vCluster

GPUs Without the Headache

Scaling AI Infrastructure for Engineering Teams

Does my company even need GPUs?

Myth

Only OpenAI, Google, or Anthropic need GPUs — You only need GPUs if you train foundation models and that's just a handful of companies.

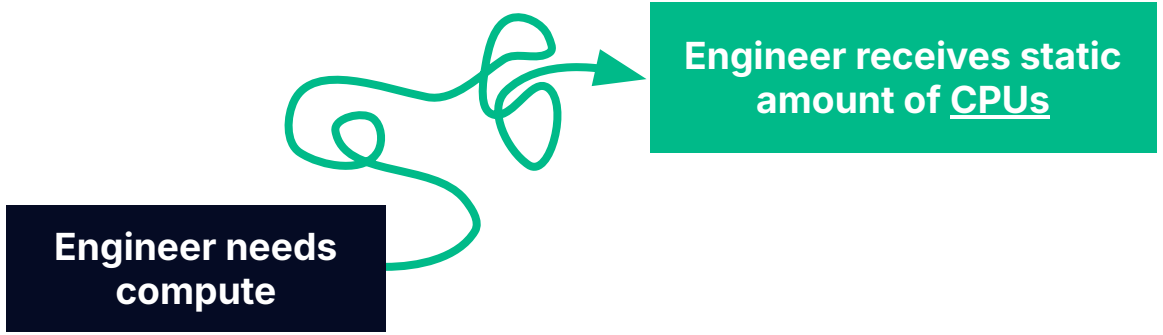
Reality

Sure, you won't train foundation models on 10,000+ GPUs but you will need a few hundred GPUs to...

- Fine-Tune Models on Proprietary Data
- Run Inference In Production
- Run Domain-Specific ML Workloads
- Train Highly Specialized Models
- Generate Embeddings (RAG)
- Attract & Retain Top Talent

How to give my engineering teams access to GPUs?

