InfoBrief, sponsored by NVIDIA
MAY 2023

REGION FOCUS: WORLDWIDE

As Businesses Pivot to AI in Production, They Need Enterprise-Grade Solutions

Siloed AI that is not integrated with the datacenter and/or the cloud is not a sustainable way to infuse AI into the organization

Peter Rutten
Research Vice-President, Infrastructure Systems, Platforms and Technologies Group,
Performance-Intensive Computing Solutions Global Research Lead, IDC

Navigating this InfoBrief

Click on titles or page numbers to navigate to each section.

Executive Summary 3

Drivers for Investing in Artificial Intelligence 4

The Shift from AI Training … to AI Inferencing 5

Stages of AI Development 6

Time to Train an AI Model 7

Time to Put an AI Model in Production 8

The Large Variety of AI Use Cases 9

Investing in GPUs Helps Speed Up the Process 10

Businesses Have Invested Aggressively in GPU-Based Servers 11

Ease of Implementation and Satisfaction Levels for GPUs Are Very High 12

The Worldwide Accelerated Server Market Has Been Growing Rapidly 13

AI Still Needs to Become Better Integrated Throughout the Organization 14

Companies Want Enterprise-Grade AI Solutions 15

AI Workloads Are Running in Various Deployment Scenarios 16

Critical Improvements Required 17

How Businesses Mitigate Challenges 18

Essential Guidance 19

About the IDC Analyst 20

Message from the Sponsor 21

Executive Summary

- Organizations expect significant benefits from AI, and IDC is seeing a major shift from AI model training to AI inferencing (taking an AI model into production), but there are challenges.
- Businesses urgently want to accelerate the time from initial planning to being in production with AI, especially the time to build an AI model; furthermore, businesses have multiple AI use cases in development or production that require near real-time processing. Investing in GPUs helps speed up these processes, and many businesses do so — but that's not enough.
- AI is still not well integrated; organizations are struggling to move from a siloed and low-efficiency environment for AI to a more streamlined and integrated environment. They are looking for greater integration with data repositories, vendor support, a full AI stack, greater security, and better management, not in the least because they plan to run their AI workloads on a combination of cloud, on-premises, and edge.
- Currently, most companies feel they need to overcome these challenges themselves through training and AI strategy development. IDC believes that organizations should also look to their vendors for complete enterprise-grade AI solutions to mitigate these challenges.

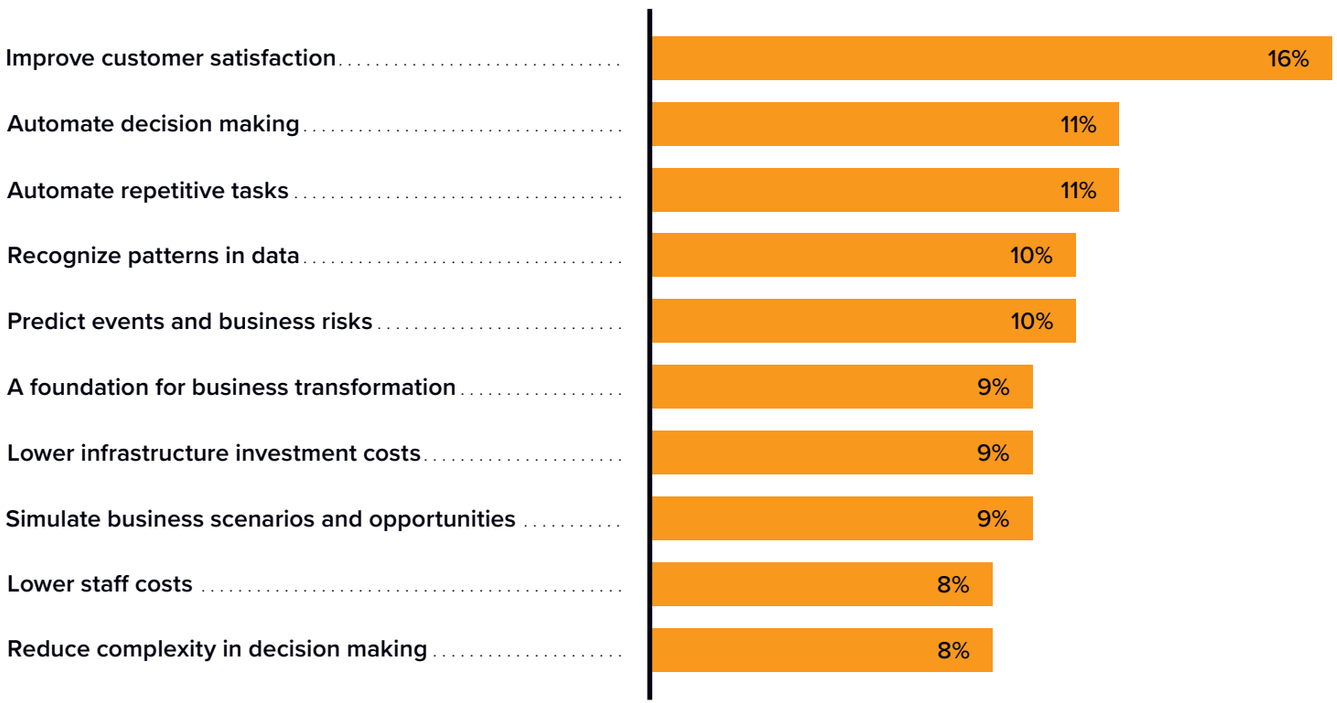
Drivers for Investing in Artificial Intelligence

Organizations expect significant benefits from AI; the most anticipated benefit is greater customer satisfaction.

But **AI model training** and the pivot to AI inferencing (taking a trained model into production) **pose significant challenges.**

Expected Benefits from Artificial Intelligence

(% of Respondents)

