BEST PRACTICE REPORT

Architecting Your Infrastructure For AI

An Architect's Point Of View

December 11, 2023

By Alvin Nguyen with Lauren Nelson, Mike Gualtieri, Tracy Woo, Michele Goetz, Naveen Chhabra, Abhijit Sunil, Brent Ellis, Kate Vella, Kara Hartig

FORRESTER®

Summary

Rapid advances inspired the world to reflect on 2023 as "the year of AI," specifically generative AI (genAI), sparking renewed interest in AI strategies and experimentation. Although data teams build these models, infrastructure professionals must serve their needs. This is no easy task because requirements vary by AI use case, with some pushing the boundaries of what existing systems can do and what infrastructure professionals have experience in managing. GenAI has created enormous infrastructure demands for data storage and processing amid chip supply shortages. This report helps infrastructure professionals make key decisions on AI infrastructure (e.g., compute, network, storage).

AI Powers The Enterprise; You Must Power AI

Rapid improvements in genAI have driven tremendous interest in AI overall. Many AI use cases push the boundaries of infrastructure in providing context and insights for data sets while addressing your specific business, adapting over time, and scaling with your organizational needs. The use cases are nothing short of inspiring: Early detection of Alzheimer's disease, real-time analytics of event data to provide hyperpersonalized experiences for attendees, automated remediation of IT incidents to reduce mean time to repair, and automatic code generation from prompts reflect the power of AI when properly implemented. The enormity of the benefits is not only driving demand for these AI capabilities but also challenging teams — especially infrastructure professionals — in:

- **Tackling potentially massive AI infrastructure sizes.** Depending on your AI use case, you may require scale that you've never needed or managed. Total investment in the largest solutions exceeds $100 million, with a count of infrastructure assets in the thousands. Why? Certain AI models can require terabytes of data and tens of thousands of GPUs to achieve the quality and speed necessary for business. Additionally, managing this AI infrastructure will push the limits of your IT professionals. Although using a cloud-based solution can alleviate some of these challenges, the need to manage this infrastructure at scale will remain. You don't require massive investment if your use case doesn't mandate it, but for those that do, the work can be very difficult.
- **Addressing supply shortages that make AI infrastructure unavailable.** The excitement around AI drives demand for GPUs and cloud AI services, both of which have supply challenges. IT infrastructure professionals may not be able to acquire these resources directly and may need to seek services through third parties like cloud platforms or AI startups, which may present other challenges. However, few have reported issues to date outside of large chip purchasers.
- **Constantly rebalancing use of AI infrastructure to ensure ROI.** Any technology investment must achieve ROI, and AI is no different. However, infrastructure costs can quickly escalate, leading enterprises to question the long-term viability of their plans. Infrastructure professionals must carefully weigh the performance gains against the incremental growth in demand and costs. Where costs are not feasible, waiting until more efficient solutions emerge may be a necessary step. For example, if you're using a cloud platform, costs are favorable for uneven, unknown, or unpredictable workloads. However, if you're always running hot, it can be extremely expensive.

- **Fine-tuning requirements of each AI solution.** Delivering infrastructure for AI is not one-size-fits-all. Multiple AI solutions are available, covering different use cases (e.g., text-to-image, text-to-speech, large language models [LLMs], open source models, fine-tuned models) for different locations (e.g., on-premises, public cloud). With so many choices, this is challenging, especially for complex workflows such as using ChatGPT to develop a prompt for Midjourney images and AI voice generators.
- **Delivering fast enough for first-mover's advantage with genAI.** In 2023, everyone is talking about and experimenting with genAI, hoping to gain first-mover's advantage. For infrastructure professionals, that means creating a system that is flexible enough to experiment with but simple enough to quickly deploy, scale, dismantle, and redeploy. This requires quickly setting up sandboxes (e.g., using SageMaker Studio in Amazon Web Services [AWS] to launch a prebuilt model) for AI experimentation and maturing them into production instances (e.g., sizing the genAI solutions for production and fine-tuning with select data).
- **Preparing for a multimodal future.** Multimodal solutions that accept images, audio, text, and numerical data are on the horizon. Preparing for AI models that work with multiple types of data requires infrastructure that works with all of them simultaneously and well. Capex investments in infrastructure cannot be too optimized for specific types of data because they vary in characteristics (e.g., average data size, access patterns), and all types will be available for these new models. This new complexity is also likely to result in more scale (e.g., maybe requiring many more training resources than today).

