Google Skills

GKE scaling

As we observe, GKE, guided by the custom compute class, will **scale up to TPU nodes** to handle the initial workload, leveraging their superior performance.

As the load increases, and if TPU capacity is reached, GKE will then **scale up to GPU nodes** making sure it’s available and cost effective.

For instance, the first pod might run on a TPU node—as we've constrained that pool to one node. Subsequent pods, due to this constraint, will then start trying to load onto L4 GPU clusters.

This dynamic scaling and intelligent hardware utilization—orchestrated by custom compute classes, GKE, and dual-container vLLM deployment—showcases the **power and flexibility of this solution.**