INTELLIGENCE UPDATE

Cloud a viable choice amidst uncertain AI returns


Dr. Owen Rogers

5 Feb 2025

Generative AI has created a surge in demand for large-scale clusters of GPUs. AI models need training, and this is vastly accelerated by using GPUs. These clusters parallel process many mathematical functions within the neural network software architectures that enable generative AI to classify and predict (see *How generative AI learns and creates using GPUs*).

However, clusters of this size are expensive and require customized data center infrastructure and highly skilled staff for installation and management. They are not easily procured or installed today, with supply chain issues further affecting their accessibility.

One of Uptime Intelligence's predictions for 2025 is that enterprises will rely on hyperscalers and cloud providers to do most AI model training, which enterprises will then fine-tune and customize (see *Five data center predictions for 2025*).

This report explores that prediction, explaining how the law of diminishing returns affects the value of fully training an AI model using dedicated infrastructure. Enterprises will need to compromise between cost and capability. The incremental cost of upgrading from shared foundation models and cloud infrastructure to bespoke models on dedicated infrastructure might not convert to realized value.

Cost and benefit drive buying decisions

The economic law of diminishing returns describes how an additional improvement fails to deliver enough benefit to justify its extra cost. While this law may be widely understood, it is rarely considered as part of the day-to-day decision-making process.

For example, Uptime Intelligence often hears from organizations that they "cannot compromise on security" or "resiliency is non-negotiable." However, in practice, all enterprises have to compromise at some point — some additions simply do not deliver enough of a benefit to justify the expense.

To accommodate AI, such compromises, too, will need to be made — most notably on data

sovereignty and model accuracy. Many IT leaders say they are not considering cloud infrastructure or third-party models for their AI needs, citing security issues or worries about model accuracy. These concerns are valid, considering how a security breach or bad decision can affect an organization's fortunes — from fines and lawsuits to reputational damage.

Ultimately, however, some operational risks have to be taken, even if they are minimized. Does developing a custom model on dedicated infrastructure resolve these concerns enough to justify the additional expense compared with the cloud?

AI clusters are expensive and complex

A dedicated cluster (i.e., a collection of servers owned and operated by an enterprise and hosted in their choice of data center) represents the most controllable and customizable basis for training AI models. Because a cluster is owned and managed by the enterprise, there is a perception of improved data security and regulatory compliance.

Enterprises can restrict data using internal controls and limit data movement to chosen geographical locations. The cluster can be customized and secured to meet the specific requirements of the enterprise without the constraints of using software or hardware configured and operated by a third party. Given these characteristics, for convenience, Uptime Intelligence has labeled the method as "best" in terms of customization and control.

Unfortunately, the cost of such clusters is prohibitive to many organizations. Such an investment might be worthwhile if the return is assured. However, the financial return on AI investments is yet to be determined.

Alternatives to dedicated infrastructure

Instead of investing in dedicated infrastructure and skills, public cloud providers and foundation model vendors offer AI capabilities without substantial capital costs. However, the use of these alternatives does require a compromise on control and customization. As such, they can provide models that are "good" or "better" rather than the "best" option achieved by dedicated infrastructure.

Public cloud

Both hyperscalers (e.g., Amazon, Google, Microsoft, Meta, Alibaba and Apple) and a new generation of cloud providers (e.g., CoreWeave and Lambda Labs) offer GPU infrastructure and platform services through the public cloud.

By using the public cloud, customers can access capability on demand, paying only for what they use. Customers consume only the needed server capacity rather than purchasing and installing a large cluster. They can use a foundation model (discussed below) as the basis for their training to simplify and accelerate development (see *Understanding AI deployment*

methods and locations).

Infrastructure as a service (IaaS) enables a customer to provision virtual machines containing GPUs via a portal or API to develop a model using cloud infrastructure. They then execute their training model on these virtual machines and download the completed model when it is finished. Cloud providers also offer platform as a service (PaaS) options that enable enterprises to make AI requests — such as translations — directly to an API, without managing any aspect of the model or underlying infrastructure.

The benefit of the public cloud is that customers do not need the capital or skills for implementation and can purchase capacity when required.

