January 2026
Business Value White Paper
Sponsored by Google Cloud

The Business Value of AI Hypercomputer

Dave McCarthy
Research Vice President,
Cloud and Edge Services,
Worldwide Infrastructure Research, IDC

Megan Szurley
Business Value Manager,
Business Value Strategy Practice, IDC

Table of Contents

Click any title to navigate directly to that page.

Business Value Highlights

Click to jump to related content. Click "Return to Highlights" to get back to this page.

353% three-year return on investment

28% lower costs

36% less time spent on coding

36% more time spent on innovation

55% more efficient IT infrastructure management teams

48% quicker deployment of compute resources

28% more efficient networking teams

27% more productive developers

37% less time spent on testing applications

Business Value Highlights continued

67% less application/workload unplanned downtime

27% lower application latency

49% quicker unplanned downtime resolution

22% quicker to go to market

38% quicker queries

40% more efficient security team

18% more productive compliance team

Executive Summary

The escalating demands of AI are creating a compelling case for strategic investment in cloud infrastructure, shifting the focus beyond conventional virtualized platforms to ones that are specifically designed for high-performance compute, storage, and networking resources at scale. This is motivating organizations to evaluate cloud providers on their ability to service the needs of general-purpose and AI workloads.

A clear market opportunity has emerged for cloud providers to offer full-stack solutions that are optimized for AI. By pre-integrating the entire hardware and software stack, these AI-native platforms and infrastructure significantly simplify the path to production and allow enterprises to capitalize on their AI investments with greater speed and at lower cost.

IDC conducted research that explored the value and benefits that organizations using AI Hypercomputer infrastructure achieved in developing and running AI and ML workloads.

Employing its specialized Business Value methodology, IDC calculated that these customers achieved benefits worth an annual average of $38.8 million per organization ($1.5 million per AI application) and a three-year ROI of 353% by:

- Accelerating AI/ML development and deployment by providing optimized hardware and software
- Improving IT efficiency and agility by reducing infrastructure complexity and providing tools such as autoscaling, dynamic scheduling, and integrated observability
- Providing flexible consumption models to optimize cost and reduce wasted IT spend
- Strengthening security and compliance with integrated features such as IAM, encryption, and anomaly detection
- Supporting rapid adaptation and responsiveness to market trends, customer demands, and business needs

Situation Overview

The AI-driven inflection point is fundamentally reshaping the landscape of enterprise cloud infrastructure. This represents a paradigm shift away from the general-purpose virtualized environments that defined the first era of cloud computing. Infrastructure architected to support generative AI and the large-scale models that power it must be built on a foundation of diverse compute, high-throughput parallel storage, and ultra-low-latency networking. The sheer scale and unique processing patterns of AI workloads, which involve massive parallelism and constant data movement, expose weaknesses in traditional cloud architectures.

This has impacts across the entire tech stack. On the compute layer, the focus shifts from standard CPUs to clusters of thousands of specialized accelerators, such as GPUs or tensor processing units (TPUs), all working in concert. To support this, storage systems must deliver massive, sustained throughput to feed these data-hungry accelerators, preventing the costly problem of "I/O starvation," where expensive compute cycles are wasted waiting for data. Perhaps most critically, the networking fabric must act as an extension of the system itself.

This technological rift is creating an urgent business imperative for modernization. IDC predicts that by 2027, the massive computational and data demands of AI will compel 80% of organizations to modernize legacy cloud environments by shifting to new platforms specifically designed for AI workloads. Organizations that fail to make this transition will find themselves at a severe competitive disadvantage, unable to train, deploy, or iterate on AI models effectively.

This shift exposes the core inadequacy of the prevailing cloud model, which has historically treated infrastructure as a collection of individual, disaggregated services. This "à la carte" approach forces teams to manually provision and stitch together separate, discrete services consisting of infrastructure (compute, storage, and networking) and platform software (orchestration engines, frameworks, and compilers). This model, while flexible, creates crippling inefficiencies for organizations without the resources and skills to manage infrastructure end to end. It introduces significant performance bottlenecks where the components are not co-engineered, and it places an immense integration and optimization burden on already scarce AI and MLOps talent.

This gap has created a critical and time-sensitive opportunity for cloud providers. The market is clearly signaling a need for options beyond a simple menu of services and the inclusion of fully integrated, full-stack systems optimized for AI. These platforms bundle the accelerators, high-speed storage, and networking fabric with the necessary

software layers — such as container orchestration, AI frameworks, and MLOps tooling — into a single, preconfigured, and pretuned solution. By offering an "AI-ready" platform and infrastructure, providers can drastically reduce complexity, eliminate performance guesswork, and ultimately accelerate the entire AI development life cycle, allowing enterprises to capture the business value of their AI initiatives faster while still supporting the needs of pre-existing applications.

Google Overview

Google Cloud's AI Hypercomputer is not a single product but rather an integrated system designed for large-scale AI. The core idea is to move beyond providing individual components (such as GPUs or TPUs) and instead offer a complete, co-designed stack where the hardware, networking, storage, and software are all optimized to work together. This holistic approach aims to maximize performance, efficiency, and reliability for the entire AI life cycle, from massive-scale model training to low-latency inference.

Benefits

By integrating all components into a single, co-designed system, AI Hypercomputer provides several key advantages:

- **Extreme performance and scalability:**
  The system is built for massive scale, with the ability to connect tens of thousands of accelerators. Technologies such as optical circuit switching (OCS) and the Jupiter network fabric provide petabit-scale bandwidth, dramatically reducing latency and bottlenecks.
- **Improved "goodput" and reliability for training:**
  AI Hypercomputer focuses on "goodput" — the measure of useful work completed. It minimizes job interruptions through automated health checks, predictive failure analysis, and multitier checkpointing. This means training runs are more likely to complete successfully without costly failures and restarts.

- **Cost and energy efficiency:**
  The system is designed for performance-per-dollar and energy efficiency across training and inference workloads. Features such as liquid cooling for next-gen TPUs and the power-saving OCS network reduce operational costs. Flexible consumption models, such as Dynamic Workload Scheduler, allow you to reserve capacity only when needed, avoiding idle, expensive hardware.
- **Developer productivity and openness:**
  The software stack integrates open source software, such as Slurm (cluster director), Google Kubernetes Engine (GKE), JAX, and PyTorch. This allows developers to use familiar tools while benefiting from the underlying hardware optimization. The system abstracts away much of the complex infrastructure management, letting teams focus on building and deploying models rather than managing hardware.

