IDC

IDC WORKBOOK

Four Considerations for AI-Ready Infrastructure Buildout

What Should You Consider Before Making AI-Ready Infrastructure Decisions?

GenAI, as a major new workload requiring adoption of AI-ready infrastructure, is entering a critical phase: extended buildout. Starting in 2024, enterprises will accelerate deployment of new AI-ready hardware and software infrastructure as they invest to drive meaningful gains in business and staff productivity as well as reimagine customer digital experiences.

We've pieced together the most crucial considerations organizations must explore before making AI-Infrastructure decisions. These considerations are not mutually all exclusive. In fact, they encompass choices of location and technologies that must be guided by the needs of the end workload. All considerations must be applied flexibly and adapted according to workload needs: public and private, on premises and off premises, and core and edge are all choices, and we expect most enterprises to adopt a mix of hybrid solutions.

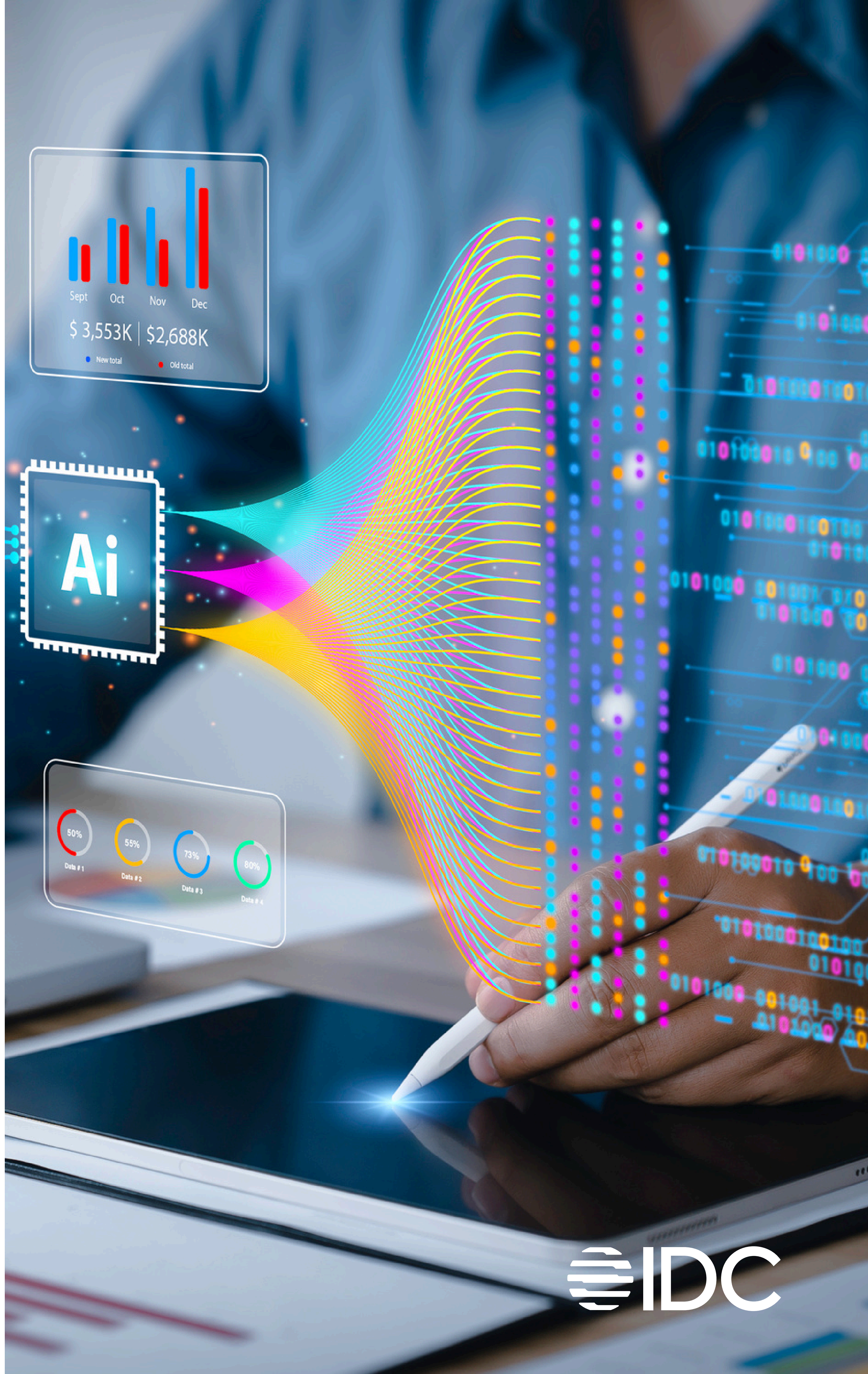
Key Considerations before Planning AI-Infrastructure Buildout:

