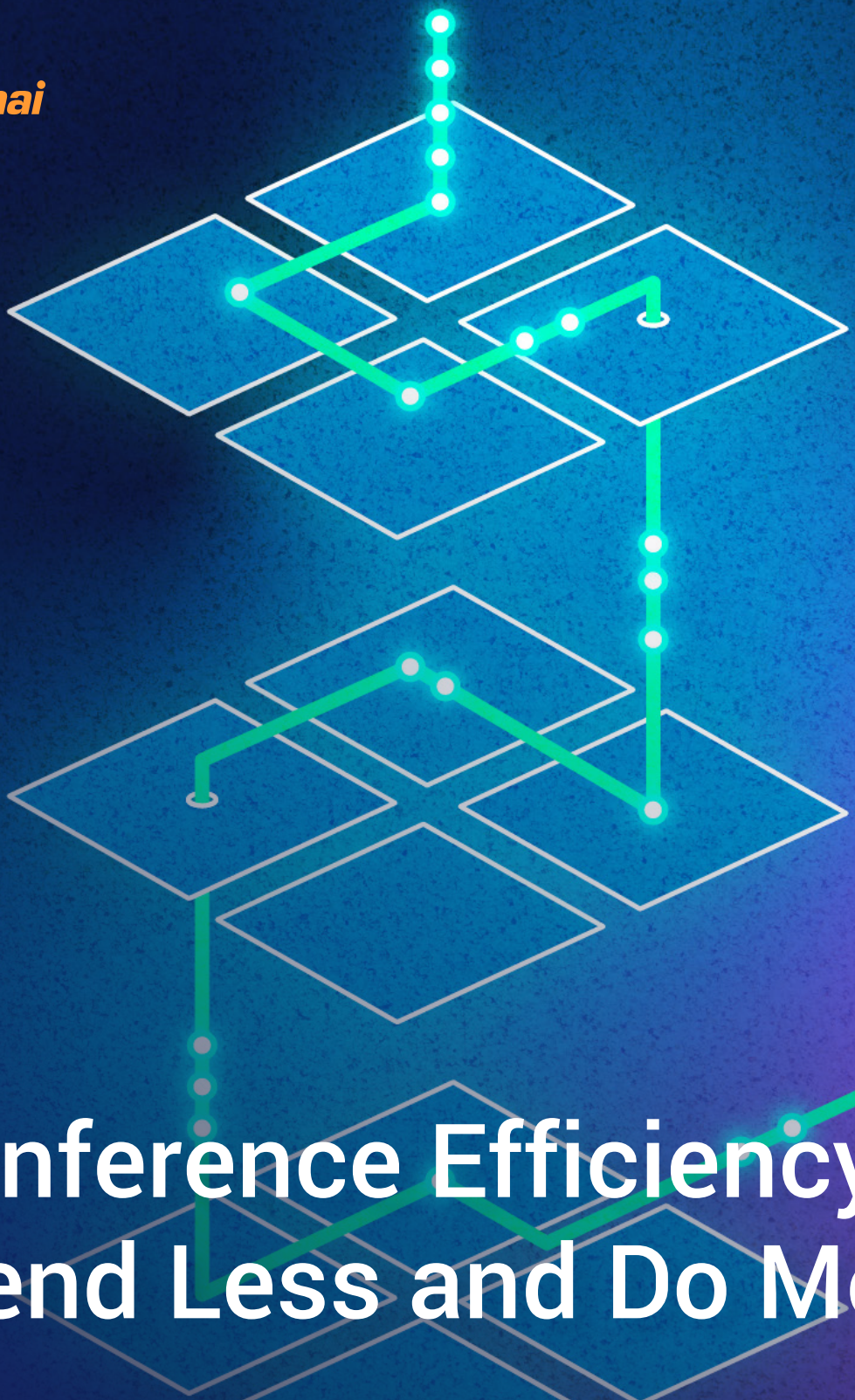
Akamai

WHITE PAPER



AI Inference Efficiency: Spend Less and Do More

Case study: Reducing Stable Diffusion inference costs

Reduce your artificial intelligence (AI) inference costs by up to 86%.This article reveals how strategic infrastructure choices can significantly impact the performance and cost of AI, as demonstrated by a case study with Stable Diffusion. Discover how Akamai outperforms hyperscalers by delivering lower latency, higher throughput, and cost savings that can fuel further innovation.