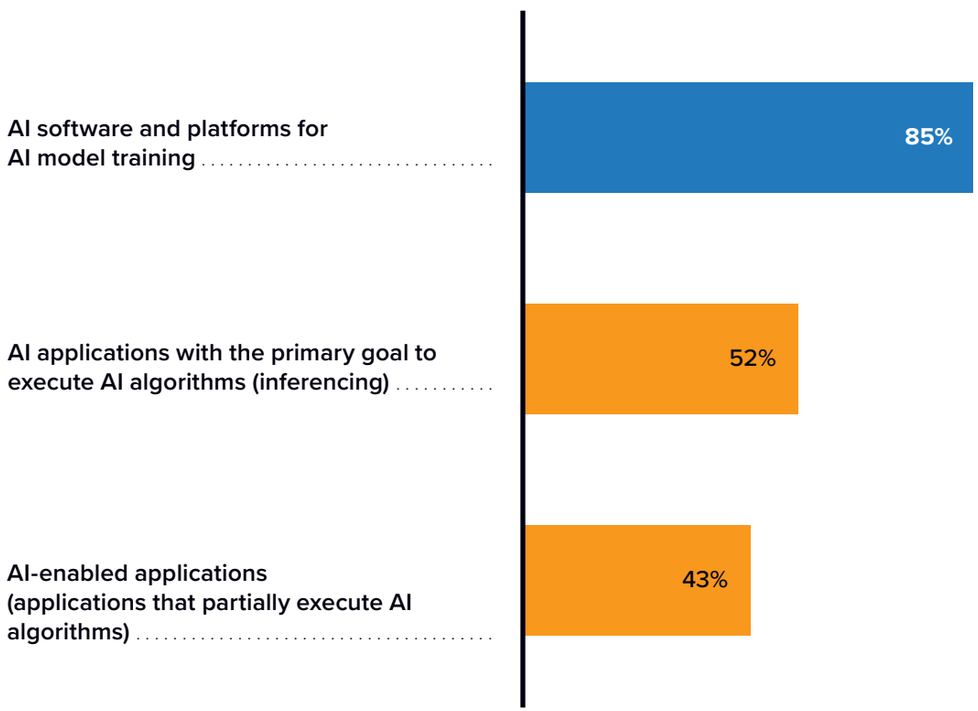


Source: IDC's *AI InfrastructureView Survey, 2021; n=2,000*

The Shift from AI Training … to AI Inferencing

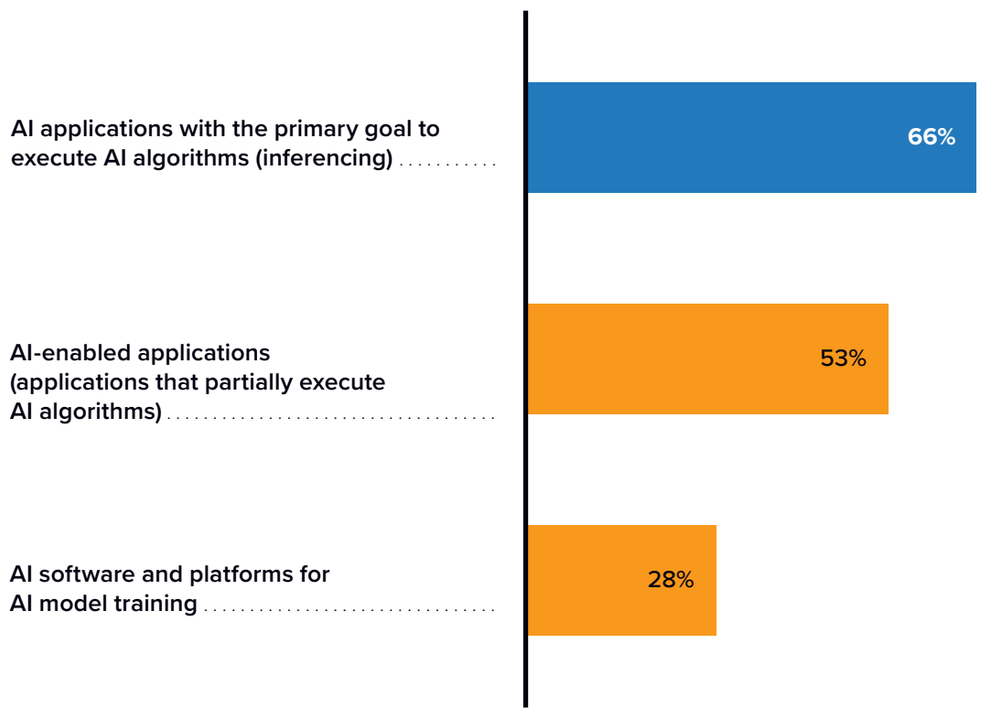
In early 2022, the majority (**85%**) of organizations were still using AI software and platforms for AI model training

(% of Respondents)



By mid-2023, the majority (**66%**) anticipates they will be focused on AI inferencing - executing AI models in production, even as they continue to train new AI models

(% of Respondents)

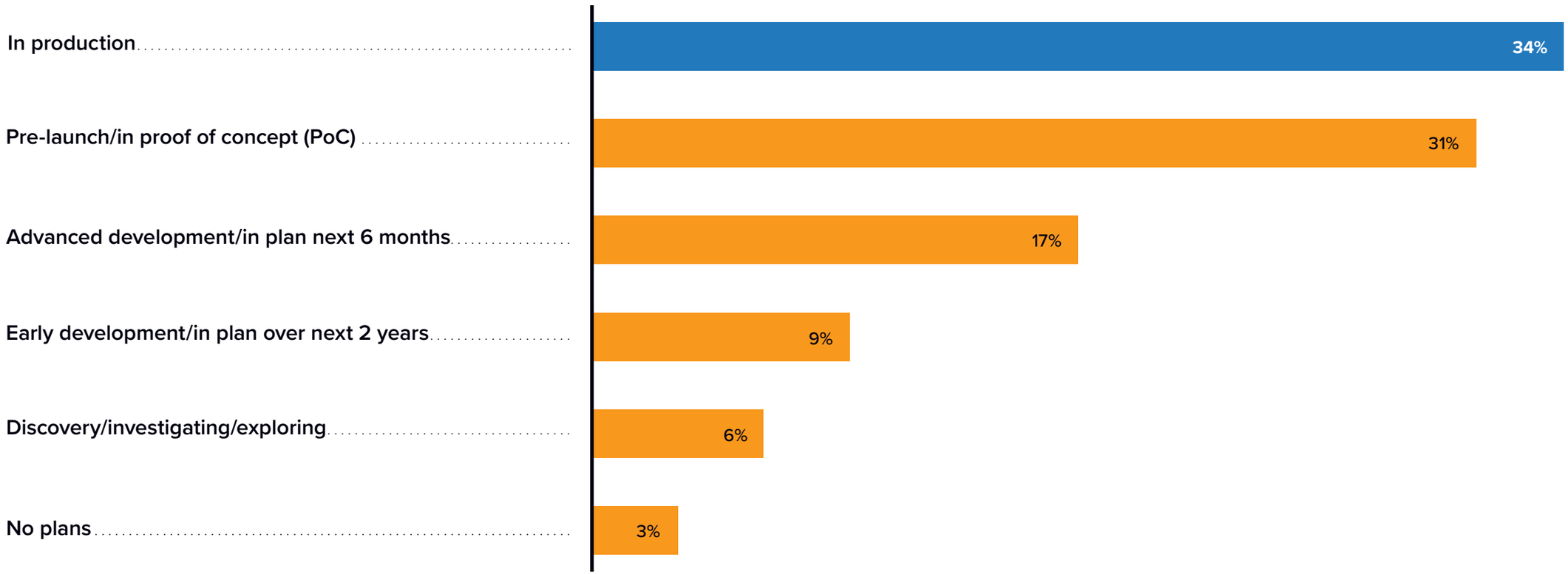


Source: IDC's *AI InfrastructureView Survey, 2021*

Stages of AI Development

Businesses urgently want to accelerate the time from initial planning to being in production with AI, but only **34%** are there today

(% of Respondents)