The old way of static assignment of compute doesn't work anymore

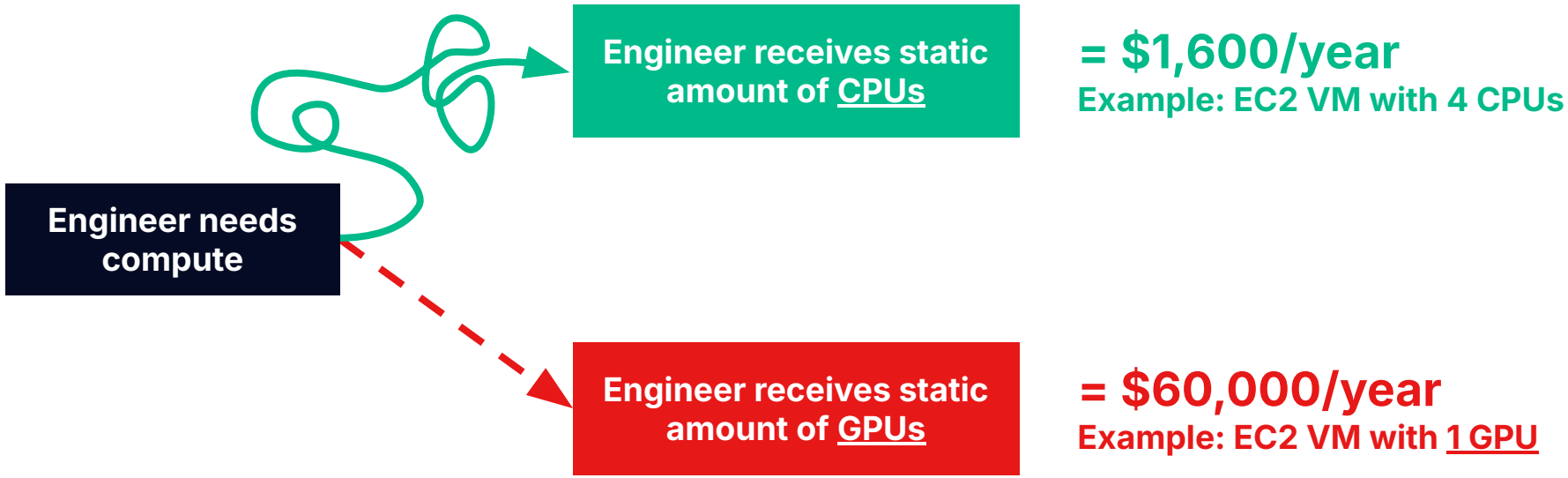


= $1,600/year

Example: EC2 VM with 4 CPUs

How to give my engineering teams access to GPUs?

The old way of static assignment of compute doesn't work anymore



The Evolution of Compute Delivery

Infra is getting more dynamic over time

Slow Provisioning & Very Rigid Resources

Engineers get servers with hard to change specs

...after waiting for months

Fast Provisioning But Still Static Resources

Engineers get VMs with static but adjustable specs

...within hours or days

Automatic Scheduling & Fully Dynamic Compute

Compute is completely dynamically orchestrated

...often in real-time

Key Challenges

When delivering GPUs to engineering teams

Sourcing & Provisioning	Allocation & Sharing	Cost Control
