Modern Infrastructure to Compete in the AI Age

Simon Piff
Research Vice President,
IDC Asia/Pacific

Pushkaraksh Shanbhag
Associate Research Director
IDC Asia/Pacific

Table of Contents

Executive Summary

- We are in the age of Digital Business, with expectations of digital revenue rising **from 14% to 37%** in the next few years.*
- Cloud serves as the foundational architecture for digital business. While cloud computing has driven this growth, a cloud-first strategy is not always economical as enterprise usage scales. As organizations mature in their cloud adoption, they begin to explore architectural models beyond the public cloud.
  - Hybrid cloud and private cloud models.
  - Workload repatriation.
  - Infrastructure modernization.
- This architectural rethink, integrating on-premises infrastructure capabilities into the enterprise hybrid cloud IT framework is particularly relevant in the context of artificial intelligence (AI), where considerations such as costs, data privacy, security, and regulatory compliance take center stage.
- At the core of this transformation is a renewed focus on data—its potential to be monetized through digital products and services, and its role as the fuel powering the AI revolution. To unlock this potential, data management must evolve beyond foundational security and integrity and embrace the growing legislation that is surrounding a range of data types across different industries.
- In the Asia/Pacific* region, the challenge of technical debt is becoming increasingly prominent. Legacy systems, though essential for critical business operations, are often not optimized for today's hybrid cloud architectures. Addressing these workloads requires a strategic blend of application modernization and robust data governance—ensuring that valuable information within these systems is accessible for monetization while carefully planning their transformation roadmap.

Ultimately organizations need to rethink their architectural strategies. Hybrid cloud now requires a modern datacentre (DC) approach, where the scale, flexibility, security and speed of cloud can be emulated as much as possible, and where required, seamless application transitions to well-managed cloud environments can take place, ensuring operational continuity and efficiency.

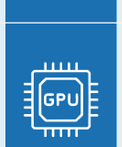
In this document we will look at the key decisions that business and technology leaders need to take into consideration when building out their on-premises environment to ensure these systems can support the evolving demands of hybridcloud and AI-driven operations.

Moving Forward, the 5 Key Advisories are:

Hybrid Cloud Architecture is the Future: Ensure all future infrastructure can support a hybrid cloud computing model.

Compute Platform Considerations: Establish platforms that bring together server, storage and data in a seamless and efficient manner.

Protect Your Data Assets: Ensuring continuous protection requires a new focus.



Define Your AI Needs: What is your specific AI demand? What alternative architectures meet the business need?

AI Workload Diversity: This is the beginning, but AI will be embedded across the organizations. Plan for tomorrow.

*Asia/Pacific includes Southeast Asia (Singapore, Malaysia, Indonesia, Thailand, Philippines and Vietnam, Japan, South Korea, India, ANZ, Mainland China and Greater China markets.

*Source: Future Enterprise Resiliency & Spending Survey Wave 6, IDC, August, 2024 (n=300 Asia/Pacific)

Aligning to the Business: The C-Suite Imperative

C-Suite Imperative: Navigating Digital Transformation and Tech Strategy Risks

CEOs are far more engaged with IT then ever before. This shift to digital business means that aligning with the C-Suite has become critical to CIO success. As a result, understanding the challenges of the business must now be the foundation upon which all IT decisions are made - alignment is critical to overall success.

Headwinds of Change: Top Risks to Executing Successful Tech Strategies:[1]

- Technology supply chain constraints.
- Managing demand for AI, within budgets.
- Managing cloud budgets.

Going Digital:

- CEOs expect 37% of revenue from digital products in the next five years, pushing C-Suite interest in leveraging technology.[2]
- Key strategic imperatives include operational efficiency, enhanced customer experience, and digital initiatives.
- AI investments currently have mindshare, but issues around technical debt and repatriation of some workloads are emerging.

Ensuring the balance between being able to invest into the desired technology, whether on or off-premises, while managing budgets is the delicate balance all organizations need to strive for.

Quote:

"You need to have a handshake—a very, very strong handshake—with the business; only then will things work, perform well, be manageable, and remain cost-optimal."

CIO, Retail & Supply Chain, South East Asia

Source: [1]Future Enterprise Resiliency & Spending Survey Wave 6, IDC, August, 2024 (n=300 Asia/Pacific); [2] IDC Worldwide CEO Survey, February 2024 (n=115 Asia/Pacific)

C-Suite Imperative: Overcoming Barriers to Move from Proof of Concept (POC) to Production

Generative AI (GenAI) is showing significant gains for those that can move from POC to production quickly, but most are still challenged by data silos, skills shortages, confusion over the multitude of options and, in Asia/Pacific, a degree of skepticism from some CEOs.

The AI Emphasis:*

- AI is seen as a key differentiator, but only **14%** of initiatives are operationalized.
- Adoption barriers include security, compliance, data management, and skills shortages.
- Only **12%** of organizations have a dedicated Chief AI Officer, highlighting a gap in AI leadership.
- The AI strategy is one of "time-to-market". Those organizations that are able to operationalize their AI investments first will accelerate away, in business success terms, from those that are struggling. IDC predicts that in 2026, over one-third of organizations will be stuck in the experimental, point-solution phase of AI experimentation, requiring a shift of focus to enterprise use cases to deliver return on investment (ROI).
- While public cloud provides a fast track for AI adoption, security and privacy concerns will lead many Asia/Pacific organizations to seek purpose built, AI-ready infrastructure that can be deployed to their own DC, edge, or colocation provider.

Quote:

"Our internal company is not prepared for AI because of platform weakness. The platform is not ready."

CIO, Manufacturing, India

Source: *IDC Worldwide CEO Survey, February 2024 (n=115 Asia/Pacific)

The CIO Challenge

The CIO Challenge: Balancing Limited Resources With Exceptional Demand

Top Trends Driving Change in Enterprise Utilization of Compute and Storage:[1]

AI: CIOs are tasked with leading AI initiatives with a focus on business-aligned projects that generate revenue.

Cost Management: Cloud repatriation is on the rise as organizations refine workload requirements and seek cost efficiency.

Addressing Technical Debt: **67%** of current on-premises applications are residing on traditional, non-cloud (virtualized) platforms and are beginning to become a drain on funding and resources. IDC predicts that by 2029, significant losses are projected due to outdated infrastructure, underscoring the need for modernization and hybrid cloud flexibility.

