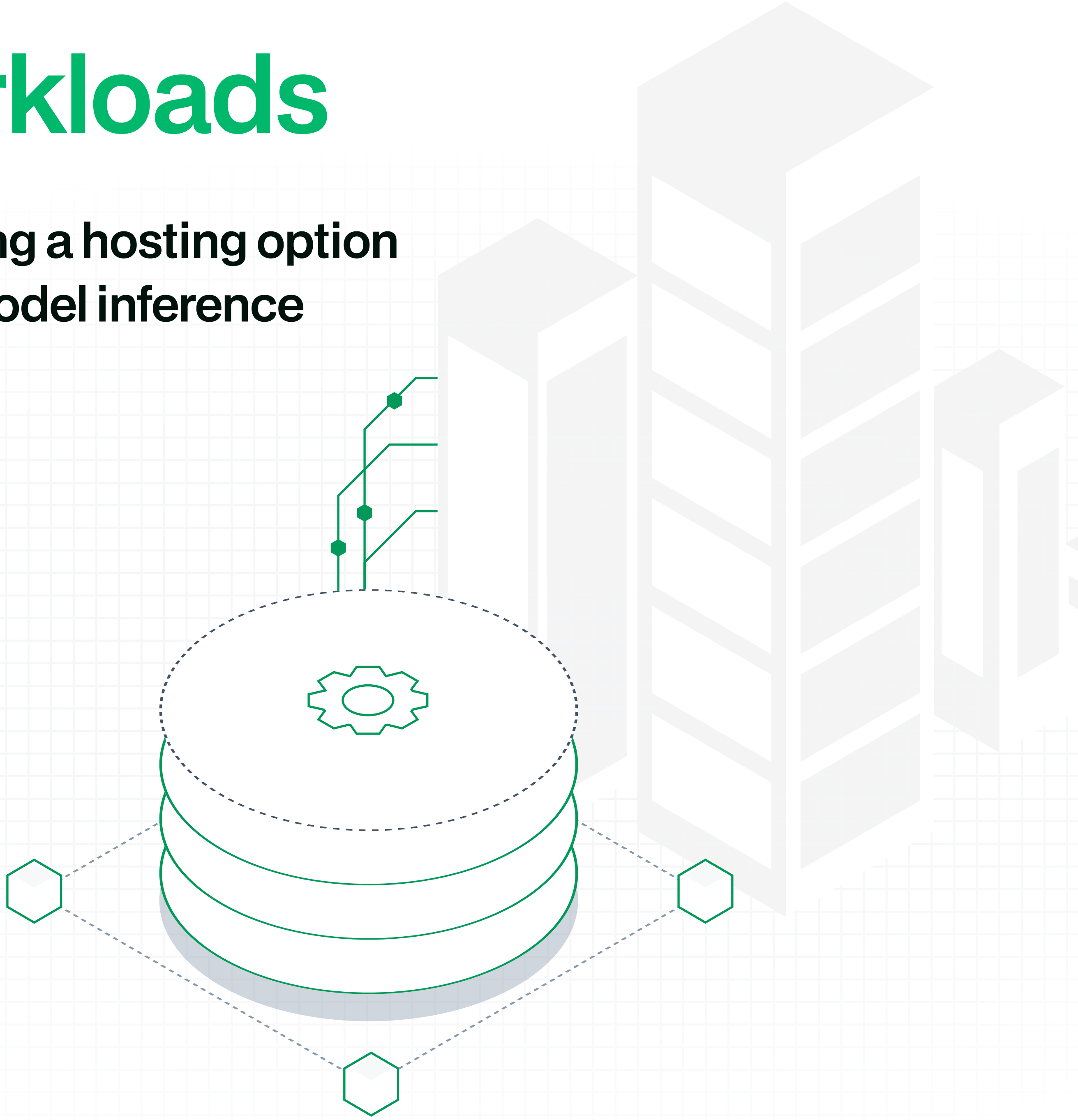
baseten

baseten.co | hi@baseten.co

Where to run your inference workloads

Choosing a hosting option for AI model inference



Introduction

From startups to enterprises, businesses of all sizes are quickly realizing that using custom, fine-tuned, or open-source AI models has become essential for building competitive products.

At the same time, using these models in production presents challenges around reliability, security, and performance. While AI providers like Anthropic and OpenAI can serve as a starting point, they fall short for enterprise-grade solutions. Organizations need greater control over model behavior, data privacy, and scalability—without sacrificing performance.

