November 4, 2024

BitOoda

BitOoda AI Research

Cloud Infrastructure for AI Workloads

Abstract

Access to services from Cloud Service Providers (CSPs) can relieve you of the need to maintain your own data centers. As competition among CSPs intensifies, companies—whether startups or large corporations new to AI—can opt for an asset-light approach by relying on cloud services. Beyond the sheer number of providers now offering cloud compute, there is also a wide variety of services available, which introduces a paradox of choice and can often lead to management making suboptimal decisions for their AI compute needs. This report will cover the various options available when using a CSP for AI compute workloads, diving deep into each type, its pros and cons, and the specific services you might consider.

Research

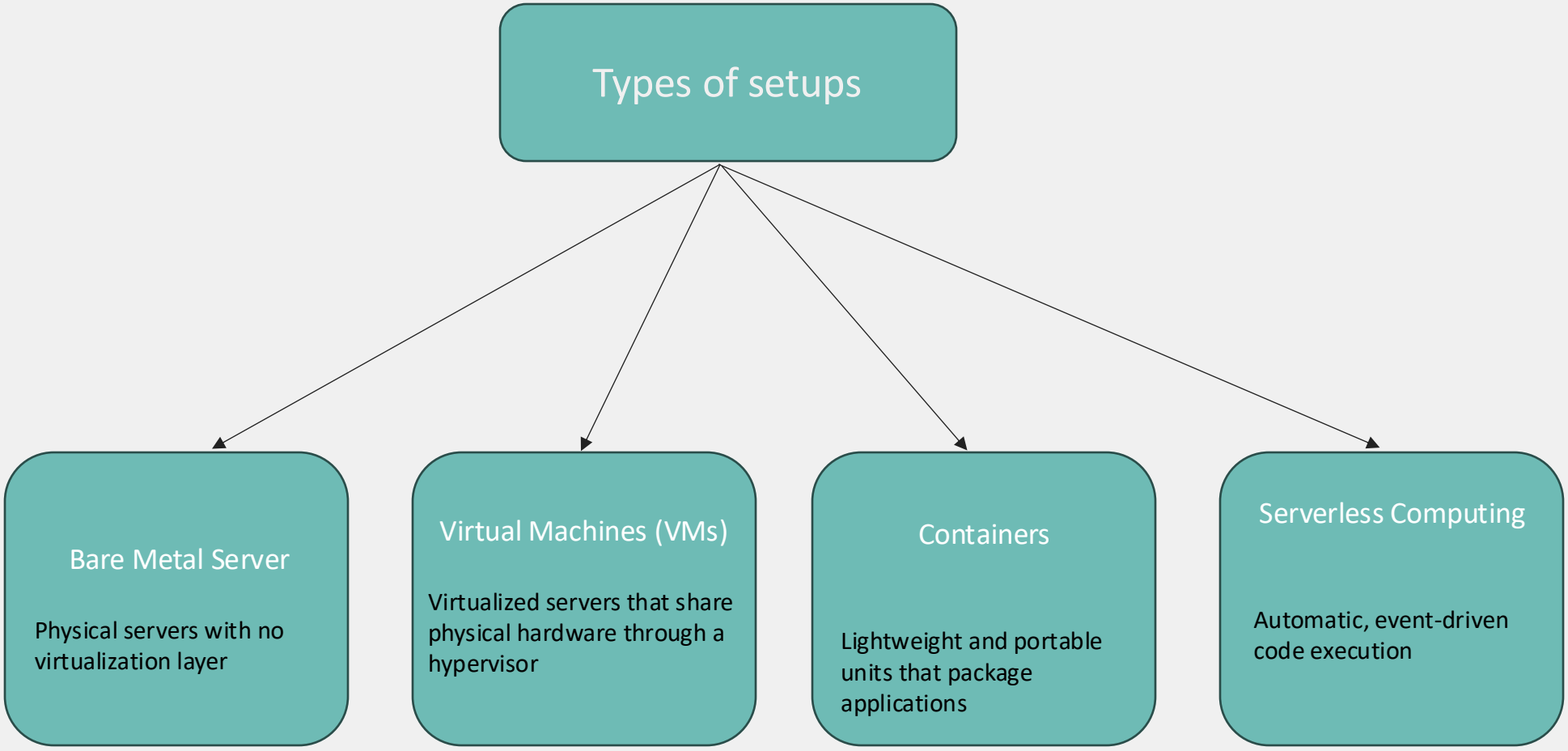
Tim Kelly, CEO & Founder
Dhyay Bhatt, Head of AI
Niraj Yagnik, Lead Developer & AI Engineer

- What are the main compute resource setups for AI workloads and why are they important?
- What makes each setup—bare metal, virtual machines, containers, and serverless—unique in its approach to AI tasks?
- How do different setups impact performance, scalability, and cost for AI applications?
- What are the pros and cons of each compute setup, and which AI workloads are they best suited for?