Options for application deployment are now many, various cloud and hosted platforms, along with traditional core and edge mean that CIOs need to deeply consider the current and future needs of the applications being deployed. With this abundance of choices comes increased complexity—an obstacle to achieving operational efficiency, which remains the top priority for CEOs.[2]

CIOs need to consider that a platform approach to infrastructure affords a greater degree of automation, leading to higher efficiencies, when sourcing server and storage from a single vendor.

Source: [1] Enterprise Workloads Infrastructure Study, 2H23 (n=750 Asia/Pacific); [2] IDC Worldwide CEO Survey, February 2024 (n=115 Asia/Pacific)

AI Tsunami: How not to get Swamped

The AI Tsunami: Overcoming Barriers to GenAI Success in Asia/Pacific

52% of Asia/Pacific organizations are either testing (**27%**) or using (**25%**) GenAI but only **22%** are so far using their own data.[1]

- IDC predicts a third of organizations will struggle to move beyond experimental AI phases without a focus on ROI-driven use cases.
- Part of the challenge is that the Asia/Pacific C-Suite remains largely unconvinced of the ROI of GenAI investments, with the current focus being on improving customer experience.

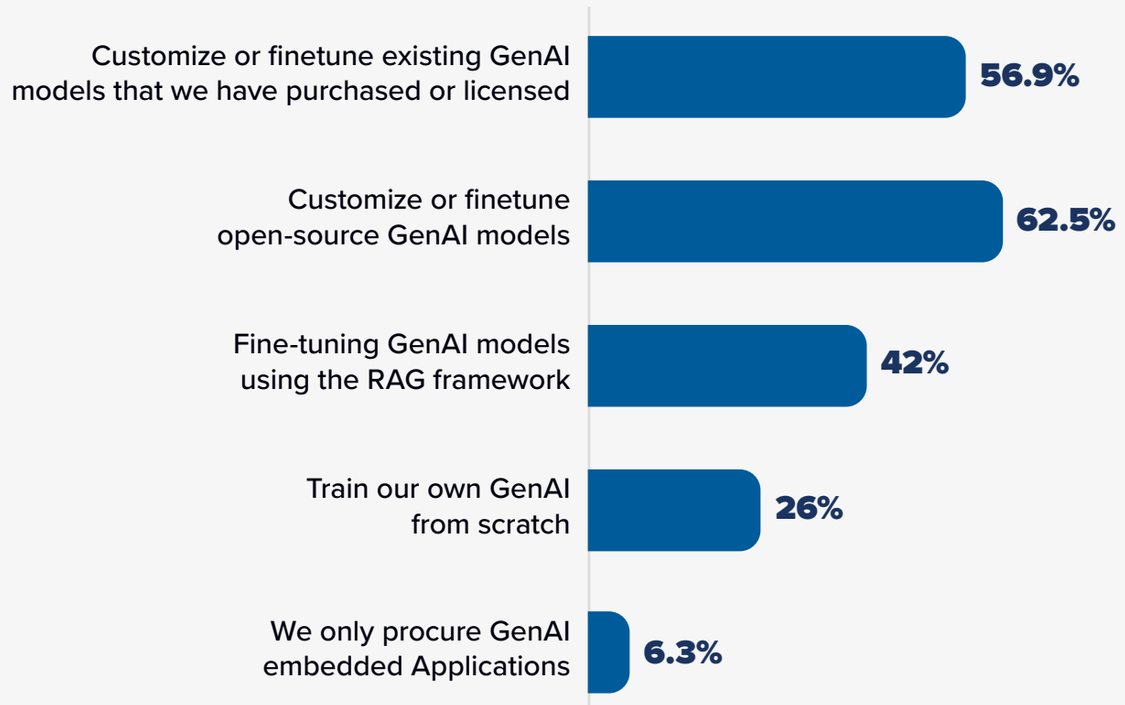
Cost and complexity are preventing organizations from unlocking higher value from their digital data and content.

- Organizations have an average of seven data silo's within their organization in-region, that are known and manged by IT, for some this number is above 20 data silo's and **78%** of organizations agree this is an inhibitor to success.[2]

A focus on fine-tuning models will delay usage.

- Fine tuning is ideal when deep understanding is required, but the majority of Asia/Pacific businesses could see earlier returns with a focus on Retrieval Augmented Generation (RAG), where integrating their own context into an existing model could deliver faster results.

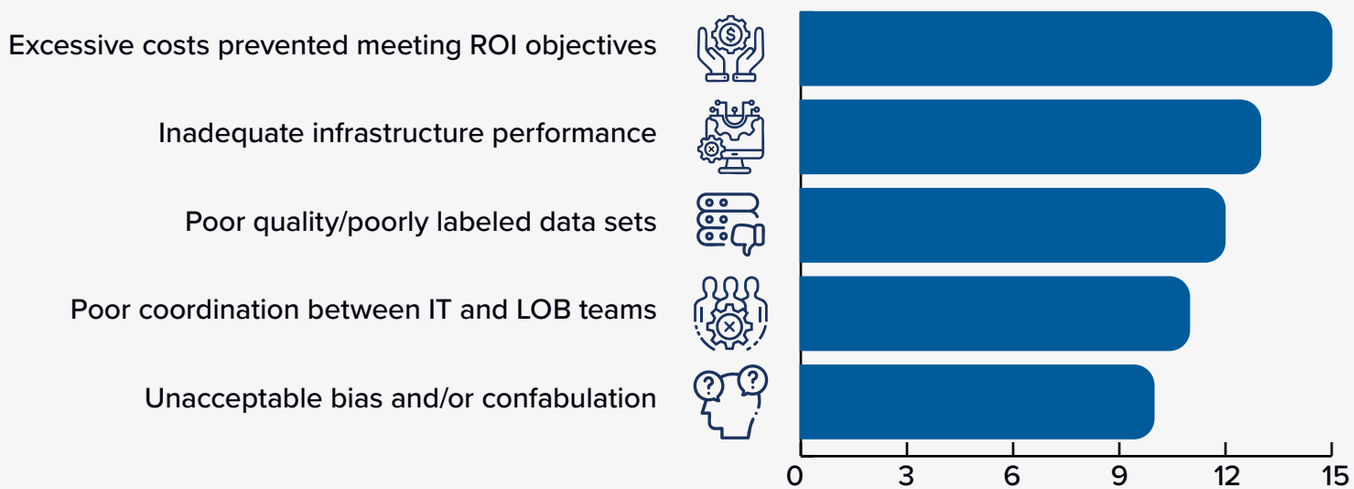
In 2023-2024, How did Your Organization Train or Tune GenAI Models?[3]