Components of AI Hypercomputer

AI Hypercomputer is built on three main pillars: hardware, software, and flexible consumption models.

AI-Optimized Hardware

This layer provides the raw power for AI workloads, with each part designed to eliminate bottlenecks.

- **Compute (accelerators):**
  AI Hypercomputer offers a choice of cutting-edge processors to suit different needs.
- **Google Cloud TPUs:**
  These are Google's custom-built ASICs, designed specifically for ML workloads. The system features Ironwood, the seventh-generation TPU, which is the most powerful and energy-efficient TPU to date.
- **NVIDIA GPUs:**
  It integrates the latest NVIDIA GPUs, including H100 Tensor Core GPUs (available in A3 VMs) and A4 VMs (based on NVIDIA's Blackwell platform). These are ideal for a wide range of AI training and inference tasks.
- **CPUs:**
  It includes Axion, Google's first custom ARM-based CPU designed for the datacenter, plus options from Intel and AMD.

- **Networking:**
  To connect thousands of accelerators, AI Hypercomputer uses Google's petabit-scale Jupiter network fabric. Inter-Chip Interconnect technology enables pods with thousands of TPUs to work together seamlessly as a single, powerful unit. Cross-Cloud Interconnect ensures high-speed, secure links between Google and other clouds.
- **Storage:**
  An integrated multitiered storage system feeds data to the accelerators at high speed.
- **Managed Lustre:**
  This is a high-performance parallel file system for demanding training workloads.
- **Hyperdisk ML:**
  This block storage service is optimized for fast ML model loading, reducing inference startup times.
- **Rapid storage and anywhere cache:**
  These are Cloud Storage solutions that colocate data with compute accelerators (in the same zone) to provide high-throughput, low-latency data access.
- **VAST data integration:**
  The VAST AI Operating System is available as a fully managed service via the Google Cloud Marketplace, providing a unified global namespace that spans both Google Cloud and on-premises environments.

Leading Software, Open Frameworks

Managing the hardware is industry-leading software, integrated with open frameworks, libraries, and compilers to make AI development, integration, and management more efficient.

Orchestration:

- **Google Kubernetes Engine:**
  The primary control plane for Kubernetes deployments, allowing you to manage massive clusters of TPUs and GPUs
- **Cluster Director:**
  A unified control pane that provides a highly scalable and fault-tolerant managed Slurm environment for distributed workloads
- **Kueue:**
  A GKE-native job-queuing system that manages batch workloads, ensuring fair sharing and efficient utilization of resources

- **ML frameworks and compilers:**
  The system provides optimized versions of popular frameworks, such as JAX, PyTorch, and Keras. Accelerated linear algebra (XLA): This is a compiler that optimizes and translates models to run with maximum efficiency on TPUs and GPUs.
- **Specialized libraries:**
  Google provides tools such as MaxText (for high-performance LLM training) and vLLM (a high-performance inference engine) that are pre-optimized for the AI Hypercomputer architecture.

Reliability and management:

- The system features automated health checks that continuously monitor hardware. If a node is predicted to fail, the system can preemptively move the workload.
- OCS technology significantly enhances network reliability by simplifying the network path and ensuring consistent, high-quality connections.
- Multitier checkpointing allows long-running training jobs to save their progress quickly and efficiently, so they can be resumed with minimal lost time in the event of a failure.

Flexible Consumption Models

AI Hypercomputer provides flexible and cost-effective access to its supercomputing capabilities. Instead of requiring massive upfront capital investment, it offers cloud-based consumption.

The Dynamic Workload Scheduler is a prime example, allowing users to choose between:

- **Scheduled:**
  Reserving a specific "slice" of the supercomputer for a future time, guaranteeing availability
- **Flex start:**
  Submitting a job that can be preempted but runs at a lower cost, ideal for non-urgent or experimental workloads; utilizes idle capacity, often called Spot VMs in the wider cloud computing landscape, offering deep discounts in exchange for the risk of preemption

For customers with stable, long-running workloads, AI Hypercomputer also offers Committed Use Discounts, which it applies when a user commits to a specific level of resource consumption. In exchange for this commitment, the effective hourly rate reduces substantially, providing a predictable and lower cost for sustained, high-volume operations.

This flexibility allows organizations of all sizes to access world-class AI infrastructure that was previously only available to the largest research institutions.

The Business Value of Google

Study Firmographics

IDC conducted research that studied the costs and benefits for organizations utilizing AI Hypercomputer to support and develop AI/ML workloads. The project included eight interviews with organizations that use Google Cloud AI Infrastructure services, such as flexible consumption models and AI software and hardware. Qualified participants had experience and knowledge about the benefits and costs of using the solutions. During the interviews, IDC asked the companies a variety of quantitative and qualitative questions about the offering's impact on their IT environments, AI development, security operations, core businesses, and costs.

For this study, IDC considered the following solutions as AI Hypercomputer services:

- **Flexible consumption models:**
  Dynamic Workload Scheduler, On Demand, Committed Use Discounts, Spot Instances
- **Software:**
  JAX, Keras, PyTorch, vLLM, Pathways on Cloud, XLA, Google Cloud Kubernetes Engine, Cluster Director, and Compute Engine
- **AI-optimized hardware**
  compute (GPUs, TPUs, CPUs), storage (block, file, object), and networking (cross-cloud, VPC)

Table 1 (below) outlines the firmographics of the organizations that participated in the study. As shown, a wide range of company sizes utilized Google Cloud AI infrastructure services — from small businesses with 200 employees to large enterprises with up to 239,500 employees. On average, these organizations operated 926 applications and managed 7,070 terabytes of data storage. In total, 6,839 IT staff supported the employees and workloads across these companies. The participant organizations represented a diverse set of industries — retail (4), healthcare, financial services, food and beverage, and software. The table also contains additional metrics.

Table 1

Firmographics of Interviewed Organizations

Firmographics	Average	Median	Minimum	Maximum
Number of employees	70,463	14,000	200	239,500
Number of IT staff	6,839	265	32	50,000
Number of business applications	926	200	12	3,000
Total storage (TB)	7,070	1,000	50	20,000
Annual revenue	$19B	$7B	$20M	$84B
Countries	United States (6), Canada, Turkey			
Industries	Retail (4), Healthcare, Financial Services, Food and Beverage, Software			

n = 8; Source: IDC Business Value In-Depth Interviews, October 2025

Choice and Use of Google

The organizations that IDC interviewed described the decision criteria involved in their selection of AI Hypercomputer to overcome a variety of business and technical challenges. Many had struggled with fragmented systems, legacy infrastructure, and inefficient resource management, which created barriers to performance and innovation. The organizations noted that the solutions would help unify disconnected environments and streamline operations. Ultimately, this would enable improved customer experiences and reduce friction across channels.

