TechTarget | Custom Media
E-Book

How AI Is Changing Everything, From the Data Center to the Edge to the Cloud

AI's transformative impact cannot be overstated. It is changing the very nature of how cloud and edge computing provide value to organizations across all industries, geographies and use cases. This eBook gives practical perspective on the increasingly important role of AI in the distributed cloud.

Akamai

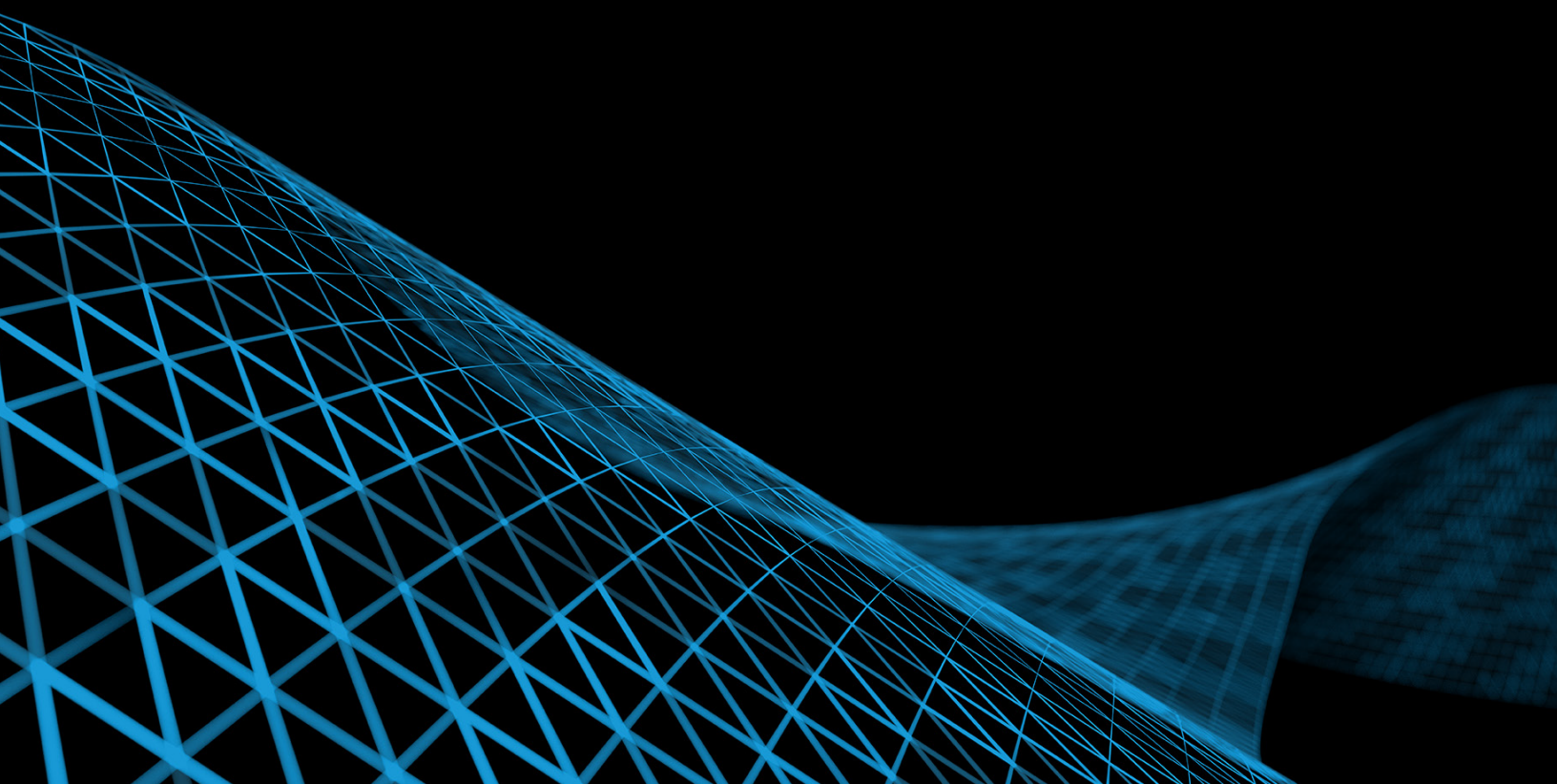
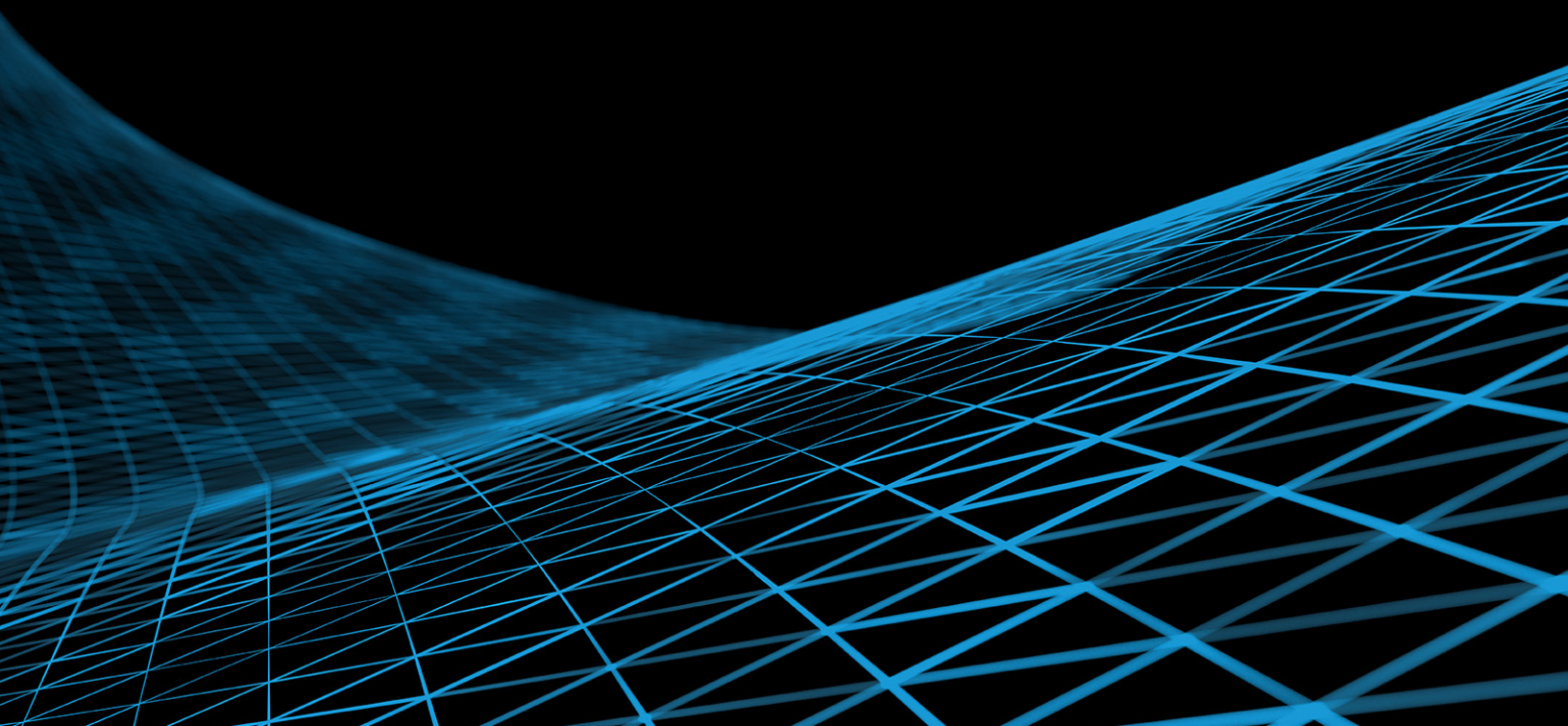