Source: [1] GenAI ARC Survey, IDC, August, 2023 (n=450 Asia/Pacific); [2] Worldwide Digital Infrastructure Sentiment Survey, IDC, June, 2024 (n=342 Asia/Pacific); [3]Asia/Pacific Data, AI, GenAI and Insights Survey (n=269)

The AI Tsunami: Implementing AI Strategy

Within Asia/Pacific the Top Barriers to AI Success are:*



Enterprise-wide AI Strategy

- Align Governance & AI Strategy
- Work Model AI Enhancement
- AI Infused Processes & Apps
- AI Platform Unification
- Eliminate Dark Data
- Hybrid Cloud Infrastructure for Scale

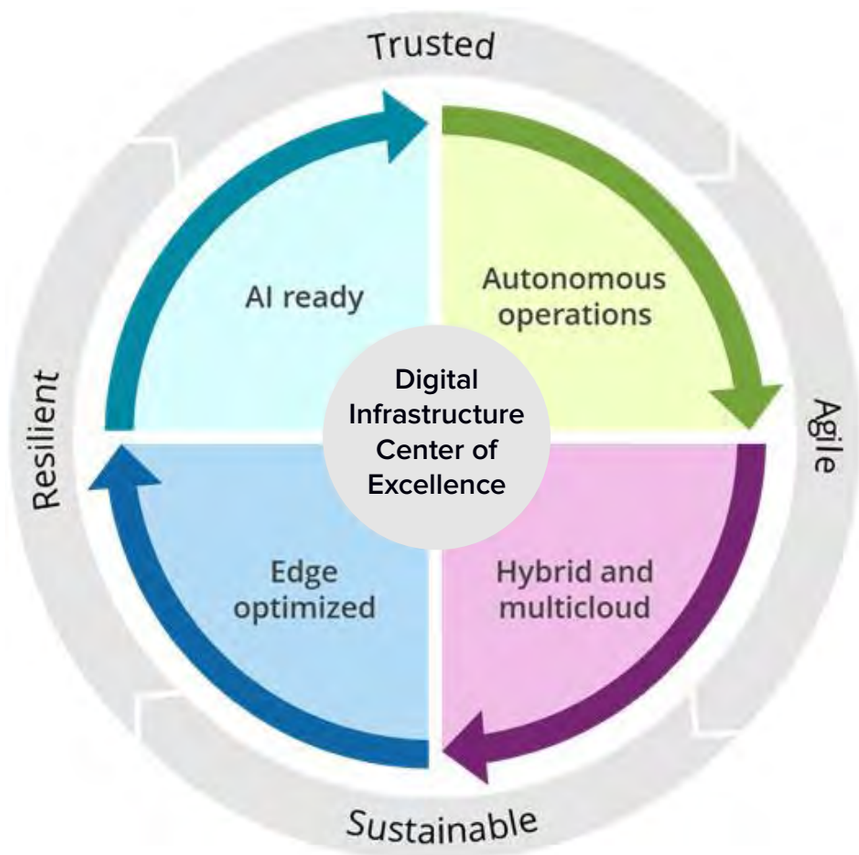
- IDC proposes organizations implement an enterprise-wide AI strategy, that can address the five main barriers to success.
- This is a cross functional team that embraces Line-of-Business (LOB) and IT, ensuring key stakeholders are engaged at the outset of programs, and allows IT to fully inform the business of issues and challenges to success.

Source: *Future Enterprise Resiliency & Spending Survey Wave 4, IDC, April, 2024 (n=300 Asia/Pacific)

Modern Infrastructure Demand

Modern Infrastructure: AI-Ready, Hybrid Multicloud, and Edge Integration

- A robust and scalable modern infrastructure that effectively addresses the immediate and future needs of the business is necessary to ensure agility, efficiency, and long-term growth.
- AI-readiness is now a business expectation, and CIOs must ensure that their platforms can support autonomous operations. By leveraging automation and AI, organizations can enhance functionality and performance without overburdening resources or compromising productivity.
- Hybrid multicloud is also a critical choice, and knowing how and when to leverage these multiple foundations is rapidly becoming a critical function of IT, driven by Infrastructure Centres of Excellence in the best performing organizations globally.
- Edge also needs to be taken into consideration. IDC believes that 20% of data used for AI will be generated at the edge,* so ensuring secure and simple integration to the core has to be taken into consideration.
- While there will be many components to this modern infrastructure, the overall guiding principals will be: trusted, agile, sustainable, and resilient.



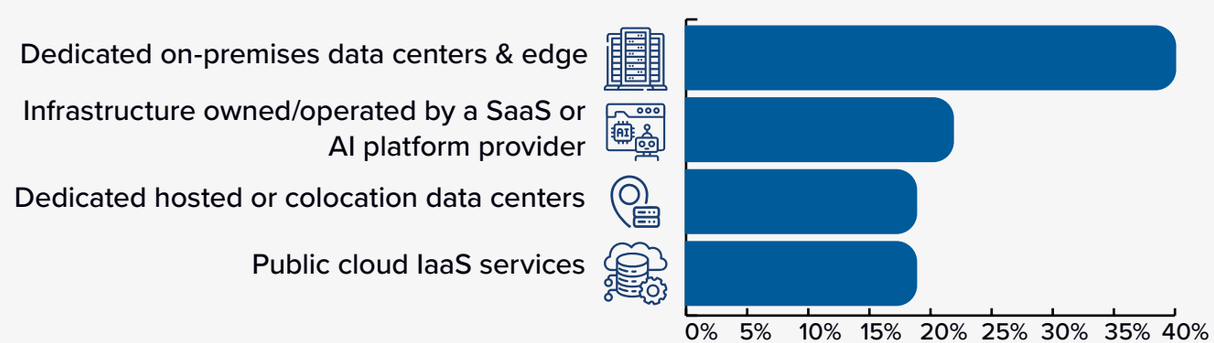
Source: *The Future of Digital Infrastructure, 2024: AI-Ready Platforms, Operating Models, and Governance March 2024, IDC #US50614724