Architecting Infrastructure For AI Solutions

Designing the right infrastructure architecture for your AI use case requires diligence in key architectural decisions and refinement of each component in your infrastructure. Many AI technologies are still emerging. Technologists supporting these endeavors should follow the guidelines that the PIES model outlines. Early on, test what the technology can do (or not do) for you and your use cases. Because this often occurs before funding, many technologists run experiments using cloud-based options or available infrastructure resources.

The refinement, optimization, and scaling of the solution will require more sophisticated architectural decisions. How do you get started? Go step-by-step through workload and environment, compute, network, and storage.

Workload And Environment: Set Your Requirements And Expectations

Building infrastructure to support AI generally addresses two key considerations: the AI workloads and the stage of the environment. These levers affect how you build for your model's needs today. As you consider more use cases, more data, multimodal solutions, and the scale to sustain your model, these infrastructure solutions must evolve. As such, many organizations lean on the flexibility of cloud to help construct a system that can morph with their models.

- **Environments.** The rapid progress of AI solutions has created a race to effectively leverage the latest models and services. To this end, developers in an organization need to have environments with specific infrastructure capabilities to address the full AI application lifecycle. Major environments include: 1) sandbox, to experiment and evaluate AI solutions; 2) development, to construct an AI product or service in an organization; 3) test, to understand the performance of the AI product or service and identify potential issues; and 4) production, to provide the AI product or service. As noted above, if you're creating a sandbox environment, use whatever is available and easy. For later stages, you'll have increasing requirements.
- **AI workloads.** The are three primary workloads to consider: 1) data preparation, 2) training, and 3) inferencing. Each workload has distinct characteristics that are important to understand to determine their specific IT infrastructure needs (see Figure 1). Collecting, exploring, cleansing, validating, and labeling data sets with significant interactivity characterizes data preparation (i.e., transforming raw data into a form for use by AI models). Feeding large amounts of data to the AI models characterizes training (i.e., supplying AI models with prepared data to teach them how to properly process information to generate accurate decisions and insights). Inferencing (i.e., leveraging AI models to interpret live data to make predictions) uses trained neural networks to infer results. Know what you're supporting to best understand its technical requirements.

Figure 1

The Three Major AI Workloads

Workloads	Definition	Examples	Infrastructure considerations*
Data preparation	Data preparation involves taking raw data and transforming it into a format that the AI model(s) can use. Steps in this workload include collection, discovery, exploration, cleansing, validation, and labeling.	• Collection of image data sets for object identification (e.g., vehicles, people, signs) • Standardization and normalization of data formats (e.g., UTC format for time) • Creation of labels or categories for specific pieces of data	• **Compute.** More computational power is needed for larger data sets and specialized computational capabilities (e.g., audio, image, video processing). Specialized hardware like DPUs can help offload work and accelerate the data path. • **Network.** For large data sets, high-bandwidth network capabilities are needed. • **Storage.** Storage capacity ties directly to data size. Storage performance should match throughput of the preparation workload. Read performance must meet or exceed the input capacity. Read and write locations should be independent storage locations.
Training	Training involves feeding AI models with (prepared) data to teach them how to correctly interpret the data and generate accurate decisions.	• Use of a data set of images and videos with labels to train an AI model how to identify vehicles, people, and signs • Use of normalized numeric data to train AI how to process and categorize tables and spreadsheet information • Use of the Common Crawl data set representing TBs of raw web page data, metadata extracts, and text extracts	• **Compute.** Computational needs for training will be high. The larger the data set is, the more computational power you need. If your IT infrastructure can keep GPUs fed with data, they are a good choice if you have access to them. Multicore CPUs are a solid choice as well but at lower performance levels. • **Network.** Bandwidth is most important for training. • **Storage.** If the training data can fit within the AI IT infrastructure memory footprint, then storage simply needs to load the training data into memory in a reasonable amount of time. High performance requires continuously feeding GPUs data (sans interruption/latency). Some architectures use high-speed memory caches or NVMe drives directly attached to GPUs to improve throughput of data to GPUs.

