML in the cloud (or in very small proofs of concept on-premises) many recognize that workload placement becomes critical as use cases mature. A sign of enterprise maturity will be recognizing when, where, and why training and/or inference models should run in the cloud, in the data center, or at the edge.

Rearchitect specifically around memory and storage throughput and performance. Of the possible hardware constraints limiting AI/ML today, including compute constraints, hardware programmability, thermal management issues and network issues, memory and storage performance/throughput rose to the top of survey respondents' concerns. Latency and bandwidth are critical factors that will need to be factored in as well.

Retrain infrastructure and operations for AI/ML. While data scientists are quickly coming up to speed with new techniques and models, most enterprises do not have the skills to implement or manage the infrastructure these models run on. Organizations must focus on training operations around massively parallel workloads. As solutions become more customized, support from integration in particular will be needed.

Refocus change and risk management for AI/ML. It's not just a hardware exercise. Many respondents called out privacy and security concerns as they grew out their AI/ML footprint. Change and risk management processes built in an era of largely static data centers and limited modelling will need to be rebuilt for the 21st century.

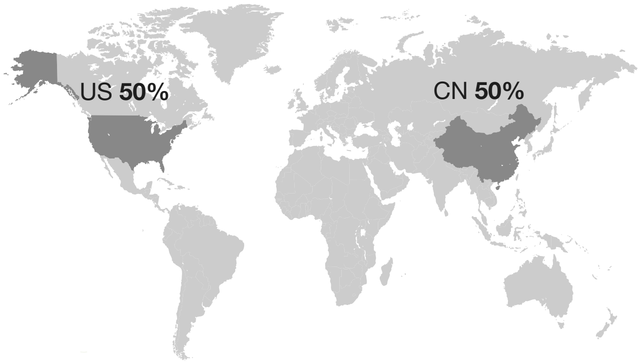
You can't buy off the shelf . . . yet. While many architects would prefer a model similar to how they buy data center servers and cloud services — picking from a selection of prechosen configurations and buying as necessary — we are not yet at that stage for AI/ML. Instead, focus on critical vendors to partner with and how the pieces come together. Over time, buying AI/ML capabilities off the shelf will become a reality.

Appendix A: Methodology

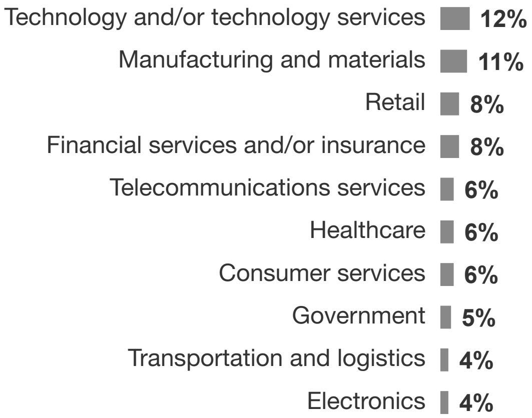
In this study, Forrester conducted an online survey of 200 organizations in the US and China to evaluate artificial intelligence and machine learning hardware architecture. Survey participants included IT and business professionals that manage architecture, systems, or strategy for complex data sets. In addition, Forrester conducted telephone interviews with three participants fitting the same description as the online participants. The study was completed in August 2018.

Appendix B: Demographics/Data

GEOGRAPHY



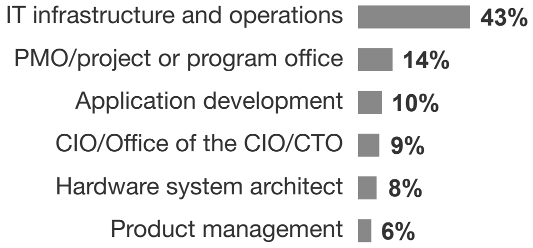
INDUSTRY (TOP 10 SHOWN)



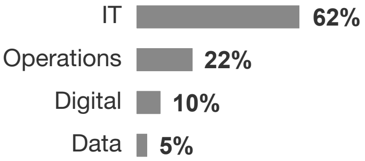
COMPANY SIZE (# OF EMPLOYEES)

30%	20,000 or more
70%	5,000 to 19,999

IT JOB FUNCTION



POSITION/DEPARTMENT



TITLE

12%	C-level executive
14%	Vice president
28%	Director
38%	Manager
9%	Project manager

Base: 200 IT and business professionals that manage architecture or strategy for complex data sets at large enterprises in the US and China.
Note: Some percentages may not total 100 due to rounding.
Source: A commissioned study conducted by Forrester Consulting on behalf of Micron, August 2018

Appendix C: Supplemental Material

RELATED FORRESTER RESEARCH

“Predictions 2019: Artificial Intelligence,” Forrester Research, Inc., November 6, 2018.

“Deep Learning: An AI Revolution Started For Courageous Enterprises,” Forrester Research, Inc., September 5, 2018.

“AI Deep Learning Workloads Demand A New Approach To Infrastructure,” Forrester Research, Inc., May 4, 2018.