Choosing the right hosting solution is key to overcoming these challenges. Where you run your inference workloads—on a cloud platform, your own virtual private cloud (VPC), or using a hybrid approach—directly impacts the effectiveness of your AI systems. This choice is particularly important when it comes to AI inference, the process of running models to generate responses or predictions from new data.

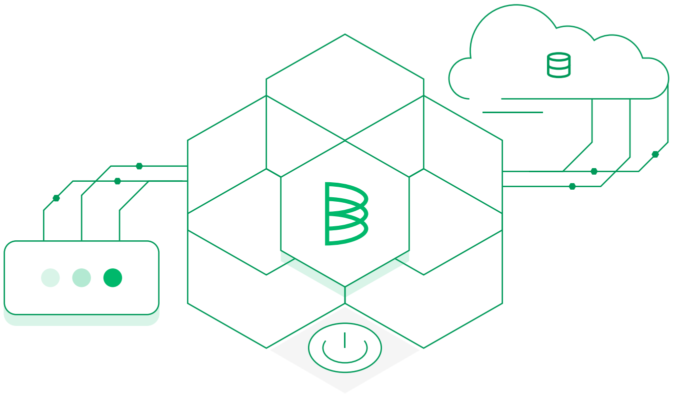
Effective inference is vital for real-time services, seamless user experiences, and data-driven decision-making. As businesses increasingly rely on AI to drive results, the need for efficient, secure, and scalable hosting solutions has never been more urgent. Companies with superior infrastructure will hold a competitive edge.

Baseten is laser-focused on providing the most performant and customizable deployment options tailored to organizational needs. Unlike other infrastructure providers, with Baseten, you can run your inference workloads in your cloud, our cloud, or both. As a result, our customers achieve lower costs, industry-leading latencies, and 100% uptime—delivering powerful user experiences.

For CIOs, CTOs, VPs of AI, and IT leaders, understanding the benefits of different hosting solutions is critical to ensuring performance, compliance, and cost-efficiency for AI-powered products. In this guide, we'll explore the differences between cloud, self-hosted, and hybrid hosting solutions, and how they can play a key role in successful AI initiatives.

baseten

baseten.co | hi@baseten.co



AI Model Inference on Baseten

Baseten is the leading machine learning inference platform for performant, reliable, and secure model inference. Trusted by companies like Bland AI, Descript, and PicnicHealth, our mission is to empower companies with the most customizable model deployment solution coupled with the lowest latency. With blazing-fast cold starts, effortless autoscaling, and heightened observability, we provide our customers with record-breaking latencies, throughput, and time to market.

Part of our mission involves offering customers the right solution for their needs, whether that's the full control of a Self-hosted setup, the convenience of Baseten Cloud, or a Hybrid approach that blends the best of both.

Understanding Baseten's hosting options: Self-hosted, Cloud, and Hybrid

Baseten Self-hosted

Baseten Self-hosted offers enterprises complete control over their AI infrastructure and data, making it ideal for organizations with stringent compliance requirements or those utilizing their existing resources.

Key advantages:

- **Data control and security:** Provides complete control over data residency, handling, and storage, ensuring consistency with security and compliance policies like GDPR, HIPAA, and other industry-specific standards.
- **Customization and integration:** Allows tailored configurations and integrations with existing enterprise systems, facilitating custom workflows.
- **Credit utilization:** Utilize your existing GPU allocation, spend commit, and credits with cloud providers like AWS and GCP.

Baseten Cloud

Baseten Cloud is designed for organizations that prioritize operational simplicity and rapid time to market. It provides a managed, scalable environment for deploying AI models, ideal for enterprises looking to minimize infrastructure costs and management while focusing on development.

Key advantages:

- **Scalability:** Offers elastic scaling to accommodate varying workload demands, ensuring that resources are available when needed.
- **Cost Efficiency:** Operates on a pay-as-you-go model, which helps manage costs effectively and particularly benefits businesses with fluctuating AI workloads.
- **Operational simplicity:** Managed services reduce the burden of maintaining and upgrading infrastructure, allowing internal teams to focus on innovation and development.

Baseten Hybrid

Baseten Hybrid combines Self-hosted and Cloud to provide ultimate flexibility. Utilize internal resources whenever they're available; seamlessly flex on Baseten Cloud whenever necessary. Spend down existing cloud commitments and gain multi-cloud flexibility with full resource management.