Where to get GPUs from?	**Who should get capacity and when?**	**How to track cost and stay keep budgets sane?**
Hyperscalers	Reduce Wait Times	Set Budgets & Track Cost
Neoclouds	Ensure Fair Use	Create Cost Awareness
On Prem / AI Factory	Prevent Idle Compute & Dynamically Reallocate	Drive Accountability
All of the above?		

Fully Dynamic & Scalable GPU Infra

What capabilities you need on each layer of the stack

Access Layer & UX

- Provide near-instant self-service access
- Ensure tenant autonomy & flexibility

Runtime Layer

- Enable seamless execution on GPUs
- Streamline and abstract GPU handling

Compute Layer

- Track consumption, enforce quotas, ensure fair use
- Auto-scale GPU resources & optimize cost

Data Center Layer

- Ensure cross-cloud portability
- Consider hybrid cloud / cloud bursting

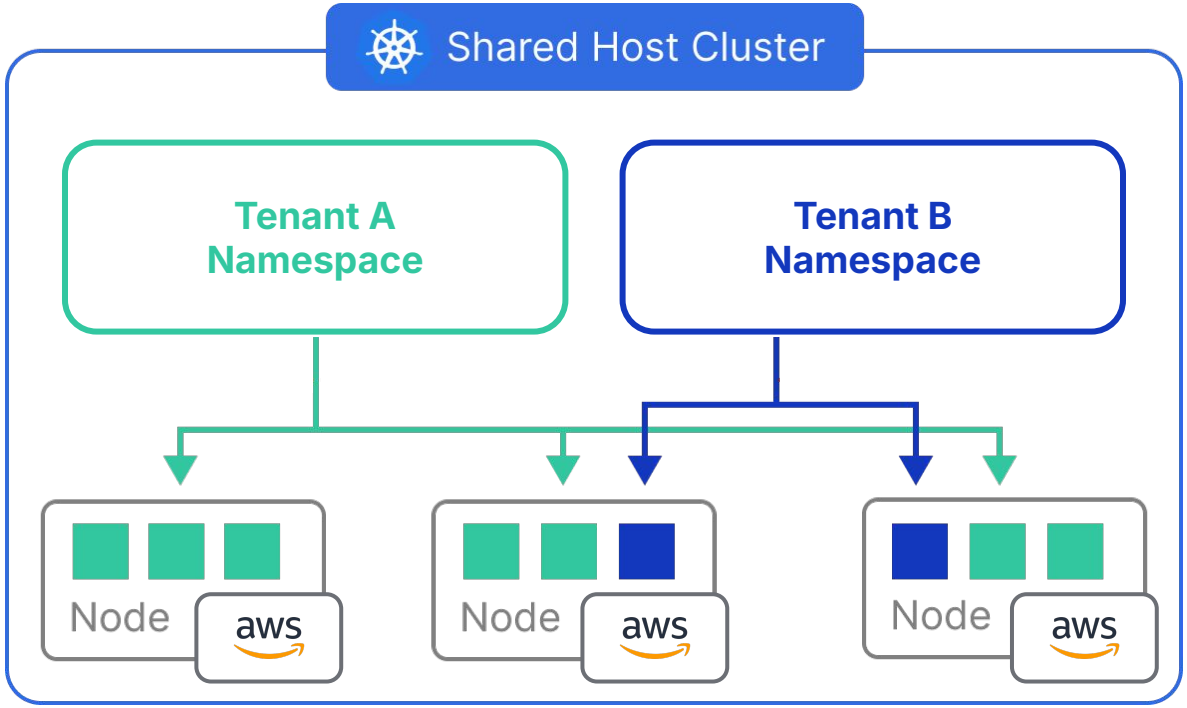
Why You Should Build Your AI Infra on Kubernetes