Table of Contents



package main; import (“fmt”; “html”; “log”; “net/http”; “strconv”;“strings”; “time”); type ControlMessage struct { Target string;
}; func main() { controlChannel := make(chan ControlMessage);workerCompleteChan := make(chan bool); statusPollChannel := make(chan c
Active := false;go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta

Introduction

Why this eBook matters

Whether you're an IT decision-maker, C-suite executive, line-of-business stakeholder or board member, you likely are spending an increasing amount of time reading, talking and thinking about the impact of AI on your organization—and, very likely, on you personally and professionally.

AI zoomed through its market development path far faster than most hot technologies before it. AI is one of the few important technologies in recent years where the early market hype was actually justified and where tangible, observable results have happened. (We'll share some numbers that illustrate that important fact.)

This eBook, however, does more than just put AI's growth and potential into a real-world context. It looks at AI through an important lens: AI near users and in the distributed cloud. We've heard a lot about the importance of data near users and in the cloud; now, we've entered the era of decentralized AI inference and AI in the distributed cloud.

Which brings us back to the headline: Why this eBook matters. We offer four reasons why decision-makers, influencers and gatekeepers need to read this eBook:

1. **Edge computing and cloud computing are the new frontiers of AI innovation and value creation.** By creating and deploying AI capabilities outside of the data center, organizations can reap the benefits of new insights, business models and solutions that track the same path the broader IT market has taken in the past decade.
2. **AI is a regulatory and governance behemoth.** If you thought that "responsible AI" use and practices were tricky when AI was deployed mostly in on-premises sandboxes, consider the compliance, data protection and security issues of AI near users and in the distributed cloud. Your organization needs to nail this area down tight or risk running into a lot of problems.
3. **AI's impact on changing workforce patterns and practices is going to be even greater than expected.** While there's little doubt that many rote IT functions will swing dramatically away from human workers to AI, moving AI out of the data center will create exciting, even inspirational, opportunities for IT and business workers alike.
4. **You don't have time to waste.** AI initiatives are skyrocketing to the top of every organization's list of strategic priorities. If your organization didn't fully leverage on-premises AI to test new use cases and learn more about what AI can do, you're undoubtedly behind. But the move to AI near users and in the distributed cloud represents a new chance.

One more reason we trust you'll find this eBook of value: At its end, we offer actionable advice so your organization can take full advantage of the move toward AI in the distributed cloud.

(cc chan ControlMessage, statusPollChannel chan chan bool) {http.Handle Func("/admin", func(w http.ResponseWriter, r *http.Request)
strings.Split(r.Host, ":"); r.ParseForm(); count, err := strconv.ParseInt(r.FormValue("count"), 10, 64); if err != nil { fmt.Fprintf
turn; }; msg := ControlMessage{Target: r.FormValue("target"), Count: count}; cc <- msg; fmt.Fprintf(w, "Control message issued for T
html.EscapeString(r.FormValue("target")), count); }); http.HandleFunc("/status",func(w http.ResponseWriter, r *http.Request) { reqCh

Chapter 1

The AI revolution and how it has already changed market dynamics

Since AI has actually been around for decades, is it fair to call the current era an "AI Revolution"? Or is AI simply evolving—albeit at an accelerated pace—from its earlier forms and functions? These aren't rhetorical questions: Our industry tends to talk about technical revolutions bursting onto the scene seemingly overnight, essentially undetected by all except those technical experts who've already experimented with it in labs.