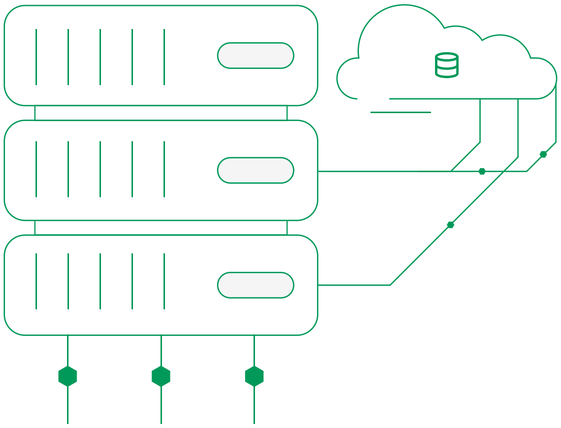
Key advantages:

- **Cloud elasticity:** Deploy workloads wherever there's capacity with zero bandwidth needed to make them compliant with AWS, GCP, or another cloud provider.
- **Cost efficiency:** Utilize internal resources whenever capacity permits and supplement them with Baseten's pay-as-you-go compute, eliminating the need for additional hardware purchases.
- **Data control and security:** Run sensitive workloads on your VPC with complete control over data handling and storage.

Baseten Self-hosted, Cloud, and Hybrid: a side-by-side comparison

The table below compares Baseten's hosting options, highlighting key factors such as cost, customization, data control, performance, scalability, and security.

Feature	Baseten Self-hosted	Baseten Cloud	Baseten Hybrid
Cost	Potentially higher up front, and lower long-term costs	Pay-as-you-go for on-demand compute	Pay-as-you-go for overflow, reducing up front capital expenditure on extra compute
Data control	Full control, with region-locked data and deployments	Partial control, with managed data security and multi-region support	Full control for self-hosted workloads, partial for spillover
Maintenance	Minimal, handled by Baseten	Minimal, handled by Baseten	Minimal, handled by Baseten
Performance optimization	Optimized for low-latency, high-throughput inference	Optimized for low-latency, high-throughput inference	Optimized for low-latency, high-throughput inference
Scalability	High, tailored scalability	Highly scalable, elastic compute	High, tailored scalability with flex capacity on Baseten cloud
Security and compliance	High, with enterprise-level custom security protocols	High, with SOC 2 Type II certification and HIPAA compliance	High, with the flexibility to adhere to custom policies or lean on our secure infrastructure
Utilization of existing cloud commits	Use credits or commits	Spend down existing cloud commits	Use credits or commits



Selecting the right hosting solution for your organization

When deciding between the control of Baseten Self-hosted, convenience of Baseten Cloud, and flexibility of Baseten Hybrid, consider the following factors:

1. **Data sensitivity and compliance:**
   For handling sensitive data and ensuring strict compliance, Baseten Self-hosted offers unparalleled control and security, whereas Baseten Hybrid offers this option for whichever deployments you choose.

2. **Operational simplicity vs. control:**
   Baseten Cloud is best for those prioritizing simplicity and reduced management overhead, while Self-hosted and Hybrid offer deeper customization and control.

3. **Cost considerations:** Evaluate the total cost of ownership, including infrastructure, maintenance, and staffing. Baseten Cloud offers commit-based discounts, whereas Self-hosted and Hybrid may offer long-term savings for those with significant infrastructure.

4. **Unpredictable traffic:** With fluid flex compute, Baseten Hybrid helps you ensure you maintain SLAs during traffic spikes. In addition to elastic compute, Baseten Cloud offers solutions like asynchronous inference for increased robustness in the face of unpredictable traffic.

5. **Performance requirements:** All three deployment options are ideal for applications demanding low latency and high throughput.

Baseten's architecture: a detailed overview

Baseten's architecture is designed to deliver robust performance, scalability, and security for AI applications. The platform comprises three key components: the management layer, control plane, and workload plane.

For Self-hosted solutions (and Hybrid solutions where certain workloads are run in a self-hosted manner), the workload plane is configured in your virtual private cloud (VPC).

1. Management layer

The management layer is the interface that allows users to interact with the Baseten platform. It provides tools for managing teams, setting access privileges, and overseeing metrics and reporting. This layer also includes a comprehensive model library, enabling easy access to various open-source AI models.

Key features:

- Role-based access control (RBAC) and single sign-on (SSO) for secure user management.
- Real-time analytics and custom reporting capabilities.
- Integration with external tools and services for extended functionality.