Additionally, several organizations sought to advance their analytics capabilities and move from historical reporting to predictive and prescriptive insights that support better decision-making and personalized services. They also noted that the ability to integrate seamlessly with existing Google technologies, flexible pricing models, and access to high-performance AI hardware contributed to their decision. Overall, they selected AI Hypercomputer to provide the scalability, speed, and intelligence needed to modernize operations and support evolving business demands.

Study participants elaborated on these and other selection criteria below:

Better customer experiences (retail):

"My company had multiple systems supporting our consumer retail environment, and they didn't communicate with each other. For example, buying online and returning in-store made you two different people in our systems, and returns could take up to nine minutes. We selected AI Hypercomputer to layer AI over these legacy tools and connect systems like parcel management, payments, and inventory to reduce that process."

AI-driven analytics (food and beverage):

"The analytics framework at my company has traditionally focused on descriptive reporting to help us understand how past promotions or initiatives performed. But, as the business grows more competitive, we need to shift toward predictive and prescriptive approaches like identifying key operational drivers of profitability, forecasting demand, and segmenting customers more effectively. This includes using AI to personalize offers based on dietary preferences and purchasing behavior, which is why we're moving toward platforms like Google Vertex AI and AI Hypercomputer."

Cost-effective integrations (retail):

"My organization was already using Google Workspace and G Suite, so we wanted something that would integrate easily with our predominantly Google tech stack. We appreciated that AI Hypercomputer offered committed-use discounts that would align infrastructure spend with our business rhythm, especially for always-on workloads like personalization and inventory prediction. Also, given our significant data and storage needs, we thought that TPU clusters would let us train multimodel recommendation engines much faster."

Legacy challenges (retail):

"My company chose AI Hypercomputer because we faced a lot of challenges with our legacy infrastructure, including scaling bottlenecks, performance issues, and limited GPU availability for redundant AI workloads. Manual resource allocation also created operational complexity and waste, as we had to assign compute and storage by hand.

On top of that, market pressures, innovation demands, and ESG goals pushed us to model faster and operate more transparently."

Better data management (healthcare):
"My hospital has so much clinical information stored and needed a solution like AI Hypercomputer to help manage that amount of data effectively."

Table 2 (below) details the aggregate environment that AI Hypercomputer was supporting at the time of the interviews. Study participants were using AI Hypercomputer to operate environments with an average of 7,630 cloud containers and 1,799 terabytes of storage. These environments supported an average of 45 AI applications, 103 AI models, and served 5,367 internal users. This scale demonstrates the system's ability to support large, complex workloads and widespread adoption across enterprise teams. The table also presents additional metrics.

Table 2
AI Hypercomputer Environment

Environment	Average	Median
Cloud containers	7,630	180
Compute clusters	764	70
TB, total of storage	1,799	200
Site/branches	3,302	45
Geographical locations (countries)	4	3
AI applications	45	12
AI models	103	10
Internal users	5,367	600

n = 8; Source: IDC Business Value In-Depth Interviews, October 2025

Business Value and Quantified Benefits

Interview participants reported that AI Hypercomputer delivered meaningful efficiency, reliability, and innovation improvements. It reduced the time and effort required to manage infrastructure, allowing teams to focus more on strategic initiatives. They also found that Google Cloud infrastructure helped their environment become faster, adaptive, and more responsive. These improvements enhanced overall performance and user experience. Additionally, integrated tools and services helped the organizations modernize their environments while maintaining continuity with existing systems. As a result, teams were able to adopt AI more quickly, experiment with new solutions, and bring innovations to market with greater speed and confidence. These benefits collectively supported a shift from operational maintenance to forward-looking development and business agility.

Interviewed customers detailed their most significant benefits achieved from using AI Hypercomputer below:

Higher efficiency (retail):

"One clear benefit is energy efficiency, which our ESG team has recognized through reduced power consumption across datacenters. From a financial perspective, we are acquiring less tangible hardware now that we've moved to hypercomputing, which shows up positively on the P&L. Operationally, users are seeing faster task execution, quicker data aggregation, and more responsive systems overall."

Improved reliability and performance (software):

"Most importantly, reliability has improved, and latency is much better with AI Hypercomputer. We can easily track our workloads and get alerts if something goes wrong. Google provides the transparency that we need to be responsive and stay on top of everything. This is critical since we run these workloads nonstop."

Better availability and integrations (retail):

"Availability and integration have been key benefits of AI Hypercomputer. We've been able to extend the life of some legacy tools while making a major pivot toward a more modern environment."

More innovation time (retail):
"AI Hypercomputer has helped reduce our time spent on infrastructure and physical capacity planning. This is due to the managed Google storage and computer that we have. It has helped us reduce management time by about 30%. This has freed up engineers for more strategic AI and ML innovation projects rather than just keeping "the lights on."

Quick AI adoption (healthcare):
"The biggest benefit of AI Hypercomputer is that we are able to adopt and use AI with speed."

Figure 1 (next page) presents IDC's calculations of the cumulative customer benefits after the adoption of AI Hypercomputer services.

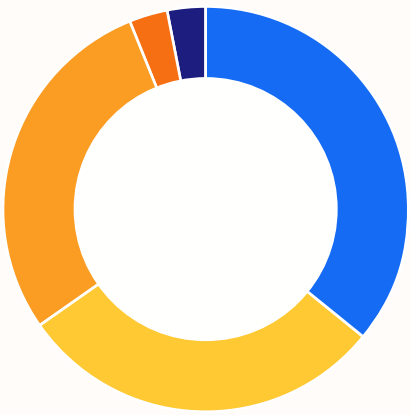
As shown, IDC quantified the average annual benefits at $38.8 million per organization ($1.5 million per AI application) in the following categories:

- **Business enablement:**
  Customers accelerated time to market, improved customer engagement, and increased revenue by using AI-powered personalization, real-time insights, and agile infrastructure.
- **Development benefits:**
  Teams built and deployed AI models faster through integrated CI/CD pipelines, scalable compute resources, and automated infrastructure provisioning.
- **IT benefits:**
  Organizations reduced infrastructure complexity and operational overhead with autoscaling, dynamic scheduling, and managed services, allowing IT staff focus more on innovation.
- **Performance benefits:**
  Workloads became more responsive and resilient due to reduced latency, faster recovery times, and predictive failure management.
- **Security and compliance benefits:**
  Integrated security features, such as encryption, IAM policies, and AI-driven anomaly detection, helped reduce risk, improve compliance, and shift teams from reactive to proactive security operations.

