Building for Success:

Optimizing Infrastructure for the GenAI Era

Matt Eastwood
SVP, WW Research
February 2024

Agenda

- GenAI and Digital First
- New Applications and Workloads
- New Edge Deployment Locations
- New Security Challenges
- The Operating Model is Hybrid

Where Must Enterprises Invest?

Which of the following are immune to budget reduction regardless of the economic environment?

1. **AI and Automation Projects**
2. **Security, Risk, & Compliance**
3. **Infrastructure & IT Operation optimization initiatives/ projects**
4. **Back-office Applications (ERP, HR, SCM)**
5. **Customer Experience Initiatives**

Source: Future Enterprise Resiliency & Spending Survey Wave 11, IDC, December 2023

GenAI Dominates the Conversation

The New York Times

Generative A.I. Can Add $4.4 Trillion in Value to Global Economy, Study Says

The report from McKinsey comes as a debate rages over the potential economic effects of A.I.-powered chatbots on labor and the economy.

THE WALL STREET JOURNAL.

Generative AI Promises an Economic Revolution. Managing the Disruption Will Be Crucial.

FT FINANCIAL TIMES

The global race to set the rules for AI

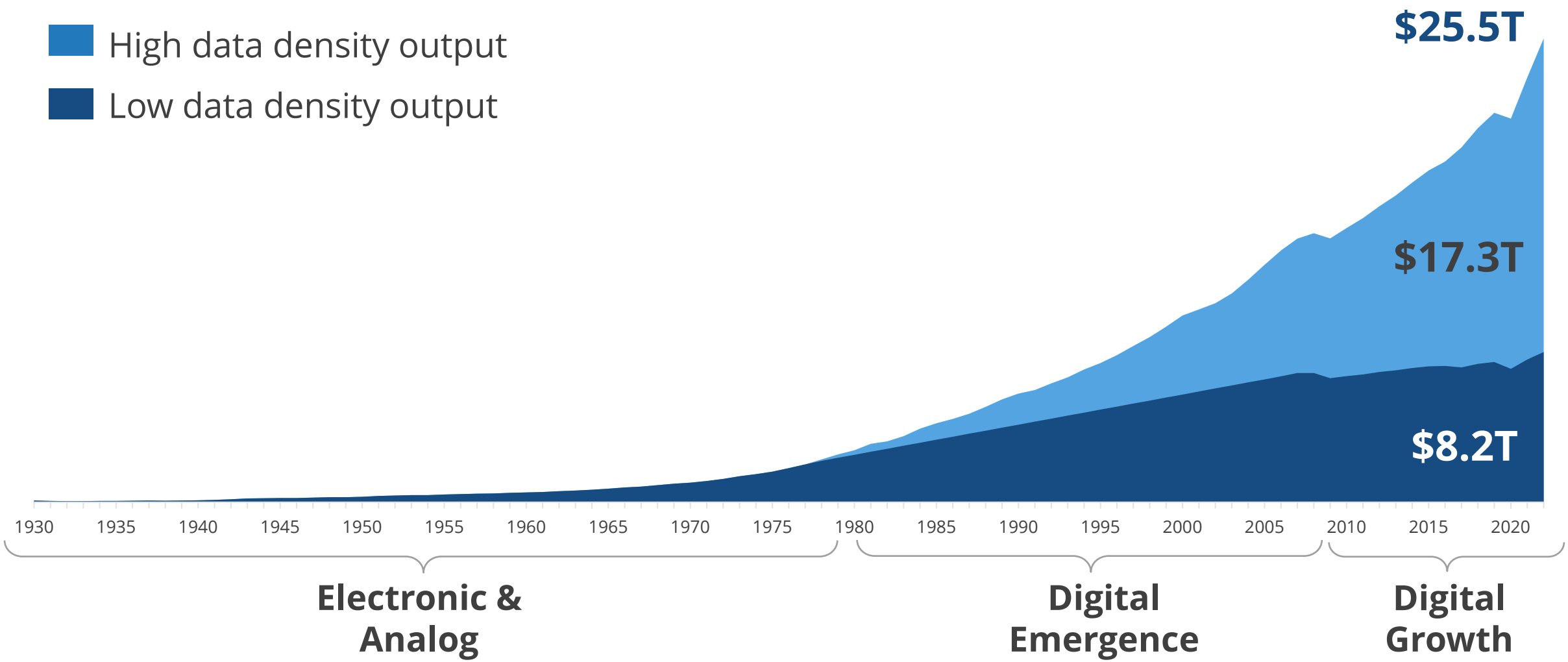
The industry and policymakers agree that the emerging technology needs regulating. But no one is quite sure how

Bloomberg

Companies Go All Out to Up Their Generative AI Game

Experts expect the new technology to transform jobs more than eliminate them.

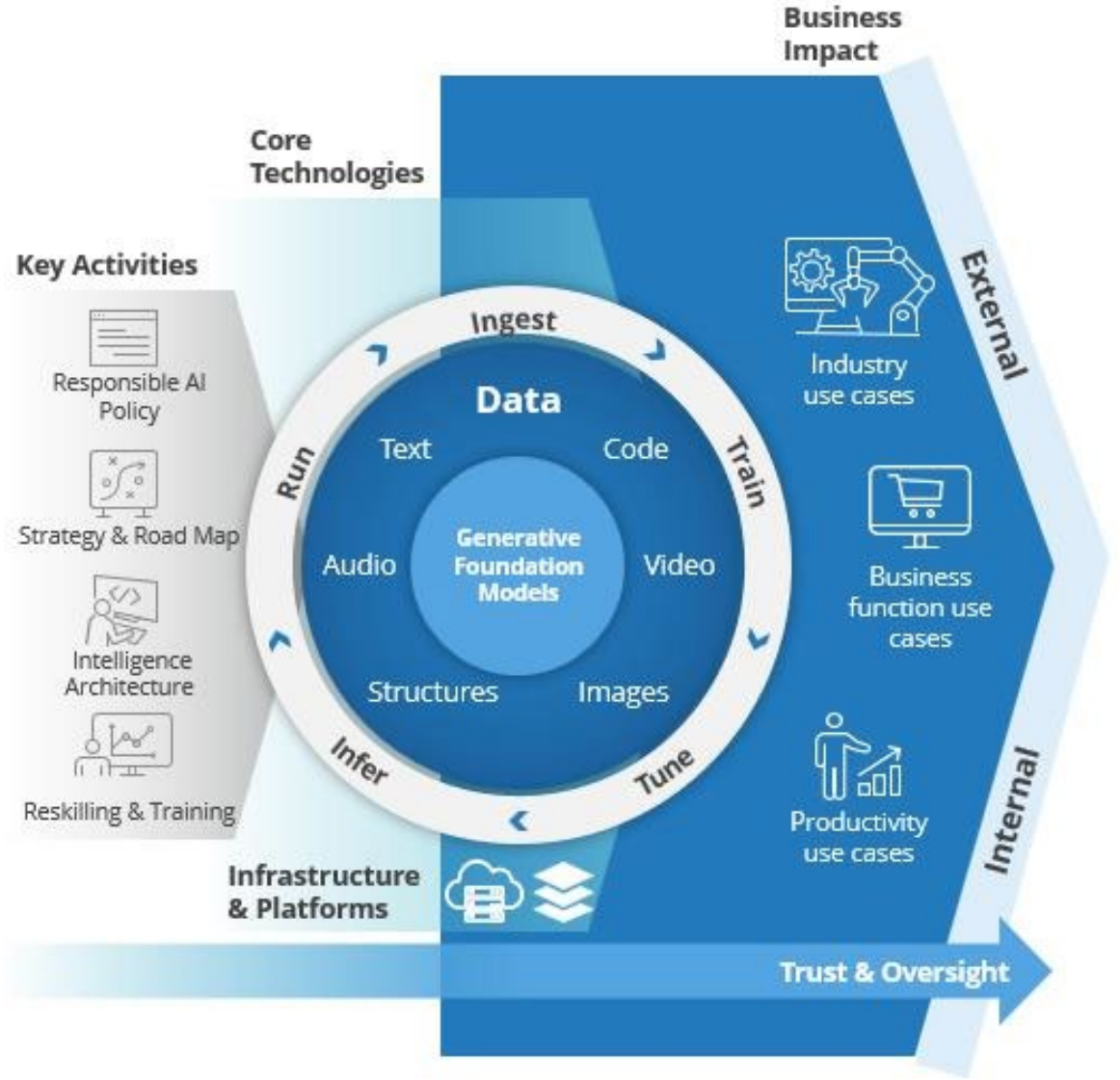
In 2022 2/3 of U.S. GDP based on data dense products & services