2. Control plane

The control plane acts as Baseten's operational hub, orchestrating the policies and configurations that govern the AI environment, as well as packaging models and engines. It manages autoscaling policies, monitors performance metrics, and provides middleware services to ensure consistent and efficient operation across different environments.

Key features:

- Autoscaling to dynamically adjust resources based on demand.
- Performance monitoring and logging tools integrated with popular platforms like Prometheus and Grafana.
- Configuration management to maintain consistency across production and development environments.

3. Workload plane

The models and engines packaged by the Baseten control plane are deployed to the workload plane, which can be set up in your cloud environment and regions as long as they have GPU availability. All model inference inputs and outputs are processed by the workload plane without entering the control plane, providing a level of separation for self-hosted solutions; inference calls will never hit Baseten’s VPC.

Key features:

- Support for containerized deployments using Docker and orchestration via Kubernetes.
- Available fractional H100 GPUs for more efficient compute resource usage.
- Compatibility with various AI frameworks and libraries, facilitating diverse model deployment.

Baseten's architecture visualized

The Baseten architecture can be visualized as a layered stack, where each layer builds upon the capabilities of the previous one.

MANAGEMENT LAYER:

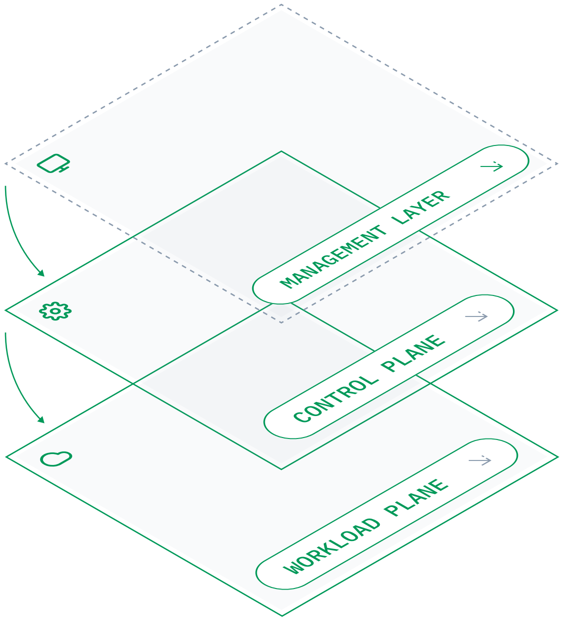
The top layer, providing user interface and management tools.

CONTROL PLANE:

The middle layer, ensuring load balancing and orchestrating policies and configurations.

WORKLOAD PLANE:

The bottom layer, managing compute resources and executing model inference.



Each layer is designed to interact seamlessly, providing a cohesive and integrated platform that supports the end-to-end lifecycle of AI applications. This modular structure allows enterprises to scale their AI operations efficiently while maintaining control over critical aspects such as data security, compliance, and performance.

On-chip performance is optimized with the best inference optimization options (TensorRT/TensorRT-LLM, vLLM, TEI, TGI, etc.) and model serving tooling (including a custom implementation of NVIDIA Triton), as well as instance and network-level improvements to cold starts and end-to-end latency. This ensures AI workloads are low-latency, high-throughput, and cost-effective.

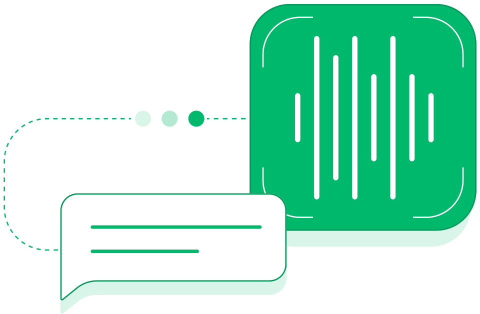
Key use cases for Baseten

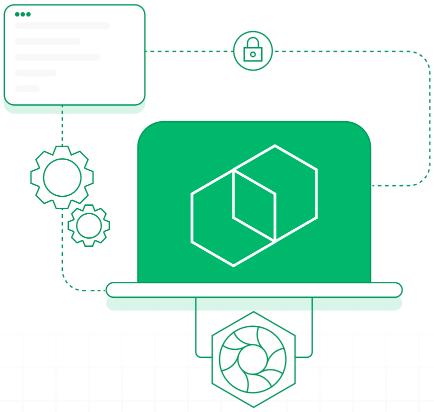
Audio transcription

Baseten optimizes transcription models for low latency and high throughput, which is essential for real-time audio processing. Self-hosted deployment ensures data privacy and compliance, which is crucial for handling sensitive audio data. Go beyond basic Whisper transcription with support for advanced features and fine-tuned variants at a fraction of the unit price OpenAI charges.