Modern Infrastructure: On-Premises GenAI Tuning and Cost Optimization

Modern Infrastructure:*

- On-premises is proving to be the main location for GenAI model tuning, which is the business-critical part of the process, since it is an ongoing effort while use of RAG captures many internal data sources, further driving an on-premises expectation.
- Critical elements of the on-premises requirements will not only be the hardware, but also the data management tools.
- Concurrently, cost management is reaching an inflection point: cloud costs are rising, technical debt is emerging, and the ROI for AI is under focus.
- Ensuring the right systems are procured, that can address not just AI, but also repatriation and modernization, is a prerequisite for success.

Primary Infrastructure Choices for GenAI Model Tuning are:*



Source: *Future Enterprise Resiliency & Spending Survey Wave 4, IDC, April, 2024 (n=300 Asia/Pacific)

Server

The Server Platform: AI and Beyond

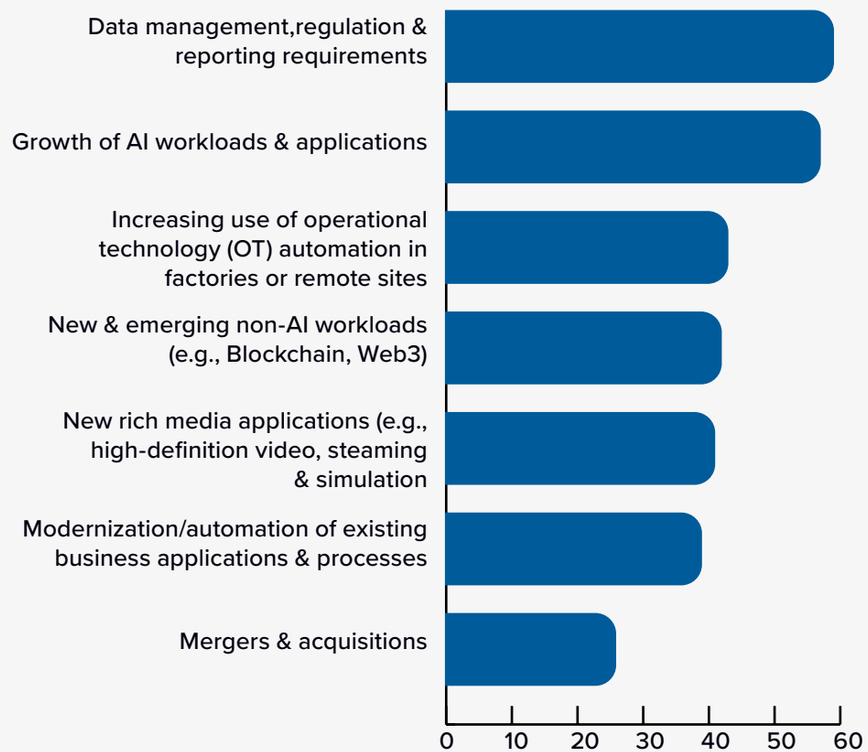
The infrastructure choices for AI are typically optimized for speed but often fail to address the organization's long-term requirements for scale, interoperability, security, and cost control.

01	Define your AI needs and leverage a range of fit-for-purpose platforms optimized for the unique needs of different AI use cases in production.
02	Establish a consistent framework for assessing workload requirements and deployment choices based on critical service level objectives (SLOs) and key performance indicators (KPIs).
03	Focus on data first when making infrastructure choices.
04	Plan for interoperable hybrid cloud architectures.
05	Be ready to continuously adapt as infrastructure and AI capabilities evolve.
06	AI systems are similar to High-Performance Computing (HPC); density of systems and cost of running need to be considered.

CIOs Currently Need to Architect With Three Things in Mind:[1]

- **AI**: While one of the most pressing needs, but it is not the entire challenge.
- **Modernization/Addressing Technical Debt**: Perhaps a more challenging issue as it calls for re-architecting while ensuring that existing performance characteristics are maintained or improved.
- **Cloud Repatriation**: Finding the right placement for workloads is becoming increasingly important, making cloud-native architecture essential.

Which of the Following Trends Will Drive the Most Change in Your Organization's Utilization of Compute and Storage Resources Over the Next Two Years?[2]



Source: [1] Harnessing Hybrid Infrastructure to Fuel AI Business at Scale: A C-Suite Playbook August 2024; [2] Worldwide Digital Infrastructure Sentiment Survey, IDC, June, 2024 (n=342 Asia/Pacific)

The Server Platform: Hybrid Cloud Capability Essential

- Aside from AI, application modernization and cloud repatriation are two areas that are growing in importance and need to be considered when architecting for the future.
- To fully leverage hybrid cloud architectures, applications need to be designed as cloud-native. For infrastructure, this means supporting containerization and leveraging virtualization.
- Currently, only **67%** of applications run in virtualized environments. While this is set to grow, organizations need to carefully consider the right virtualization platforms.*
- Further, to address skills shortages and drive efficiency, server platforms must allow for extensive automation, be secure and deliver on the C-Suite's environmental, social, and governance (ESG) expectations.

Most Important Challenges Organizations Need to Address to Maximize the Business Value Delivered by IT Infrastructure are:

Lack of knowledge about accelerated computing for AI.

Lack of up-to-date IT and cloud operations staff, skills and automation.

Poor hybrid cloud and multicloud interoperability.

37% of organizations are planning some form of cloud-repatriation and the top 10 workloads are as follows:*

Unstructured Database
Text & Media Analytics
Engineering/Technical Applications
Networking & Security
Supply Chain Management Applications
Content Applications
HR/Human Capital Management
Infrastructure such as File & Print, Systems Management
Content Delivery
Industry/Company-specific Business Applications

- After AI, cloud (especially hybrid cloud) presents a significant challenge for organizations. As a result, CIOs must consider a server platform specifically designed to meet the demands of hybrid cloud environments. Today, vendors offer a variety of embedded AI solutions for IT Operations (AIOps), including AI-driven, cloud-based applications for proactive management and predictive analytics. These solutions help predict, prevent, and resolve infrastructure incidents faster—before they impact service availability.
- Leveraging such innovation aligns to key goals of the C-Suite.