Source: Forrester Research, Inc. Unauthorized reproduction, citation, or distribution prohibited.

Workloads	Definition	Examples	Infrastructure considerations*
Inferencing	Inferencing is the process of using live data with an AI model to make predictions.	• Autonomous vehicle navigation • Facial recognition • Content generation based on prompts	• **Compute.** Computational demands are high, but the use of specialized processors (e.g., inference accelerators and custom programmable hardware like FPGAs and ASICs) offer options beyond GPUs and CPUs, with the need for additional skills and experience in your organization to support them. • **Network.** Low latency is most important for inferencing workloads. • **Storage.** Capacity is high (largest scale to low TBs) but not extreme (less than full-frame HD video). Focus on fast file and object storage for RAG as well as fast block or object storage for vector databases working alongside the model.

*Networks: You're not going to change your network by AI workload. However, expect the demand on your networks to vary by stage. To support AI work, teams should consider developing a fabric of high-bandwidth, lossless, low-latency, scalable, multitenant networks. Data size and amounts can vary dramatically at various times, pushing the limits of networking fabrics. The use of streaming telemetry can offer real-time monitoring, analytics, and visualization of network data, offering insights into network performance, traffic patterns, and resource utilization.

Source: Forrester Research, Inc. Unauthorized reproduction, citation, or distribution prohibited.

Compute: Get What You Can; Cloud Can Make It Easier

Compute is the engine for AI performance, whether it is data analysis, coordination of data flow, or creation of new data content. However, chip supply constraints may mean working with what is available and maximizing the use of those resources. This can limit the types of AI models and environments that your organization can deploy: Without enough compute resources, larger LLMs may not be viable unless restricted to sandbox environments where lower performance levels are more acceptable.

- **Processor type.** At a basic level, you have a choice of processor: GPU (high power across compute/network/storage but at a shortage), tensor processing unit (TPU) (high power but limited to Google Cloud Platform), general CPU (less performant for the power/space), programmable logic devices like app-specific integrated circuits (ASICs) and field programmable gate arrays (FPGAs) (highly optimized but at a high cost with specialized skills requirements), and new classes of AI compute components like inference-focused processors (specialized but limited to the function). Base your choice of processor on the scale of your AI ambitions, what you can acquire, and your internal expertise — along with the AI model and environment type (see Figure 2). But again, you may need to adjust your plans based on cost or availability of the processor type. Leveraging cloud increases

your level of flexibility and optimization.

- **Compute location.** Your data, data storage, and AI infrastructure dictate your compute location. To maximize performance and minimize cost, colocate your AI compute infrastructure and data because latency and egress can be painful. The compute location will also depend on your data: Store IP and regulated data where you can secure and monitor them and large data sets where you can grow and access them cheaply. With current genAI models, size and supply constraints make cloud the preferred location for many organizations.
- **Sustainability.** Without proper planning, your AI solutions may face power/space limitations or conflict with your environmental sustainability plans. The use of high-power processors like GPUs demand more power, space, and cooling to maximize their lifespan. This impacts data centers' location and design (e.g., floor space, power density, cooling, resource availability, energy requirements). Although fulfilling these requirements can yield benefits (e.g., improved scalability, performance, ROI), most organizations can't handle this additional commitment and look to the cloud.