Source: U.S. Bureau of Economic Analysis, "Value Added by Industry" (accessed Wednesday, March 1, 2023)
2021 U.S. Data Valuation by Industry Vertical, IDC #US49939922, December 2022, updated with full year 2022 estimates

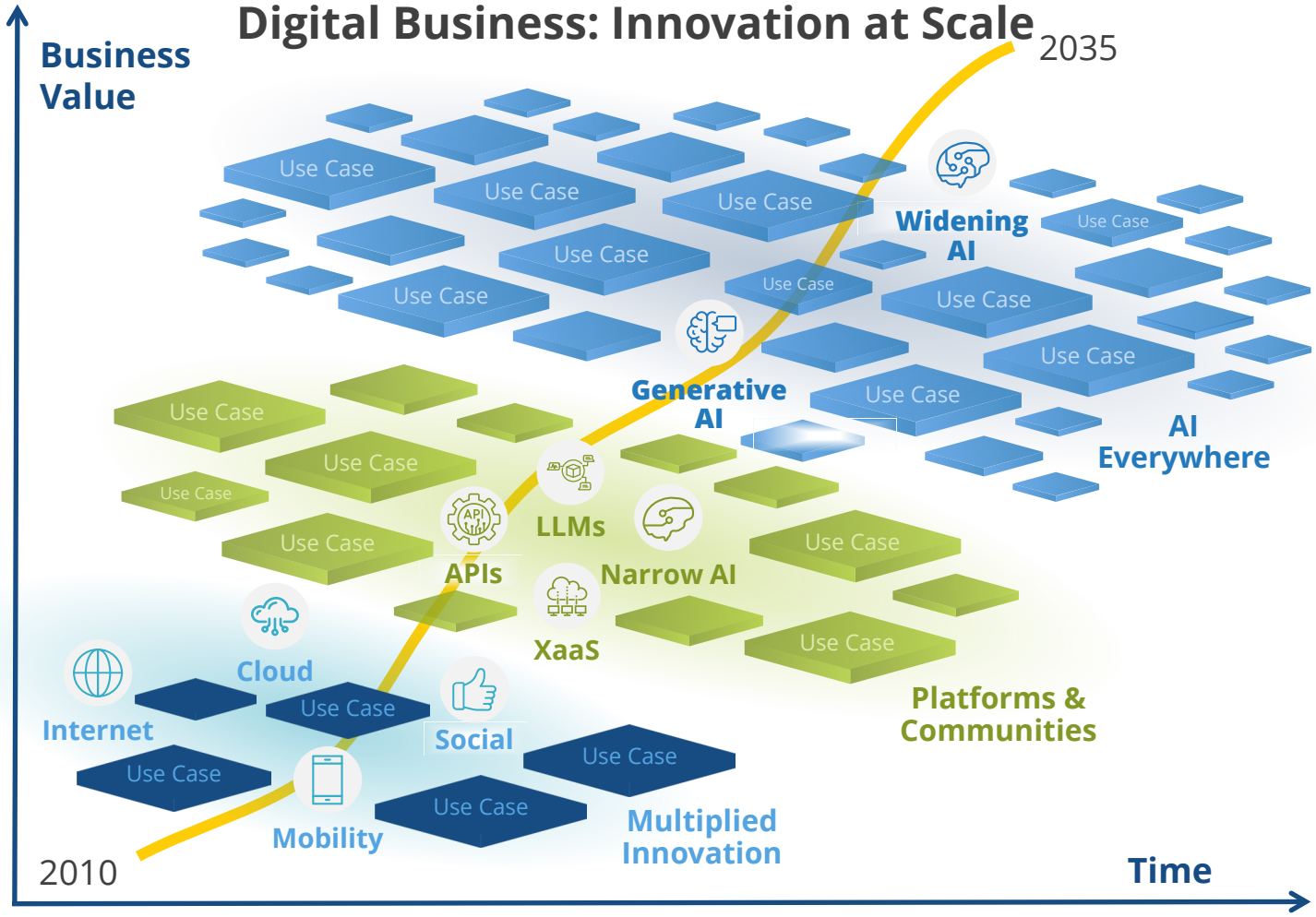
The Path to Generative AI Impact

- **Key Players**: Foundational models from Google, Meta, Microsoft, Amazon, and the open-source community contribute significantly.
- **Cloud-Based Solutions**: Mainly used for building large models, driving demand for cloud services.
- **Fine-Tuning and Data Sensitivity**: Enterprises explore fine-tuning pretrained models on prem to address data sensitivity, cost efficiency, and capacity concerns.
- **Energy Considerations**: Varying energy requirements for models and applications; model training is power-intensive; different media types impact consumption.
- **Datacenter Evolution**: Datacenters adapt for high-density GenAI racks with GPUs and TPUs.
- **Location Considerations**: Centralized locations favored for model training and inferencing.



AI Everywhere

- **AI Journey**. From narrow to widening AI
- **Intelligence Architecture**. A data-centric platform underpinning the enterprise
- **Digital Operations At Scale**. Cost-effective digital infrastructure for AI workloads
- **Skills**. Attracting and reskilling talent for transformed work models
- **Trust**. An upfront focus on trust