AI costs a lot

AI is driving significant IT spending, but are organizations investing wisely? Many companies waste resources, particularly during inference, because of operational inefficiencies. This diminishes the return on investment (ROI) of AI projects and could discourage future AI adoption. These wasted resources represent a missed opportunity to fund innovation, experimentation, and new initiatives.

The inference challenge: Cost vs. desired outcome

While much of the initial AI hype focused on training large models, the reality is that more than 80% of computational demand comes from inference tasks. Organizations are under pressure to deliver on the promises that AI makes, but impulsive investments driven by hype can lead to significant cost overruns. Balancing ambition with efficiency, particularly in inference, is essential.

Understanding what "good" looks like is crucial. Organizations need to balance cost with desired outcomes such as low latency, fast inference times, high accuracy, and even sustainability goals. This requires careful planning, optimization, and monitoring throughout the AI model lifecycle.

A case study: Stable Diffusion inference with GPUs

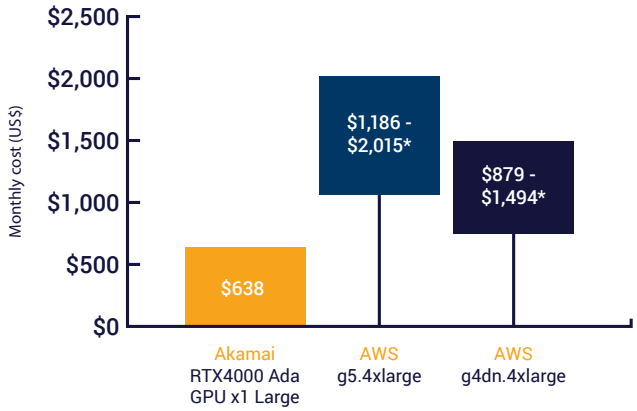
Stable Diffusion, a powerful image generation model, exemplifies the complexities of optimizing inference. Its performance relies on various factors, including CPU, RAM, GPU, VRAM, Disk I/O, and networking. Thorough testing and robust monitoring tools are essential to identify and address bottlenecks.

Benchmark setup

This benchmark compares running Stable Diffusion XL between the recommended Amazon instances and Akamai Cloud with similar VM sizes. Image sizes generated were 512x512 pixels. Three key metrics were measured:

1. **Latency** — This is an end-user experience measurement. It measures the length of time from when the prompt is submitted to when the image is returned.
2. **Throughput** — This is a measure of how many images can be generated in a given period.
3. **Iteration** — This is a drilldown of throughput and measures the time for a single iteration in a given period. A higher number of iterations per image corresponds to a higher level of detail in the generated image, but also requires more processing time.

The following instance types were tested:



* Pricing varies based on the region of deployment

The tests were run in September 2024 and validated in December 2024.

Results

Latency

Running a Stable Diffusion XL model on an Akamai RTX4000 instance results in a decrease in latency by 15.0% vs. the AWS A10g instance, and a decrease in latency by 62.8% vs. the AWS T4 instance (Figure 1).

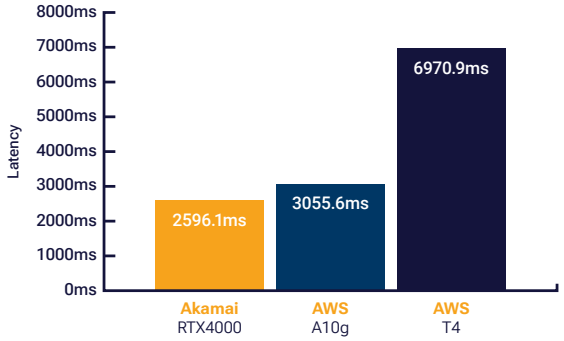


Fig. 1: Latency (lower is better)

Throughput

Running a Stable Diffusion XL model on an Akamai RTX4000 instance results in an increase of throughput by 29.4% vs. the AWS A10g instance, and an increase in throughput by 314.3% vs. the AWS T4 (Figure 2).