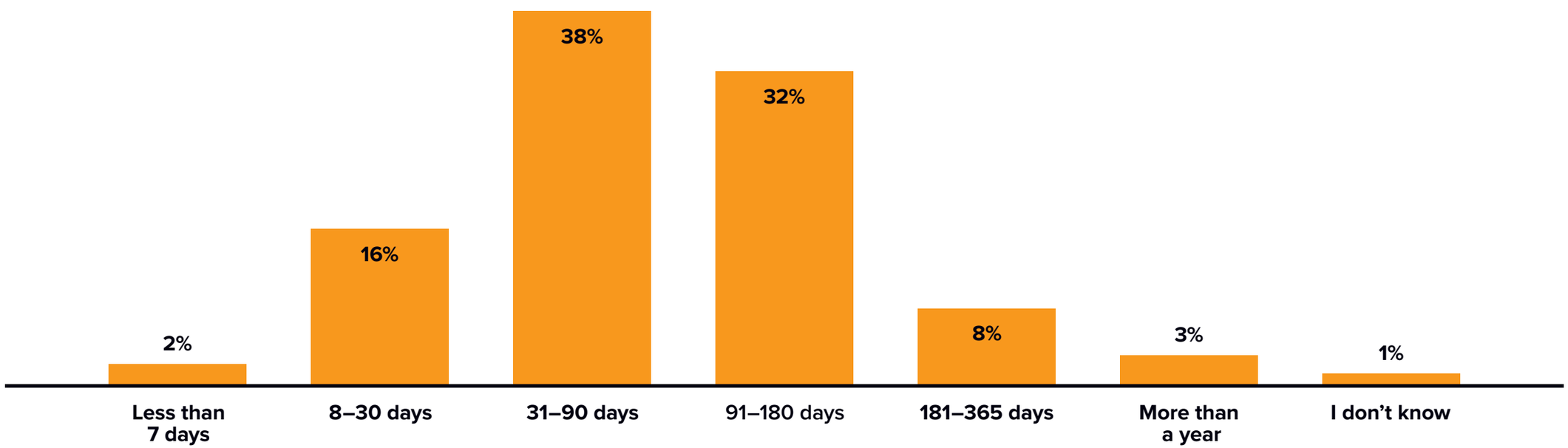
Source: IDC's *AI InfrastructureView Survey, 2021*

Time to Train an AI Model

The time to build an AI model especially needs to be addressed; for almost half of organizations this takes more than **90 days**.

Time Spent Building an AI Model

(% of Respondents)



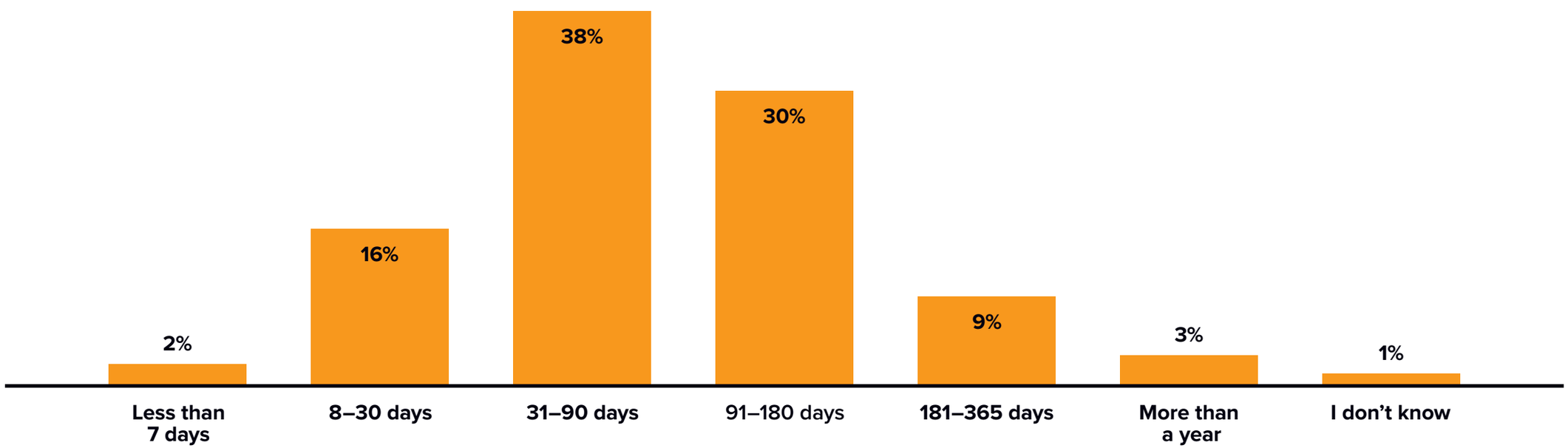
Source: IDC's *AI InfrastructureView Survey, 2021*

Time to Put an AI Model in Production

Similarly, the time to put an AI model into production takes too long.

Time Spent Taking a Completed AI Model into Production

(% of Respondents)



Source: IDC's *AI InfrastructureView Survey, 2021*

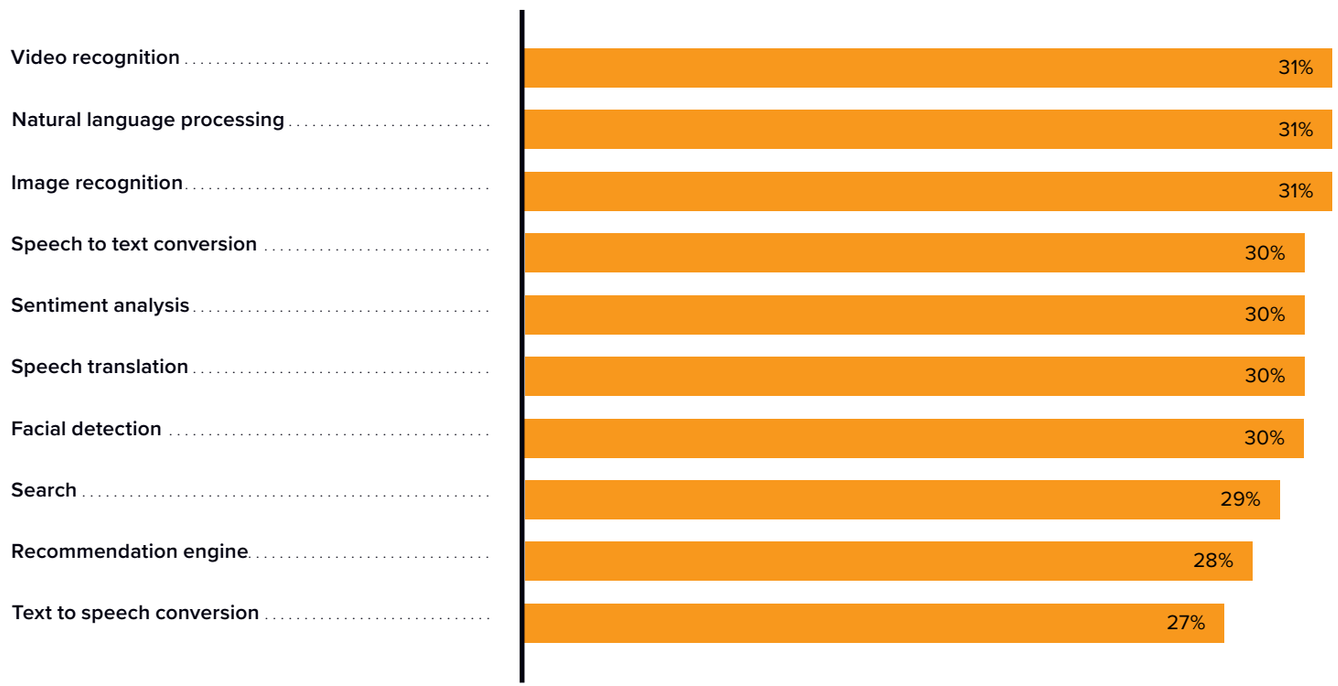
The Large Variety of AI Use Cases

There are many use cases for AI that are enabling innovative processes and products.