Generative AI: Top Business Benefits

Expanding Labor Productivity

Personalizing Customer Experience

Accelerating R&D

Emerging New Business Models

AI Everywhere Readiness Model

Strategy and Roadmap

Align the C-Suite to update your AI roadmap

Trusted Enterprise

Navigate the new dimensions of trust

Intelligence Architecture

Underpin the organization with an enterprise intelligence architecture

Skills

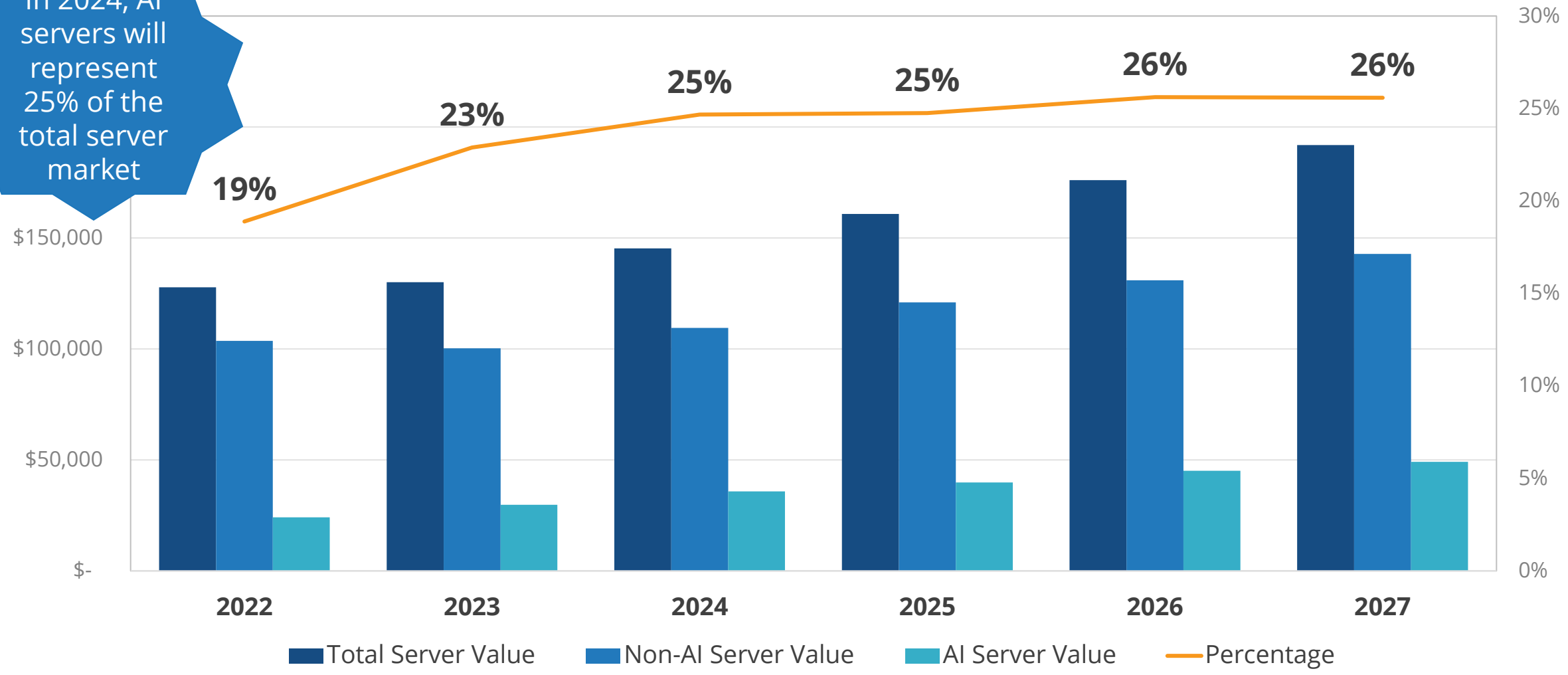
Prepare for required roles and skills

Digital Infrastructure

Attain GenAI productivity improvements in IT

Worldwide AI Server Revenue Trend as a Percentage of Total Worldwide Server, $M

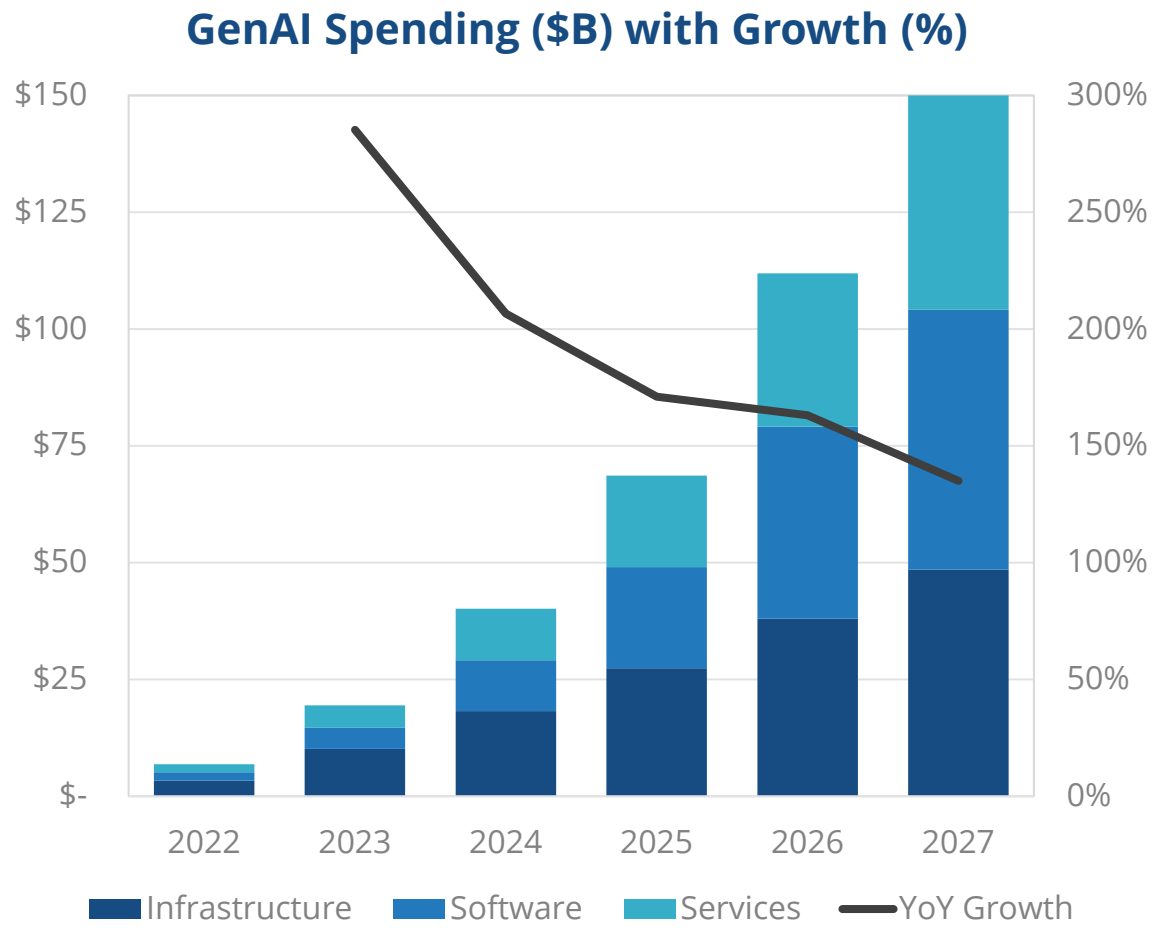
In 2024, AI servers will represent 25% of the total server market



Source: IDC, Nov. 2023

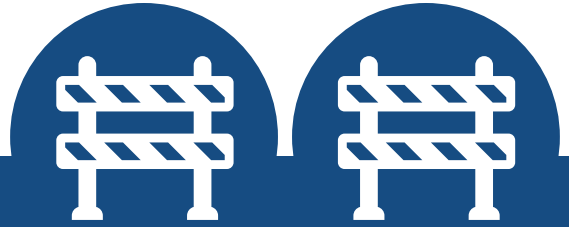
GenAI Will Drive 8% Increase in Infrastructure Spending

Source: IDC FERS Wave 7, August 2023 (n=883)