Source: *Enterprise Workloads Infrastructure Study, 2H23 (n=750 Asia/Pacific)

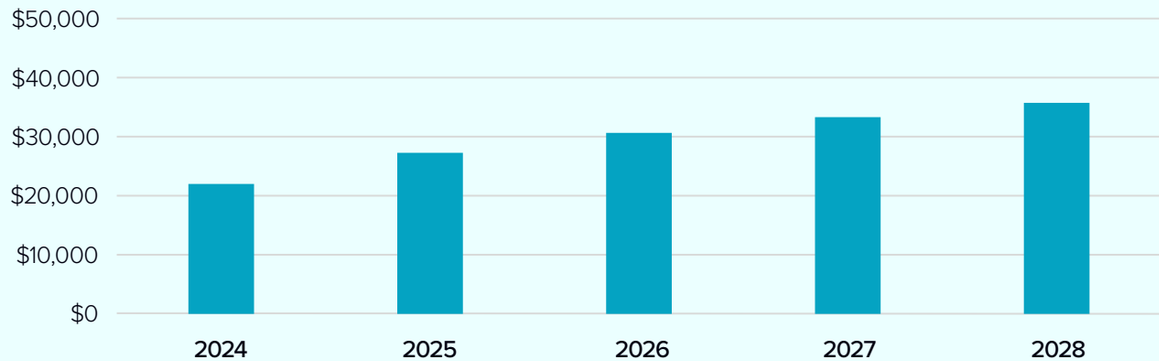
The Server Platform: Growth of Edge

What is Edge? Edge is a distributed computing paradigm that includes infrastructure and applications outside of centralized, dedicated and cloud datacenters located as close as necessary to where data is generated and consumed.

Adoption Trends:[1] Across Asia/Pacific, 41.3% of organizations are currently using edge technology, reflecting its growing importance in business operations. Further, edge is delivering results for many organizations as 45% of organizations report that half or more of their edge projects have delivered a return on investment.

Market Growth:[2] IDC forecasts Asia/Pacific edge computing hardware spending to reach $36 billion by 2028 at a CAGR of 20.5%, driven by demand for low-latency processing and improved operational efficiency.

Edge Computing Hardware Spending ($M)



Key Factors Driving Edge Deployment are:[3]

- Upgrading existing deployments
- Implementing new digitally enhanced business operations
- Implementing new digital processes in the field
- Creating customer experiences and engagement
- Business expansion into new regions

While upgrading existing deployment remains the top priority, the demand for new operations and processes is driving the majority of edge workloads.

Source: [1]IDC Future Enterprise Connectivity Survey 2024, (n=250 Asia/Pacific); [2] IDC Worldwide Edge Spending Guide - Forecast 2024, Aug (V2 2024), Asia Pacific; [3] IDC Edgeview 2023 (n=800 Worldwide)

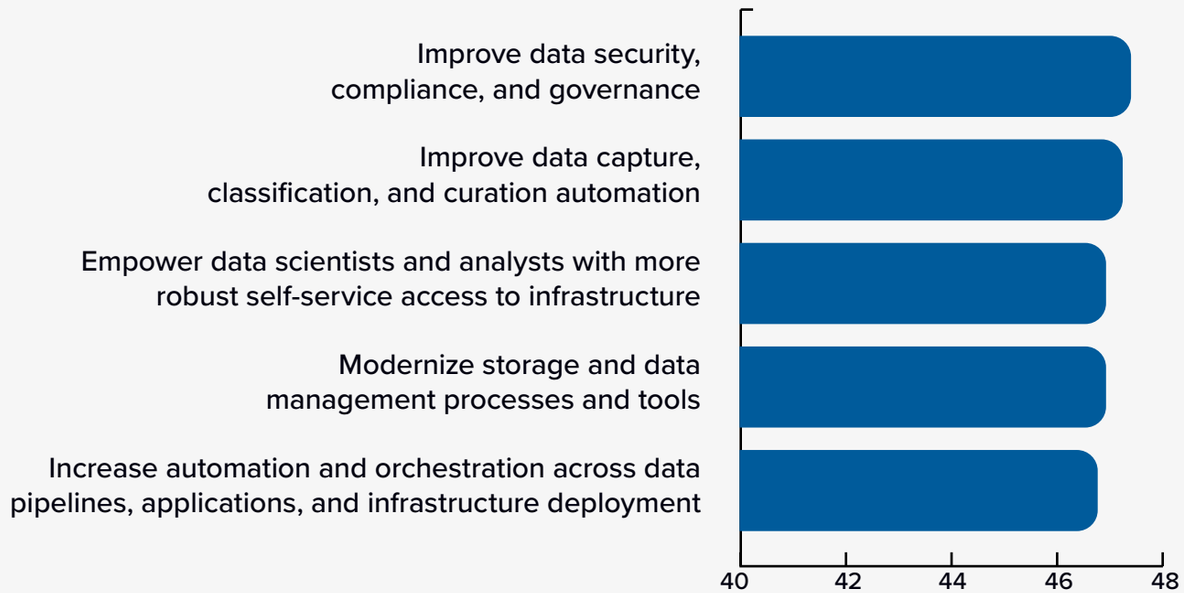
Storage

Storage: Making the Right Platform Choices

The current modern infrastructure will demand storage systems that can deliver on a range of features that aside from 'AI optimized', would include support for:

- Business-critical transactional workloads
- Virtualised applications
- Database workloads
- Modern containerized apps
- Multi-protocol file storage including cloud integration
- Scaling emerging workload performance
- Zero trust architectures
- Simplifying hybrid cloud
- Accelerating application delivery
- Optimizing workload placement
- Edge data capture and security

Top Five Actions to Ensure IT Infrastructure Supports Data-Driven Business and AI Needs*

