Balancing Power and Performance: Opportunities for Success Amidst the Rising Power Demands of Generative AI

KAISER ASSOCIATES

The growing use of Generative AI requires more power and higher density from training and inference infrastructure, raising serious energy concerns. Rising power costs, pressure to reduce carbon footprints, and the emergence of energy bottlenecks will necessitate a shift in how GenAI models are designed and deployed, requiring innovation from GenAI platform vendors, semiconductor manufacturers, and data center developers.

Distributed GenAI Training

GenAI platform vendors will need to distribute training across hyperscale centers, edge locations, and endpoints, aligning with available distributed energy resources while enhancing model customizability and user experiences

- **Embrace decentralized federated learning** by training models across distributed devices or servers to reduce the load on centralized data centers
- **Invest in edge infrastructure** closer to major metros to accelerate model training and distribute power consumption
- **Optimize the allocation of compute, storage, and network resources** between data centers, edge data centers, and user devices

Efficient Model Design

Semiconductor players and GenAI platform vendors should prioritize developing and implementing smaller, more efficient models by utilizing techniques like model compression and quantization

- **Invest in R&D to innovate in model and data compression** to minimize storage and accelerate data transfer
- **Deploy a suite of efficient, task-specific mini-models** to address diverse industry needs, thereby minimizing the computational burden on larger, energy-hungry LLMs
- **Develop specialized AI accelerators and chips** optimized for running smaller models at the edge, enabling efficient federated learning

Data Center Design & Development

Data center developers and GenAI platform vendors should invest in distributed energy generation, implementing energy-efficient infrastructure, and using advanced cooling technologies

- **Prioritize ESG-focused data center development**, integrating social and environmental considerations into site selection and construction
- **Investigate innovative data center designs** that leverage natural resources for cooling, such as earth sheltered or underwater facilities
- **Implement advanced cooling technologies**, such as liquid cooling, immersion cooling, and free air cooling
- **Invest in specialized AI hardware accelerators** that are designed for energy efficiency and optimized for specific model architectures

Solutions for High-Value Use Cases

GenAI platform vendors should focus on developing tailored solutions for high-value use cases, enhancing model energy efficiency and guiding customer usage to areas with substantial ROI

- **Prioritize high-impact GenAI use cases** and identify the minimum model parameters required for optimal performance
- **Develop specialized GenAI models** tailored to specific high-value business use cases to optimize efficiency and resource utilization
- **Implement adaptive model scaling** to dynamically adjust model size and complexity based on real-time resource availability and task demands, maximizing efficiency and ensuring optimal performance

For over 40 years, Kaiser Associates has employed robust methodologies to craft innovative go-to-market strategies. Our TMT practice is experienced in empowering clients to navigate and seize strategic opportunities in the rapidly expanding field of GenAI.

Connect with Kaiser's Technology Practice experts to learn more about how Kaiser can support your business:

Joe Kestel
VICE PRESIDENT
JKESTEL@KAISERASSOCIATES.COM

Azulina Green
VICE PRESIDENT
AGREEN@KAISERASSOCIATES.COM

Rodger Heidgerken
SR. MANAGER
RHEIDGERKEN@KAISERASSOCIATES.COM

Source: Kaiser Research and Analysis