AUGUST 2023

IT & DATA SCIENCE

The Data Scientist's Survival Guide to the GPU Shortage

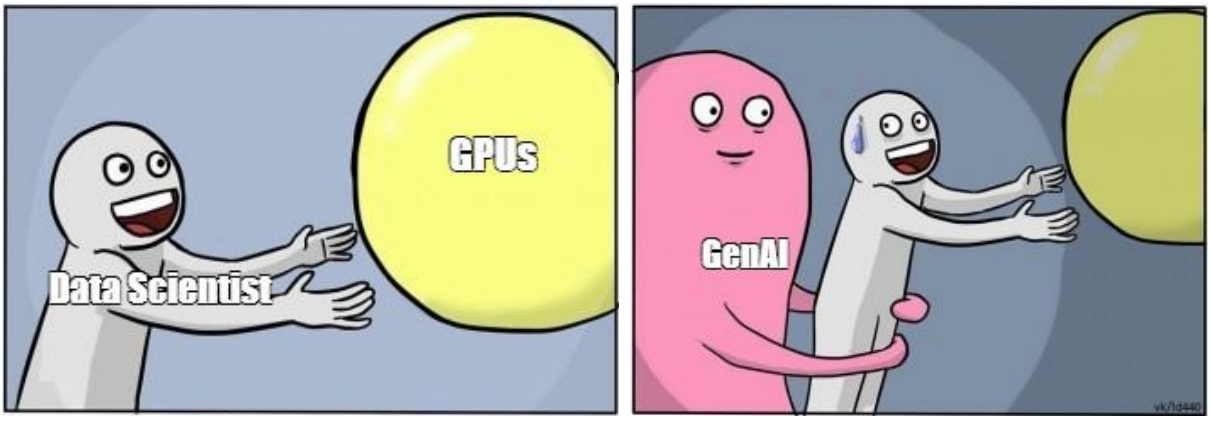
Table of contents

Introduction

The past year has seen the spectacular rise of generative AI, and with it the specter of a worsening GPU shortage. For data scientists, this isn't just a trending news item—without affordable and reliable access to these critical resources, data scientists face many challenges that hinder their productivity and impact their everyday work. In this post, we'll dissect some of those challenges, and offer some strategies to mitigate the effects.

The GPU shortage: How did we get here?

This actually isn't a new phenomenon. Back in 2017, crypto's sudden growth unleashed a surge in demand for compute resources, causing a GPU shortage. Today, the shortage is back, and this time it might be worse thanks to the meteoric rise of generative AI. GenAI's major breakthroughs are undoubtedly impressive, but they also create a tug-of-war with other types of data and computer scientists over GPU resources.



The current crisis also stands apart because the unprecedented growth in the semiconductor industry over the past few years has hindered supply chains, complicating firms' efforts to source the necessary components for GPU production. This, plus the ever-escalating demand brought on by innovations like ChatGPT compound the data science community's fears about a worsening shortage.

Some estimates suggest that more than 10,000 Nvidia GPUs were used in training ChatGPT alone. The projection is that the demand for GPUs will only surge as these services continue to expand. This GPU crunch poses a significant hurdle for large language models like ChatGPT, impeding the rollout of novel features and services. Solutions might lie in ramping up production capacity and exploring alternative methods for training and running large language models.

The silver lining is that the cryptocurrency boom, a past driver of GPU shortage, seems to have fizzled out (for now). Still, the fast growth of AI means data scientists will need to adapt to do more with less GPU resources, at least in the near future.