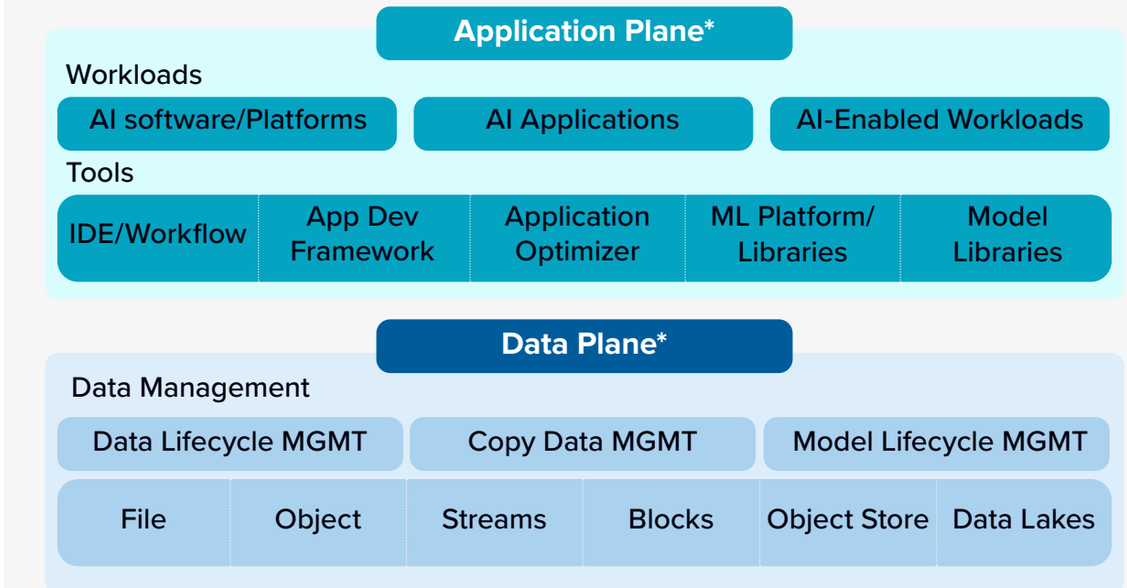


- The security and integrity of this data as it relates to ongoing business resiliency and the regulatory environment cannot be ignored. With **37%** of revenue being digital today, and many data types subject to a growing range of regulations, digital businesses need to protect data to ensure ongoing operations.*
- As more data is generated at the edge, the use of AI inferencing becomes more relevant to the timeliness of incoming data. Inferencing at the edge and relaying results will require an edge data storage and protection strategy to be in place.

Source: *Worldwide Digital Infrastructure Sentiment Survey, IDC, June, 2024 (n=342 Asia/Pacific)

Storage: Creating the Organizational Data Plane

- Many businesses make the mistake of treating AI as a standalone workload rather than a complex set of new, extended and interconnected applications and data. Adoption of an architecture that defines both application and data planes help create clarity.
- Today's AI-driven workloads present a very different input/output profile to traditional applications, or even HPC systems. AI workloads need to be able to perform many small, random reads simultaneously against massively scalable data sets, requiring, in many cases, very low latency as well as high read and write throughput, and need to efficiently handle metadata-intensive operations.
- Many of these AI workloads are mission critical so enterprises need systems that are easy to back up, provide flexible degrees of resilience, and can be restored quickly and easily.



- AI can be a standalone workload in some cases, or as is often the case it could be an incremental enhancement to an existing enterprise workload, or embedded functionality within a new enterprise workload.
- The elements of an organization's AI application plane and AI data plane play a key role in determining the infrastructure stack that will deliver the best business outcome and return on investment concurrently.

For many organizations, implementing a data lakehouse for AI will be a strong starting point to enable the data governance to be centralized.

Source: *Harnessing Hybrid Infrastructure to Fuel AI Business at Scale: A C-Suite Playbook August 2024

Data Protection

Data Replication and Protection: Delivering Resilience

As digital revenues increase, the importance of data integrity, data availability and data security is rapidly becoming a significant concern for the C-Suite.

Considering that more than half of all organizations in Asia/Pacific experienced some form of ransomware attack in 2023 and 2024, data protection systems need to be able to provide:*

- The speed of protection for significantly large datasets, but also offer near real-time snapshots to protect from ever-present cyber threats, such as ransomware.
- Restoring data needs to be equally fast, and to achieve that, there needs to be inbuilt deduplication technology to ensure maximum efficiency.
- Protection for modern workloads in on-premises, virtual, and multicloud environments, including in-cloud and cloud-native applications.
- Data protection designed for performance, efficiency, and scale.

Data protection is no longer merely about securing access to information, it is about ensuring uninterrupted business operations as well as compliance with regulatory demands.

Source: *Future Enterprise Resiliency & Spending Survey Wave 6, IDC, August 2023 (n=300 Asia/Pacific)

Data Protection: Securing the Foundations of Digital Business

The #1 action that will ensure IT infrastructure can fully support data-driven business requirements including AI is to secure the data.[1]

Considering the repatriation list of applications, it emerges that these applications are a valuable source of information that AI will need to draw from, so ensuring that these sources are equally protected as AI algorithms and output is now a top priority.

While security teams understand the need to protect the overall environment, specific data security protocols need to be in-place to ensure that data resilience is not optional, but an embedded process across all applications.

Ensuring that dedicated data protections can deliver on the requisite levels of business resilience will separate the leaders from the rest. Not simply protection, but integrity and recovery are crucial aspects of any data protection strategy.

- As per IDC research, two of the top three factors regarding the payment of a ransom when confronted with a ransomware attack were insufficient air gapping and lack of immutable backups.[2]
- Ensuring air-gapped backups do not contain malware is now something AI is capable of ensuring.
- Data, and therefore, business resiliency needs to be assured.

Leading workloads to be repatriated from public cloud

- Unstructured Database
- Text & Media Analytics
- Engineering/Technical applications
- Networking & Security
- Supply Chain Management Applications
- Content Applications
- HR/Human Capital Management
- Infrastructure Such as File & Print, Systems Management,
- Content Delivery
- Industry/Company-specific Business Applications

Source: [1]Worldwide Digital Infrastructure Sentiment Survey, IDC, June, 2024 (n=342 Asia/Pacific); [2]Future Enterprise Resiliency & Spending Survey Wave 11, IDC, December 2023 (n=300 Asia/Pacific)

Industry Spotlights

CIO Perspectives from Manufacturing, Financial Services, Retail, Healthcare, and Education

Digital Infrastructure Priorities and Perspectives in Manufacturing

Tech Priorities for the Next Year[1]

#1 — Implementing AI/GenAI — **33%**

#2 — Modernizing core business applications — **33%**

#3 — Investing in tech to meet sustainability goals — **29%**

Challenges[1]

C-Suite respondents in the manufacturing sector identified **"security and risk management" (43%)** and **"lack of digital skills" (34%)** as key hurdles to completing digital initiatives.