Figure 1

Average Annual Benefits Per Organization

($ per interviewed organization)



- $14,025,155 Business enablement benefits
- $11,375,997 Development benefits
- $11,104,300 IT benefits
- $1,198,077 Performance benefits
- $1,137,064 Security and compliance benefits

$38.8 millions

n = 8; Source: IDC Business Value In-Depth Interviews, October 2025

IT Benefits of AI Hypercomputer

Interview participants shared that AI Hypercomputer significantly enhanced the cost-effectiveness and agility of their IT operations. By enabling faster provisioning of compute resources, the system helped reduce delays and allowed teams to respond rapidly to evolving business needs. Importantly, it provided participants with the ability to launch and scale AI workloads without manual infrastructure provisioning. This helped streamline operations and eliminate common bottlenecks. Built-in tools for traffic control and security further simplified network management and strengthened data protection. Participants also reported measurable reductions in training and inference costs and lower storage expenses. These combined improvements allowed IT teams to shift their focus from routine maintenance to strategic innovation, fostering more responsive and efficient technology environments.

Study participants offered these detailed comments:

Better workflows and scaling (retail):

"In IT, it comes down to measurable results. Deploying predictive leak detection for our data scientists used to take seven weeks due to hardware procurement, provisioning,

and approvals, but now with Vertex AI and GKE, we can do it in two days. This faster rollout reduces operational risk, allows us to scale instantly, and delivers significant cost savings."

Cost reduction (retail):
"Training costs per model are now 40% lower per cycle on Google Cloud TPUs compared to our previous hybrid setup, and inference costs per query have dropped by 40%–45%. This has resulted in tens of millions in savings. Additionally, storage costs are down by 15%–20%, and infrastructure provisioning time has decreased by 20%."

Reduced costs with Hyperdisk (software):
"My company is saving money using the Google Cloud Hyperdisk solution; it has only a one-minute bootup. Beforehand, the bootup process took 25 minutes."

Network traffic control (software):
"The machine learning infrastructure is completely private, with no external traffic coming in. All communications happen through messaging, so they are not exposed to the internet. This setup gives us strong control over network traffic and enhances security across our workloads."

IDC first validated these findings by quantifying the impact on IT infrastructure management teams. Participants reported that AI Hypercomputer streamlined operations through features such as autoscaling, dynamic scheduling, deployment and integration simplification, and built-in observability. **These capabilities enabled IT teams to work more efficiently and dedicate 36% more time to innovation. The enhanced functionality also allowed teams to deploy new compute resources 48% faster, which improved their responsiveness to business needs.**

Supporting these statements, one participant noted, *"IT is no longer spending time on maintenance and support, which lets us focus more on core applications and customer experiences. We do not have to worry about infrastructure limitations like CPU or storage because AI Hypercomputer automatically monitors workloads and adjusts resources as needed. That flexibility and automation have been a major benefit to our operations."*

As illustrated in **Table 3 (next page),** these improvements enabled IT Infrastructure teams to work with 55% greater efficiency, meaning they spent 149,343 fewer hours annually managing their environment in comparison to their previous methods. This freed up the time of highly skilled individuals to focus on other important business initiatives and drive innovation. IDC valued this efficiency gain at $7.9 million per year.

Table 3

IT Infrastructure Team Efficiency Gain

Efficiency Gain	Before AI Hypercomputer	With AI Hypercomputer	Difference	Benefit
Hours worked per year	271,016	121,674	149,343	55%
Value of staff time per year	$14,415,768	$6,472,018	$7,943,750	55%

n = 8; Source: IDC Business Value In-Depth Interviews, October 2025

IDC next looked at the impact of AI Hypercomputer infrastructure on network team efficiency levels. These teams significantly benefited from Google Cloud's integrated VPC controls, automated traffic filtering, and predictive monitoring. These functions and integrations helped the team manage network security and performance more efficiently. **Table 4 (below)** illustrates that participating organizations spent 23,148 fewer hours annually managing their network with AI Hypercomputer in comparison to their previous methods. IDC calculated this to be a 28% efficiency gain valued at $1.2 million per year.

Table 4

Networking Team Efficiency Gain

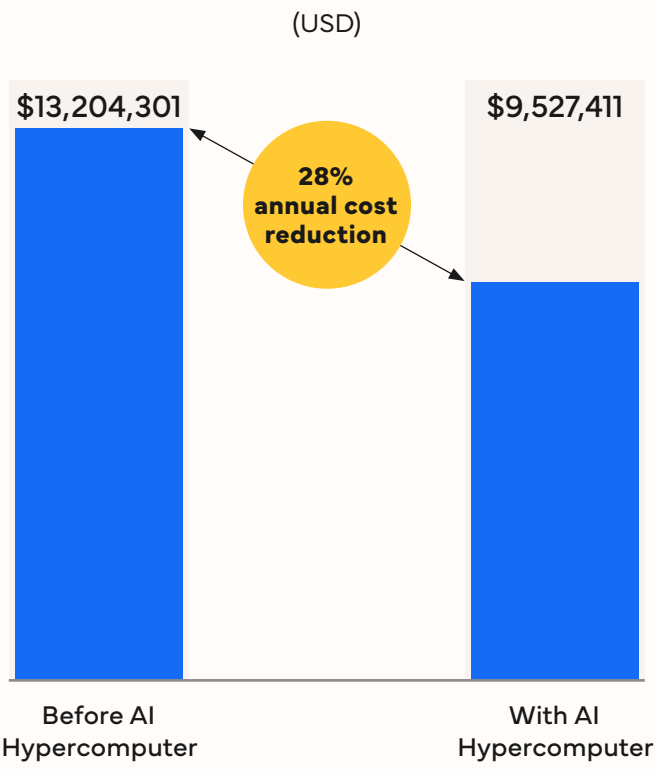
Efficiency Gain	Before AI Hypercomputer	With AI Hypercomputer	Difference	Benefit
Hours worked per year	83,031	59,883	23,148	28%
Value of staff time per year	$4,416,568	$3,185,283	$1,231,286	28%

n = 8; Source: IDC Business Value In-Depth Interviews, October 2025

IDC also noted that interview participants experienced substantial cost benefits after adopting AI Hypercomputer. By embracing a consumption-based model, on-prem organizations were able to reduce expenses related to provisioning, hardware, electricity, and datacenter space. For everyone else, flexible consumption options, including spot instances and committed use discounts, enabled the companies to better align infrastructure spending with actual usage patterns. This minimized waste and improved budget predictability. The ability to scale resources on demand and eliminate overprovisioning further supported cost optimization and made the system a financially strategic choice for managing AI workloads. **Figure 2 (below)** demonstrates that these efficiencies enabled participants to reduce IT spend by 28% annually.