Day-to-day challenges for data scientists

For data scientists (and anyone else who relies on these resources to do their work), the GPU shortage creates a bottleneck and throttles innovation. The obvious solution—spinning up more GPUs—is usually impractical due to their prohibitive expense and, of course, the shortage itself. The result is a more challenging, frustrating job for data scientists. Here's what it can look like in reality:

- **Training or production batch workloads with on-demand resources**
  The scarcity of GPUs in the cloud (combined with high demand) poses a significant challenge for data scientists. They must constantly search for available GPUs across various regions and clouds whenever they need to run a job—a tedious and time-consuming task. It's particularly painful when they are training or fine-tuning models and running multiple jobs in parallel. Moreover, in scenarios where ML teams handle scheduled production workloads such as batch inference or model re-training, automating GPU provisioning and developing tools to automatically locate GPUs in different regions and clouds can be complex and burdensome.

- **Scaling high-end, interconnected GPUs**
  Scaling a single job to tens or hundreds of high-end, interconnected GPUs to do things like distributed model training can be problematic. This is because high-end GPUs are not as abundant as lower-end ones, and availability can often be constrained. This is especially true when close proximity between the machines is required for attaining maximum performance. Even large cloud providers may struggle to provide this level of resources, especially when many clients are demanding the same resources simultaneously, such as during a GPU shortage.

- **Availability during peak hours**
  An organization wishing to deploy their model on GPU and support demand for tens or hundreds of GPUs at peak hours will likely encounter significant hurdles. Unlike CPUs, GPUs are less elastic. In other words, the supply of GPUs doesn't scale up and down as quickly or easily in response to demand. This is often due to their high cost, more complex setup, and scarcity, resulting in slower scaling rates and lower overall availability. This is true anytime, but especially when peak hours require the provisioning of many more GPUs.

Tips and tricks for mitigating the effects of the shortage

Luckily, there are ways to mitigate these problems, and lessen the impact that the GPU shortage will have:

Reserve resources
Instead of relying on on-demand, opt for reserved instances in the cloud, which are typically a third of the price of on-demand resources. Once you secure reserved instances, ensure efficient use by creating pools and queues. Creating an offline queue can be a smart strategy to manage GPU usage efficiently, especially when tasks are not time-sensitive and can be run during off-peak hours or when the necessary resources become available. Tasks can be prioritized based on their urgency, ensuring that the most critical jobs get done first.

Switch to previous-gen GPUs or to GPUs specialized for inference
An effective solution to the scarcity of GPUs in the cloud is to utilize previous-generation GPUs for low-priority training jobs or batch inference tasks, where the speed or service level agreement (SLA) is not critical. This approach maximizes existing resources without straining the availability of high-demand GPUs. Additionally, for real-time inference, specialized GPUs like T4 or A10 can be used, as they offer superior availability and cost efficiency, ensuring smooth and responsive real-time inference while optimizing resource utilization.

Maximize the resources you DO have
Leverage an AI management platform to automate resource management and work orchestration. Using these platforms empowers data scientists by helping them to share resources and maximize limited GPU resources. For example, by enabling fractional allocation of GPU resources, data scientists can right-size their GPU workloads, and over provision GPUs. This is mainly effective for inference workloads and times where data scientists need to spin up Jupyter notebooks (or other IDE tools) to build and debug their code and need only a fraction of a GPU.

Resource pools, where you can group together multiple GPUs and then allocate them to different projects or teams based on priority are also helpful. This approach stretches the resources you do have further by reducing the idle time of GPUs and allowing data scientists or inference workloads to share GPUs and exploit idle resources.

The bottom line

While the GPU shortage poses a big challenge for today's data scientist, not all hope is lost. By using reserved rather than on-demand resources, switching to low-end carefully choosing the right generation and type of GPU, and optimizing existing GPU usage via an AI management platform, data scientists can be sure that their GPU resources are utilized to the fullest extent.

About Run:ai

Run:ai is an AI management platform for MLOps, Data Science, and DevOps teams. In addition to helping these teams access and utilize their GPU resources more effectively, it also has a powerful set of features that can abstract infrastructure complexities and simplify the process of training and deploying models. With or without a GPU shortage, Run:ai enables data scientists to focus on innovation without having to worry about resource limitations.

Read more about how Run:ai supports data scientists here

www.run.ai/runai-for-data-science