Organizations have several of them in development or production, with AI applications calling on not one but several AI models, often in near real time.
This requires a well orchestrated high-performance environment.

Most Popular AI Use Cases

(% of Respondents)



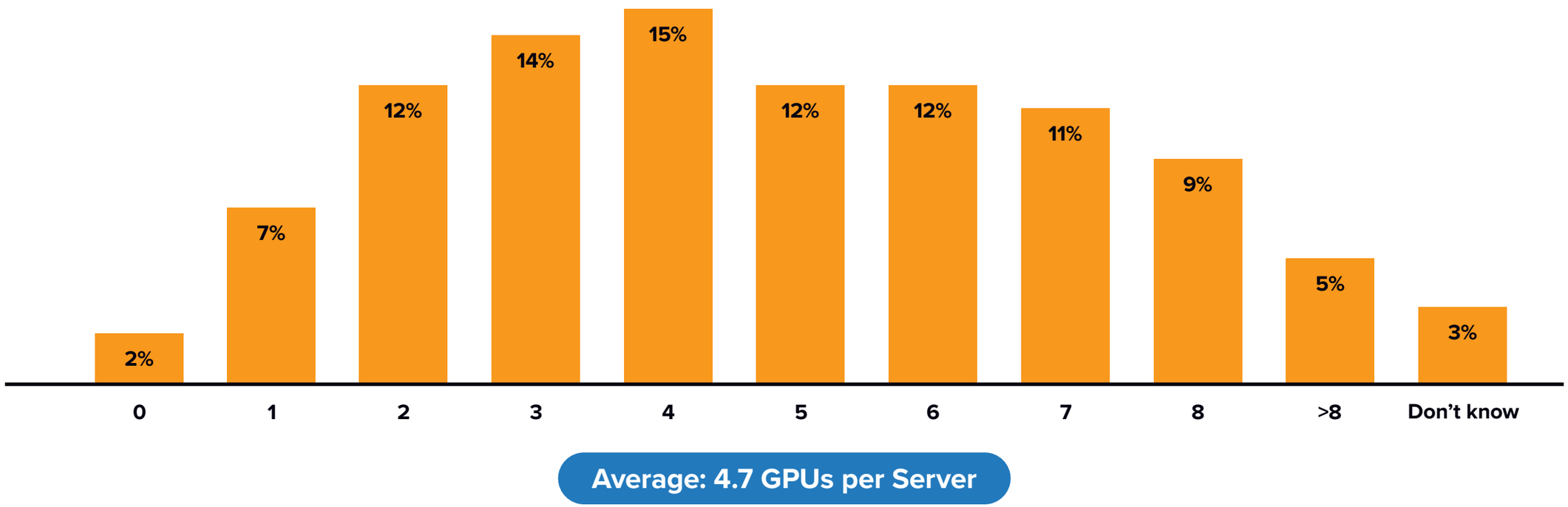
Source: IDC's *AI InfrastructureView Survey, 2021*

Investing in GPUs Helps Speed Up the Process

Time spent on training an AI model and taking it into production is directly correlated to compute capacity – which is why businesses are investing heavily in GPU acceleration. Today, businesses run servers for AI with on average **4.7 GPUs per server.**

Number of GPUs per Server for AI

(% of Respondents)



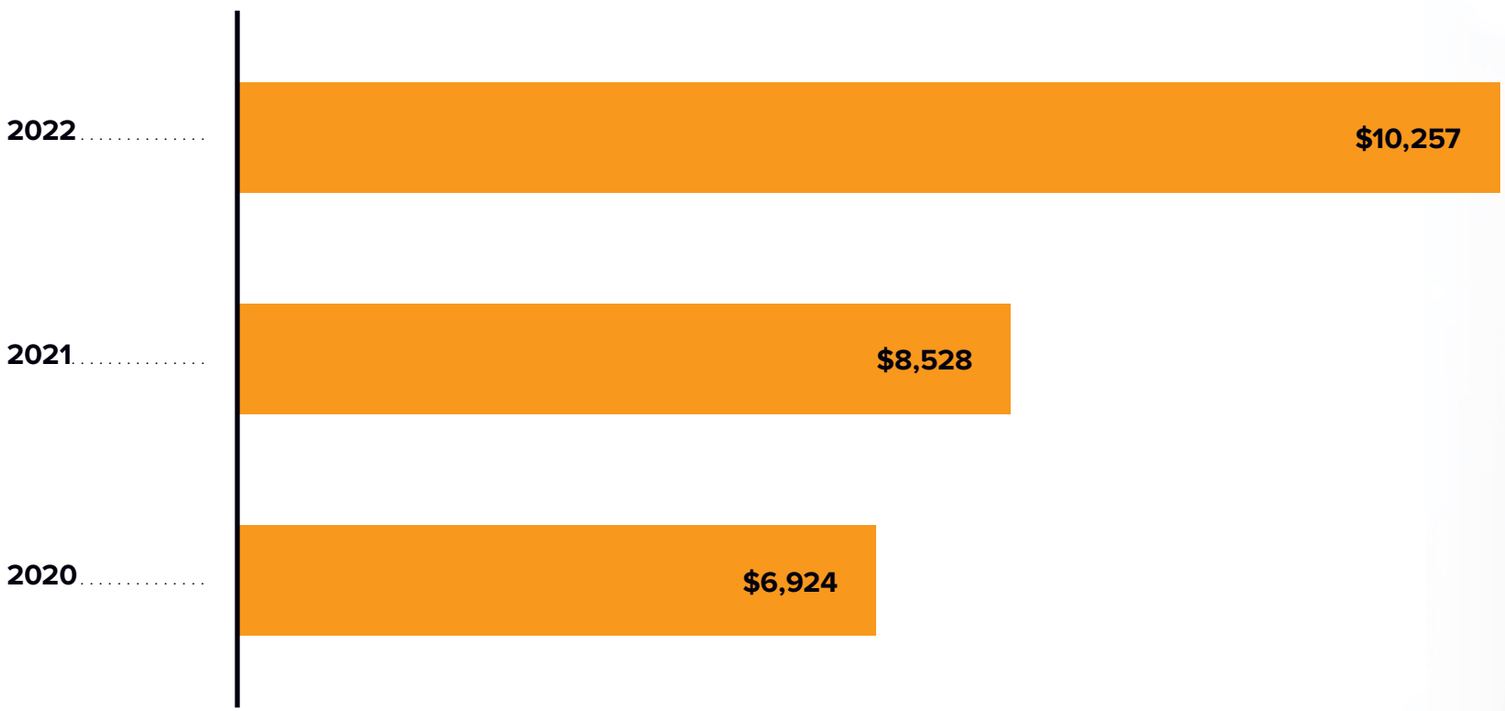
Average: 4.7 GPUs per Server

Source: IDC's *AI InfrastructureView Survey, 2021*

Businesses Have Invested Aggressively in GPU-Based Servers

Worldwide Investments in GPU-Based Servers, $M (2020–2022)