Text-to-speech

Generate speech in record time, and gain flawless reliability–even during peak demand. Baseten works with companies generating speech from text to ensure real-time conversations and human-level experiences. And with our HIPAA-compliant platform, you get the data privacy and compliance guarantees that you need to handle sensitive audio data in medical domains.



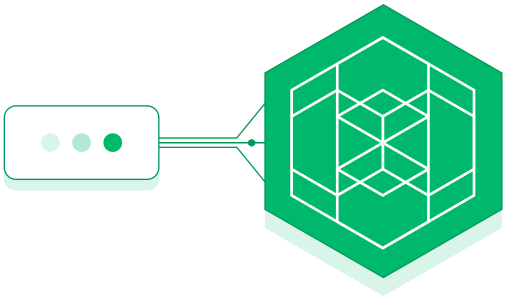


Compound AI

Complex AI applications like real-time phone calling and automated agents rely on multiple AI models and business logic. Baseten Chains simplifies building compound AI systems by removing boilerplate code, ensuring type-safe interfaces, and optimizing GPU usage to reduce latency. Self-hosting ensures that every step of your complex pipelines has the same security protocols in place and is fully integrated with your existing systems.

Image generation

Baseten provides high-performance, cost-efficient infrastructure for image generation, delivering low latency and high throughput. Self-hosting ensures the secure handling of proprietary image data, which is crucial for enterprises with strict compliance requirements.

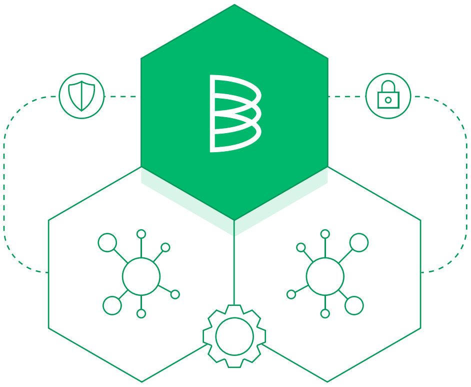


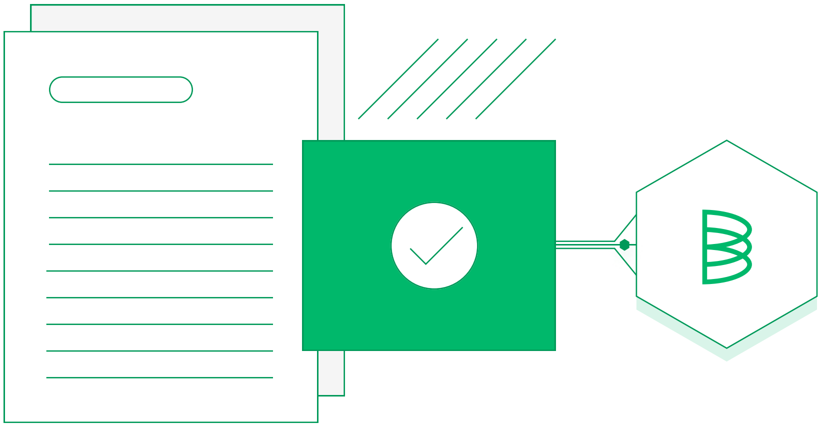
Proprietary models

Baseten supports secure deployments of proprietary models that achieve similar performance to leading public model endpoints. Data governance, compliance, reliability, and control don't have to come at the expense of industry-leading performance.

AI development platform

Baseten offers a cohesive platform for deploying proprietary, fine-tuned, and third-party models. It simplifies infrastructure management, enhances developer productivity, and ensures secure, compliant AI deployments within existing tech stacks.





Conclusion

Choosing the right hosting option is crucial for the success of your AI initiatives. Baseten's flexible solutions—whether Self-hosted for control and customization, Cloud for simplicity and scalability, or Hybrid for ultimate flexibility—cater to diverse enterprise needs. By carefully evaluating your specific requirements against the features and benefits of these hosting options, you can make an informed decision that aligns with your organizational goals.

For personalized guidance or more information, please contact Baseten’s team of experts.

Ready to see what Baseten can do for you?

We’d love to hear how you’re handling inference in production and discuss how Baseten can help your company excel. Reach out to us at sales@baseten.co!

Further reading:

- Case studies →
- Technical documentation →
- Webinars and events →