My code minimizes API calls through caching. There is a cache (dictionary) in Checker to store GPU types by zone. There is another that stores specific GPU quotas by region. This reduces Google API calls since repeated calls can be drawn from the cache.

Initially, my code took long (40+ minutes). For the creation/deletion of VMs, I concluded that parallel execution would greatly optimize the script—an important concept drawn from distributed systems that would be helpful for a large job. Concurrent operations greatly sped up the testing. One way to improve efficiency would be to alter batch sizes.

I'm also unsure if my GCP account allows requests for more than 1 GPU, and I had only requested that in us-regions. Because of these limitations, I know there are calls would definitely fail. However, I left them in the script because the requirements of this assignment include iterating through all regions and zones.

To compare the methods of listing the available gpu types by zone, checking for available quota by region, and creating/deleting the VM, I found that listing the gpu types was fastest, followed by checking the quota, and then creating/deleting the VM. Creating/deleting the VM had the most API calls. In the future, I would look at the available gpu types before creating and deleting the VM.

Zone	GPU Available	GPU Allocated Successfully	Reason for Failure	Time Taken
us-east1-b	No	No	GPU Not Available	1078.58 ms
us-east1-b	No	Yes		1465.34 ms
us-east1-c	No	Yes		1507.97 ms
us-east1-c	No	No	GPU Not Available	800.05 ms
us-east1-d	No	Yes		1411.90 ms
us-east1-d	No	No	GPU Not Available	1062.35 ms
us-east4-c	No	Yes		3752.75 ms
us-east4-c	No	No	GPU Not Available	1037.34 ms
us-east4-b	No	Yes		1533.88 ms
us-east4-b	No	No	API Error	384.66 ms
us-east4-a	No	Yes		1569.37 ms
us-east4-a	No	No	GPU Not Available	1021.18 ms
us-central1-c	No	Yes		1626.06 ms
us-central1-a	No	No	GPU Not Available	921.97 ms
us-central1-a	No	Yes		1711.02 ms
us-central1-f	No	Yes		1256.68 ms
us-central1-f	No	No	API Error	509.53 ms
us-central1-b	No	Yes		1536.31 ms
us-central1-b	No	No	No GPUs Available	0.00 ms

I had to reference ChatGPT for help with the VM configurations and parallel processing. It was also incredibly helpful with logging errors to help catch what was wrong. I had to reference google Cloud documentation for the different regions, zones, and gpu types. It had to be very exact.