Vendor Capabilities[2]

"Deep knowledge of (manufacturing) industry"(54%) and "strong data governance and security" (49%) were the characteristics respondents in the manufacturing sector valued highly in their technology providers.

Digital Infrastructure[3]

63% of respondents in the manufacturing sector reported acquiring new **servers** in the last 6 months, and **59%** procured new **storage**.

Workloads[3]

AI workloads and apps (73%), data management and regulatory requirements (61%), and OT/automation in factories and remote sites (54%) are the top workloads expected to drive on-premises infrastructure investments over the next year in manufacturing

Source: [1]WW C-Suite Tech Survey May 2024 (n=117 for Manufacturing); Source: [2]Enterprise Workloads Infrastructure Study, 2H23 (n=317 for Manufacturing); [3]Worldwide Digital Infrastructure Sentiment Survey (n= 131 for Manufacturing)

Peer Insights From an IT Decision Maker in the Manufacturing Sector

Key Insights

1. **Sectoral Challenges:**
   Legacy IT processes and systems, talent acquisition and retention, and enhancing speed of delivery.
2. **Infrastructure Procurement Priorities:**
   Hybrid cloud and Opex focus, performance optimization under variable load and value maximization from IT investments.
3. **Tech Investment Focus:**
   Modernization and tech debt retirement, cloud consumption models, and data analytics and AI.

> "Budgetary constraint is a key challenge (for manufacturing), along with digitalization and legacy modernization. We have executed a comprehensive risk assessment across the organization and leveraged it to make risk-mitigated decisions on our technology strategy."

– CIO of an India-based manufacturing organization

Digital Infrastructure Priorities and Perspectives in Financial Services

Tech Priorities for the Next Year[1]

#1 — Improving customer-focused digital experiences — **45%**

#2 — Investing in strengthening security — **42%**

#3 — Implementing AI/GenAI — **41%**

Challenges[1]

C-Suite respondents in the financial sector identified **"security and risk management" (59%)** and **"inability to scale due to organizational silos" (33%)** as key hurdles to completing digital initiatives.

Vendor Capabilities[2]

"Strong data governance and security" (53%), **"helping deliver business outcomes" (43%)**, and **"helping us create a more innovative culture" (39%)** were the characteristics respondents in the financial sector valued highly in their technology providers.

Digital Infrastructure[3]

66% of respondents in the financial sector reported acquiring new **servers** in the last 6 months, and **56%** procured new **storage.**

Workloads[3]

AI workloads and apps (68%), data management and regulatory requirements (62%), and **automation of customer processes (55%)** are the top workloads expected to drive on-premises infrastructure investments over the next year in financial services.

Source: [1]WW C-Suite Tech Survey May 2024 (n=105 for BFSI); [2] Enterprise Workloads Infrastructure Study, 2H23 (n=261 for BFSI); [3] Worldwide Digital Infrastructure Sentiment Survey (n= 140 for BFSI)

Peer Insights From an IT Decision Maker in Financial Services

Key Insights

1. **Sectoral Challenges:**
   Legacy modernization and DC exit, standardization of IT systems and processes, and effective tracking and allocation of infrastructure costs.
2. **Infrastructure Procurement Priorities:**
   Business and operational innovation, Opex focused procurement, and operational agility.
3. **Tech Investment Focus:**
   Modernization and tech debt retirement, cloud consumption models, and data analytics and AI.

> "We want to know our customer better and have a better journey with us. It isn't about better technology, but about organizing and understanding the customer data better."

– CTO of a financial services major with pan-Asia presence

Digital Infrastructure Priorities and Perspectives in Retail

Tech Priorities for the Next Year[1]

#1 — 64%

Accelerating shift to cloud consumption models

#2 — 59%

Building out agile development capabilities to deliver new digital services quicker

#3 — 55%

Improving technology staff competencies via inhouse training, upskilling, and talent recruitment initiatives

Challenges[1]

C-Suite respondents in the retail sector identified **"lack of digital skills"** (**34%**) and **"inability to scale due to organizational silos"** (**31%**) as key hurdles to completing digital initiatives.

Digital Infrastructure[2]

55% of respondents in the retail sector reported acquiring new **servers** in the last 6 months, and **62%** procured new **storage**.

Vendor Capabilities[1]

"Deep knowledge of (retail) industry" (**43%**) and **"helping deliver measurable business outcomes"** (**41%**) were the characteristics respondents in the retail sector valued highly in their technology providers.

Workloads[2]

CRM applications (**22%**) and **Dev tools/apps** (**21%**) are the top workloads expected to drive on-premises infrastructure investments over the next year in retail.

Source: [1]WW C-Suite Tech Survey May 2024 (n=44 for Retail); [2]Enterprise Workloads Infrastructure Study, 2H23 (n=181 for Retail)

Peer Insights From an IT Decision Maker in the Retail Sector

Key Insights

1. **Sectoral Challenges:**
   Business growth, business agility, and supply chain stability and resilience.
2. **Infrastructure Procurement Priorities:**
   Technology-service integration, long-term ROI, and breadth of services capabilities.
3. **Tech Investment Focus:**
   Real-time data analytics and visualization, scenario-based data applications, and cybersecurity & regulatory compliance.

> "The application of cloud-native technologies has significantly enhanced business stability, supporting long-term sustainable growth, while enhanced real-time supply chain monitoring and optimization have reduced the impact of disruptions and improved overall operational efficiency."