Public Vs. Private AI →

Location →

System Architecture →

Technology →



Public Vs. Private AI Infrastructure

An appreciation of the unique benefits/challenges associated with selecting public AI-ready infrastructure offerings (public AI) versus private AI-ready infrastructure (private AI) depends upon adoption of a suitable framework for assessing AI-ready infrastructure choices. For virtually all enterprises, both types of infrastructure will be required to achieve the long-term benefits of broad GenAI adoption. Enterprises should be guided by the required balance among the foundational AI-ready infrastructure building blocks: compute (processing), memory, storage, security, and networking technologies required to supporting evolving GenAI and broader AI.

Insights into the outlook for each of these core building blocks are relevant for enterprise IT teams as we are seeing significant advances across technology roadmaps, platform innovation, and deployment/consumption models in every area. These insights will make it easier to implement cost-efficient, secure, resilient, and high-performance infrastructure stacks for handling end-to-end AI/ML workflows and be suited to the needs of a diverse range of industry verticals, use cases, and IT price points.

In the coming years, IDC recommends that enterprises pay special attention to major changes in how emerging technologies will be deployed in infrastructure, how technology is designed/bundled and delivered based on infrastructure location, and how the fundamental architectures of the systems deploying are altered based on use of new technologies.

Public AI

Public AI is the use of cloud provider–delivered compute, storage, and network resources plus AI frameworks and relevant cloud PaaS and network services capabilities that enterprise IT practitioners use to action enterprise-specific AI/ML workflows. Examples of public AI framework elements include AI platforms such as AWS Bedrock, Azure AI, Google Vertex AI, and IBM watsonx.

CRUCIAL CONSIDERATIONS

Determining when to opt for public AI-ready infrastructure depends on numerous factors:

- Consider the scale and scope of AI projects
- Evaluate the upfront budget and long-term cost-effectiveness of public AI infrastructure
- Assess the level of data sensitivity and regulatory compliance requirements
- Consider the expertise and resources available within your organization

Assessing the provider's AI infrastructure and platform offerings in terms of scalability, interpretability, and integration with existing workloads is crucial. These will determine the required level of interoperability with existing tools:

- Access to expertise
- Community and support
- Global reach
- Complementary infrastructure services

Private AI

Private AI is the use of enterprise datacenter infrastructure and AI framework capabilities by enterprise IT practitioners for actioning enterprise-specific AI/ML workflows. The datacenter assets could be hosted at interconnect providers or in enterprise-owned facilities. When using private AI, enterprises will also need to look for AI platforms that support hybrid deployment options that allow them to govern model development and use. Several solutions that support this approach are now available or in preview.

CRUCIAL CONSIDERATIONS

Elements to consider before selecting Private AI:

- Private infrastructure is perceived to be safer
- Control over data residency, compliance requirements, and the ability to customize security protocols
- Research and development of proprietary algorithms or handling highly classified information that demand absolute control and isolation
- Higher initial costs
- The need for highly skilled maintenance and management personnel
- Potentially, a much slower ability to scale compared with public AI solutions

Key next steps when selecting private AI: scalability, interpretability, and integration:

- **Competitive advantage**: Having control over an AI Infrastructure can provide a competitive advantage by enabling faster innovation and better integration with existing systems, allowing the ability to leverage proprietary data more effectively
- **Customization and flexibility**: Having a private environment offers the flexibility to customize AI models and algorithms to suit the organization's unique needs
- **Data sovereignty**: Organizations in highly regulated industries or with sensitive data, opt for private AI infrastructure to maintain sovereignty and ensure compliance with data protection regulations

Location, Location, Location

Systems in core enterprise datacenters, while the dominant domain for non-GenAI-centric AI processing to date, will not be the only significant location for all forms of AI processing in the future. The emergence of public cloud–based AI platforms is becoming an increasingly attractive option for organizations that are not prepared to make the capital investments required for sustained operation and enhancement of private AI–ready infrastructure.

IDC predicts that by 2025, 70% of enterprises will form strategic ties with cloud providers for GenAI platforms, developer tools, and infrastructure, requiring new corporate controls for data and cost governance.

Dedicated cloud environments will be the preferred choice for organizations with a high sensitivity to privacy concerns or situations where more control over the environment is desired. The ability to deploy AI-ready infrastructure on premises can also positively impact costs by limiting data movement and reducing or eliminating data transfer fees associated with the public cloud.