(US$)



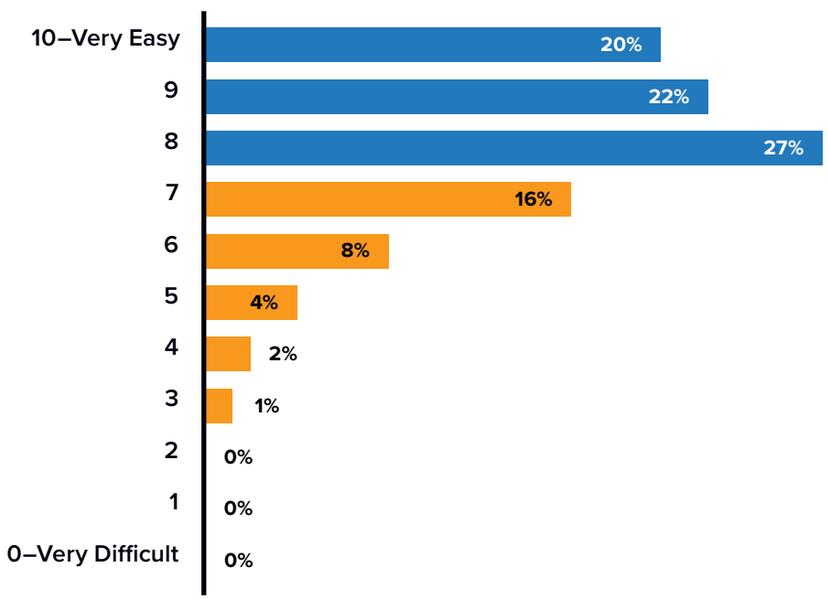
Source: IDC's *AI InfrastructureView Survey, 2021*

Ease of Implementation and Satisfaction Levels for GPUs Are Very High

A large majority of organizations find GPUs easy or very easy to implement and the satisfaction levels with GPUs for acceleration are extremely high.

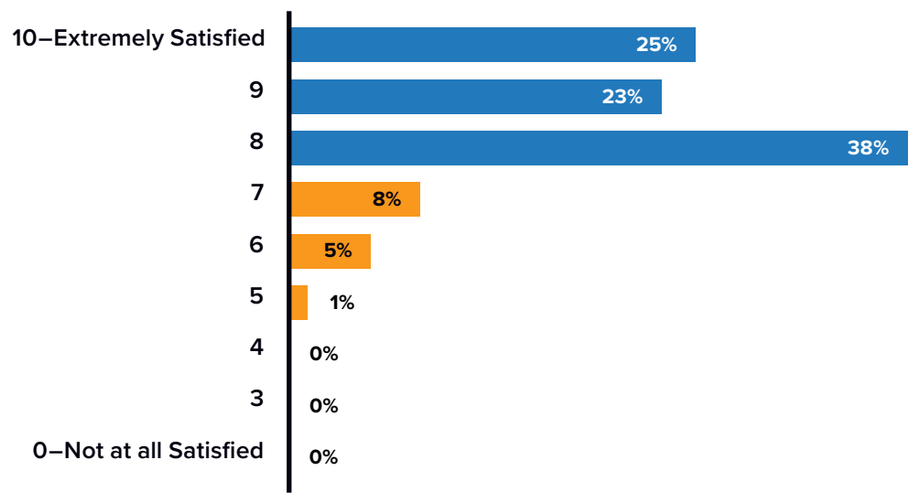
Ease of GPU Implementation (1–10)

(Rating 1–10)



Satisfaction Levels with GPUs (1–10)

(Rating 1–10)



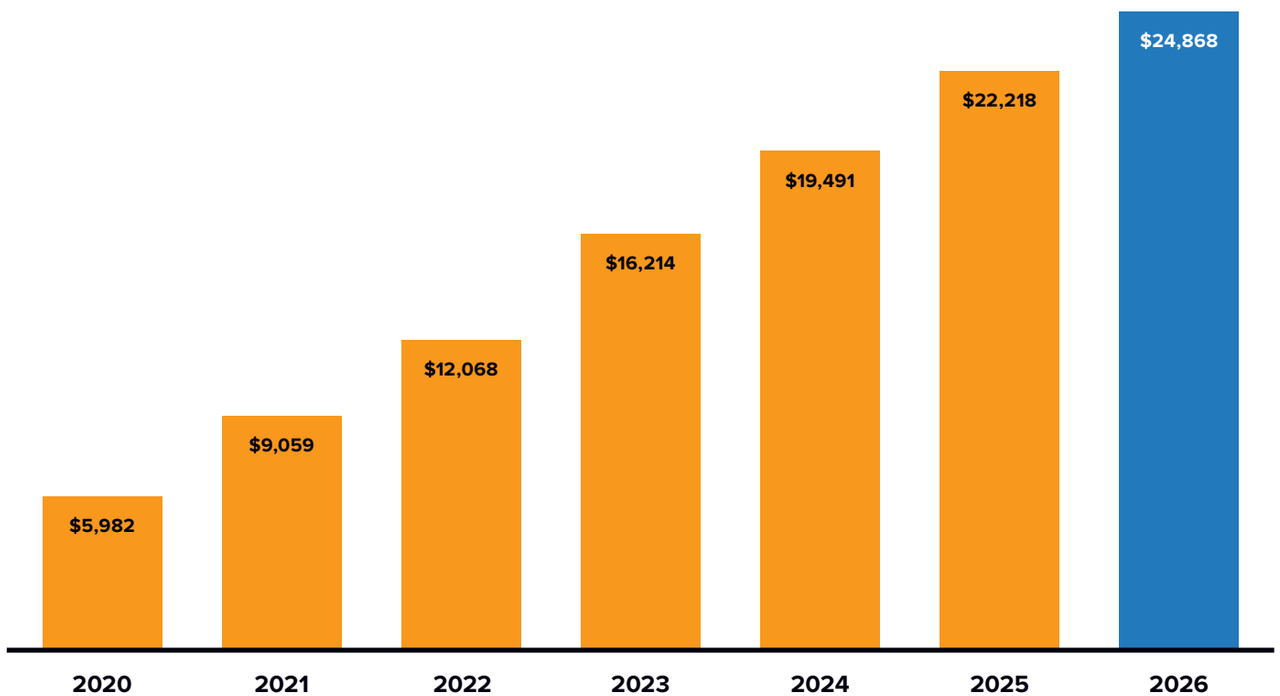
Source: IDC's *AI InfrastructureView Survey, 2021*

The Worldwide Accelerated Server Market Has Been Growing Rapidly

As a result of GPU implementation and satisfaction, **the worldwide accelerated server market is projected to reach $25B by 2026**, with all major server vendors offering comprehensive AI server solutions that leverage multiple GPUs.