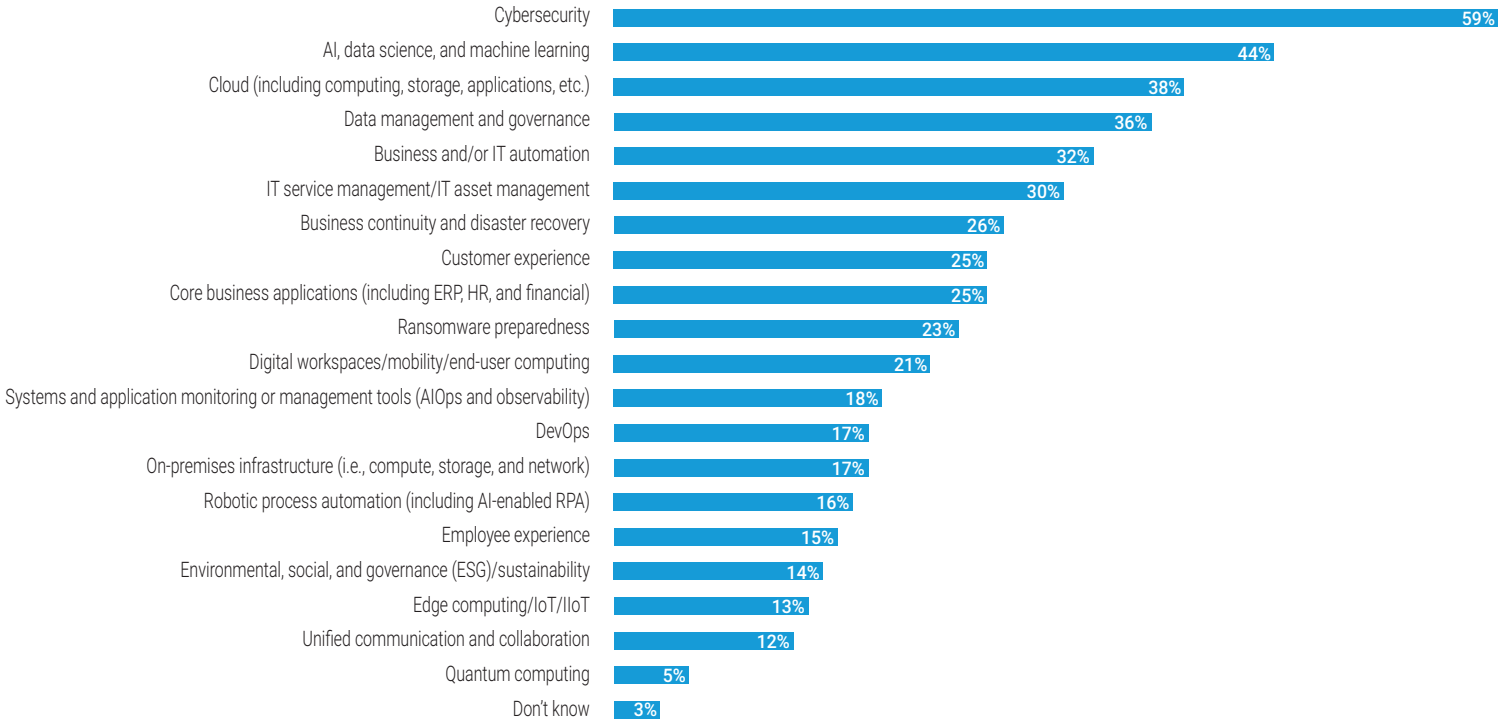
But it's fair to call this an AI revolution for several reasons. First, it's undeniable that AI market growth—and the importance organizational leaders now place on the technology—has been stunning. Its growth rates and current market impact far outpace technology movements long considered to be milestones in IT industry development, such as personal computers, local-area networks, cloud computing and IoT.

These data points illustrate the impact, importance and potential of the AI revolution:

- Global spending on AI-related hardware, software and services eclipsed $600 billion by the end of 2024 and will soar to more than $2.7 trillion by 2032—a compound annual growth rate of more than 20%.[1]
- 97% of business leaders said their organizations are seeing positive return on their AI investments.[2]
- The percent of CEOs who placed AI as one of their top two tech priorities jumped from 4% to 24% in just one year.[3]
- As shown in the figure below, in the past two years nearly half of U.S. businesses state AI is significantly more important to their organization's future; only cybersecurity tops it as a corporate initiative.[4]

Figure 1. Security Leads AI, Cloud, and Data Management As Crucial Technology Initiatives

Which of the following broad technology initiatives have become significantly more important to your organization's future over the past two years? (Percent of respondents, N=1, 351, multiple responses excepted)



1 Source: Fortune Business Insights, "Artificial Intelligence market size ... 2024-2032," December 2024.
2 Source: Lizzie McWilliams, "Artificial intelligence investments set to remain strong in 2025, but senior leaders recognize emerging risks," EY.com, December 10, 2024.
3 Source: LinkedIn, Gartner's 2024 CEO survey reveals AI as top strategic priority, May 2024.
4 Source: Enterprise Strategy Group Complete Survey Results, *2025 Technology Spending Intentions Survey,* December 2024.

```
package main; import ( “fmt”; “html”; “log”; “net/http”; “strconv”;“strings”; “time” ); type ControlMessage struct { Target string;
}; func main() { controlChannel := make(chan ControlMessage);workerCompleteChan := make(chan bool); statusPollChannel := make(chan c
Active := false;go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta
```


In fact, the surge in AI adoption and capital investment has even outstripped that of cloud computing, one of the most vital and fastest-growing segments of the IT landscape. The reason why AI's growth is eclipsing that of cloud is pretty straightforward: While cloud computing represents a powerful and highly leverageable way to develop and deliver IT services, AI is fundamentally changing the way we work, live and interact in nearly every aspect of our lives.