Figure 2

Annual IT Cost Reductions/Avoidance



n = 8; Source: IDC Business Value In-Depth Interviews, October 2025

Application, ML, and AI Development Benefits of AI Hypercomputer

Interview participants shared that AI Hypercomputer significantly enhanced developer productivity and flexibility. The system allowed teams to experiment more freely by enabling rapid provisioning and deprovisioning of compute clusters. This helped reduce early development costs and supported iterative testing. Interviewed organizations also found that developers were able to scale applications efficiently and deploy models with greater speed because of integrated tools and containerized environments. It also streamlined workflows, which sped build and test cycles and allowed for more frequent releases. These capabilities empowered developers to focus on innovation and deliver high-performance AI applications without the constraints of traditional infrastructure.

Participants detailed these benefits below:

Ability to experiment (food and beverage):

"Scaling up to the level of compute we need would be cost-prohibitive with another solution. With AI Hypercomputer, we can spin up and shut down clusters instantly, which stops billing right away. This cuts out a lot of the early development costs and lets us experiment quickly. It removes the need for hiring network engineers or building out datacenters and gives us the flexibility to run high-compute jobs only when needed."

Less expensive AI deployments (retail):

"With Google Cloud TPUs, new AI applications are 40% cheaper to deploy. What used to take four months and up to $2 million per market can now be launched in six weeks at a significantly lower cost."

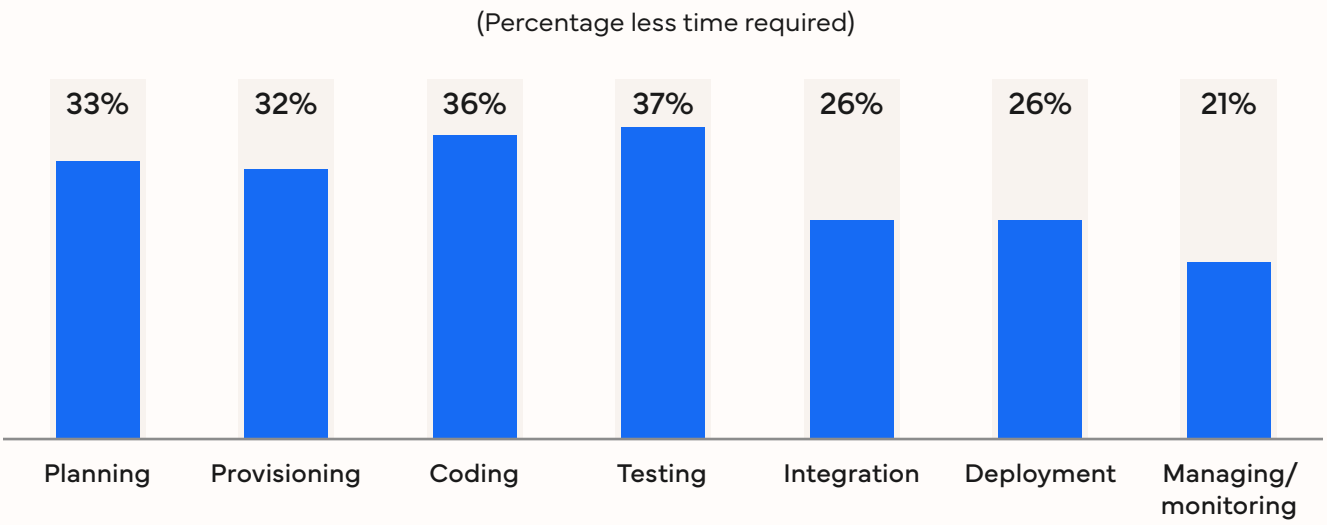
Quicker development cycles (retail):

"For our development teams, source code pushing has increased, and the test cycle is faster when building projects. With AI Hypercomputer, build times before pushing to the CI/CD pipeline are quicker, and we can support large-scale integration and multiple incident releases at once."

Keeping in mind the comments above, **Figure 3 (next page)** demonstrates the tangible benefits of AI Hypercomputer infrastructure for developers across the development cycle in the form of key performance indicators (KPIs). As shown, Google Cloud was particularly impactful on time requirements during planning (33%), coding (36%), testing (37%), and provisioning (32%) processes. Participants related these gains to integrated CI/CD pipelines, containerization, and automated infrastructure scaling.

Figure 3

DevOps-Related KPIs



n = 8; Source: IDC Business Value In-Depth Interviews, October 2025

As a result of these improvements across the development cycle, developers at interviewed organizations also achieved a notable productivity gain of 27% **(Table 5, below).** In other words, this enhancement meant that teams of 518.3 FTEs could work at the equivalent productivity level of having 140.5 additional FTEs on staff. IDC quantified this productivity gain at approximately $14 million per year.

Table 5

Developer Productivity Gain

Productivity Gain	Before AI Hypercomputer	With AI Hypercomputer	Difference	Benefit
Equivalent productivity level, FTEs	518.3	658.8	140.5	27%
Value of staff time per year	$51,827,269	$65,879,971	$14,052,702	27%

n = 8; Source: IDC Business Value In-Depth Interviews, October 2025

Security and Compliance Benefits of AI Hypercomputer

IDC also noted that AI Hypercomputer strengthened security posture and improved compliance management at interviewed organizations. It enabled faster detection and response to threats by analyzing large volumes of network traffic and system logs in real time. Built-in encryption and multiregion replication helped protect sensitive data at rest and in transit, while automated anomaly detection reduced the risk of misconfigurations and unauthorized access. These capabilities allowed teams to shift from reactive incident handling to proactive policy development and audit readiness. Organizations also benefited from improved data governance and streamlined compliance processes, particularly in regulated industries where privacy and risk management were critical.

Participants offered these comments regarding security and compliance:

Quicker detection and response (retail):

"Security has definitely improved with AI Hypercomputer because it can analyze massive amounts of network traffic and system logs in real time, much faster than before. Malware detection, encryption and decryption, and incident response are all quicker when using AI-powered models."

Risk reduction (retail):

"In the past, we had misconfigurations and failed updates, but now with AI Hypercomputer's multiregion replication, encryption at rest and in transit, and AI-driven anomaly detection, risky activity is automatically prevented or flagged. Because of this, staff time has shifted from reactive firefighting to proactive threat modeling, policy creation, and audits, which has led to a strong business impact."