Compute Resource Allocation Options

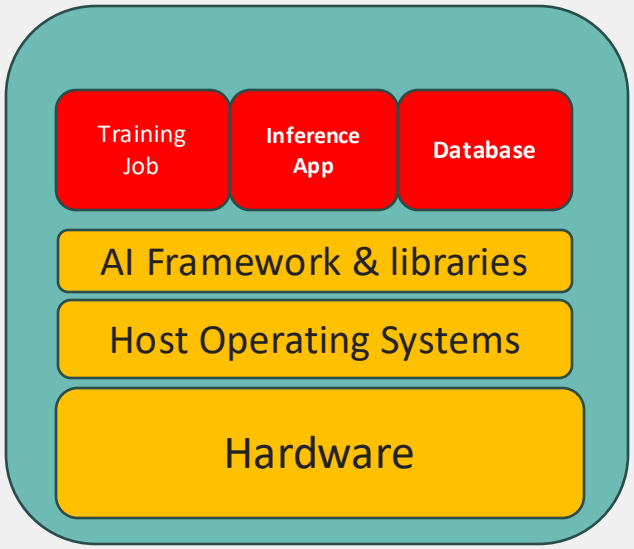
Types of Compute Resource Setups

When choosing infrastructure for AI workloads, one of the first considerations is how resources are allocated. Here are the main types of compute resource setups



Types of Compute Resource Setups

Bare Metal Servers



Bare metal servers are physical servers dedicated to a **single tenant** without any virtualization layer. They provide direct access to hardware resources.

In a bare metal setup, applications like training, inference, and databases run directly on dedicated hardware without virtualization, providing high-performance, **direct access to resources**. This setup is ideal for intensive AI tasks, such as l**arge scale LLM training**, requiring maximum processing power, customization, and low latency.

Pros:

- **High Performance:** Direct hardware access, no overhead.
- **Customizable:** Tailored configurations for AI tasks.
- **Low Latency:** Ideal for real-time applications.

Cons:

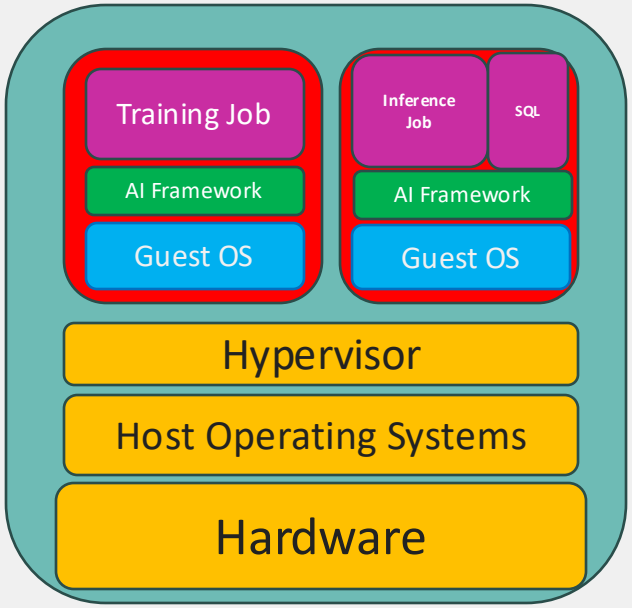
- **Costly:** Higher expense for dedicated resources.
- **Limited Scalability:** Slower to scale compared to virtual solutions.
- **More Maintenance:** Requires manual management and updates.

Options you could consider:

- **Oracle Cloud Infrastructure (OCI)**: Offers bare metal instances with NVIDIA GPUs (H100, L40S, A100) tailored for AI and machine learning, with plans to add H200 and Blackwell GPUs, as well as AMD Instinct MI300X GPUs.
- **NVIDIA GH200 Superchip**: Oracle provides the NVIDIA GH200 Grace Hopper Superchip for efficient large language model (LLM) inference.
- **IBM Cloud**: Provides customizable bare metal servers with NVIDIA GPUs and Intel/AMD CPUs, allowing specific GPU selection and integration with IBM Watson for enhanced AI capabilities.