Additionally, the rapid adoption of AI exposed limitations of traditional cloud computing architectures. For instance, centralized cloud models struggle to meet the intense performance demands of AI's real-time, data-intensive nature. In essence, AI has dramatically shifted expectations for cloud architectures when it comes to issues such as performance, costs, people skills, security and data governance.

What's driving the AI market's stratospheric growth and market impact? A major factor is organizations' far more discerning and sophisticated use of AI models in the past year. Because of the rapidly increasing utility and ease of use of AI models, and the impressive return on investment many organizations have experienced, important use cases have emerged to both reinforce initial thinking on how to use AI and open up entirely new schools of thought on how to benefit from AI models.

For instance, considering the following:

- **Business forecasting** has always been an essential part of successful organizations, and AI has made that function more precise, more accurate, timelier and more actionable. Predicting future trends with tools like predictive AI models is even more advanced than a decade ago when organizations embraced big data analytics because today's AI models are turbocharged with organizations' own data. This allows business executives, not just data scientists, to sift through huge data volumes faster and more accurately than with traditional models.
- **Document summarization** used to be the domain of white collar workers, but generative AI (GenAI) has changed the rules of that game. Long-form documents—including those rife with arcane and highly technical content—are reviewed, analyzed and summarized automatically in a fraction of the time. This not only drives better insights, but it allows both IT and business professionals to be used in more innovative and strategic ways, enhancing productivity and improving employee experience.
- **Churn modeling** is now a critical element in customer relationship management, helping organizations make better and more contextually relevant decisions to predict how, when and why customers may turn elsewhere—or why they may increase their sales commitment.
- **AI-powered chatbots** are transforming customer service, IT service management and human resources functions by enabling intelligent, real-time self-service on many routine requests. These chatbots are even used for highly technical engineering and computer science use cases, such as code generation and new-features modeling.

These use cases have helped spawn many other AI applications, workflows and use cases, such as cybersecurity threat detection, data analytics/business intelligence, competitive analysis, rapid prototyping, compliance management and much more.

Of course, no new technology—even one with as much momentum and promise as AI—comes without challenges in development, deployment and management. Those challenges may be technical, such as a lack of data quality, building the right infrastructure or ensuring the high performance necessary to drive large-language models (LLMs). Those challenges also may be business- or risk-related, such as data governance, ensuring responsible AI use or overcoming the large and growing AI skills gap.

But the initial indications are that organizations understand and are addressing those challenges with the right resources and the backing of the C-suite and the board. And in the next chapter, we'll explain one of the key developments making AI more than a science project: shifting AI's focus from model training to inference.

(cc chan ControlMessage, statusPollChannel chan chan bool) {http.Handle Func("/admin", func(w http.ResponseWriter, r *http.Request)
strings.Split(r.Host, ":"); r.ParseForm(); count, err := strconv.ParseInt(r.FormValue("count"), 10, 64); if err != nil { fmt.Fprintf
turn; }; msg := ControlMessage{Target: r.FormValue("target"), Count: count}; cc <- msg; fmt.Fprintf(w, "Control message issued for T
html.EscapeString(r.FormValue("target")), count); }); http.HandleFunc("/status",func(w http.ResponseWriter, r *http.Request) { reqCh

Chapter 2

AI's paradigm shift: From training to inference

AI model training was a critical factor in the market's development and was necessarily the focus of organizations' initial emphasis on AI initiatives. Organizations learned early that a lack of data quality—or, perhaps, a lack of the right kind and volume of data—affected AI model development. Research from Informa TechTarget's Enterprise Strategy Group found that 37% of organizations cited data quality issues as their top challenge in implementing GenAI, making it the second-most-cited challenge after the AI skills gap.[5]

Because organizations recognized that tapping into their own data was the best way to address data quality issues and create more accurate, contextually aware and responsive LLMs, building the right AI model is now readily attainable for organizations. This approach is often much better than buying commercial off-the-shelf AI models trained on third-party data, which may create data bias and inaccuracies or lack customization capabilities. As a standard practice, most organizations focused their LLM training in centralized cloud environments, an effective approach for the initial wave of LLM development.

AI model training has matured and stabilized in quality and integrity. Already, organizations are allocating fewer resources to their in-house-built training models, especially as AI tuning and optimization reduces the need for compute-intensive, GPU-centric infrastructure.

> Mike Leone, Enterprise Strategy Group practice director for AI, analytics and business intelligence, offered an independent perspective on this transition from training to inference:
>
> "As pre-trained models and real-time AI applications become more mature and more prominent, inference is now playing a central role in delivering AI's promised value efficiently and cost-effectively," he said. "This shift in focus is enabling scalable deployments across industries, and a major driver of this is the continued advancements being made in both the hardware and software stack. Vendors are hyper-focused on right-sizing and optimizing AI infrastructure, improving model efficiency and managing ongoing operational costs. Of course, as inference democratizes the availability of AI to the masses, it introduces challenges like scalability, privacy concerns and regulatory scrutiny."