To address the power and cooling demands of dense AI-ready infrastructure, colocation is expected to house a larger share of dedicated infrastructure over time.

In addition to public cloud and dedicated core infrastructure, edge computing has a role to play in overall AI architectures. IDC defines edge computing as a distributed computing paradigm where infrastructure and workloads are placed closer to where data is generated and consumed. As more data is created outside of datacenter environments, it is logical that edge-optimized AI-ready infrastructure will follow. This is already taking place in federated learning environments, where AI models are trained on local data and then shared externally to create composite models. Edge AI-ready infrastructure will also become necessary as applications scale to increase the performance of inference tasks and reduce network traffic to centralized systems.

System Architectures Adapting for GenAI

Under both private AI and public AI frameworks, future system architectures will adapt for technology pooling, intrasystem efficiency, and the local needs of the workload. The move from fixed building blocks to flexible, on-demand pools of resources is often referred to as composable infrastructure.

Composable infrastructure is poised to revolutionize architectures for GenAI and broader AI, unlocking its full potential and boosting its effectiveness in several ways. First, it dismantles the rigid, siloed architectures of traditional datacenters, replacing them with a dynamic pool of shared resources like processing power, storage, and specialized hardware like GPUs. This allows AI tasks to access the exact resources they need, precisely when they need them. When training a massive language model with composable infrastructure, it is possible to dynamically scale up its processing power on the fly, optimizing resource utilization and accelerating training times.

Second, composability fosters an "AI building block" approach. By breaking down AI systems into modular, reusable components, it empowers developers to mix and match pretrained models, algorithms, and data pipelines effortlessly. This opens doors to rapid experimentation, faster innovation, and easier adaptation to changing needs. Instead of starting from scratch each time, developers can leverage existing elements, leading to quicker deployment of effective AI solutions across diverse use cases. Composable infrastructure transforms AI development from a monolithic endeavor into an agile, adaptable symphony of reusable components, ultimately making it more efficient, cost effective, and impactful.

Technology-Specific Characteristics to Consider

Compute

Core systems are designed for high-performance tasks like training complex models, requiring powerful CPUs, GPUs, and FPGAs. In contrast, edge systems will prioritize efficiency over raw power due to resource constraints, leading to the use of low-power CPUs or specialized AI silicon for inferencing operations.

Security

AI often deals with highly sensitive data, including personally identifiable information (PII), healthcare records, and financial data. Protecting this data is crucial for ethical and legal reasons. AI models themselves require protection. These algorithms can hold valuable intellectual property and be vulnerable to theft or manipulation. Robust security measures like encryption and access controls are essential to safeguard models and prevent unauthorized access or tampering.

Storage

There’s a driving interest in using a combination of object and file storage for AI operations. This aligns with data pipelines that must access data stored in both formats. Vendors are responding to this need by creating systems that merge storage types to mitigate the process of moving data between systems. There are also new cloud services that create file interfaces into high-performance object storage tiers.

Networking

Networking is an integral part of all cost- and performance-efficient private AI and public AI solutions. Networking is critical for data movement that minimizes lag time to the GPUs, within and between infrastructure systems. Further, the need for intelligence of processing, security, and performance (high throughput, low jitter, and low latency) within the networks makes networking critical to the effective implementation of AI-ready public and private cloud infrastructures.

Fabrics

Key value propositions of GenAI network fabrics include minimizing model training times and efficient handling of unpredictable network transients in latency, jitter, and congestion hotspots throughout the leaf and spine network — thereby, minimizing packet loss and accelerating AI/ML model training completion times.

Multicloud Networking

Multicloud networking is enabling business-critical technology capabilities that enterprise buyers require and are willing to pay for, accelerating their AI and digital transformation journeys. This is for composing, delivering, and monetizing a dynamic mix of legacy and modern IT services and applications, securely, and with consistent performance, across public and private AI-ready cloud infrastructures.

Generally, we observe that Technology Leaders across industry verticals and use cases are building out for AI models in hybrid clouds relying on mixed deployments of compute-, storage-, and network-intensive technologies often in dedicated back-end network clusters, with meaningful cost and performance efficiencies. **The key differentiator will be which organizations invested in the governance and AI orchestrations solutions that enable use of hybrid AI-ready infrastructure by design, not just by circumstance.**

Interested in learning more?

Program Subscribers can learn more about the important factors that Technology Leaders must consider before making decisions around building out infrastructure to be AI-Ready with IDC's research document, "Key Considerations for Making AI-Ready Infrastructure Decisions".

If you would like to learn more about how IDC can help you on your Journey to AI Readiness, contact us today.

Contact Us

May 2024