Compliance and risk management (financial services):

"Being a financial company, it is critical to have data privacy, data security, compliance, and risk management. AI Hypercomputer has the right framework and infrastructure to support all our regulatory and compliance requirements. Its broad range of solutions across infrastructure, networking, AI, storage, and analytics helps us operate efficiently and stay ahead with innovation."

Reduction in manual work (retail):

"My company has achieved meaningful cost efficiency, including operational savings and reduced compliance penalties. AI-based imaging and natural language processing have replaced manual compliance inspection teams, which has really increased speed."

To quantify the impact of AI Hypercomputer, IDC started with the security operations team at interviewed organizations. This team appreciated that Google Cloud provided multiregion replication, real-time threat detection, encryption, AI-driven anomaly detection, and rapid response. These features and functions enabled the teams to shift from reactive response to proactive, strategy-driven security methods.

Importantly, they also had a significant impact on the efficiency level of this very important team. As **Table 6 (below)** shows, AI Hypercomputer enabled the team to spend 21,127 fewer hours annually managing the security of their environment. This time saving resulted in a significant efficiency gain of 40%, which IDC valued at nearly $1.3 million annually.

Table 6

Security Operations Team Efficiency Gain

Efficiency Gain	Before AI Hypercomputer	With AI Hypercomputer	Difference	Benefit
Hours worked per year	60,952	36,825	24,127	**40%**
Value of staff time per year	$3,242,105	$1,958,772	$1,283,333	**40%**

n = 8; Source: IDC Business Value In-Depth Interviews, October 2025

Next, IDC examined the compliance team within the interviewed organizations. This team benefited from Google Cloud services' provision of zero trust networking, IAM policies, encryption, data labeling, and real-time reporting. These critical functions helped participants shift from manual oversight to faster audit-ready operations. **Table 7 (next page)** illustrates that these improvements resulted in an 18% productivity gain, which IDC valued at $121,275 per year.

Table 7

Compliance Team Productivity Gain

Productivity Gain	Before AI Hypercomputer	With AI Hypercomputer	Difference	Benefit
Equivalent productivity level, FTEs	9.8	11.5	1.7	18%
Value of staff time per year	$686,000	$807,275	$121,275	18%

n = 8; Source: IDC Business Value In-Depth Interviews, October 2025

Performance Benefits of AI Hypercomputer

Across organizations, the use of AI Hypercomputer led to noticeable improvements in application and workload performance. Teams observed faster response times and reduced latency, which enhanced the user experience and supported more efficient operations. The system's ability to scale resources elastically and manage surges in demand helped maintain consistent performance during peak periods. Participants also found that integrated monitoring and logging tools streamlined troubleshooting and improved system observability. This allowed for quicker identification and resolution of issues. These capabilities contributed to greater system resilience and ensured that applications remained stable and responsive, even under complex and high-volume conditions.

The customers supplied the following quotes regarding performance:

Reduced latency (retail):

"AI Hypercomputer has helped latency drop by 50% globally, and we are seeing elastic scaling and faster deployments. Native integration with cloud monitoring, cloud logging, and Vertex AI pipelines has reduced troubleshooting time for our SREs and improved observability. The flexible consumption model and built-in cross-region redundancy have improved utilization efficiency and resilience across our Google tech stack."

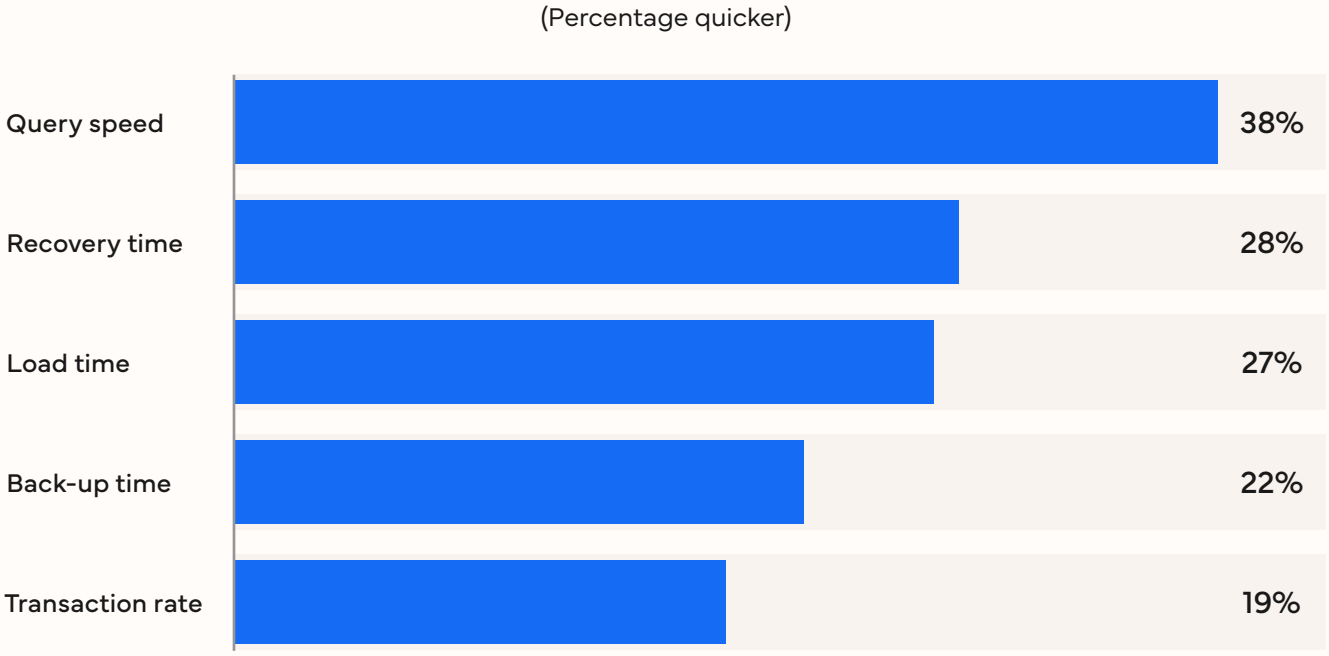
Less downtime (financial services):

"AI Hypercomputer is very stable, and we have experienced very limited downtime. In the past, we used to have 55 applications go down. Google's AI Hypercomputer has had a big impact; we have seen a 40% improvement in downtime."