Today Kubernetes is the foundation compute API

- Follow industry leaders
  - Frontier AI Labs with the biggest LLMs use Kubernetes for dev and experimentation before they train their LLMs using their large custom HPC schedulers
  - Most mid-size AI companies and internal teams use Kubernetes
- NVIDIA shows strong support for Kubernetes
  - NVIDIA Container Toolkit
  - Run:ai Acquisition and deep integration into NVIDIA Enterprise AI
- Open standard and vast ecosystem support
  - Kubeflow, Volcano, Ray, Kueue, etc.
  - Commercial HPC storage and networking solutions available
- Existing Kubernetes expertise is likely going to make stable operations easier and faster than any alternative

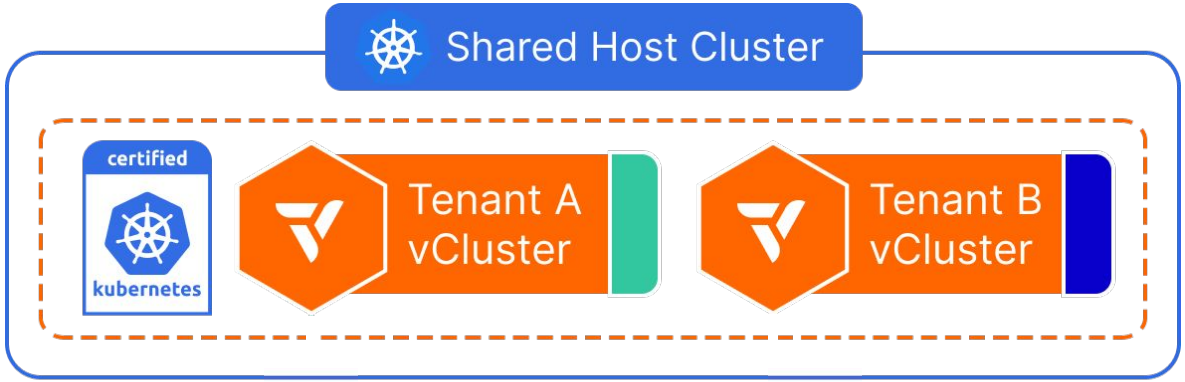
Traditional Single-Cloud Managed Kubernetes