Figure 2

Processor Types By Performance, Availability, Cost, And Scalability

Compute	Performance: memory bandwidth	Availability	Cost	Scalability	Examples
GPUs	High: TB/s	Low	High	High	NVIDIA
TPUs	High: TB/s	Medium	High	High	Google
CPUs: general	Medium: GB/s	Medium	Medium	Medium	Advanced Micro Devices, Intel
CPUs: customized	Medium: GB/s	Medium	High	Medium	IBM
Programmable logic: ASICs, FPGAs	High: GB/s to TB/s	Low	High	High	Infineon, STMicroelectonics, Xilinx (Advanced Micro Devices)

Source: Forrester Research, Inc. Unauthorized reproduction, citation, or distribution prohibited.

Network: Connect The Pieces Of Your AI Infrastructure

Complex AI solutions can require high bandwidth and low latency, which can be a challenge to accommodate. Your network architecture depends almost entirely on your deployment model. If you're using an AI service on public cloud, your team will have little control over networking decisions. If you're custom-building a hybrid model connected to your on-premises data center, you'll likely lean on your existing WAN connection. In some cases, you may need to further enhance that connection. If it is on-premises, you will likely lean heavily on the general networking connections in place or work to deploy more sophisticated solutions if it is a high priority and you've got the networking skills to handle it. You will need to consider (see Figure 3):

- **Network type.** At the most basic level, you are deciding whether or not to enhance your networking. With a managed public cloud service, the choice is already made and outside of scope. With a hybrid model, you can either lean on your existing WAN connection or enhance it with additional capacity. With on-

premises, you must decide whether or not to enhance your connection. Juniper Networks and Cisco Systems offer sophisticated low-latency options, which organizations require. For solutions that need to incorporate data from remote locations, satellite and cellular may be your only options.

- **Network path length and throughput.** As you start to architect the specifics of your AI networking solution for hybrid or on-premises, path length and throughput can pull focus. You can lean on techniques to enhance each, but they can escalate cost or complexity in ways your organization can't handle. Some include working with larger bus sizes, remote memory access, memory buses in GPUs, multicore and multiprocessor systems, as well as colocated processors, memory, and storage. Another way to reduce network path length is decreasing the frequency of traveling that length. To enable high throughout, solutions like RDMA over Converged Ethernet and AWS's Elastic Fabric Adapter can help, especially in high-performance computing use cases. RDMA allows memory access without interrupting the processor or OS.
- **Security.** Don't add in security after you've architected and optimized your networking plans for AI; your security systems will drastically diminish performance. Security solutions inspect everything and will slow down your network connections. Using a Zero Trust architecture (i.e., one that embeds security throughout) will impact your latency and bandwidth less dramatically. In a cloud environment, you will similarly experience less diminished performance due to security tooling. However, for everyone else, consider this major factor from the start.

Figure 3

Network Options By Performance, Availability, Cost, And Scalability

Network	Performance: network bandwidth	Availability	Cost	Scalability	Examples
Memory buses	High: 1,000+ GB/s	High	High	Low	Advanced Micro Devices, Intel, NVIDIA
Specialized networks	High: 400+ GB/s	Medium	High	High	InfiniBand, RDMA over Converged Ethernet
General networks	Medium: 10+ GB/s	High	Medium	High	Ethernet

Source: Forrester Research, Inc. Unauthorized reproduction, citation, or distribution prohibited.

Storage: You Can Always Use More

For AI solutions, data is critical. The relevance and accuracy of AI output highly correlate with the size and quality of the training data. As a result, data sizes for training sets and output of AI solutions have experienced tremendous growth, which increases the importance of storage to house data for training, output, and historical context. Today, file and object are the most common for AI use cases, with both providing simple, scalable, and resilient storage platforms. Block-level storage platforms can provide higher levels of performance but need a storage management layer to ensure that performance and reliability scale with demand. Low-access archival storage also has value for record and audit purposes. Quality will be the responsibility of your AI partners, but they might add storage requirements to help address quality issues (see Figure 4).

