### **Deployment paths for Qdrant on AWS**

| **Option** | **What it is** | **When it shines** | **You run / patch** | **Monthly cost driver** | **Portability** |
| --- | --- | --- | --- | --- | --- |
| **1. EC2 + Docker / Compose** | Single VM (or ASG) with the official Docker image | Dev, PoC, small workloads (<50 M vectors) | OS, Qdrant upgrades, backups, scaling | EC2 + gp3/EBS | ⭐⭐⭐ (lift-and-shift AMI) |
| **2. ECS Fargate service** | Qdrant container as a task behind an ALB | You like “serverless”, don’t need multi-AZ replication yet | Task definition, snapshots | vCPU + GiB-RAM-hrs, EFS | ⭐⭐ |
| \*\*3. EKS (Kubernetes) with the **official Helm chart** | StatefulSet + persistent EBS; supports clustering, sharding, replicas | Long-term prod, need HA, want the same YAML in GKE later | Helm upgrades; K8s handles rolling restarts | EKS control-plane + node groups + EBS | **⭐⭐⭐⭐ (cloud-agnostic)** |
| **4. AWS Marketplace AMI** | One-click single-node image published by Qdrant | Fastest test drive for non-Docker teams | Just the EC2 OS | Same as EC2 | ⭐ |
| \*\*5. **Qdrant Cloud (managed SaaS)** | Fully-managed cluster provisioned **inside AWS**; free 1 GB tier | No-ops, SLA, auto-scaling, built-in monitoring | Nothing—vendor patches & scales | Pay per vCPU/RAM/disk (starts free) | ⭐⭐ (runs on AWS or GCP) |

Sources: Helm chart repo [github.com](https://github.com/qdrant/qdrant-helm?utm_source=chatgpt.com), Docker/Compose guide [qdrant.tech](https://qdrant.tech/documentation/guides/installation/?utm_source=chatgpt.com), AWS Marketplace listing [aws.amazon.com](https://aws.amazon.com/marketplace/pp/prodview-rtphb42tydtzg?utm_source=chatgpt.com), Qdrant Cloud docs & pricing [qdrant.tech](https://qdrant.tech/documentation/cloud/create-cluster/?utm_source=chatgpt.com)[qdrant.tech](https://qdrant.tech/pricing/?utm_source=chatgpt.com).

### **How to choose**

| **Decision factor** | **Lean EC2 / ECS** | **Lean EKS** | **Lean Qdrant Cloud** |
| --- | --- | --- | --- |
| **You need cloud-agnostic parity with future GCP** | Manual or Terraform-ised EC2, but *rework* later | **Same Helm chart runs on GKE** → zero drift | Runs on both clouds, but you lose infra control |
| **Ops bandwidth** | You patch & monitor OS/Qdrant | You still own upgrades, but rolling & zero-downtime via K8s | Vendor handles patching, HA, metrics |
| **Scale target** | Single-AZ up to ~200 GB SSD | Multi-node, multi-AZ, sharding & replicas | Elastic; cluster slider in UI |
| **Spend** | Cheapest raw EC2 + gp3 | +15-20 % for K8s control plane | Bill for managed CPU/RAM/disk; no EC2 |
| **SLA / backup** | DIY EBS snapshots / ASG | DIY snapshots, but StatefulSet safety nets | Built-in daily snapshots & SLA |

### **Recommended roadmap**

1. **Stage / QA** – spin a **single-node EC2** with Docker Compose (10 min).  
    *Pros*: identical API, cheapest; you already have the compose file.

**Production** – deploy the **Helm chart on EKS**:  
  
 bash  
CopyEdit  
helm repo add qdrant https://qdrant.github.io/qdrant-helm

helm install vecdb qdrant/qdrant \

--set persistence.size=500Gi \

--set service.type=LoadBalancer \

--set replicas=3 \

--set nodeSelector."eks\\.amazonaws\\.com/nodegroup"=sorage

t

* + Shards + replicas give HA; EBS gp3 volumes snapshot to S3 automatically.
  + The same chart runs on GKE when you spend the GCP credits—true portability.

1. **If ops becomes a bottleneck** – migrate to **Qdrant Cloud’s “Hybrid”** mode.  
    You create an EKS cluster (or GKE) but let Qdrant-Cloud’s control plane manage upgrades and monitoring; pricing starts at ≈ $0.014 / vCPU hour, and you keep data inside your AWS account.

### **Key implementation notes**

* **Open ports**: 6333 (REST) and 6334 (gRPC). Use an NLB or ALB with TLS.
* **Security**: enable Qdrant’s built-in API key or use ALB’s OIDC auth.
* **Back-ups**: EBS Snapshot or qdrant.backup CLI → S3; automate with AWS Backup.
* **Monitoring**: Helm chart exposes Prometheus metrics; pipe into AMP + Grafana.
* **Scaling**: in EKS, cluster-autoscaler adds nodes; Qdrant shards rebalance online.

#### **Bottom line**

* **Fastest path today**: keep your single-node EC2 while validating search.
* **Best long-term, cloud-agnostic**: EKS + official Helm chart.
* **Least-ops**: Qdrant Cloud (managed) on AWS region of your choice.

Pick the tier that matches your ops budget and traffic forecast—you can move up the ladder without changing your application code, because the Qdrant HTTP/gRPC APIs stay identical across all options.