Worldwide Accelerated Server Infrastructure Value ($M) for AI

(US$)



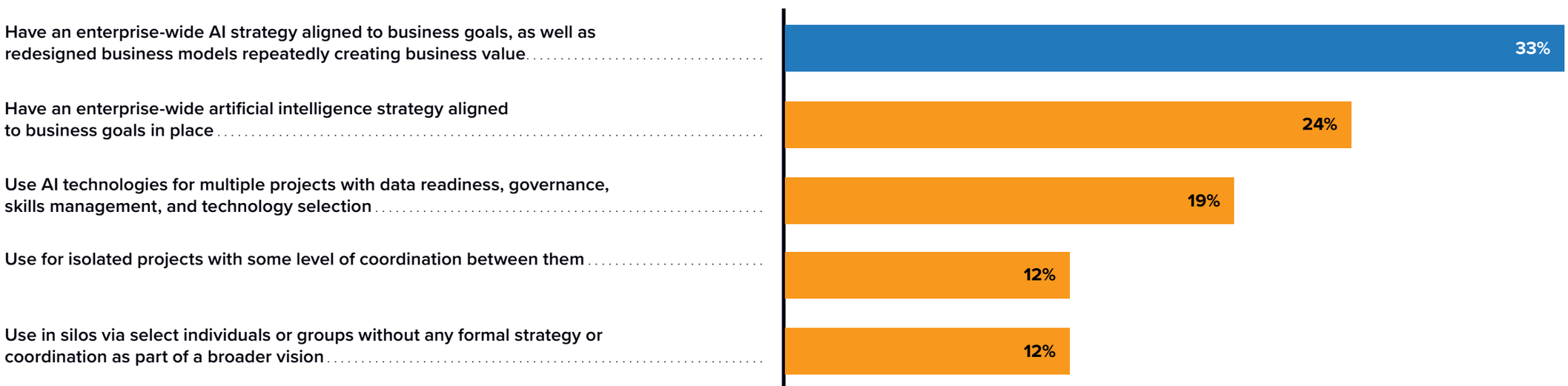
Source: IDC's *AI InfrastructureView Survey, 2021*

AI Still Needs to Become Better Integrated Throughout the Organization

Many organizations are struggling to move from a siloed and low-efficiency environment for AI to a more streamlined and integrated environment. Only **33%** have an enterprise-wide AI strategy aligned to business goals and redesigned business models to repeatedly create business value and maximize efficiency.

Which of the following best describes the maturity of adoption of AI technologies among your organization?

(% of Respondents)



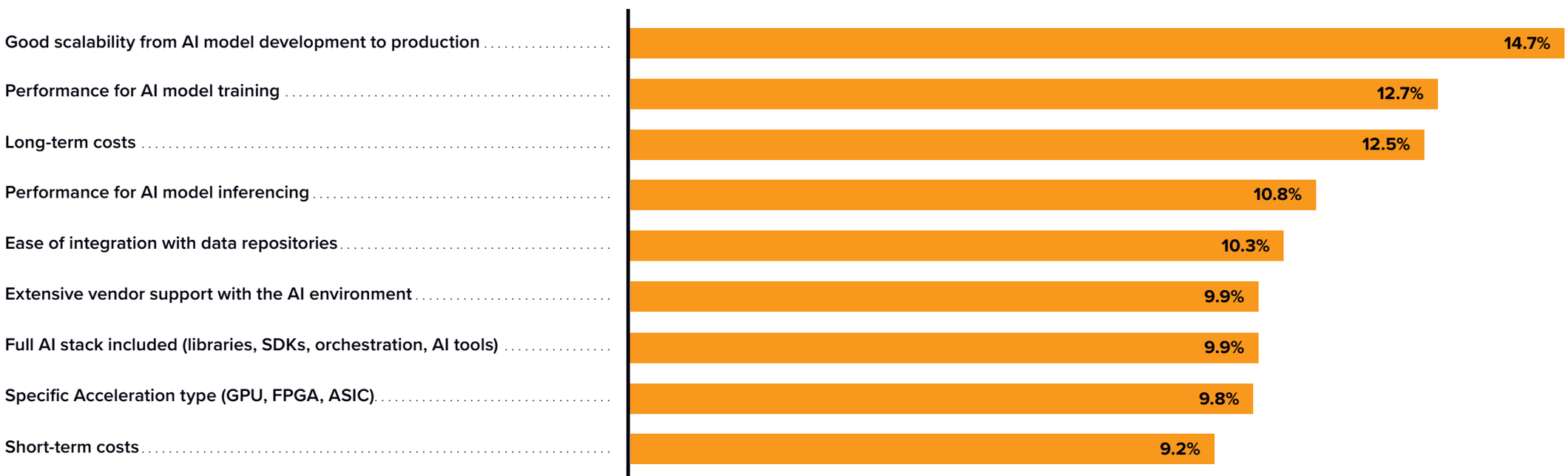
Source: IDC's *AI InfrastructureView Survey, 2021*

Companies Want Enterprise-Grade AI Solutions

Businesses are increasingly looking for more than just performance, scalability, and low cost for their AI infrastructure. They want AI infrastructure to be delivered with solutions to easily integrate with data repositories, extensive vendor support, and a full AI stack on top of the acceleration solution they choose.