- **Quantity of data.** Larger AI solutions, such as GPT4, utilize tremendous amounts of data with a related increase in size of metadata/categorization to improve the quality of data in use. Although the fastest storage technologies (e.g., remote flash memory, storage caches, NVMe storage, solid-state drives) will best maximize bandwidth, increase input/output operations per second, and reduce latency, cost is a concern at these data sizes. Hard disk drives can be a valid solution but require architectural and implementation decisions to minimize performance

issues.

- **Proximity of storage to compute.** The closer storage is located to the computational elements, the better the performance is. This makes in-memory storage solutions the best if the data can fit in memory. For certain use cases where immediate responses are essential, such as interactive AI models, prioritize the proximity of storage to compute. If the use case doesn't demand immediate response (e.g., training, noninteractive context generation), higher-performance solid state, solid state, and hard disk are viable options. Caching or tiering data across different storage types can work in specific applications, like training, but requires expertise from your personnel and performance tuning to the AI solution.
- **Historical context.** As AI models become more accepted, the need to reproduce past results becomes important for many reasons, whether it is operational, compliance-related, or legal in nature. This requires storing the state of an AI solution, which means that all data (e.g., training, fine-tuning) and the model (e.g., weights) that define the AI responses need to be saved. With current AI model sizes, this requires large storage capacities, which backup and archival solutions best address because storage capacity is more important than performance.

Figure 4

Storage Options By Performance, Availability, Cost, And Scalability

Storage	Performance: I/O bandwidth	Availability	Cost	Scalability	Examples
Memory optimized	High: 10-100 GB/s	High	High	Low	Micron, Samsung, NVIDIA
NVMe	High: 10+ GB/s	High	High	Medium	Intel, Samsung, Western Digital
SSD	Medium: 10+ GB/s	High	High	Medium	Intel, Samsung, Western Digital
HDD	Low: 1+ GB/s	High	Medium	Medium	Western Digital
Tape	Low: 1+ GB/s	High	Low	Low	Hewlett Packard Enterprise, IBM, Quantum

Source: Forrester Research, Inc. Unauthorized reproduction, citation, or distribution prohibited.

Make Your Key Infrastructure Decisions By Workload

Infrastructure professionals, specifically architects, need to make several decisions to determine the AI infrastructure path for their organization. The promise of and interest in AI has created a supply-versus-demand issue that will take time to resolve. In the short term, supply constraints on IT infrastructure components tied to AI promote the use of cloud AI services. Unless your organization or your partners have access to those key AI infrastructure components right now or have insight into how your model will evolve, making such a significant capex purchase early in your AI journey is unadvisable. Even if your organization can acquire AI infrastructure assets, the scale that AI is reaching requires technology professionals to become more effective and efficient in how they manage it. Companies that repatriate their AI solutions due to cost have mature forecasting skills and confidence in their long-term AI infrastructure requirements. To get started, Forrester recommends that infrastructure professionals:

- **Select the right types of partners to fill skill and technology gaps.** Not every organization has established AI competencies. If you're jumping in on the genAI buzz, but your organizational AI IQ is low, engage a partner to bolster these shortcomings quickly. Partners come in all shapes and sizes. They can provide access to AI services or infrastructure that you may otherwise not get. They might enhance data science skills or data architectural skills. Figure out which categories of partners will help catapult your organization forward. Many of these partners will be familiar; as existing partners (e.g., Amazon, Microsoft, Google, global system integrators, consultancies), all are building out value streams and competencies to help you on that journey. Many even span outside their normal scope of work and provide operational efficiencies if they already support your workloads, data, or online presence.
- **Establish the AI models for creation and scope/scale of the product.** As you start to craft an infrastructure plan to support your needs, start with the model and intended scope. What AI use cases does your organization pursue? Which AI solutions does your organization consider? Answering these questions will inform how comprehensive the AI tech stack needs to be, how much AI infrastructure needs to be acquired, and what AI products and services can integrate into existing products and services or new ones. IT infrastructure professionals who get involved in these discussions early will be better prepared to address long-term needs and react to sudden changes in AI plans. This can also start realistic conversations about budget limitations relative to AI aspirations.