Types of Compute Resource Setups

Virtual Machine



Virtualized servers run on a **hypervisor**, allowing multiple virtual machines (VMs) to share the **same physical hardware.**

A hypervisor is software that creates and manages virtual machines by **allowing multiple operating systems** to share a single physical host's hardware resources.

This setup provides a balance between performance, flexibility, and cost-effectiveness, making it suitable for a **wide range of AI workloads.** Virtualized environments offer isolation between applications while enabling flexible resource allocation, ideal for tasks like **inference** or **moderate-scale model training.**

Pros:

- **Flexible Scaling:** VMs can easily be resized or added to meet changing AI workload demands.
- **Cost-Effective:** Shared hardware reduces costs
- **Isolation for Multi-Tenancy:** VMs provide isolation, allowing multiple AI applications to run securely

Cons:

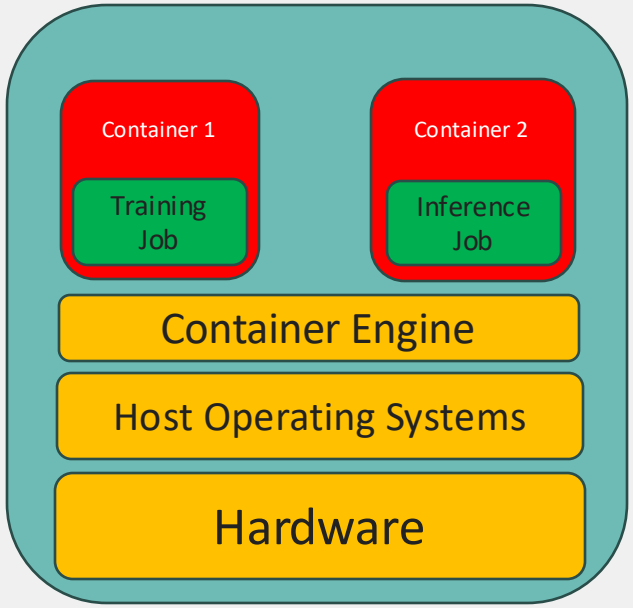
- **Performance Overhead**: Virtualization introduces latency
- **Resource Contention**: Shared resources can cause performance issues due to other VMs.
- **Complex Management**: Running multiple VMs for AI workloads demands careful monitoring.

Options you could consider:

- **Amazon EC2:** Provides a variety of GPU-optimized VM types, suitable for both training and inference AI tasks.
- **Microsoft Azure Virtual Machines:** Supports GPU-accelerated VMs for high-performance AI applications, offering both Linux and Windows environments.
- **Google Compute Engine:** Offers customizable VMs with GPU support, ideal for scaling AI inference and moderate model training.

Types of Compute Resource Setups

Containers



Containers bundle applications and everything they **need to run in small, separate units.** This setup makes it easy to **deploy and scale** AI models quickly and reliably across different environments. Containers work especially well for systems with multiple small tasks (microservices) and for running AI tasks, as they use resources efficiently and can scale up or down quickly.

Containers run applications (like training and inference jobs) in isolated units on shared hardware. A **Container Engine** manages these containers on top of the **Host Operating System**, enabling efficient and consistent operation across multiple environments.

Pros:

- **Portability:** Ensures consistent environments across development, testing, and production.
- **Efficient Resource Use:** Shares the host OS, reducing overhead and making containers lightweight.
- **Rapid Scaling:** Ideal for quickly scaling AI inference to meet demand.

Cons:

- **Security Risks:** Shared host OS kernel can introduce vulnerabilities.
- **Orchestration Complexity:** Requires tools like Kubernetes, adding management overhead.
- **Limited for Long Training:** Best for short-lived or inference tasks, not long-running training jobs.

Options you could consider:
Managed Kubernetes services like **Amazon EKS (Elastic Kubernetes Service), Azure Kubernetes Service,** and **Google Kubernetes Engine** offer scalable container orchestration with AI integration, supporting GPU-based inference and machine learning model deployment.

Next Slide:
Some services provide ML-specific, container-based offerings, tailored for the unique demands of AI workloads.