What Businesses Want in AI Infrastructure

(% of Respondents)



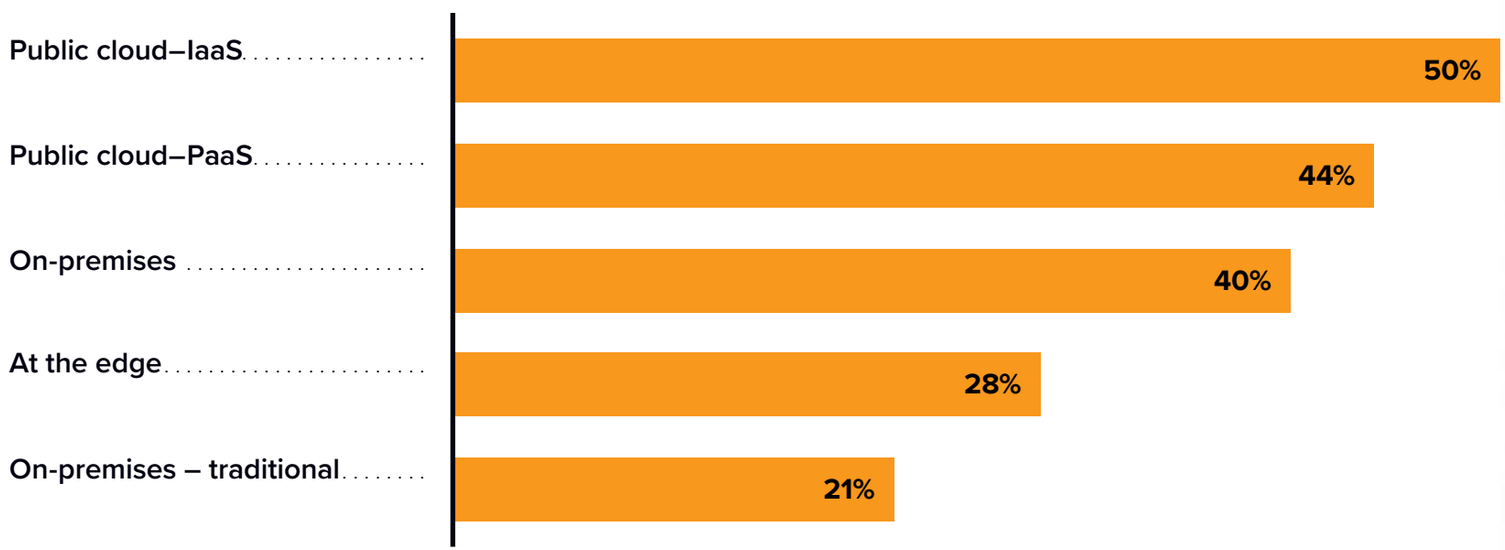
Source: IDC's *AI InfrastructureView Survey, 2021*

AI Workloads Are Running in Various Deployment Scenarios

Businesses plan to run their AI workloads not just on-premises and not just in the cloud but in a combination of cloud, on-premises, and edge – this further drives them to look for end-to-end enterprise-grade solutions across these deployment scenarios.

Locations for AI Workloads

(% of Respondents)



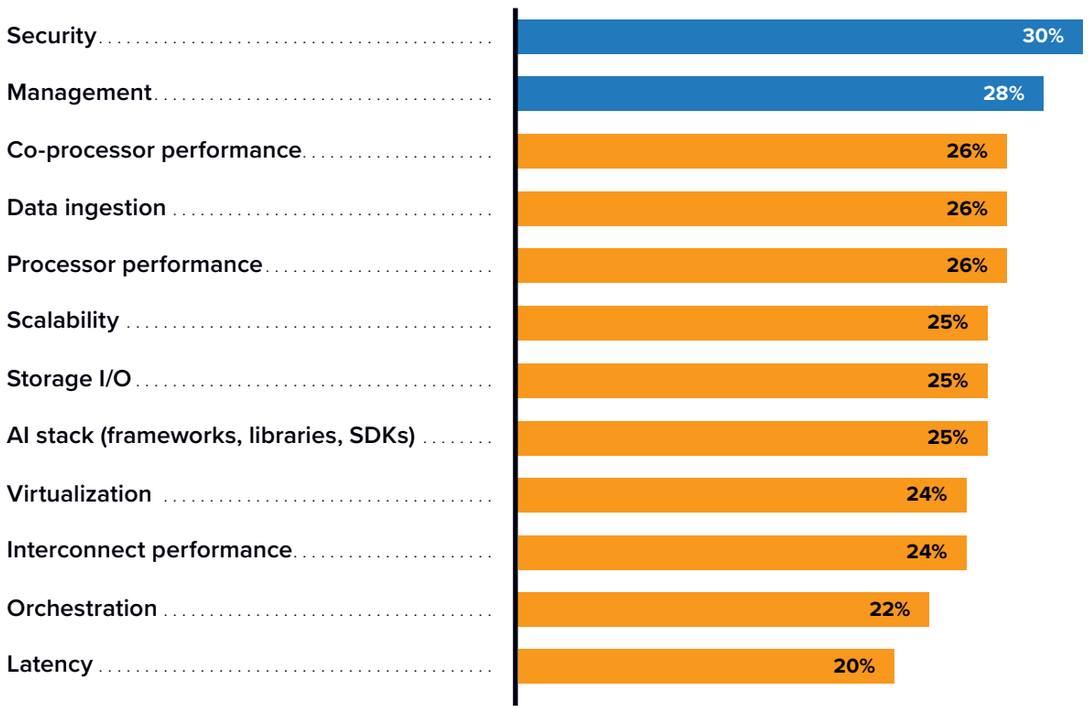
Source: IDC's *AI InfrastructureView Survey, 2021*

Critical Improvements Required

As they improve their AI performance, businesses are also looking for critical improvements across the AI stack, such as **better security and management**; **71%** have compliance, security, and privacy requirements for AI.

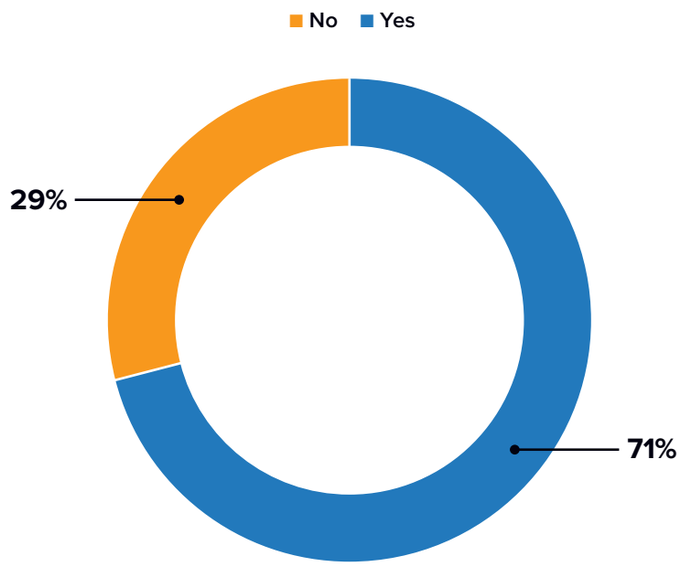
Areas for Improvement

(% of Respondents)



Businesses with Compliance, Security, or Privacy Requirements for AI

(% of Respondents)



