Summary Insights: Infrastructure Considerations for Generative AI

RFG Perspective: Unlike the "cloud first" directives, which were viewed as saving operational costs, a "generative AI first" directive will add to one's data center operation's Capex and OpEx. Generative AI (genAI) models are computationally intensive, necessitating robust processors, large datasets, and sophisticated data handling. Those AI "engines" may be located on-prem, or at a colocation or cloud provider site, but they are there, and will not be inexpensive.

The recent versions of ChatGPT and other genAI models have advanced exponentially over prior attempts. They now offer the potential to become a force multiplier for productivity – playing a variety of roles by supporting work being done by different personas within the company. Among other capabilities, genAI will soon enable enterprises to easily and quickly respond to voice, text, chat, and/or on-screen interactions when business clients need assistance with a transaction – and today's genAI algorithms can even generate code. In the future, one can expect generative AI to do even more: For example, we can foresee a time when genAI will be able to proactively deliver personalized insights and advice at the critical moments of a real-time business transaction. For large enterprises with sensitive data, these systems will need to be developed and housed in-house -- not in the cloud -- with their own data and data models, for the sake of security, data integrity and data privacy.

But what is the cost required to deliver this outcome, especially if the data needed for the search depends upon data that the enterprise will not allow to be stored outside of the corporation? It has been estimated that the cost to train a model such as GPT-3 (not the new, expanded GPT-4 version) ran upwards of $4 million (USD). Enterprises will have to commit significant computing power, funding, network bandwidth, space, skill, time, and other scarce resources if they hope to leverage information, transforming it into actionable knowledge and improved products and services by using genAI.

INTRODUCTION

From a business perspective, Generative AI's implementation requires a strong business case that is grounded in clear strategic objectives. Potential benefits, including improved operational efficiency, enhanced customer experience, and the generation of new revenue streams, must be balanced against the costs and risks of implementing genAI. These operating expenses encompass not only the direct infrastructure or cloud service costs, but also the ongoing expenses of system maintenance, AI model training and updating, and potential security measures.

Unlike traditional software that follows preset algorithms, genAI utilizes machine learning algorithms to generate outputs based on patterns they recognize in input data. However, the

46 Kent Hills Lane, Wilton, CT. 06897; (203) 429 8951;
http://www.rfgonline.com/; *Contact: inquiry@rfgonline.com*

implementation of these advanced AI models requires specific infrastructure capabilities and significant investment. One component crucial to success is the need to include AI accelerators, such as ASICs, FPGAs, and neuromorphic processors. These technologies, designed to expedite AI model training and inference processes, can significantly streamline genAI implementation and reduce the infrastructure costs.

The Panel

The panelists on the call were:

- Val Bercovici – CEO, Click
- Richard Kuzma - Senior Product Manager, Cerebras Systems
- David Follett - Founder/CEO, Lewis Rhodes Labs

Investment Considerations

The initial financial investment in AI is typically significant, with the development cycle likely exceeding one year. That is why it is essential to plan early and thoroughly, acknowledging that AI deployment is not a quick-fix solution but rather a strategic, long-term investment. IT executives must also understand that the complexity of the technology means that it requires significant computing power (CPUs and GPUs) whether on-premises or off-premises, extensive data sets for training the AI, and expert knowledge to maintain and adjust the system.

For infrastructure executives, the first concern should revolve around the cost of the substantial upfront investment in hardware (GPUs or cloud computing services), software (machine learning platforms), and talent (data scientists and AI specialists). According to NVIDIA, the supercomputer developed for OpenAI GPT-3 was a single system with more than 285,000 CPU cores, 10,000 GPUs and 400 gigabits per second of network connectivity for each GPU server.

Scaling up from the first NVIDIA DGX to the tenth to 2000+ GPUs becomes incredibly complex and managing the infrastructure can be daunting. Experience has shown that the research teams are 40 to 50 or more people just to manage all the GPUs.

Cerebras, an NVIDIA competitor, has cut the largest silicon wafer – it is dinner-plate-sized rather than postage-stamp-sized. It contains 850,000 cores with a design that is optimized for sparse linear algebra. A total of 2.6 trillion transistors is on the chips, along with 40 gigabits of on-chip memory and very large memory bandwidth within the data fabric. All this hardware fits inside a single rack, versus the many racks required for a standard CPU/GPU configuration. This example shows how proximity and dense packing are needed to optimize overall AI performance.

Timelines for Deployment

The deployment timeline for genAI is another crucial aspect for executives to consider. The development and integration of genAI systems are often lengthy, taking anywhere from several months to a few years, depending on the complexity of the application. Training the model, integrating it with existing systems, and ensuring that its performance meets the necessary standards are very time-consuming processes.

One of the biggest challenges in testing out the model prior to production is ensuring that not only the business rules are met, but also that the business values and guardrails have been included and tested. Too often, enterprises have put the genAI models into usage only to find that the results are misleading, incorrect, or – in the worst cases – presenting what are now known as "hallucinations" – where the reference is not based on real data.