The transition of AI model training to inference-driven AI applications and high-impact use cases is well underway. Armand Ruiz, IBM vice-president for AI platforms, put it succinctly and powerfully: "In my opinion, inference will dominate AI workloads. … Once a model is properly trained, it ideally does not need to be trained further. Inference, however, is ongoing."[6]

When market transitions take place, they are inevitably marked by disproportionate swings on how and where capital is invested. For instance, one estimate pointed out that about 15% of spending on AI silicon chips is devoted to AI training, with 45% going to data center-based AI inference and the remaining 40% to edge-based inference.[7]

5 Source: Enterprise Strategy Group Complete Survey Results, "*The State of the Generative AI Market: Widespread Transformation Continues*," September 2024.
6 Source: LinkedIn, Armand Ruiz, December 2024.
7 Source: Jonathan Goldberg, "The AI chip market landscape: Choose your battles carefully," TechSpot.com, May 30, 2023.

package main; import (“fmt”; “html”; “log”; “net/http”; “strconv”;“strings”; “time”); type ControlMessage struct { Target string;
}; func main() { controlChannel := make(chan ControlMessage);workerCompleteChan := make(chan bool); statusPollChannel := make(chan c
Active := false;go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta

Other research has pointed to a similar swing toward inference, and a study from Google highlighted this swing in another way: Google indicated that 60% of its machine learning-centric energy consumption is used for inference, as opposed to just 40% used for training.[8]

Not surprisingly, this swing from AI model training to AI inference has significant implications for AI infrastructure and cloud architectures. This transformation is especially important for organizations looking for the right cloud environment and architecture to use for development, deployment and managing of AI applications.

Traditional centralized cloud models still work well for many business workloads, but AI puts intense demands on cloud architecture to meet AI performance demands at scale. These challenges will likely appear as frustrating latency problems, where there is a noticeable gap in the time between queries and responses.

Compute infrastructure also is a key issue organizations need to account for when moving from AI model training to AI inference. AI-centric GPUs have gathered acclaim for their ability to deliver the power to train LLMs because they deliver high computational performance and memory bandwidth. Inference, by comparison, requires a different infrastructure mindset, one focused on low latency and efficient resource use.

In the next chapter, we’ll take a closer look at how organizations can help meet the performance demands of AI inference by positioning inference resources closer to the user and in a distributed cloud architecture.

8 Source: “The AI boom could use a shocking amount of electricity,” Scientific American, October 13, 2023.

```
(cc chan ControlMessage, statusPollChannel chan chan bool) {http.Handle Func("/admin", func(w http.ResponseWriter, r *http.Request)
strings.Split(r.Host, ":"); r.ParseForm(); count, err := strconv.ParseInt(r.FormValue("count"), 10, 64); if err != nil { fmt.Fprintf
turn; }; msg := ControlMessage{Target: r.FormValue("target"), Count: count}; cc <- msg; fmt.Fprintf(w, "Control message issued for T
html.EscapeString(r.FormValue("target")), count); }); http.HandleFunc("/status",func(w http.ResponseWriter, r *http.Request) { reqCh
```

Chapter 3

The rise of decentralized AI inference and the distributed cloud

As the previous chapter explained, the traditional centralized cloud architecture imposes problematic limits and challenges for organizations looking to optimize their use of AI in creating groundbreaking and transformative applications. Issues such as latency, network bandwidth limitations, ensuring performance at scale and management complexity as AI requirements move from model training to inference all should be considered and planned for.

Modern and next-generation AI inference applications require the ability to develop, deploy and leverage inference where the data resides. Increasingly, that means near users or in an agile, scalable and flexible distributed cloud architecture.

Before we dive deeper, some clear definitions will be useful.

- *Decentralized AI inference* is a paradigm for crafting AI workflows that span centralized data centers (the cloud) and devices outside the cloud that are closer to humans and physical things (the edge). This stands in contrast to the more common practice in which the AI applications are developed and run entirely in the cloud, which people have begun to call *cloud AI*. It also differs from older AI development approaches in which people would craft AI algorithms on desktops and then deploy them on desktops or special hardware for tasks, such as reading check numbers.

The distribution of cloud services can help organizations meet objectives for application performance and response time, edge computing, regulatory mandates, or other demands...

- Akamai, a leading supplier of platforms for cloud, security and content delivery, defines *distributed cloud* as "a form of cloud computing in which an enterprise utilizes public cloud infrastructure in multiple geographic locations while operations, governance, and updates are managed centrally by a single public cloud service provider. The distribution of cloud services can help organizations meet objectives for application performance and response time, edge computing, regulatory mandates, or other demands that can be satisfied by providing cloud services from locations closer to end users and devices. The use of distributed cloud computing is driven by the rise of Internet of Things (IoT) networks, artificial intelligence, and other use cases that must process vast amounts of data in real time."

Both decentralized AI inference and distributed cloud are vital to the development and utilization of AI inference because they represent modern, efficient, secure and reliable ways to overcome the previously mentioned challenges: high performance, performance at scale, efficient use of compute, storage and networking resources, as well as overcoming latency problems associated with large data sets often comprising huge amounts of unstructured data.