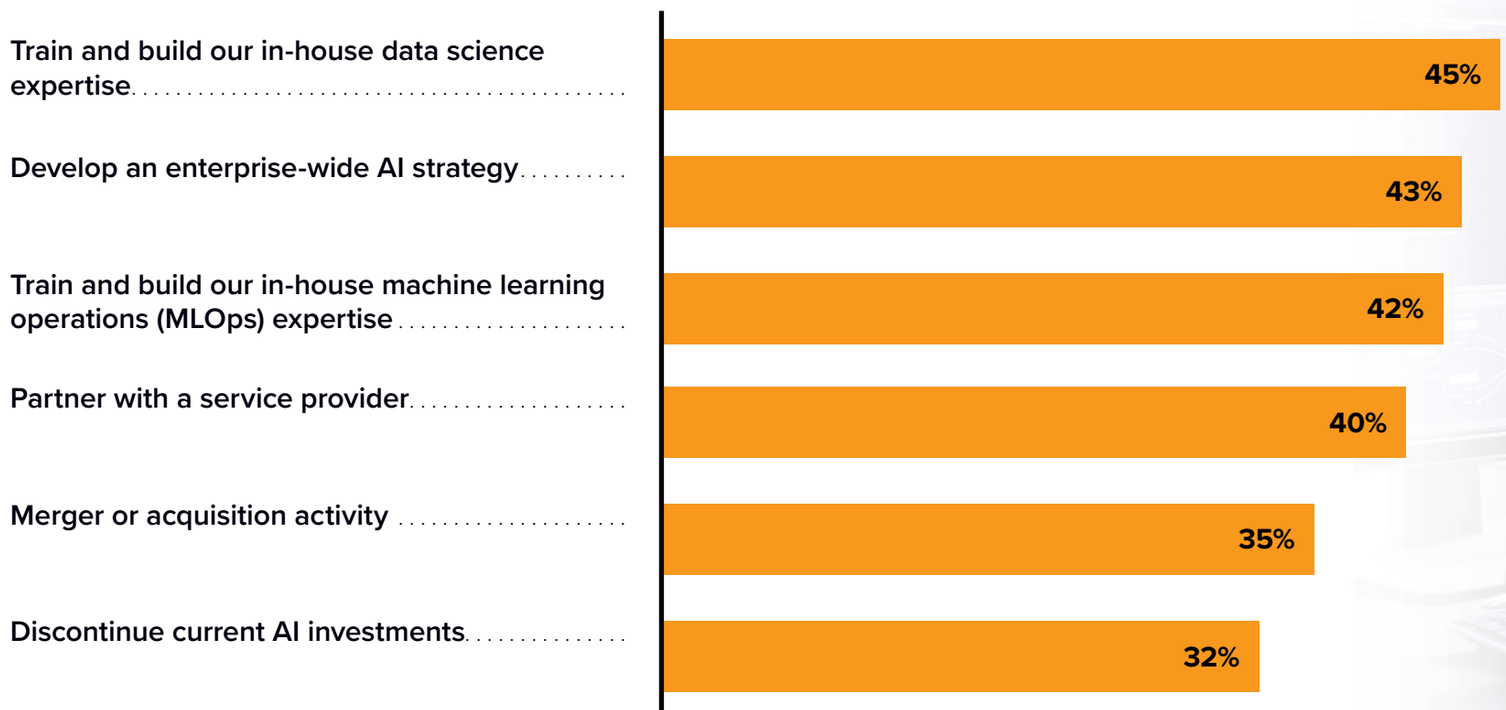
Source: IDC's *AI InfrastructureView Survey, 2021*

How Businesses Mitigate Challenges

Currently, most companies feel they need to overcome the challenges themselves through training and AI strategy development.

What Companies Are Doing to Address Challenges

(% of Respondents)



Source: IDC's *AI InfrastructureView Survey, 2021*

Essential Guidance

IDC believes that organizations should also look to their vendors for help with their AI challenges. A complete enterprise-grade solution should provide:

Compute and acceleration for high performance

Abstraction layers (bare metal, virtualization, containerization, orchestration)

Popular AI frameworks

AI libraries and SDKs

Critical security features

Data management solutions

Cloud integration capabilities

About the IDC Analyst

Peter Rutten
Research Vice President, Infrastructure Systems, Platforms and Technologies Group;
Global Research Lead, Performance-Intensive Computing Solutions, IDC

Peter Rutten is a research vice president within IDC's Worldwide Infrastructure Practice, covering research on computing platforms. He is IDC's global research lead on performance-intensive computing solutions and use cases. This includes research on high-performance computing (HPC), artificial intelligence (AI), and big data and analytics (BDA) infrastructure and associated solution stacks. In this role, Peter leads three IDC programs: High-Performance Computing Trends and Strategies, High-Performance Computing as a Service, and Infrastructure Trends and Strategies: Artificial Intelligence and Analytics. His coverage of performance-intensive computing includes supercomputing as well as institutional and mainstream high-performance computing; high-end, accelerated, in-memory, and heterogeneous computing infrastructure systems, platforms, and technologies.

More about Peter Rutten

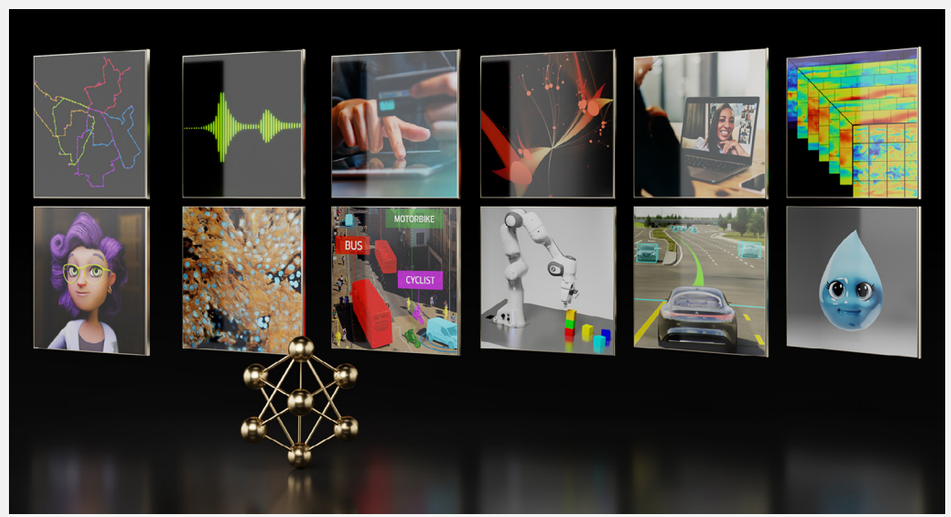
Message from the Sponsor

NVIDIA

The Accelerated AI Platform

Solve new challenges while increasing operational efficiency with NVIDIA AI Enterprise, an end-to-end, secure, cloud-native suite of AI software.

- Reduce development time from months to hours.
- Ease the transition from pilot to production with guidance from NVIDIA Enterprise Support.
- Develop and deploy anywhere on the cloud, data center, and edge.
- Industry leading performance for increased efficiency and costs savings.
- Kick-start your AI journey with immediate, short-term access to NVIDIA AI Enterprise – for free.



GET STARTED

IDC Custom Solutions

This publication was produced by IDC Custom Solutions. As a premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications, and consumer technology markets, IDC's Custom Solutions group helps clients plan, market, sell, and succeed in the global marketplace. We create actionable market intelligence and influential content marketing programs that yield measurable results.

IDC

IDC Research, Inc.
140 Kendrick Street, Building B, Needham, MA 02494, USA
T +1 508 872 8200

@idc @idc idc.com