Foundation models

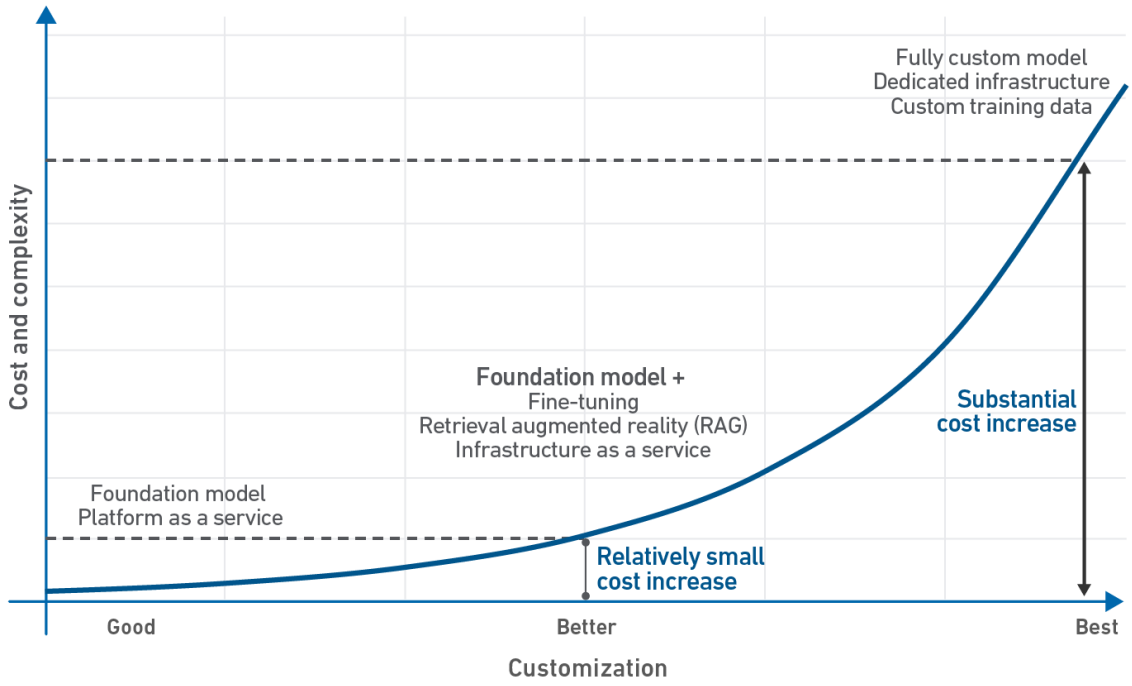
Hyperscalers, startups and the open-source community offer foundation models — pre-trained software from a third party. Many organizations already use these models to avoid training costs and complexity.

A foundation model can be fine-tuned to meet a particular use case. As most of the model's training has already been done in advance, fine-tuning will usually not require a dedicated cluster because it would not be sufficiently utilized over its life to be cost-effective compared with IaaS. Other features, such as Retrieval Augmented Generation (RAG), may improve the model without significant (or any) training. RAG enables a large language model to retrieve external data to resolve a query without retraining the model.

Trained once and for all

Figure 1 shows how the conceptual law of diminishing returns applies to the cost of customizing AI training deployment models as the level of customization varies. As mentioned, the "best" option is training a model from scratch using dedicated infrastructure.

Figure 1 Law of diminishing returns regarding AI training



UPTIME INSTITUTE 2025

A less expensive “good” alternative is to use a platform as a service or a pre-trained foundation model.

Only the most industrious enterprises will use the “best” dedicated infrastructure enough to make it more cost-effective than the “good” cloud alternative (see *Sweat dedicated GPU clusters to beat the cloud on cost*). As such, most enterprises should expect to pay a premium for dedicated infrastructure.

Cost-effective cloud implementations use shared, general-purpose models that are not designed for specific enterprise use cases. However, these foundation capabilities can be inexpensively customized using tools to improve the model for specific enterprise requirements — making them the “better” option.

A foundation model provides a baseline level of general-purpose capability at a relatively low cost. Capabilities, such as RAG and fine-tuning, can substantially improve the relevancy of these models at a relatively low cost. These costs are low because only a small number of additional resources are needed to tweak a general-purpose model so that it can be more purposeful.

The cost jump from “good” to “better” is small; the incremental cost is likely justified if some value is realized. However, the leap from “better” to “best” is substantial. What remains unclear is whether the additional cost is worthwhile.

Is paying more for dedicated worth it?