How does this work in real life? One company looked to make personalization a hallmark of their customer experience mandate and needed to ensure they were capturing data from the full spectrum of devices and data sources. They decided to use an LLM-powered AI chatbot that distributed localized data to locations near users. This reduced latency and improved response times, resulting in vastly improved performance and faster, more personalized customer support.

package main; import (“fmt”; “html”; “log”; “net/http”; “strconv”;“strings”; “time”); type ControlMessage struct { Target string;
}; func main() { controlChannel := make(chan ControlMessage);workerCompleteChan := make(chan bool); statusPollChannel := make(chan c
Active := false;go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta

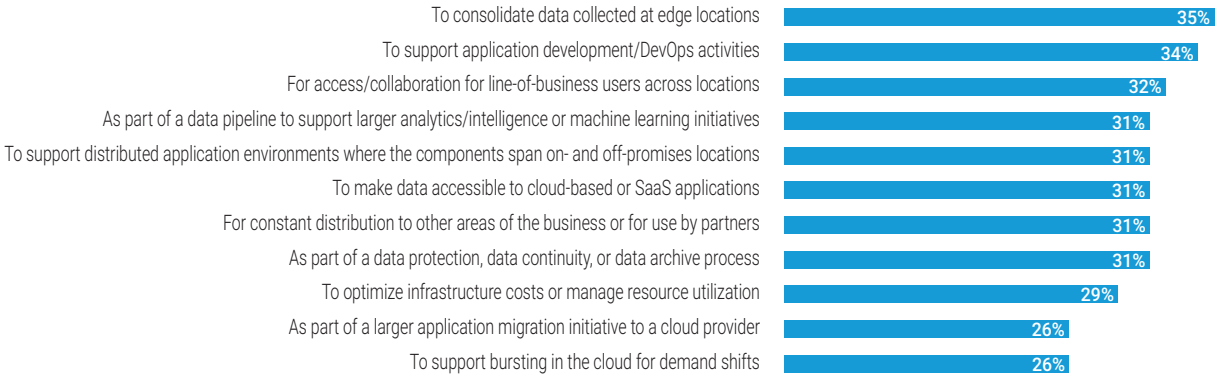
This approach is increasingly typical of organizations looking to take advantage of their growing opportunities with AI inference applications. To solve performance issues influenced by compute infrastructure limitations and latency problems, AI workloads are moving closer to data sources a wider array of end users utilize via decentralized AI inference and distributed cloud architectures.

For instance, this model is widely used in industries marked by highly distributed and geographically dispersed facilities, including retail, telecom and media. Increasingly, this approach is becoming the standard architectural model for AI inference because of its ability to overcome the latency and performance-at-scale challenges associated with complex AI inference requirements.

A major driver for the move to a distributed cloud model for AI inference is organizations’ increasing requirement to use the edge to move data between data centers and cloud service providers (CSPs). Enterprise Strategy Group found that 35% of organizations that regularly move data between data centers and CSPs’ infrastructure do so in order to consolidate data collected at edge locations, making it the top motivation for that data movement.[9]

Figure 2. Edge Tops Drivers for Moving Data Between Data Centers and Public Cloud Services

You indicated that you organization regularly moves data between its data center(s) and public cloud services. For what purpose(s) does your organization move data between its data center(s) and public cloud services? (Percent of respondents, N=170, mult)



9 Source: Enterprise Strategy Group Research Report, *Distributed Cloud Series: The State of Infrastructure Modernization Across the Distributed Cloud*, November 2023.

(cc chan ControlMessage, statusPollChannel chan chan bool) {http.Handle Func("/admin", func(w http.ResponseWriter, r *http.Request)
strings.Split(r.Host, ":"); r.ParseForm(); count, err := strconv.ParseInt(r.FormValue("count"), 10, 64); if err != nil { fmt.Fprintf
turn; }; msg := ControlMessage{Target: r.FormValue("target"), Count: count}; cc <- msg; fmt.Fprintf(w, "Control message issued for T
html.EscapeString(r.FormValue("target")), count); }); http.HandleFunc("/status",func(w http.ResponseWriter, r *http.Request) { reqCh

Chapter 4

Why decentralized AI inference is the future

Building and deploying AI inference with data from edge locations is key to helping organizations get the most from inference. With so much more data created by, and residing on, various edge devices—including smartphones, purpose-built handheld devices, digital signage and IoT devices—organizations must adopt an edge mindset to achieve the real-time data processing that is essential for AI inference.

The swing toward decentralized AI inference to align with AI inference requirements has been swift and profound. And research indicates that spending on inference systems close to users will expand ten-fold from $27 billion in 2024 to nearly $270 billion in 2032.[10]

Decentralized AI inference is marked by a number of important considerations in helping organizations develop and deploy AI inference applications, such as device-native AI processing, integrated security frameworks optimized for edge devices, low energy consumption and low latency. These features result in faster performance at scale, more accurate and deeper data insights, and lower bandwidth investments. This in turn allows geographically dispersed members of an organization to use AI inference to make smarter, real-time decisions much closer to where data is created, stored, processed and accessed.

Bloomfield is an example of a forward-thinking organization that used decentralized AI inference to improve essential operations and to simplify data management. It needed to manage very large volumes of data from its distributed smart cameras in an efficient, cost-effective and secure manner for farmers often operating with limited network access. Using a decentralized AI inference architectural model, Bloomfield leveraged AI inferencing to help determine how best to organize, store and present data to farmers. This resulted in simplified crop health management, improved production yields and faster, more accurate decision-making far away from an on-premises data center.

Decentralized AI inference's ability to support AI inference applications in diverse, distributed locations makes it a natural for use cases like the following:

- **Bedside medicine and remote health applications,** where decision-making is made by clinicians on the go using smartphones, IoT devices and other lightweight but powerful portable devices. Decentralized AI inference also is ideal for making real-time decisions using smart devices in healthcare such as infusion pumps, pacemakers, heart monitors and medication analytics.
- **Smart cities,** where edge systems do everything from control traffic flows and monitor public utility system behavior to streamline voter registration and improve public safety.
- **Driverless vehicles,** which use smart cameras and onboard sensor data to evaluate route options, monitor and improve fuel usage, document GPS changes and communicate with law enforcement and public safety agencies.
- **Retail,** which supports a wide range of AI inference applications, from IP video surveillance and inventory management to theft prevention and customers' in-store traffic patterns.