The adoption of AI Hypercomputer has led to measurable improvements in performance across several key indicators. As **Figure 4 (below)** indicates, study participants reported that they experienced a 38% increase in query speed, a 28% improvement in recovery time, and a 27% reduction in load times as a direct result of using the system. Driving these gains were optimized compute resources, lower latency, and faster access to data through integrated AI infrastructure. As a result, businesses operated with greater speed and reliability, which enabled them to better serve their customers.

Figure 4

Performance-Related KPIs



n = 8; Source: IDC Business Value In-Depth Interviews, October 2025

In addition to the strong KPIs noted above, IDC also found that AI Hypercomputer services had a strong impact on unplanned application-related downtime regarding end-user productivity levels. With features such as live migration for compute, multiregion failover, and AI-driven predictive failure management, the system helped reduce the frequency of unplanned downtime outages by 67% while improving the time required to resolve them by 49% **(Table 8, next page).**

Considered together, these two improvements enabled 83% less end-user productivity loss, which IDC valued at $354,977 per year. The table below also presents additional granular metrics.

Table 8

Unplanned Application Downtime — End-User Impact

Impact	Before AI Hypercomputer	With AI Hypercomputer	Difference	Benefit
Number of outages per year	26.8	8.8	18.0	67%
Time to resolve per outage, hours	6.8	3.5	3.4	49%
Users impacted by downtime	250	250	N/A	N/A
Percentage of productivity loss factor	25%	25%	N/A	N/A
Total number of FTE impacted per year	6.1	1.0	5.1	83%
Value of lost productive time per year	$426,255	$71,277	$354,977	83%

n = 8; Source: IDC Business Value In-Depth Interviews, October 2025

Also, regarding unplanned application downtime, IDC noted that AI Hypercomputer significantly reduced revenue loss associated with outages. **Table 9 (next page)** depicts that by reducing the number of outages annually by 67% and reducing the revenue loss per outage by 40%, interviewed organizations avoided net revenue losses of approximately $1.1 million per year.

Table 9

Unplanned Application Downtime — Revenue Impact

Impact	Before AI Hypercomputer	With AI Hypercomputer	Difference	Benefit
Number of outages per year	26.8	8.8	18.0	67%
Percentage revenue-impacting outages	78%	78%	N/A	N/A
Average revenue loss per outage	$445,833	$268,750	$177,083	40%
Value of annual net revenue loss	$1,402,630	$277,630	$1,125,000	80%

n = 8; Source: IDC Business Value In-Depth Interviews, October 2025

Business Enablement Benefits of AI Hypercomputer

Interview participants described AI Hypercomputer as a catalyst for business agility and innovation. The system empowered organizations to accelerate product launches and respond more effectively to market demands. Teams were able to personalize customer experiences, streamline operations, and improve supply chain efficiency, all of which contributed to stronger engagement and resilience. The ease of scaling and integrating AI capabilities allowed businesses to pivot quickly, seize new opportunities, and maintain a competitive edge. Many participants also noted that working with advanced technologies inspired greater enthusiasm and creativity among their teams and fostered a culture of innovation. This often positioned their organizations ahead of industry peers in terms of modernization.

Study participants offered these comments:

Faster to market (retail):

"Speed to market and launch has been a major benefit of using AI Hypercomputer. Real-time recognition engines have helped drive higher revenue, and customer

engagement has increased. We have also improved supply chain simulations, which has optimized inventory, reduced costs in reverse logistics, and made the business more resilient."

Improved customer experience (retail):
"We've seen a much greater level of personalization and stronger consumer engagement with AI Hypercomputer. Store staff now have more time to focus on selling instead of handling administrative tasks. That shift has helped improve efficiency and the customer experience."

Competitive edge (financial services):
"Google has certainly enabled us to go to market faster, and its AI Hypercomputer is the best we have seen so far. Google stands out with top-tier machine learning integration and gives us a clear competitive edge. Our experience has been excellent, and our five-year strategy is to make Google Cloud our number 1 hyperscaler."

Higher levels of innovation (financial services):
"My organization has become very innovative with AI Hypercomputer. We have already done some good work in the last year, and as a result, we think that we are probably much ahead of our competitors in technology innovations, technology adoption, and technology modernization."

IDC found that AI Hypercomputer had a substantial impact on revenue generation among the companies interviewed. The solutions enabled quicker product launches and improved customer experience through personalization. These improvements enabled participants to gain a competitive edge with real-time insights and respond to market demands with agility. Interview participants emphasized that Google Cloud infrastructure gave them the ability to scale AI workloads efficiently and integrate them across business functions, which ultimately contributed to greater business agility and innovation.

In terms of business enablement KPIs, organizations reported:

- **22% quicker to go to market with new applications, products, or services**
- **27% less application latency that impacted their customers and workforce**

Table 10 (next page) shows that participants attributed over $17.3 million in additional net revenue to their deployment of AI Hypercomputer, which reflected the system's role in driving measurable business outcomes.

Table 10

Business Enablement — Higher Revenue

Revenue	With AI Hypercomputer	Per AI Application
Total additional gross revenue per year	**$115,501,281**	**$2,595,534**
Assumed operating margin	15%	15%
Total additional net revenue, IDC's model	**$17,325,192**	**$389,330**

n = 8; Source: IDC Business Value In-Depth Interviews, October 2025

ROI Summary

IDC calculated the average ROI to sum up the financial and business-related benefits for study participants' use of AI Hypercomputer. IDC calculated that these companies achieved three-year discounted benefits worth an average of $91 million per organization through improved staff efficiency, performance improvements, cost savings, and business enablement **(Table 11, next page).** These benefits are in comparison with the total three-year discounted costs of $20.1 million per organization. These levels of benefits and investment costs resulted in an average three-year ROI of 353% with a payback period of 8.8 months.

Table 11

Three-Year ROI Analysis

ROI Analysis	Label 1	Label 2
Discounted benefits	$90,998,900	$2,044,919
Discounted investment	$20,107,000	$451,843
Net present value	$70,891,900	$1,593,076
ROI	**353%**	**353%**
Payback	**8.8 months**	**8.8 months**
Discount factor	12%	12%

n = 8; Source: IDC Business Value In-Depth Interviews, October 2025

Challenges/Opportunities

Despite its technical advantages, Google Cloud must navigate intense competition from other cloud providers that are also significantly investing in AI infrastructure. Additionally, TPUs — a key differentiator — have lower adoption in the broader computing market compared with GPUs.