Source: IDC #US51539723, Dec. 2023

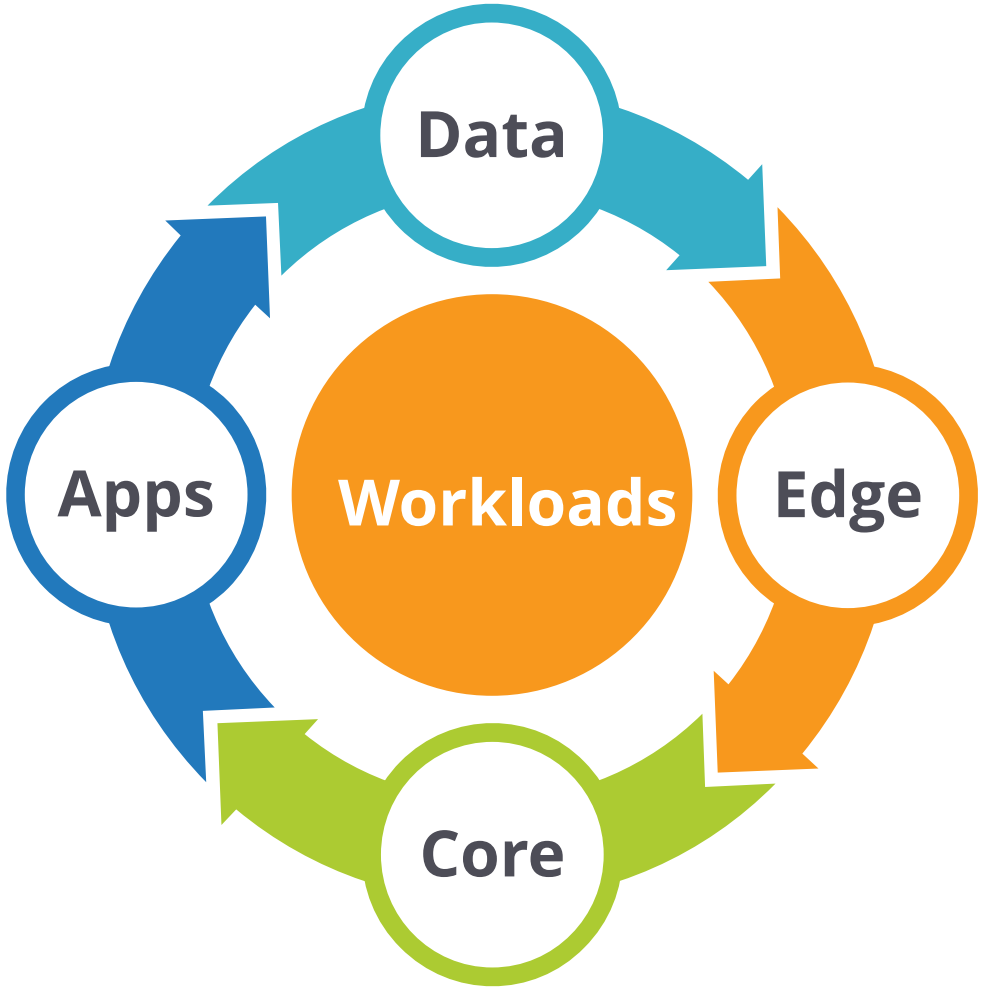
Becoming Digital-First – Major Infrastructure Barriers



39%	35%	34%	32%	32%
cost and complexity of supporting multiple generations of infrastructure is a major barrier from achieving resiliency goals	**Shifting more workloads to cloud** or colocation and not upgrading own datacenters	**Lack of sufficient automation and analytics** to effectively optimize infrastructure	**growing volumes of data, and diversity** is challenging, disrupting data storage, integration, and management	**Supporting older legacy systems** challenge compatibility, security, cost, maintenance, and scalability

N=778, Base=Respondents indicated the barriers to achieving organization's digital infrastructure resiliency goals over the next two years Note: Managed by IDC's Quantitative Research Group; Data weighted by country GDP (500+ emp size); Use caution when interpreting small sample sizes. Source: Future Enterprise Resiliency & Spending Survey - Wave 2, IDC, March, 2022

Continuum of interconnected Workloads



1 Billion
Applications deployed by 2027

3x Data
Data Volume,
Data Variety
Data Velocity

57% of Apps
Will shift platforms Next year

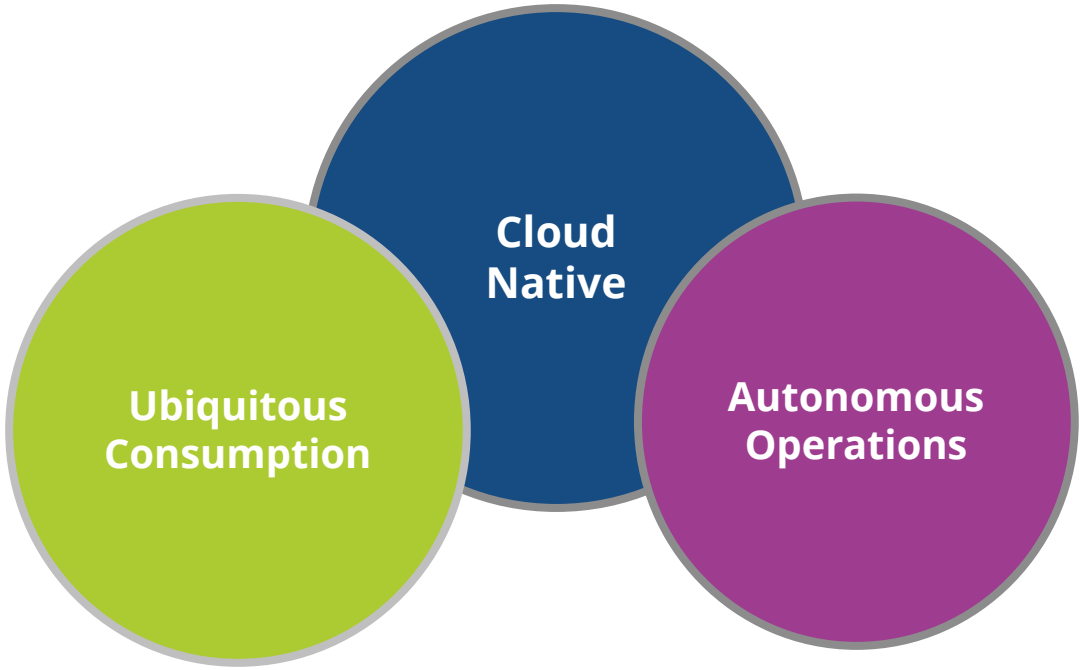
35% of Apps
Using some form of AI today

Key Infrastructure Qualities for Supporting Digital-First Journey

Digital-first organizations need to be resilient, as well as able to optimize and innovate simultaneously