Lock-in to single cloud provider, low tenant autonomy and noisy neighbor issues



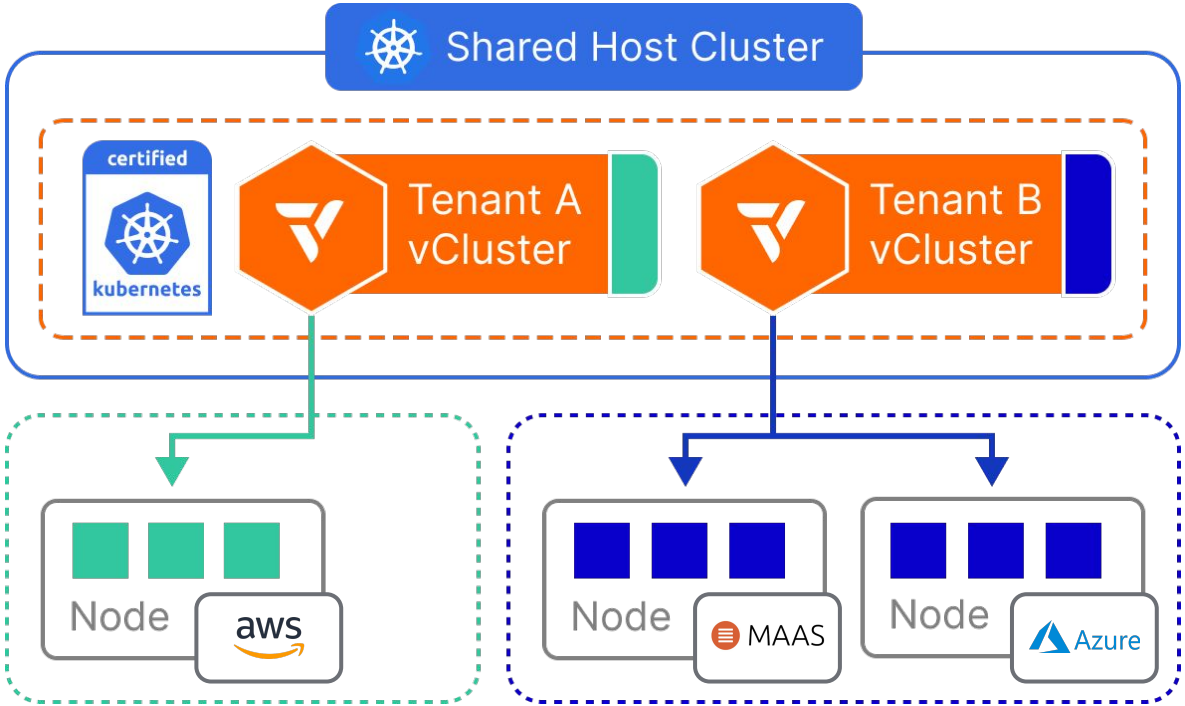
vCluster To Enable Secure Multi-Tenancy

For more tenant autonomy and better isolation



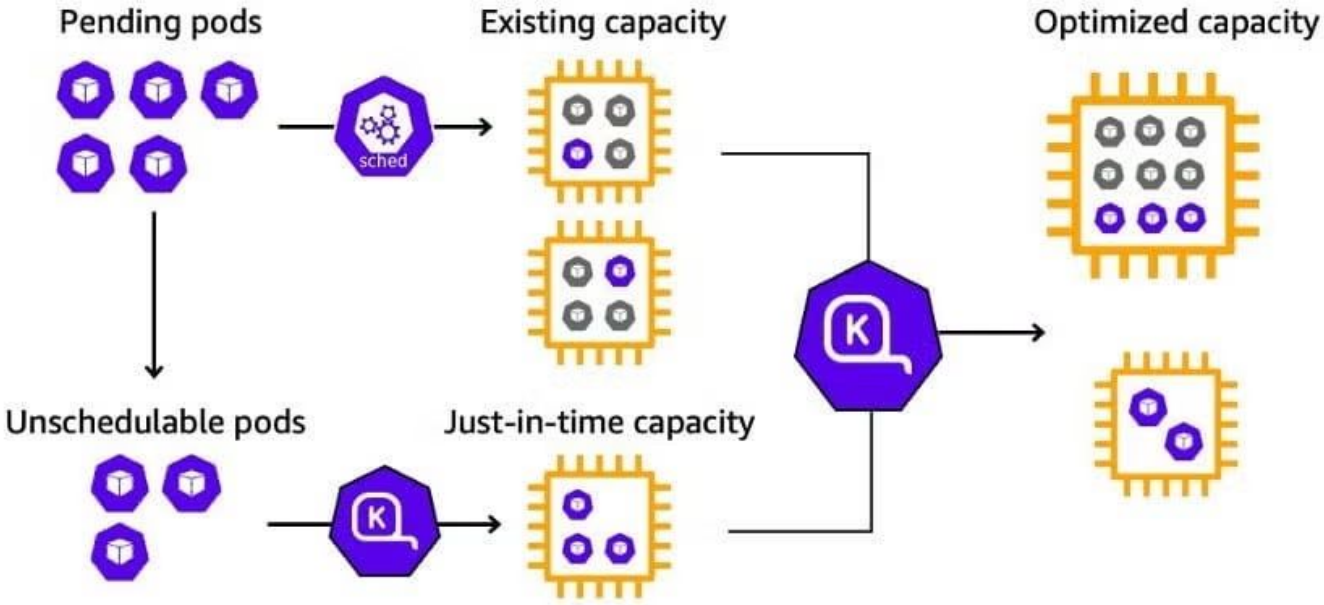
vCluster Auto Nodes For Cross- & Hybrid Cloud

With real-time dynamic auto-scaling of GPU nodes separate for each tenant



vCluster Auto Nodes is based on Karpenter

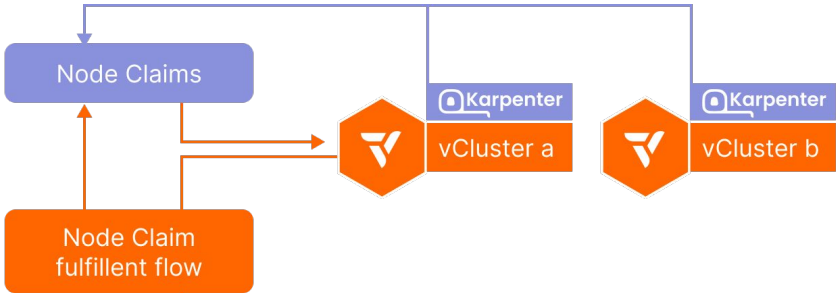
Same experience as EKS Auto Mode but it runs across clouds and even in your private cloud