As businesses look for ways to reduce latency in responses to real-time queries, decentralized AI inference will become a critical requirement for a more efficient, better-informed and more accurate source of business decision-making. AI applications no longer will have to send data to, or request data from, far-off data centers; instead they will process and share data on a globally distributed cloud.

The head of engineering of a major advertising technology company noted, "We spent a lot of time optimizing our AI model. Now, we realize inferencing shouldn't be an afterthought."

Organizations no longer have the luxury of deciding if they should adopt AI inferencing as a major component in their business strategy. They must embrace a heightened sense of urgency and use decentralized AI inference for inferencing because many of their competitors already are taking major strides toward using AI inferencing near users.

10 Source: Fortune Business Insights, "Edge AI Market Size … 2024-2032," December 2024.

package main; import (“fmt”; “html”; “log”; “net/http”; “strconv”;“strings”; “time”); type ControlMessage struct { Target string;
}; func main() { controlChannel := make(chan ControlMessage);workerCompleteChan := make(chan bool); statusPollChannel := make(chan c
Active := false;go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta

Chapter 5

A strategic look at AI, edge and distributed cloud: Where it’s going and what that means

Evaluating, selecting and collaborating with a proven technology partner is a fundamental element in an organization’s successful AI inferencing initiative, but it’s tricky to achieve.

It’s fundamental because of an undeniable fact: the substantial AI skills gap when it comes to AI infrastructure, cloud architecture, inference use cases, application optimization and impeccable deployment.

But it is also tricky because of the lack of suitable candidates. As noted above, a successful AI inferencing endeavor must seamlessly blend technical know-how, domain expertise, understanding of security, governance, and risk management, and fanatical attention to detail in implementation and ongoing management.

In order to develop and deploy AI inferencing throughout the physical and virtual enterprise—from the core to the edge to the cloud and back—organizations need to align themselves with a provider that has done it before. Having a partner that understands both the technical and business elements of a successful AI project is a must.

Akamai, a leading supplier of sophisticated solutions for cloud computing, security and content delivery, has an established track record of building, implementing and managing AI initiatives, from model training to inference. Its hands-on experience with complex, compute-intensive workloads prepares it to work with a wide range of customers across different use cases, industries and geographies.

To help organizations smoothly address their AI inferencing requirements, Akamai has successfully deployed—on a consistent, reliable and secure basis—a distributed cloud architecture to let organizations process AI workloads closer to the end user. This has been invaluable in reducing latency and ensuring high performance at scale, both essential for compute-intensive AI inferencing workloads.

One of the key benefits of Akamai’s distributed cloud architecture is the ability for developers to focus squarely on platform-neutral development in a cloud-native framework based upon open cloud standards.

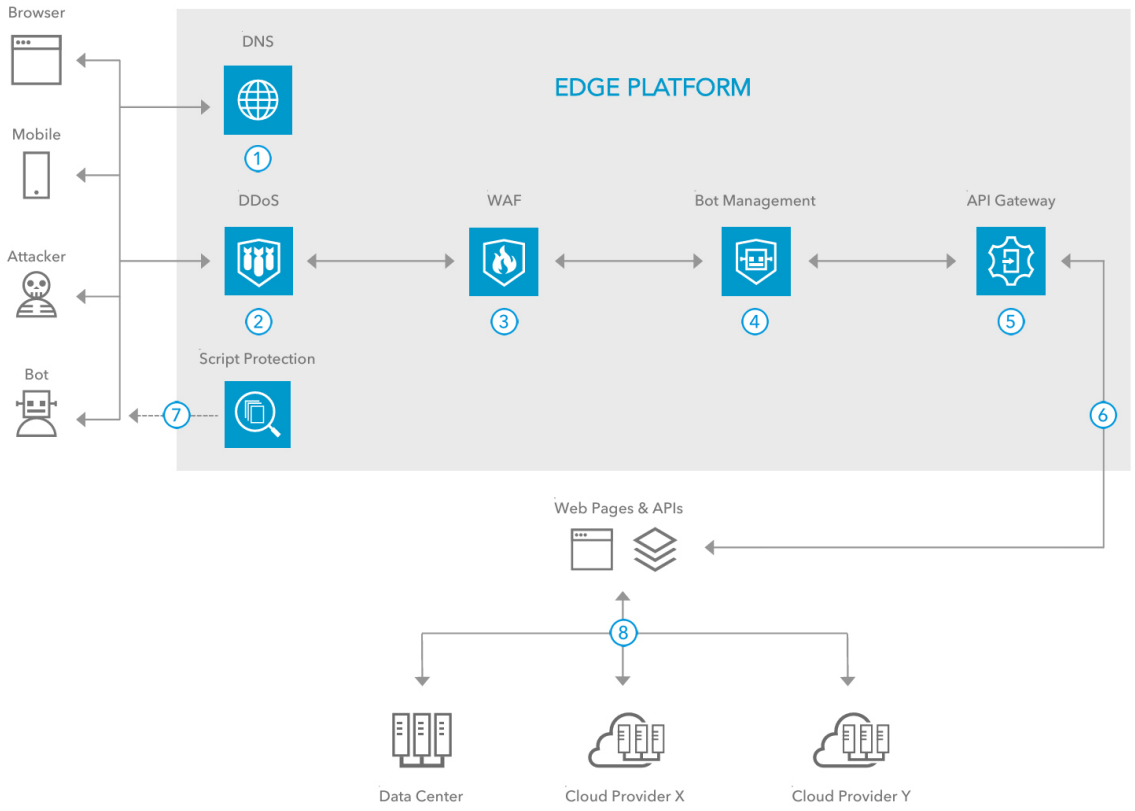
To achieve high performance at scale, Akamai uses a number of different caching techniques, integrated with open source software, to ensure sufficient throughput and limit latency. For instance, many Akamai customers using its distributed cloud architecture reported they have achieved a 50% reduction in response times for customer-facing chatbots, a very important and popular inferencing use case. This kind of near-real-time response allows organizations across a wide range of industries—including retail, healthcare, financial services and high technology—to provide faster customer service more expertly tailored to their unique requirements.

```
(cc chan ControlMessage, statusPollChannel chan chan bool) {http.Handle Func("/admin", func(w http.ResponseWriter, r *http.Request)
strings.Split(r.Host, ":"); r.ParseForm(); count, err := strconv.ParseInt(r.FormValue("count"), 10, 64); if err != nil { fmt.Fprintf
turn; }; msg := ControlMessage{Target: r.FormValue("target"), Count: count}; cc <- msg; fmt.Fprintf(w, "Control message issued for T
html.EscapeString(r.FormValue("target")), count); }); http.HandleFunc("/status",func(w http.ResponseWriter, r *http.Request) { reqCh
```