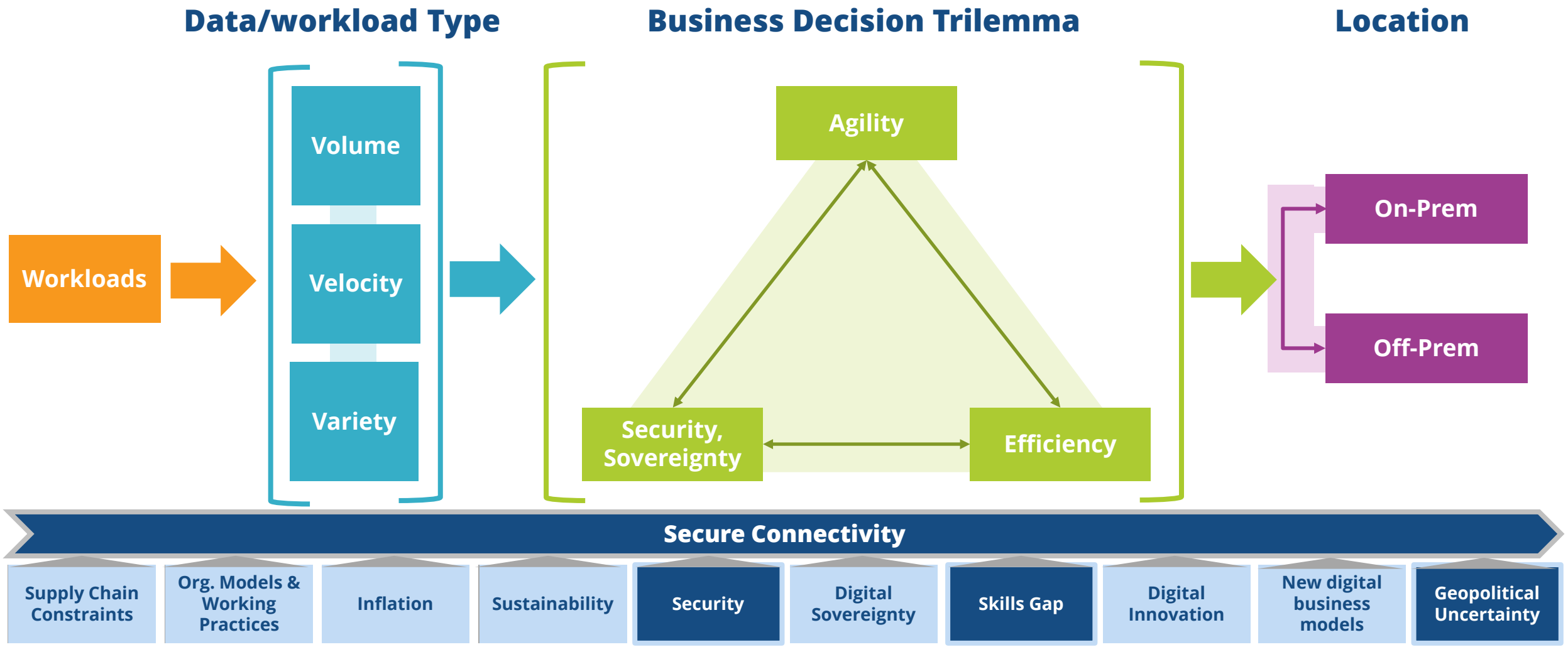
Digital-First Requirements

- Digital-first requires access to IT in all areas where business is conducted and where people and data reside
- Greater automation and ability to scale quickly
- Resilience of IT service as business and operations are underpinned by technology



Optimization	Innovation	Resiliency
Global Reach	Distributed, local resources	aaS Consumption
Secure Connectivity	Autonomous Operations	Cloud Adjacency

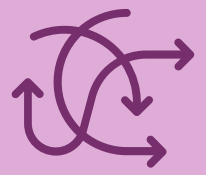
Digital-First Organization Workload Decision Tree



Workload Takeaways

Buyers have different barriers to digitally transform and to manage that transformation on an ongoing basis – understanding their business will be critical to help make decisions on the best location for their workloads

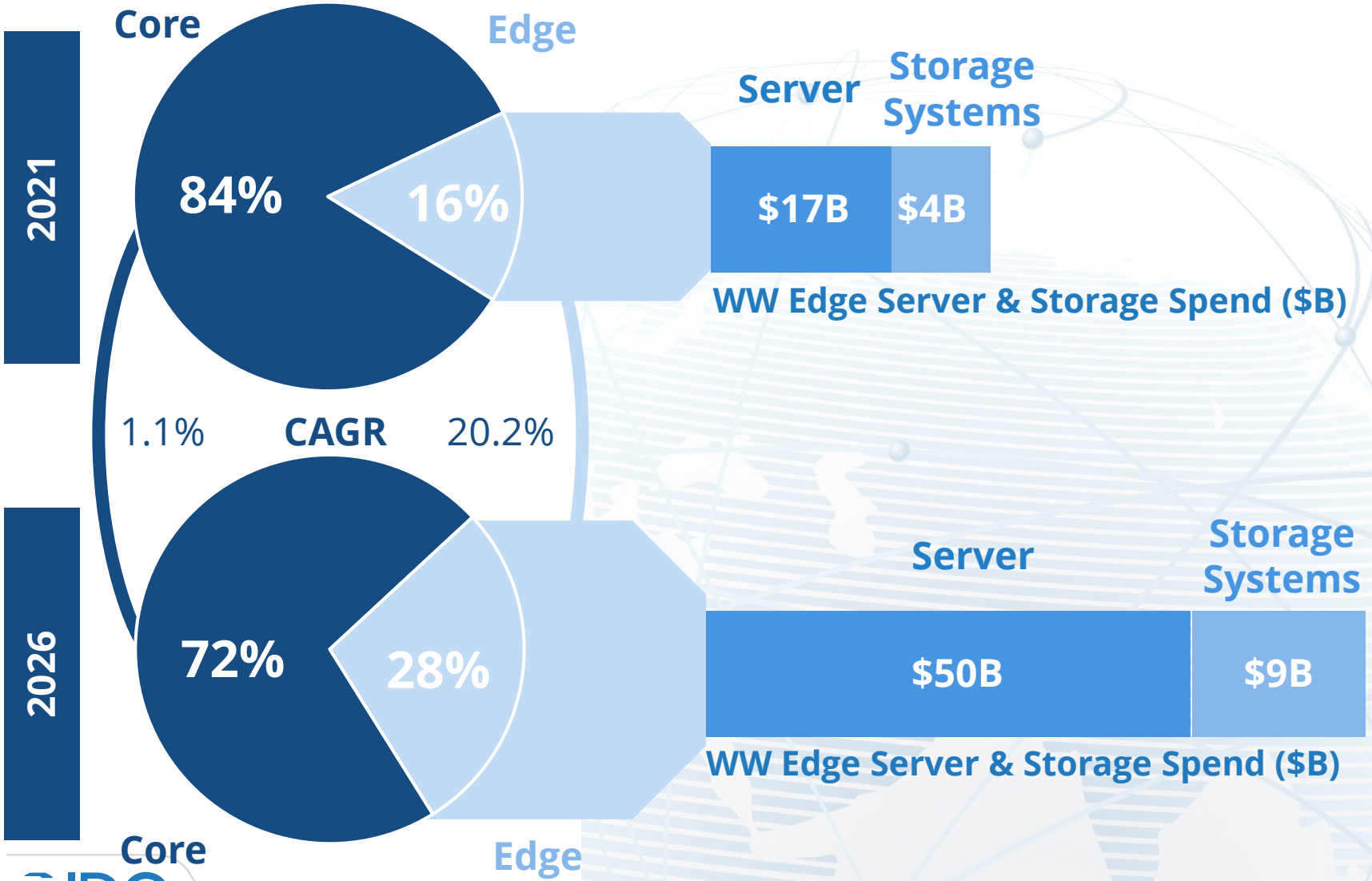
Application modernization – Applications and workloads will need to be cloud capable, to scale and perform like the infrastructure



Workloads will migrate to the edge – connectivity and transparent cost and management of workloads will be key

Worldwide Edge IT Growth

Spend on Edge is expected to grow 20x faster than core with biggest increase in edge servers