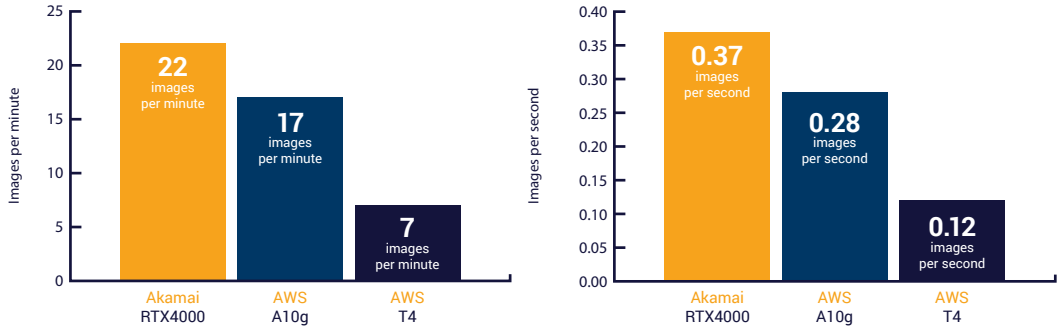


Fig. 2: Throughput (higher is better)

Iteration speed

The iteration speed results are similar to the throughput results, but iterations are an especially important parameter for Stable Diffusion since it affects the quality of the output (Figure 3).

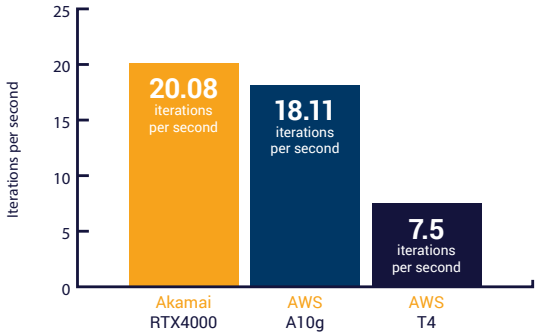
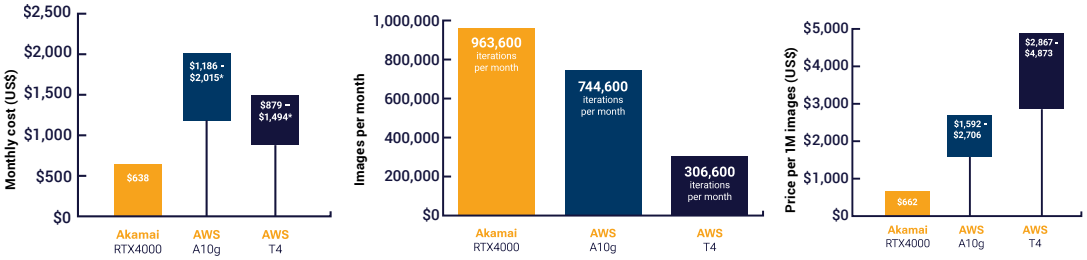


Fig. 3: Iteration speed (higher is better)

Cost

Running a Stable Diffusion XL model on an Akamai RTX4000 instance costs 58.4% to 75.5% less per million images vs. the AWS A10g instance, and 76.9% to 86.4% less per million images vs. the AWS T4 (Figure 4). Note that although the AWS T4 instance has a lower monthly cost, the poor performance for this model offsets any cost benefit.



* Pricing varies based on the region of deployment

Fig. 4: Cost

Outcome

This case study demonstrates how infrastructure choices can significantly impact performance and cost. Akamai RTX4000 delivers lower latency, higher throughput, and lower cost per million images compared with AWS instances. Such optimizations provide opportunities to:

- **Enhance the project:** Pass cost savings to customers, improve image quality by allowing for higher iteration counts, or add new features.
- **Fuel innovation:** Fund new initiatives and experiments within the organization.

Additional optimization for inference

Beyond infrastructure choices, numerous techniques can optimize inference:

- **Model optimization:** Quantization, knowledge distillation, and sparsification can reduce model size and complexity, enabling execution on less expensive hardware. For example, Akamai partner Neural Magic works with organizations to create sparse models while minimizing loss of accuracy.
- **Continuous monitoring and analysis:** Tracking resource utilization helps identify bottlenecks and optimize resource allocation.
- **Automated scaling and load balancing:** Adjusting resources dynamically on the basis of demand ensures efficiency.

Remember to balance optimization efforts with the specific objectives of the AI application.

Conclusion

AI offers immense potential, but realizing that potential requires a strategic and cost-conscious approach. Leaders must guide AI investments strategically to avoid costly inefficiencies driven by hype. By prioritizing inference optimization and adopting a data-driven approach, organizations can maximize their AI ROI, accelerate innovation, and achieve their desired outcomes.