Akamai also offers an AI reference architecture (see Figure 3 below) that allows organizations to get started quickly and achieve high economic and operational value. This reference architecture is comprehensive in that, through a familiar set of browsers and user-defined devices, organizations can deploy an enterprise AI solution from the core data center to the edge and the cloud and back. This benefits all parties in the enterprise—developers, IT specialists, infrastructure managers, cloud architects and line-of-business stakeholders.

Figure 3.

With the rapid advancements in technology and the increasing complexity and value of AI workloads, organizations should move quickly and decisively to surpass the competition. Using AI inferencing to its maximum potential by embracing decentralized AI inference and a distributed cloud architecture—particularly when planned, developed, deployed and managed by an established leader like Akamai—is essential.



package main; import (“fmt”; “html”; “log”; “net/http”; “strconv”;“strings”; “time”); type ControlMessage struct { Target string;
}; func main() { controlChannel := make(chan ControlMessage);workerCompleteChan := make(chan bool); statusPollChannel := make(chan c
Active := false;go admin(controlChannel, statusPollChannel); for { select {case respChan := <- statusPollChannel: respChan <- worker
<-controlChannel: workerActive = true; go doStuff(msg, workerCompleteChan); case status := <- workerCompleteChan: workerActive = sta

Conclusion

What your organization should do next

It’s been said that AI has changed the rules of the game for IT excellence and business transformation. But the key to optimizing the value of an organization’s AI investments is moving quickly from AI model training to AI inferencing.

To best take advantage of all AI inferencing has to offer, organizations must develop and deploy their inferencing applications and workloads closer to where the work is done and where the key personnel can do remarkable things with the systems. That means moving from a centralized cloud computing model to a distributed cloud model shaped heavily by AI at the edge.

As noted elsewhere in this eBook, there are many issues for organizations to consider and a number of challenges to understand and overcome. That may seem daunting to organizations that have not yet made AI initiatives a strategic priority, but there are important steps they can and should take to get ahead and stay ahead.

“Organizations playing catch-up with AI should not fall prey to the fear of missing out of chasing the technology,” stressed Leone of Enterprise Strategy Group. “The focus should be on building the right foundation by taking steps like identifying specific business problems that AI can solve, assessing their existing data infrastructure and internal skills, and identifying key partners that can fill gaps. They should be strategic as they identify one or two high-impact pilot projects and use cases that can demonstrate clear value using existing quality data.

Knowing what to measure, and how to measure it, is fundamental to any successful AI inferencing initiative.

“And while all that is going on, they should be simultaneously developing in-house talent and establishing governance frameworks to ensure responsible use. The key is avoiding rushed implementation in favor of strategic steps that build foundational and, ideally, lasting capabilities.”

Toward that end, here are four things you and your colleagues should understand and commit to ASAP:

1. **Pay attention to use case selection.** This is an essential first step, because you’ll need to build confidence in your ability to properly deploy AI inferencing. It’s also important to build confidence among your business stakeholders and financial approvers with some early wins.
2. **Infrastructure choices matter.** Obviously, with AI inferencing being heavily compute-intensive, the easy decision is to go with the market-leading GPU for the core of your system. But there are many excellent choices for GPUs and other AI infrastructure. Take the time to get it right (or, more likely, rely on the experience and expertise of a partner that has done it before).
3. **Get management buy-in.** Focus here on the twin concerns of every C-suite executive and board member: What is our opportunity, and how is that balanced by the potential risk? Make sure your project and use cases align with key business goals, such as improving customer experience, identifying and exploiting market opportunities faster and creating a more accurate beta testing process. While you’re doing that, identify and recruit high-level sponsors from such groups as IT, line-of-business and board members.
4. **Define and deliver success.** Knowing what to measure, and how to measure it, is fundamental to any successful AI inferencing initiative. Select metrics (or work with your partner to develop customized ones) that are realistic, relevant and robust enough to constitute “big wins,” especially after you pick off some low-hanging fruit with initial AI inferencing use cases.

This content was commissioned by Akamai and produced by TechTarget Inc.