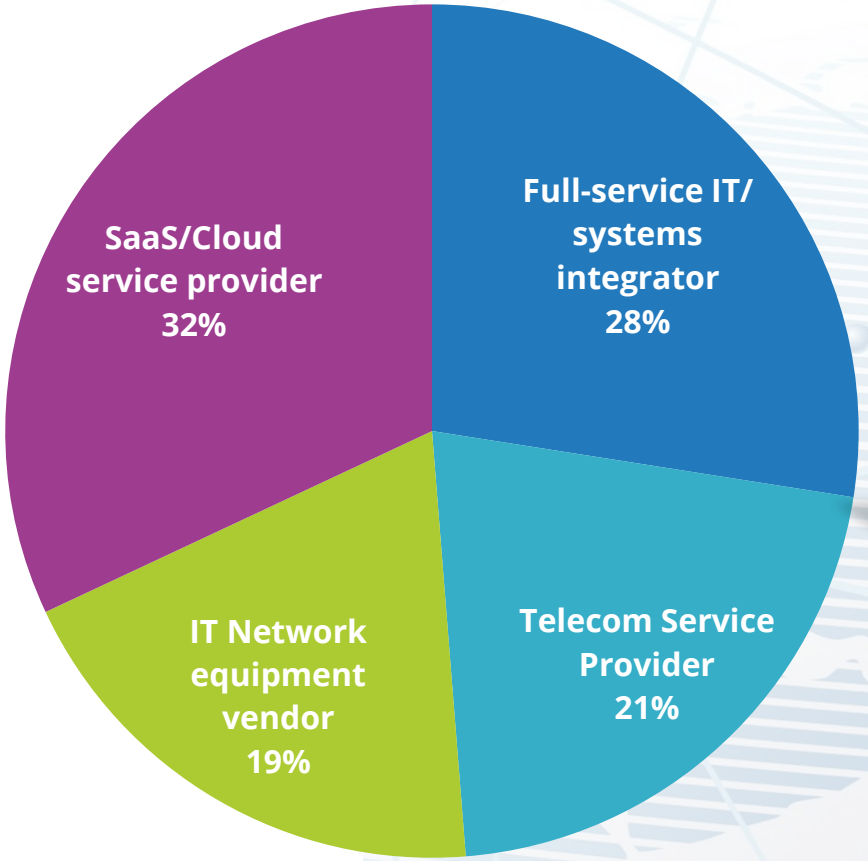


- Digital-first business requires IT service in new locations to support innovation, use cases
- Heightened awareness of data sovereignty and compliance needs. Location and infrastructure placement are part of resilience planning

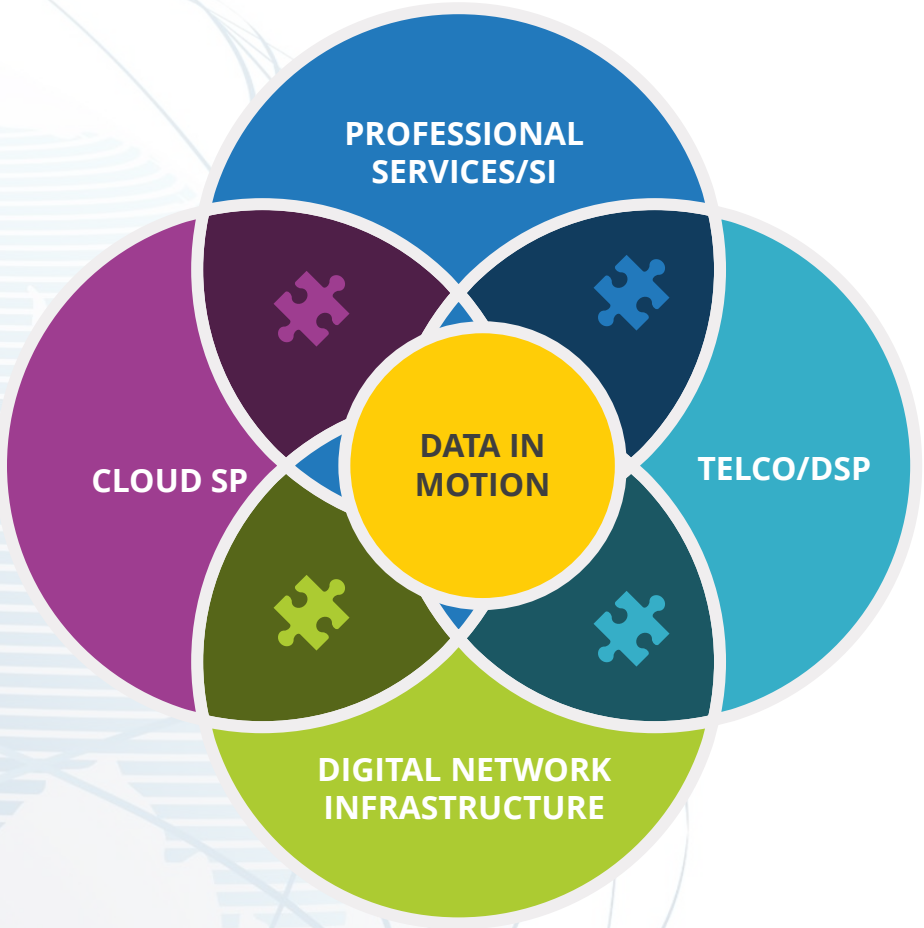
Ecosystem Revolution

Divide around connectivity necessitates a connected ecosystem to achieve positive outcomes

Primary type of service provider organizations use to help business stay connected