– CIO of a leading retail entity with pan-Asia operations

Digital Infrastructure Priorities and Perspectives in Healthcare

Tech Priorities for the Next Year[1]

#1 — Investing in security technologies — **37%**

#2 — Improving customer-focused digital experiences — **36%**

#3 — Improving digital skills across the organization — **33%**

Challenges[1]

C-Suite respondents in the healthcare sector identified **"security and risk management" (46%)** and **"lack of buy-in from front-line employees" (34%)** as key hurdles to completing digital initiatives.

Vendor Capabilities[2]

"Deep knowledge of (healthcare) industry" (46%), **"strong data governance and security" (43%)**, and **"responsiveness to client requests" (40%)** were the characteristics respondents in the healthcare sector valued highly in their technology providers.

Digital Infrastructure[3]

53% of respondents in the healthcare sector reported acquiring new **servers** in the last 6 months, and **50%** procured new **storage**.

Workloads[3]

AI workloads and apps (68%), data management and regulatory requirements (56%), and **odernization/ automation (46%)** are the top workloads expected to drive on-premises infrastructure investments over the next year in healthcare.

Source: [1]WW C-Suite Tech Survey May 2024 (n=99 for Healthcare); [2] Enterprise Workloads Infrastructure Study, 2H23 (n=116 for Healthcare); [3] Worldwide Digital Infrastructure Sentiment Survey (n= 129 for Healthcare)

Peer Insights From an IT Decision Maker in the Healthcare Sector

Key Insights

1. **Sectoral Challenges:**
   Legacy infrastructure, limited cloud adoption, and data privacy requirements when using AI.
2. **Infrastructure Procurement Priorities:**
   Opex based infra consumption, moving from on-premises infrastructure to colocation facilities, and cloud.
3. **Tech Investment Focus:**
   Automation and AI, facilities management, diagnostic assistance for clinicians.

> "Hospital and healthcare companies want to move from onsite DC to colocation for access to skills, connectivity and Opex models. Managing this within the patient data privacy framework, and ensuing medical practitioner productivity is the main challenge."

– CIO of a large Korean hospital

Digital Infrastructure Priorities and Perspectives in Education

Tech Priorities for the Next Year[1]

#1 — Accelerating shift to cloud consumption models — **33%**

#2 — Modernizing core business applications — **32%**

#3 — Improving IT support — **31%**

Challenges[1]

C-Suite respondents in the education sector identified **"security and risk management concerns" (52%)** and **"insufficient business case" (41%)** as the top hurdles to completing digital initiatives.

Vendor Capabilities[1]

"Having a strong data governance and security practice" (66%) and **"helping deliver measurable business outcomes" (46%)** were the characteristics respondents in the education sector most valued in their technology providers.

Digital Infrastructure[2]

61% of respondents in the education sector reported acquiring new **servers** in the last 6 months, and **64%** procured new **storage**.

Workloads[2]

Structured databases/data management (19%) is the top workload expected to drive on-premises infrastructure investments over the next year in education.

Source: [1]WW C-Suite Tech Survey May 2024 (n=43 for Education); [2] Enterprise Workloads Infrastructure Study, 2H23 (n=134 for Education)

Peer Insights from an IT Decision Maker in the Education Sector

Key Insights

1. **Sectoral Challenges:** Cybersecurity, cost management, and digitalization of learning experience.
2. **Infrastructure Procurement Priorities:** Architectural fitment, pricing and total cost of operations, and service and support.
3. **Tech Investment Focus:** Cybersecurity, AI-led automation and analytics, and pay-by-use consumption models.

> "Our goal is to provide the best student experience to attract and retain the best and brightest. Another key priority is to make them market ready by creating programs that are tuned to market needs by incorporating feedback from the industry in the shortest time. We aim to provide a personalized experience across learning, consumption of the learning, and applying that learning in a job."

– CIO of a leading Australian university

Essential Guidance

Why Fit-for-Purpose Infrastructure Matters: Modern Infrastructure Needs to Embrace AI, but Cannot Ignore Existing and Non-AI Workloads.

Hybrid Cloud Works Best

- To reduce latency and time to value, the infrastructure stack must move closer to where data lives.
- Span shared and dedicated cloud, hosted and self-operated facilities, including remote edge locations and endpoints, and be bound by a common control plane and provide end-to-end observability.

Compute Platform Considerations

- Integration, ease of operations, and increased levels of automation need to be the main selection criteria for new infrastructure.
- Manageability across edge, core and to cloud demands a cloud-native approach.
- New observability demands for performance, and security, need consideration.

Protect the Data Assets

- Data is the fuel for business and for AI, creating new challenges and opportunities for storage strategies.
- Have a clear data management strategy that encompasses, creation, ingestion, classification, and protection.

Define Your AI Needs

- Optimize computing and storage to extract timely, relevant, and actionable insights from a wide variety data sources.
- Fit-for-purpose choices required - many AI POCs fail because of inadequate infrastructure strategies and planning.

AI Workload Diversity

- AI workloads could be standalone or embedded and could be generative, predictive, and interpretative in nature.
- They could involve one or more model related activities such as training, fine-tuning, and inferencing.

While AI is the pressing demand for today, over time it will become a utility, and the shift in focus will be on applications that deliver business value. Ensuring the support infrastructure for these applications to future-proof is the essence of a modern infrastructure strategy.

Message from the Sponsors

In the fast-evolving landscape of AI and multicloud computing, modern technology infrastructure becomes more important than ever. As the IT estate becomes more distributed across clouds and sprawling edge locations, having a centralized operational hub is critical to handle intensive demands from emerging and traditional workloads. Dell Technologies and NVIDIA can help you design technology infrastructure that is smart, flexible, and resilient to meet the needs of today and tomorrow.

Dell Technologies and NVIDIA have a long-standing partnership with more than 25 years of joint innovation, focused on accelerating innovation and delivering cutting-edge platforms, solutions, and software that enable transformative results for our joint customers.

The Dell AI Factory with NVIDIA delivers a comprehensive portfolio of AI technologies, validated and turnkey solutions with expert services, speeding AI adoption to accelerate your AI-powered use cases, integrate data and workflows, and enable you to design your own AI journey for repeatable, scalable outcomes.

Dell Technologies and NVIDIA work together to accelerate your journey from possible to proven with a complete portfolio of innovative technologies, a comprehensive suite of professional services, and an extensive network of partners.

To learn more

Visit our website

IDC Custom Solutions

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

idc.com @idc @idc