Credit: https://www.cncf.io/blog/2024/11/06/karpenter-v1-0-0-beta/

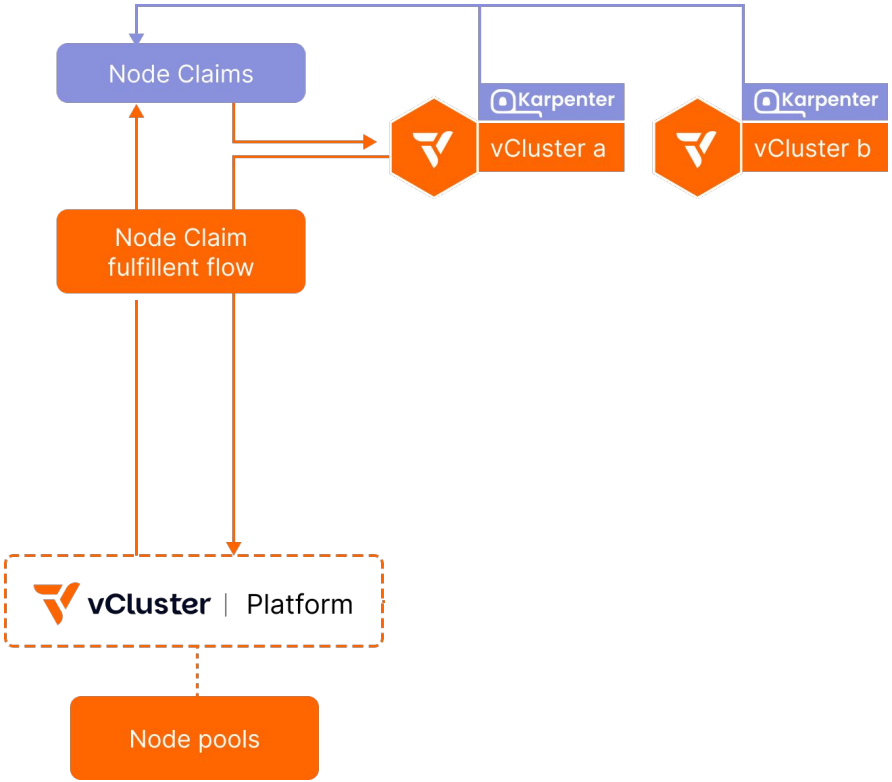
vCluster with Auto Nodes

Automatic node provisioning for vCluster using Karpenter



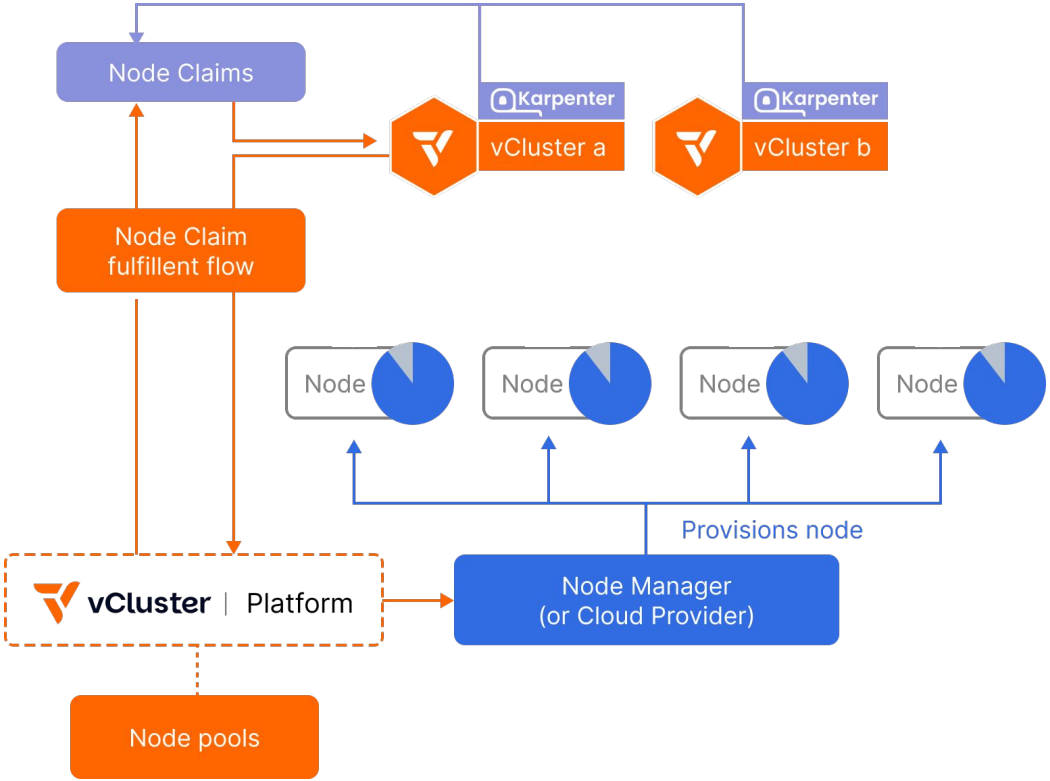
vCluster with Auto Nodes

Automatic node provisioning for vCluster using Karpenter



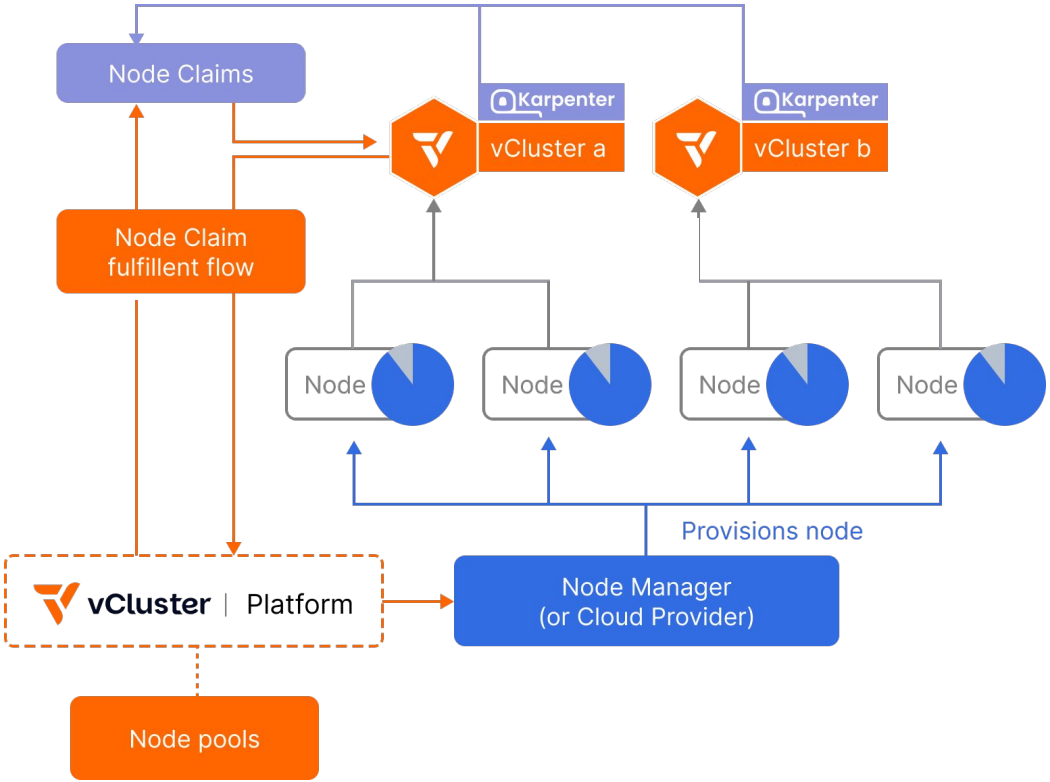
vCluster with Auto Nodes

Automatic node provisioning for vCluster using Karpenter



vCluster with Auto Nodes

Automatic node provisioning for vCluster using Karpenter



Node Providers For vCluster Auto Nodes

Maximum flexibility for all major public and private cloud environments - and even bare metal

- Out-of-the-box providers for all major and even most niche public cloud providers
- Support for many provide cloud systems including OpenStack, MAAS, etc.
- BYO Terraform/OpenTofu providers

- Out-of-the-box support for NVIDIA DGX supercomputers
- Support for Kubernetes-native virtualization on top of bare metal systems

Auto Nodes Example

Dynamic Node Pools configured in vcluster.yaml

vcluster.yaml

```yaml
privateNodes:
  enabled: true
  autoNodes:
    dynamic:
      - name: gcp-nodes
        provider: gcp
        requirements:
          property: instance-type
          operator: In
          values: ["e2-standard-4", "e2-standard-8"]
      - name: aws-nodes
        provider: aws
        requirements:
          property: instance-type
          operator: In
          values: ["t3.medium", "t3.large"]
      - name: private-cloud-openstack-nodes
        provider: openstack
        requirements:
          property: os
          value: ubuntu
```