Implications for the Connectedness Ecosystem



Source: IDC's *2021 Future of Connectedness Survey, July 2021 n=607*

Entering the World of Cybersecurity in 2023

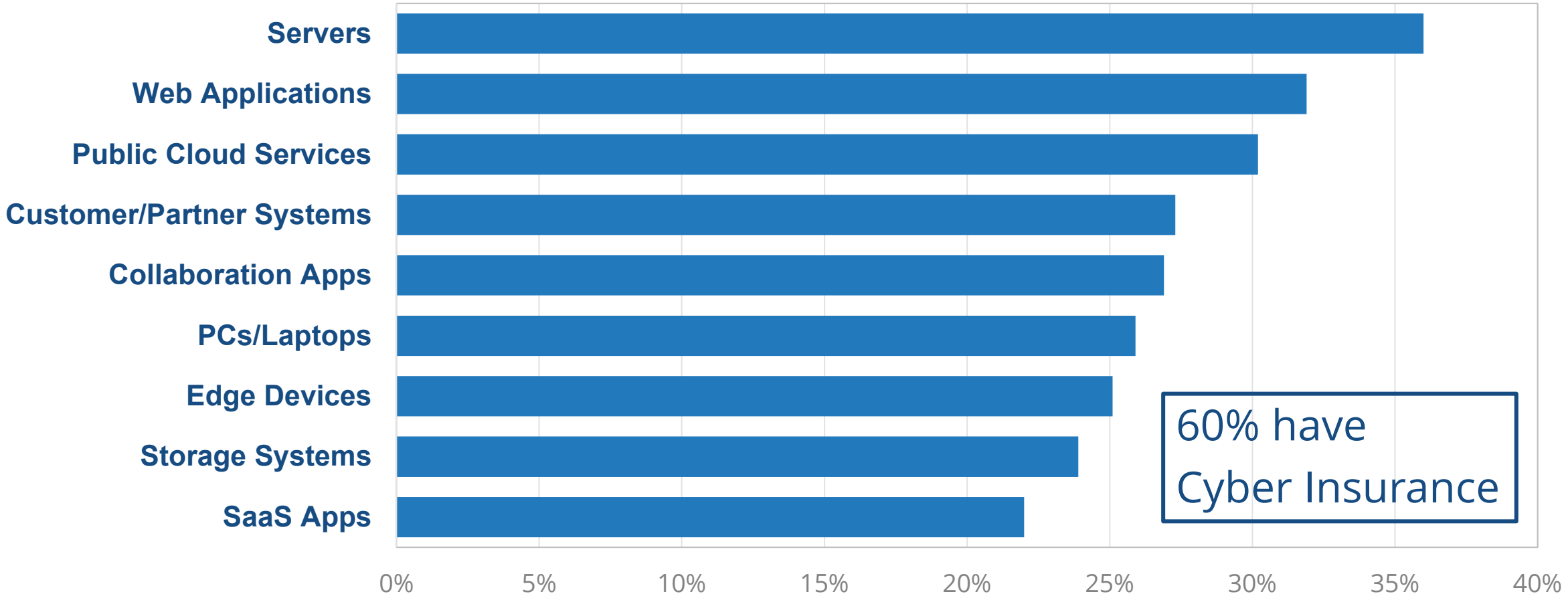
Ransomware Attacks

Q. If your organization paid a ransom in the past 12 months to regain access to systems or data, how much was paid?

Source: Future Enterprise Resiliency & Spending Survey Wave 11 (N=550), IDC, December 2023

Ransomware Impact

Q. Overall, which of the following were directly impacted by ransomware attacks over the past 12 months?

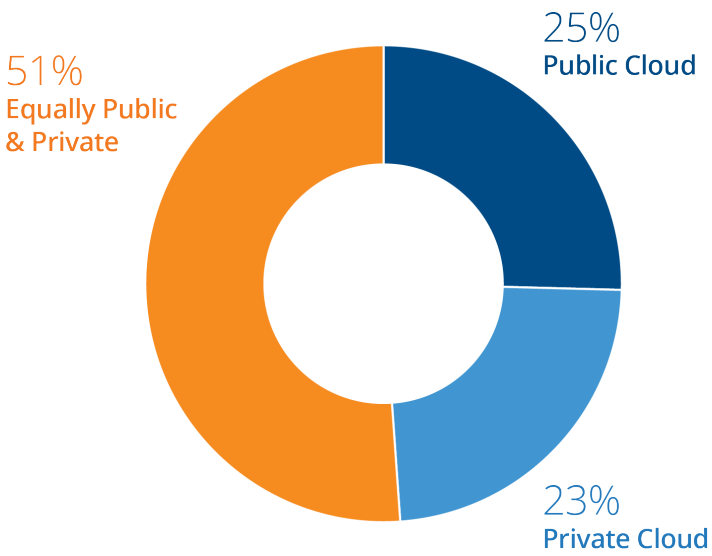


Source: Future Enterprise Resiliency & Spending Survey Wave 11 (N=550), IDC, December 2023

Multi-cloud goals drive the hybrid focus:

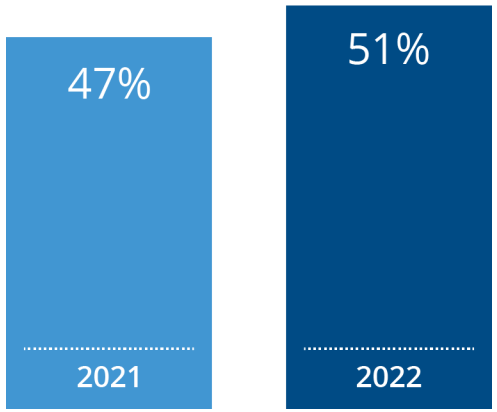
Interoperability between cloud environments continues to grow year-on year

Most important for meeting business goals in the next 5 years



Level of Interoperability Between Clouds		
On Prem Private Cloud	**41%**	Hosted Private Cloud
On Prem Private Cloud	**40%**	Public Cloud
Hosted Private Cloud	**36%**	Public Cloud
Public Cloud	**40%**	Public Cloud

Preference for a single cloud architecture or software platform that can run consistently across multiple different hardware infrastructures



*** % is percent of total respondents**

More than half of companies now say they consider **Public Cloud** and **Private Cloud** as equally important for meeting future business goals

Interoperability is increasing across all clouds, with fewer companies preferring single cloud environments, though companies are less likely to be connecting **Hosted Private Cloud** with their **Public Clouds**

N = 1,700 Source: QC1 , QC2, QC5 3Q22 Cloud Pulse Survey, September 2022, IDC

Vendors must listen to buyers when it comes to cost:

Cloud costs and ROI influence cloud buyers' purchasing decisions

Most important for cloud investments in the next two years

51%

of cloud buyers say **cost management/containment** is highly important when it comes to making cloud investment decisions over the next two years

Cost containment becomes even more important as organizations adopt multi-cloud or hybrid cloud strategies

56% of cloud buyers are now seeking the ability to negotiate price across multiple vendors

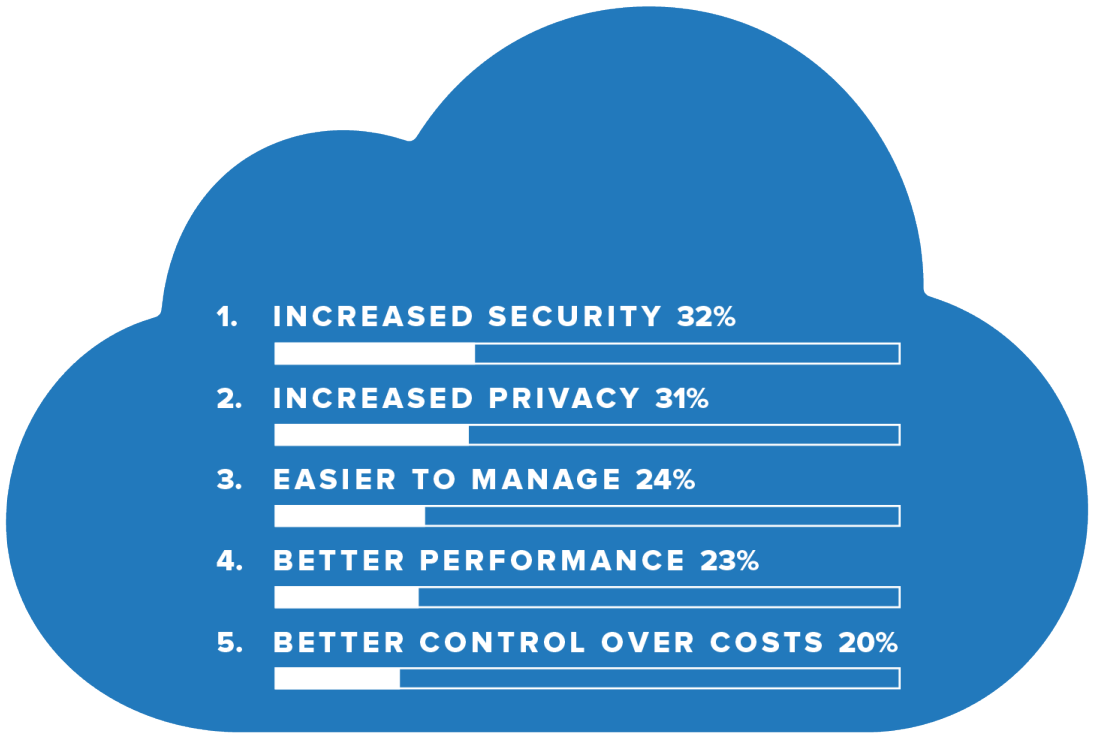
50% of cloud buyers now relate cloud ROI directly to business outcomes