Initial Considerations

The implementation and training of genAI models involves a series of steps that demand a focused approach and a thorough understanding of genAI principles, guardrails, and business needs. The approach towards implementing and training in genAI should be strategic, comprehensive, and iterative. The following steps outline a comprehensive approach:

1. **Utilizing User-friendly AI Services:** Today, major cloud service providers (CSPs) like Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform (GCP) offer pre-trained AI models and machine learning services that are easy to use and integrate. Services like Microsoft Bing chat, equipped with the power of GPT-4 from OpenAI, can serve as an excellent starting point for enterprise testing. There are alternative options available to GPT-3 and GPT-4. While not technically "open source," there are several competing models that can be downloaded and used for free commercial purposes. These models can also be self-hosted and fine-tuned with proprietary data, allowing organizations to maintain control and privacy without the need to send their data to a third party. However, these models unmodified may not be trained to address the types of inferences that the enterprise is seeking to perform.

2. **Training and Upskilling:** There is a requirement for a level of understanding and competence in using genAI services and interpreting their output. Online training programs, webinars, and even the documentation and learning resources can be effective in upskilling a team of administrators, developers, and users. These roles, or personas, come to AI with quite different skill-sets and expectations about results.

There is no one master model. That is why it is important to train and constrain the use of each model to a specific problem at hand, or else the inferences could mislead or be incorrect. Bad or incorrect results would likely lead to some form of damage to the business – either to the business' reputation or to its revenue and profitability. Enterprises need to train their models for more than just the business rules – they must instill business values and guardrails to ensure that responses do not violate ethical responsibilities, intellectual property (IP) rights or any international standards regarding data protection and data privacy (e.g., the updated EU-U.S. Trans-Atlantic Data Privacy Framework and the GDPR in the European Union).

3. **Sandbox Environment:** Setting up a sandbox environment for the team to experiment with genAI models is highly recommended. This environment would act as a "test box" for the models, allowing the team to learn, make mistakes, and improve without affecting the operational infrastructure.

4. **Identify Initial Projects:** Enterprises should start with projects that have a relatively low risk but potentially result in better business results. One lesson learned is that a set of smaller optimized models will be more cost-effective and more accurate than a small number of generalized models. Lines of business (LOBs) should select internal projects where the use of genAI can improve efficiency or effectiveness. The learning from these initial projects could then be used to tackle more complex, high-stakes projects. Building out a massive model initially poses a higher risk of failure. The end-user organization should identify a high-value application within the business that is well-defined. For such cases, the results can be validated and the risk to the business would be minimal. Then, once the model is fully tested in these proofs of concept (PoCs), one can expand the model and iterate, resulting in a more widely used software platform for genAI.

5. **Invest in the Necessary Infrastructure:** GenAI models require significant computing resources, especially for training and production systems. Infrastructure executives must make the line of business and development executives aware of the infrastructure costs and they must ensure the development, testing, and production systems are fully funded throughout the standard corporate planning period.

6. **Ethical and Security Considerations:** Given that genAI can create realistic and convincing content, it is crucial to consider the ethical implications of its use. Also, all "hallucinations" should be analyzed and corrected. Developers, testers and users should be responsible for accuracy checks to ensure usability. Similarly, data privacy and security concerns must be addressed. For instance, care must be taken while interacting with cloud or public platforms that sensitive or personal data should not be shared with any external third-party business partners.

7. **Iterate and Improve:** As the organization gains more experience with genAI models, inferences, and prompts, it's important to continuously improve and refine the models, processes, and techniques based on feedback and results.

Timelines for Implementation

The timeline to establish operational genAI capabilities using external services can be implemented fairly quickly. However, enterprises that require specialized genAI models must understand that using corporate or sensitive data may take more than one year to implement.

1. **Fully Managed AI Services and APIs:** The most expedient way to begin using genAI is to leverage fully managed services like ChatGPT, or the GPT-4.0 API. These services provide pre-trained models, enabling seamless integration of AI capabilities into applications without the need for individual model training. The infrastructure for these services is managed by the major CSPs. In such a scenario, a basic application can be operational in days or weeks, because the infrastructure team already has rudimentary familiarity with the API or service used.

2. **Building In-House GenAI Infrastructure:** If the organization decides to build its own genAI capabilities, by assembling the required hardware, then developing the suitable AI model, training the model with appropriate data, and integrating the model into the desired application may take several months to more than one year of preparation and effort. The planning and installation of the on-premises infrastructure for AI applications, particularly the high-power density server racks for GPU clusters alone, could take many months. IT executives will need to work with their organization's Facilities staff to address power requirements, cooling considerations, and other infrastructure modifications to accommodate high-performance computing (HPC) designated for use with genAI.

