About GKE Scalability

AUTOPILOT (/KUBERNETES-ENGINE/DOCS/CONCEPTS/AUTOPILOT-OVERVIEW)

STANDARD (/KUBERNETES-ENGINE/DOCS/CONCEPTS/TYPES-OF-CLUSTERS)

This page provides a set of recommendations for planning, architecting, deploying, scaling, and operating large workloads on Google Kubernetes Engine (GKE) clusters. We recommend you follow these recommendations to keep your scaling workloads within service-level objectives (SLOs) (https://landing.google.com/sre/sre-book/chapters/service-level-objectives).

Available recommendations for scalability

Before planning and designing a GKE architecture, map parameters specific to your workload (for example the number of active users, expected response time, required compute resources) with the resources used by Kubernetes (such as Pods, Services, and 'CustomResourceDefinition'). With this information mapped, review the GKE scalability recommendations.

The scalability recommendations are divided based in the following planning scopes:

- **Plan for scalability**: To learn about the general best practices for designing your workloads and clusters for reliable performance when running on both small and large clusters. These recommendations are useful for architects, platform administrators, and Kubernetes developers. To learn more, see Plan for scalability (/kubernetes-engine/docs/concepts/planning-scalability).
- **Plan for large-size GKE clusters**: To learn how to plan to run very big-size GKE clusters. Learn about known limits of Kubernetes and GKE and ways to avoid reaching them. These recommendations are useful for architects and platform administrators. To learn more, see Plan for large GKE clusters (/kubernetes-engine/docs/concepts/planning-large-clusters).
- **Plan for large workloads**: To learn how to plan architectures that run large Kubernetes workloads on GKE. It covers recommendations for distributing the workload among projects and clusters, and adjusting these workload required quotas. These recommendations are useful for architects and platform administrators. To learn more, see Plan for large workloads (/kubernetes-engine/docs/concepts/planning-large-workloads).

These scalability recommendations are general to GKE and are applicable to both GKE Standard and GKE Autopilot modes. GKE Autopilot provisions and manages the cluster's underlying infrastructure for you, therefore some recommendations are not applicable.

Caution: Test your planned cluster configuration before its implementation. Some design decisions might include fixed parameters, for example, CIDRs definition. Changing these parameters on existing clusters is not available and it requires cluster recreation.

What's next?

- Plan for scalability (/kubernetes-engine/docs/concepts/planning-scalability).
- Plan for large GKE clusters (/kubernetes-engine/docs/concepts/planning-large-clusters)
- Plan for large workloads (/kubernetes-engine/docs/concepts/planning-large-workloads)
- See our episodes about building large GKE clusters (https://www.youtube.com/watch?v=542XwAPKh4g).

Last updated 2023-11-27 UTC.