- **Identify the timeframe for the work and where it sits among overall priorities.** The timeframe for delivering your AI products and services will determine availability of AI infrastructure. With supply constraints, cloud AI service availability concerns, and limited AI solution implementations, focus on experimentation, evaluation, or smaller-scale AI models. Use of good, but not the best, AI infrastructure may limit performance and scalability, consume more power, require more space, or cost more. Determinations will also heavily depend on the project's prioritization relative to other investments (e.g., business requirements, customer experience), which will inform the level of commitment to, speed of delivery for, and investment in the AI products and services. This affects timelines, trade-offs, and even levels of attention that IT infrastructure professionals need to give.
- **Define what success looks like.** Knowing what the desired business outcomes are helps organizations determine what AI products and services need to be developed and operationalized. Whether the organization is after innovation or differentiation or just keeping pace with the market helps IT infrastructure professionals determine the AI infrastructure components and level of attention needed. IT infrastructure professionals can use this information to clarify partnerships, AI products and solutions for focus, and impact of AI trends and market changes on AI choices for their organization. Also important is understanding what your counterparts working on the model need from the infrastructure to meet business outcomes and what key performance indicators they've created to measure their success. This may spark conversations around budget/performance trade-offs.
- **Determine if there are location-specific requirements.** The location options depend on what AI products and services (e.g., on-premises, cloud, edge, hybrid) you develop and what is available. Supply chain constraints may limit on-premises AI infrastructure components like GPUs. Similarly, the data you're working with may dictate location and reveal architectural challenges due to these limitations. And do not forget the space and utility needs to support the demands of AI: Your existing data centers may have space, power, and/or cooling constraints that will eliminate the option of deploying your AI workloads in them.
- **Architect a plan that integrates each component effectively.** The supply chain constraints mean that each component is a precious resource. Long-term success with AI is predicated on extracting the maximum value from these resources in a sustainable manner. Replacing infrastructure assets too often increases costs and waste. Leveraging infrastructure assets for too long creates technical debt and negatively impacts performance and customer experience. This is a complicated balancing act that will tax the skills and experience of IT technology professionals.

- **Set your day-two operations plan.** Working with new and leading-edge technologies will be demanding on IT operations. Questions about how to use the AI solutions and issues with the AI platforms as the organization gets up to speed are natural. As organizations race to develop their AI products and services, set up yours for success by enabling IT operations to accelerate the adoption of AI platforms through education, providing accelerators to bring employees and customers up to speed, and adding help desk resources to ensure that they can rapidly address the overflow of questions and issues.

FORRESTER®

We help business and technology leaders use customer obsession to accelerate growth.

FORRESTER.COM

Obsessed With Customer Obsession

At Forrester, customer obsession is at the core of everything we do. We're on your side and by your side to help you become more customer obsessed.

Research

Accelerate your impact on the market with a proven path to growth.

- Customer and market dynamics
- Curated tools and frameworks
- Objective advice
- Hands-on guidance

Learn more.

Consulting

Implement modern strategies that align and empower teams.

- In-depth strategic projects
- Webinars, speeches, and workshops
- Custom content

Learn more.

Events

Develop fresh perspectives, draw inspiration from leaders, and network with peers.

- Thought leadership, frameworks, and models
- One-on-ones with peers and analysts
- In-person and virtual experiences

Learn more.

FOLLOW FORRESTER

Contact Us

Contact Forrester at www.forrester.com/contactus. For information on hard-copy or electronic reprints, please contact your Account Team or reprints@forrester.com. We offer quantity discounts and special pricing for academic and nonprofit institutions.

Forrester Research, Inc., 60 Acorn Park Drive, Cambridge, MA 02140 USA
Tel: +1 617-613-6000 | Fax: +1 617-613-5000 | forrester.com