To address these challenges, Google Cloud must emphasize how AI Hypercomputer is a system for all AI workloads, demonstrating how its integrated networking and storage improve performance across both TPU- and GPU-based workloads. The company must also continue to invest in making TPUs easier to use and interoperable with popular frameworks.

Conclusion

AI workloads are outpacing legacy cloud capabilities, exposing bottlenecks in compute, storage, and networking. Enterprises need integrated, full-stack solutions to meet escalating demands.

Google Cloud's AI Hypercomputer unifies hardware and software, eliminating integration headaches and boosting performance. Compute, storage, and networking are co-designed for AI, ensuring that data flows seamlessly and accelerators run at full speed. Flexible consumption models — such as dynamic scheduling and spot instances — let organizations scale up or down, optimizing costs. Developers benefit from integrated CI/CD pipelines, container orchestration, and open frameworks, which speed up experimentation and deployment. Built-in security and compliance features, such as encryption, IAM, and real-time anomaly detection, help teams shift from reactive fixes to proactive governance. This holistic approach streamlines operations, reduces manual work, and empowers teams to focus on innovation.

IDC's research shows that Google Cloud's AI Hypercomputer delivers $38.8 million in annual value per organization with a three-year ROI of 353% and payback in under nine months by simplifying infrastructure, accelerating AI adoption and enabling rapid innovation.

Appendix A: Methodology

IDC utilized its standard ROI methodology for this project. This methodology is based on gathering data from current users of Google as the foundation for the model.

Based on interviews with organizations using Google, IDC performed a three-step process to calculate the ROI and payback period:

1. **IDC gathered quantitative benefit information during the interviews using a before-and-after assessment of the impact of Google.** In this study, the benefits included IT cost reductions and avoidances, staff time savings and productivity benefits, and revenue gains.
2. **IDC created a complete investment (three-year total cost analysis) profile based on the interviews.** Investments go beyond the initial and annual costs of using Google and can include additional costs related to migrations, planning, consulting, and staff or user training.
3. **IDC calculated the ROI and payback period.** IDC conducted a depreciated cash flow analysis of the benefits and investments for the organizations' use of Google over a three-year period. ROI is the ratio of the net present value and the discounted investment. The payback period is the point at which cumulative benefits equal the initial investment.

IDC bases the payback period and ROI calculations on several assumptions, which are summarized as follows:

- Time values multiplied by burdened salary (salary + 28% for benefits and overhead) quantify the efficiency and productivity savings. For this analysis, IDC used assumptions of an average fully loaded $100,000 per year salary for IT staff members and an average fully loaded salary of $70,000 for non-IT staff members. IDC assumes that employees work 1,880 hours per year (47 weeks x 40 hours).
- IDC calculated the net present value of the three-year savings by subtracting the amount that the organizations would have realized by investing the original sum in an instrument yielding a 12% return to allow for the missed opportunity cost. This accounts for both the assumed cost of money and the assumed rate of return.
- Further, because Google requires a deployment period, the full benefits of the solution are not available during deployment. To capture this reality, IDC pro rates the benefits on a monthly basis and then subtracts the deployment time from the first-year savings.

Note: All numbers in this document may not be exact due to rounding.

Appendix B: Specific Calculations — Benefits from the Use of AI Hypercomputer

Table 12

Specific Calculations: Benefits from the Use of AI Hypercomputer

Category of Value	Average Quantitative Benefit	15% Margin Applied	Calculated Average Annual Value*
Annual cost reductions	$3,676,890 in annual IT cost reductions	No	$3,676,890
Network team efficiency gain	28% higher efficiency worth 12.3 FTEs, $100,000 salary	No	$996,755
IT infrastructure team — admin and management efficiency gains	55% higher efficiency worth 79.4 FTEs, $100,000 salary	No	$6,430,655
IT unplanned downtime, revenue benefit	$1,125,000 in annual net revenue loss avoided	Yes	$910,714
Unplanned downtime, productivity benefit	83% productivity saved, worth 5.1 FTEs, $70,000 salary	No	$287,363
Security operations team efficiency gain	40% higher efficiency worth 12.8 FTEs, $100,000 salary	No	$1,038,889
Compliance team productivity gain	18% higher productivity worth 1.7 FTEs, $70,000 salary	No	$98,175
DevOps productivity gain	27% higher productivity worth 140.5 FTEs, $100,000 salary	No	$11,375,997
Business enablement — end-user productivity gain	$17,325,192 in total additional net revenue	Yes	$14,025,155
Total average annual benefits	**$38.8M per organization per year**		

*includes 6.9 months deployment time in year 1; n = 8; Source: IDC Business Value In-Depth Interviews, October 2025

About the IDC Analysts

Dave McCarthy

Research Vice President,
Cloud and Edge Services, Worldwide Infrastructure Research, IDC

Dave McCarthy is research vice president within IDC's Worldwide Infrastructure Research organization and global research lead for the Cloud and Edge Services practice. McCarthy leads a team of analysts covering research on shared (public) cloud, dedicated (private) cloud, and edge deployments, services, adoption trends, vendor strategies, and market dynamics. Benefiting both technology suppliers and IT decision-makers, McCarthy's insights delve into ways in which hybrid and distributed cloud platforms provide the foundation for next-generation workloads, enabling organizations to innovate faster, automate operations, and achieve digital resiliency.

More about Dave McCarthy →

Megan Szurley

Business Value Manager, Business Value Strategy Practice, IDC

Megan Szurley is manager for the Business Value Strategy practice, responsible for creating custom business value research that determines the ROI and cost savings for enterprise technology products. Szurley's research focuses on the financial and operational impact of these products for organizations once deployed and in production. Prior to joining the Business Value Strategy practice, Szurley was a consulting manager within IDC's Custom Solutions division, delivering consultative support across every stage of the business life cycle: business planning and budgeting, sales and marketing, and performance measurement. In her position, Szurley partners with IDC analyst teams to support deliverables that focus on thought leadership, business value, custom analytics, buyer behavior, and content marketing. These customized deliverables are often derived from primary research and yield content marketing, market models, and customer insights.

More about Megan Szurley →

IDC Custom Solutions

IDC Custom Solutions produced this publication. The opinion, analysis, and research results presented herein are drawn from more detailed research and analysis that IDC independently conducted and published, unless specific vendor sponsorship is noted. IDC Custom Solutions makes IDC content available in a wide range of formats for distribution by various companies.

idc.com @idc @idc

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications, and consumer technology markets. With more than 1,300 analysts worldwide, IDC offers global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries. IDC's analysis and insight helps IT professionals, business executives, and the investment community to make fact-based technology decisions and to achieve their key business objectives.