46 Kent Hills Lane, Wilton, CT. 06897; (203) 429 8951;
http://www.rfgonline.com/; *Contact: inquiry@rfgonline.com*

The GenAI Process

The integration and effective utilization of genAI models is a multi-step process: data collection, data labeling, model training, and inference analysis.

1. **Data Collection:** The first step in any genAI project is the data collection of massive amounts of data that is representative of the problem the model aims to solve. While some data models can be trained using .5 TB of storage, most large language models (LLMs) are built using multiple terabytes or petabytes of data.

2. **Data Labeling:** Once the data is collected, it must be labeled. These models will have billions – and possibly trillions – of labeled parameters. OpenAI's GPT-3 LLM has 175 billion parameters, and the company's latest model – GPT-4 – is purported to have one trillion parameters. This labeled dataset serves as the ground truth that the model learns from during training.

3. **Model Training:** After the data is labeled, it is fed into the AI Large Language Model for training. During this process, the model learns patterns in the data that correlate with the labels. It adjusts its internal parameters to minimize the difference between its predictions and the true labels. This training process requires significant computational resources, especially for complex models like GPT-4. Part of the training process is to identify and eliminate unwanted biases and "hallucinations." Hallucinations – findings that are not real – happen because LLMs, in their most-vanilla form, do not have an internal state representation of the world. There is no concept of fact. The models are predicting the next word based on what they have seen so far in their "data piles"— so, it is a statistical estimate.

4. **Inference:** Once the model is trained, it is ready to make predictions on new, unseen data. This process is called inference. During inference, the model uses the patterns it learned during training to predict the output for new data instances.

Neuromorphic processors and other AI accelerators like ASICs (Application-Specific Integrated Circuits) and FPGAs (Field Programmable Gate Arrays) are designed to enhance the efficiency of these processes. Neuromorphic processors mimic the structure of the human brain, making these types of processors particularly adept at handling tasks associated with AI and machine learning. Their architecture allows for parallel processing and lower power consumption, which can speed up the training process and make real-time inference more feasible.

46 Kent Hills Lane, Wilton, CT. 06897; (203) 429 8951;
http://www.rfgonline.com/; *Contact: inquiry@rfgonline.com*

ASICs are customized for a specific application or task, in this case, AI processing. They are optimized to execute AI-related tasks faster and more efficiently than general-purpose CPUs or GPUs. However, ASICs lack the flexibility of CPUs and GPUs because they are hard-wired to perform specific tasks.

FPGAs are somewhat of a middle ground. They are programmable like CPUs and GPUs, but they can also be highly optimized like ASICs. FPGAs can be reprogrammed to handle different tasks efficiently, which makes them a flexible choice for AI workloads that may change over time.

In essence, these technologies can accelerate the AI model training and inference process, making it feasible to implement more complex models and handle larger datasets. However, each of these technologies comes with its own set of trade-offs, and the choice between these tradeoffs will depend on the specific requirements and constraints of the AI project.

SUMMARY

The integration of GenAI into business operations requires a comprehensive examination of the enterprise's infrastructure and business model. Inclusion of genAI into the business model will be quite costly, which means that a detailed business model needs to be evaluated before a commitment is made to expend the infrastructure investment to support genAI.

From an infrastructure perspective, the foremost requirement is raw computational power. GenAI models like GPT-4 are computationally intensive, requiring powerful processing units (CPUs and GPUs) that will power the model training and inference processes.

RFG POV: IT executives must address the choice between in-house and cloud-based models based on considerations such as confidentiality, cost, privacy, scalability, security, and data sovereignty. In-house models afford greater control and potential security. However, they necessitate considerable investment in time and resources -- both initially and for on-going operations. In contrast, cloud-based models offer scalability and speed, leveraging already-prepared models, and enabling businesses to tap into high-grade GPU clusters without upfront capital expenditure to replicate similar clusters.

While cloud-based models can be brought up more quickly, business, and ethical considerations must be evaluated as part of the business case for using genAI. Business and IT executives must consider that the use of AI brings with it a raft of ethical and legal considerations, including data privacy, security, and the risk of algorithmic bias. Therefore, companies must develop robust governance frameworks to manage these risks and to comply with regulatory requirements.

46 Kent Hills Lane, Wilton, CT. 06897; (203) 429 8951;
http://www.rfgonline.com/; *Contact: inquiry@rfgonline.com*

In summary, the successful integration of genAI into a business requires careful planning, significant investment, and a willingness to embrace new ways of working. By weighing the potential benefits against the costs and the risks, business and IT executives can determine which genAI models must be deployed in-house and which can be delivered from third-party providers.

Additional relevant research and consulting services are available. Interested readers should contact Client Services to arrange further discussion or interview with Cal Braunstein, CEO and Executive Director of Research. Jean S. Bozman, President of Cloud Architects llc, contributed to this report.

46 Kent Hills Lane, Wilton, CT. 06897; (203) 429 8951;
http://www.rfgonline.com/; *Contact: inquiry@rfgonline.com*