Types of Compute Resource Setups

Managed AI Platforms for AI Workloads (Containers – 2)

When selecting infrastructure for AI workloads, it's crucial to consider how resources are allocated. Managed AI platforms **offer end-to-end environments** for building, training, and deploying machine learning models, **leveraging containers behind the scenes** to provide a fully managed experience. This approach allows organizations to focus on AI development without the burden of managing underlying infrastructure.

Pros:

- **End-to-End Solution:** Supports the entire machine learning workflow, from data preparation to model deployment.
- **Scalability:** Automatically scales compute resources based on training and inference needs, optimizing costs and performance.
- **Ease of Use:** Managed services reduce infrastructure setup, allowing data scientists to focus on model development and experimentation.

Cons

- **Vendor Lock-In:** Heavy reliance on a single cloud service provider's ecosystem, which can limit flexibility and portability.
- **Higher Cost at Scale:** Managed services can be more expensive, especially for large-scale training or deployment.
- **Limited Customization:** Less control over underlying infrastructure compared to DIY container solutions.

Options you could consider:

- **AWS SageMaker:** Provides tools for building, training, tuning, and deploying machine learning models. It includes SageMaker Studio, SageMaker Autopilot for AutoML, and support for custom containers.
- **Google Vertex AI:** Offers a unified platform for AI development with tools for data labeling, model training, and deployment. Vertex AI integrates with Google's Tensor Processing Units (TPUs) for accelerated training.
- **Azure Machine Learning:** Provides an integrated workspace for managing machine learning workflows, with tools for automated ML, responsible AI, and deployment to various environments, including Kubernetes.

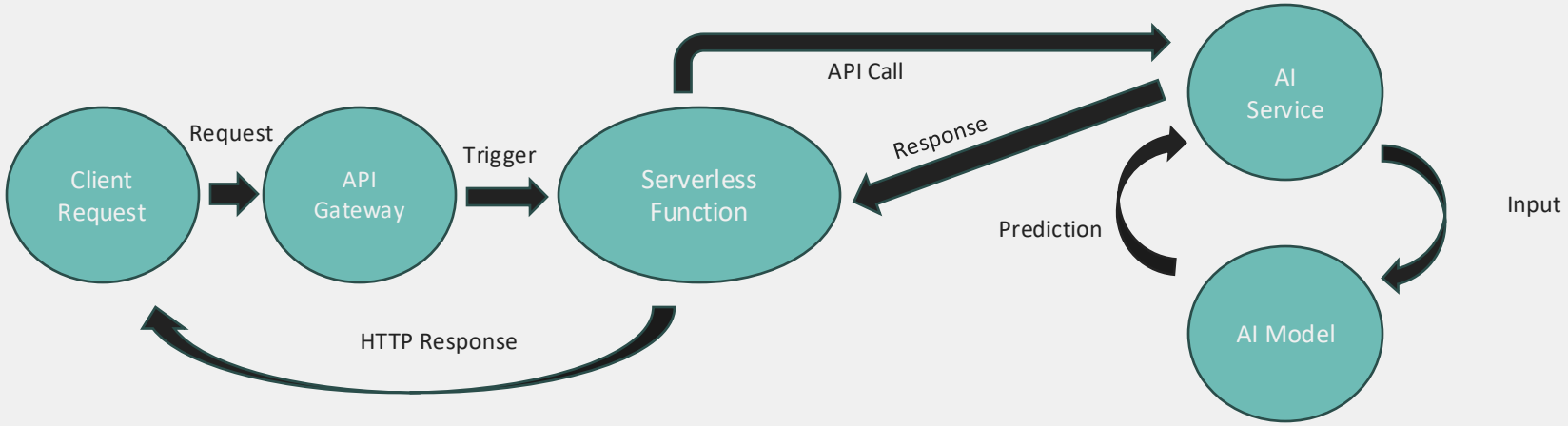
Types of Compute Resource Setups

Serverless Compute

Serverless computing offers an **event-driven, on-demand approach** where compute resources are allocated automatically in response to workload needs. This setup allows developers to run code **without managing servers**, making it ideal for lightweight, burstable AI tasks such as **inference, data preprocessing, or real-time processing.**

Options you could consider:

- **Amazon SageMaker Serverless Inference:** Allows serverless deployment and scaling of ML models, with automatic compute resource management.
- **Azure Functions:** Microsoft's serverless service supports deploying AI models for inference, integrated within Azure's AI offerings.
- **Hugging Face NVIDIA NIM API:** Offers serverless inference for enterprise users on Hugging Face, using NVIDIA DGX Cloud for accelerated compute.
- **OpenAI API:** Provides serverless access to language models and AI capabilities through an easy-to-use API.