A significant barrier to the adoption of cloud AI is data sovereignty and the need to protect and secure confidential data in a compliant manner. Many argue that data sovereignty or regulations

prevent the use of the cloud for certain workloads. A major benefit of dedicated infrastructure is the reassurance that data is held in an enterprise's choice of data center, which it fully controls, owns and operates. Having control over where and how data is managed gives a sense of security and compliance.

The challenge for enterprises is determining whether the added reassurance of dedicated infrastructure provides a real return on its substantial premium over the “better” option. Many large organizations — from financial services to healthcare — already use the public cloud to hold sensitive data. To secure data, an organization may encrypt data at rest and in transit, configure appropriate access controls, such as security groups, and set up alerts and monitoring. Many cloud providers have data centers approved for government use. It is unreasonable to view the cloud as inherently insecure or non-compliant, considering its broad use across many industries.

Although dedicated infrastructure gives reassurance that data is being stored and processed at a particular location, it is not necessarily more secure or compliant than the cloud. Arguably, an application running on a properly secured cloud platform may be more compliant with regulations than one hosted on dedicated equipment in a private data center that has not been configured correctly.

Compromises can be made to reduce data sensitivity, such as redacting or anonymizing customer-identifiable information before training. However, such compromises may reduce the value and accuracy of the model. That said, they might be worthwhile compared with the considerable investment required for a dedicated cluster (or the missed opportunities from deciding not to pursue the potential at all). There is always a risk of a security breach, regardless of where the data is located. A premium to use dedicated infrastructure does not necessarily translate into a more secure enterprise.

Another concern raised by enterprises involves hallucinations, where an AI model generates incorrect, misleading or fictitious information. However, it is unclear whether upgrading from “better” to “best” will fix such problems. An upgrade might alleviate the issue because an enterprise controls the training data entirely; however, an improvement is not guaranteed. AI models are complex and unexpected responses can still occur.

Ultimately, it is difficult to prevent hallucinations and incorrect information entirely. Enterprises will need to acknowledge that even the most rigorously trained AI model will make mistakes, and its output must be treated with a level of caution. The risk of an AI mistake having an impact on the business can never be reduced to zero.

The challenge for enterprises is quantifying how the investment in dedicated infrastructure improves outcomes. Cloud-based training using foundation models provides easy experimentation and customization at a low cost. However, this comes with the downside of entrusting potentially sensitive or valuable data to third parties.

On the other hand, dedicated infrastructure implies better data control and may be more cost-

effective than the cloud in some cases (see *Sweat dedicated GPU clusters to beat the cloud on cost*). However, it requires significant investment and long-term commitment to AI, without guaranteeing returns in terms of model accuracy, performance or data security.

Today, a company can make a substantial investment in AI infrastructure and custom development only to find that the delivered model is only slightly better than a third-party model fine-tuned on the cloud. Unfortunately, this may only become apparent after the investment has been made. The outcome is unpredictable.

The Uptime Intelligence View

Dedicated infrastructure and custom models are significantly more complex to deploy than public cloud and pre-trained foundation models — and those considering it may not have the specialist AI skills for execution. If not fully utilized, dedicated infrastructure can be an expensive option. Dedicated options may appear more secure, accurate and controllable than shared options, but this is not necessarily true. Considering that the difference in cost between cloud and dedicated AI is likely to be substantial for most, no organization can afford to ignore the cloud as a viable option for AI workloads. An investment in dedicated infrastructure and custom development should be highly scrutinized before committing.

ABOUT THE AUTHOR

Dr. Owen Rogers

Dr. Owen Rogers is Uptime Institute's Sr. Research Director of Cloud Computing. Dr. Rogers has been analyzing the economics of cloud for over a decade as a product manager, a PhD candidate and an industry analyst. Rogers covers all areas of cloud, including economics, sustainability, hybrid infrastructure, quantum computing and edge.

orogers@uptimeinstitute.com

About Uptime Institute

Uptime Institute is the Global Digital Infrastructure Authority. Its Tier Standard is the IT industry's most trusted and adopted global standard for the proper design, construction, and operation of data centers – the backbone of the digital economy. For over 25 years, the company has served as the standard for data center reliability, sustainability, and efficiency, providing customers assurance that their digital infrastructure can perform at a level that is consistent with their business needs across a wide array of operating conditions.

With its data center Tier Standard & Certifications, Management & Operations reviews, broad range of related risk and performance assessments, and accredited educational curriculum completed by over 10,000 data center professionals, Uptime Institute has helped thousands of companies, in over 100 countries to optimize critical IT assets while managing costs, resources, and efficiency.