Application performance and reliability are the #1 concern cloud buyers have when making new investments. Concerns pertaining to cost are becoming more relevant as cost-saving initiatives are brought in across companies seeking safe navigation through current macroeconomic headwinds

N = 1,351 Source: QSP9 1Q23 Cloud Pulse Survey, March 2023, IDC

Hybrid cloud combines the benefits of scalability, privacy and security: It also provides the ability to make cost-conscious cloud decisions

Top 5 reasons for deploying private and public clouds



Private Cloud

Public Cloud

N = 1,299 Source QC5 N= 1,294 Source QC3 1Q23 Cloud Pulse Survey, March 2023, IDC

Hybrid changes the cloud services landscape:

Migration, management and planning services become much more important

Most bought cloud or hosted services

Today	In 12 months
1 **Backup & Recovery (28%)**	1 **Cloud Migration Services (22%)**
2 **Security (26%)**	2 **Business Systems Integration Services (22%)**
3 **Hosted Private Cloud (26%)**	3 **Cloud Strategy Consultation Service (21%)**
4 Cloud Migration Services (25%)	4 Cloud Platform Application Recommendation Service (25%)
5 Business Systems Integration Services (25%)	5 **Cloud Training Services (20%) NEW**
6 Cloud Platform Application Recommendation Service (25%)	6 **Multicloud Management Tools (18%) NEW**
7 Cloud Strategy Consultation Service (25%)	7 Dedicated Servers (18%)
8 Antivirus (23%)	8 IaaS/ PaaS Servers/ Storage (18%)
9 IaaS/ PaaS Servers/Storage (23%)	9 **9. End-to-End App Management (18%) NEW**
10 Web Application Firewall (23%)	10 **Functions-as-a-Service (17%) NEW**

29% of cloud buyers already work with **10 or more** service providers for cloud or hosted services. This will grow to **34%** in the next two years.

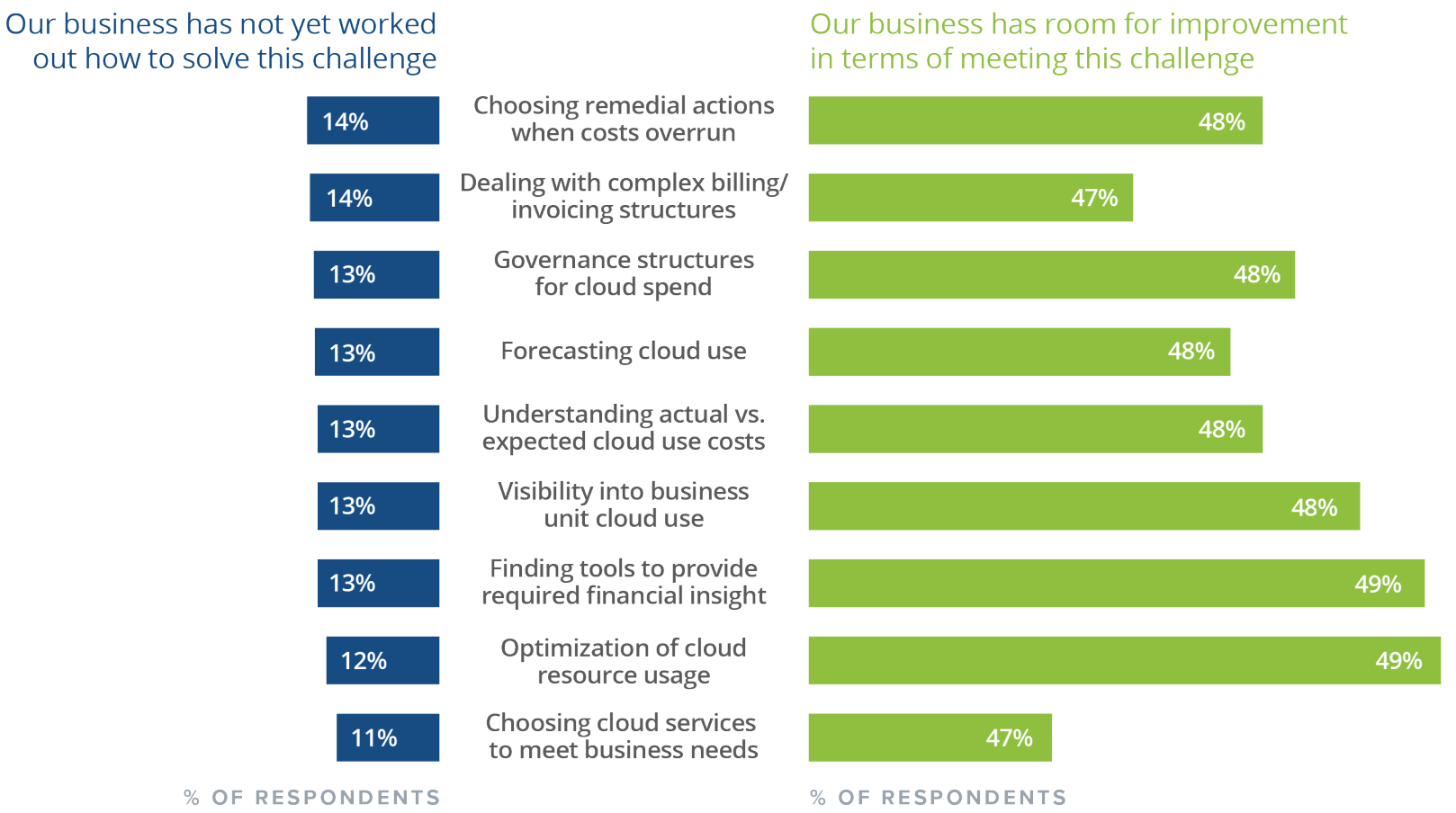
The number of companies doing **hybrid cloud** already working with 10+ providers is already **33%** and will grow to **39%** in two years.

N = 1,351 Source QSP10, QSP11, QSP12 1Q23 Cloud Pulse Survey, March 2023, IDC

Only 40% of businesses have a handle on costs:

Vendors should invest more time on capacity planning, governance and billing

Perception of current cloud cost/optimization-related issues



We are highly adept...

1. At choosing cloud services to meet business needs

2. At dealing with complex billing/invoicing structures

3. At understanding actual versus cloud use costs

n = 1,350 Source: QM12 4Q22 Cloud Pulse Survey, November 2022, IDC

Predictions

Infrastructure

With GenAI as a catalyst, by 2027, 40% of enterprises will rely on interwoven IT architectures across cloud, core, and edge to support dynamic, location-agnostic workflow priorities.

Architecture

By 2026, the strategic importance of GenAI will force a market-driven reset of the current single vendor–dominated coprocessor market for GenAI chips and drive end-user system prices down by 25%.

Ransomware

By 2028, 75% of IT organizations will implement AI-driven anomaly detection built into infrastructure components to thwart never-before-seen ransomware attack types

Operations

By 2025, G2000 that invest in training IT staff to effectively use GenAI prompts for IT operational tools will improve existing ITOps team productivity by 30% due to greater cross-persona collaboration.

Source: Worldwide Future of Digital Infrastructure 2024 Predictions, *IDC #50401023 Oct 2023*

Matt Eastwood
meastwood@idc.com
508.935.4503
@MattEastwood

IDC.com

linkedin.com/MaEastwood

MattEastwood

blogs.idc.com