Sample Serverless Workflow:

- Client makes an HTTP request
- API Gateway triggers the serverless function and function makes an API call to the AI service
- AI service sends input to the model
- Model returns its prediction
- Results flow back as responses

Summarizing the Services

What's best for you?

Compute Setup	Compute Platform	Scaling	Applications	Unique Features	Examples of Services
Bare Metal	Physical, dedicated servers	Manual scaling	High-performance AI training, HPC applications	Maximum performance, full hardware control	IBM Cloud Bare Metal, Oracle Bare Metal, Packet
Virtual Machines (VMs)	Hypervisor-based VMs	Flexible with auto-scaling	Flexible AI workloads, multi-tenancy	Cost-effective with shared hardware, isolated environment	Amazon EC2, Google Compute Engine, Azure VMs
Containers	Docker/ Kubernetes clusters	Rapid horizontal scaling	Scalable AI inference, microservices	Lightweight, portable, and consistent across environments	Amazon EKS, Azure AKS, Google GKE
Serverless	Cloud provider functions	Automatic scaling	Event-driven AI tasks, burstable workloads	Cost-effective, pay-as-you-go, minimal management	AWS Lambda, Google Cloud Functions, Azure Functions

Disclosures

Purpose

This research is only for the clients of BitOoda. This research is not intended to constitute an offer, solicitation, or invitation for any securities and may not be distributed into jurisdictions where it is unlawful to do so. For additional disclosures and information, please contact a BitOoda representative at info@bitooda.io.

Analyst Certification

Niraj Yagnik and Dhyay Bhatt, the primary research analysts of this report, hereby certifies that all of the views expressed in this report accurately reflect their personal views, which have not been influenced by considerations of the firm's business or client relationships.

Conflicts of Interest

This research contains the views, opinions, and recommendations of BitOoda. This report is intended for research and educational purposes only. We are not compensated in any way based upon any specific view or recommendation.

General Disclosures

Any information ("Information") provided by BitOoda Holdings, Inc., BitOoda Digital, LLC, BitOoda Technologies, LLC or Ooda Commodities, LLC and its affiliated or related companies (collectively, "BitOoda"), either in this publication or document, in any other communication, or on or through http://www.bitooda.io/, including any information regarding proposed transactions or trading strategies, is for informational purposes only and is provided without charge. BitOoda is not and does not act as a fiduciary or adviser, or in any similar capacity, in providing the Information, and the Information may not be relied upon as investment, financial, legal, tax, regulatory, or any other type of advice. The Information is being distributed as part of BitOoda's sales and marketing efforts as an introducing broker and is incidental to its business as such. BitOoda seeks to earn execution fees when its clients execute transactions using its brokerage services. BitOoda makes no representations or warranties (express or implied) regarding, nor shall it have any responsibility or liability for the accuracy, adequacy, timeliness or completeness of, the Information, and no representation is made or is to be implied that the Information will remain unchanged. BitOoda undertakes no duty to amend, correct, update, or otherwise supplement the Information.

The Information has not been prepared or tailored to address, and may not be suitable or appropriate for the particular financial needs, circumstances or requirements of any person, and it should not be the basis for making any investment or transaction decision. The Information is not a recommendation to engage in any transaction. The digital asset industry is subject to a range of inherent risks, including but not limited to: price volatility, limited liquidity, limited and incomplete information regarding certain instruments, products, or digital assets, and a still emerging and evolving regulatory environment. The past performance of any instruments, products or digital assets addressed in the Information is not a guide to future performance, nor is it a reliable indicator of future results or performance.

All derivatives brokerage is conducted by Ooda Commodities, LLC a member of NFA and subject to NFA's regulatory oversight and examinations. However, you should be aware that NFA does not have regulatory oversight authority over underlying or spot virtual currency products or transactions or virtual currency exchanges, custodians or markets.

BitOoda Technologies, LLC is a member of FINRA.

"BitOoda", "BitOoda Difficulty", "BitOoda Hash", "BitOoda Compute", and the BitOoda logo are trademarks of BitOoda Holdings, Inc.