Google TPU

Fernando Lima
Isabella Caselli
Rodrigo Michelassi

November 18, 2024

Abstract

Na era do desenvolvimento de sistemas baseados em Inteligência Artificial, se faz necessário o uso de máquinas super potentes, capazes de processar dados e realizar operações matemáticas de maneira extremamente rápida. Modelos de Machine Learning podem levar horas, até mesmo dias, para serem treinados, devido principalmente a operações como produto interno entre matrizes e a enorme quantidade de dados que são usados, trazendo um prejuízo não apenas de tempo, mas também energético, ambiental e sobretudo lucrativo. Nesse artigo, iremos tratar brevemente sobre a utilização de Cloud TPUs, unidades de processamento de tensores do Google Cloud, que atuam na otimização do treinamento de modelos de aprendizado de máquina, e que se tornou indispensável na academia e na indústria, para todos estudiosos e profissionais da área.

1 Introdução

2 History

2.1 Tensores

2.2 Modelos de Aprendizado de Máquina

3 Arquitetura da TPU

4 TPU vs GPU

5 TensorFlow

6 Cloud TPU v5p

Como discutido anteriormente, LLMs são os novos modelos mais explorados no mercado e na academia. Todavia, esses modelos utilizam dados de toda (ou quase toda) a internet, que cresce ainda mais exponencialmente. Todos os dias há milhões de novos dados sendo gerados. Dessa forma, é importante que o hardware acompanhe o crescimento na quantidade de dados disponíveis, de forma que seja possível possuir poder computacional suficiente para treinar modelos como esses.

Com isso em vista, a evolução das TPUs deve ser rápida, portanto, em 2023, a Google apresentou, em 2023, a v5p, a TPU mais poderosa da empresa. Com esse lançamento, a Google prometeu a entrega de uma performance até 2.8 vezes mais rápida, utilizada para alimentar ecossistemas internos da multinacional, como o Youtube, Android e Gmail.

Essa TPU se diferencia das demais, pois é focada em desempenho possível, sem levar em consideração a facilidade operacional. Um grande diferencial é a forma como a TPU lida com operações em ponto flutuante. Por dados oficiais da Google, essa arquitetura traz 8960 chips e três vezes mais memória HBM.

Mas no fim, o que significam esses valores? A Google promete que a nova TPU consegue escalar o tempo de treinamento, sendo até $4X$ mais rápida que TPUs mais baratas, como a $V4$, devido ao dobramento no tamanho de operações em ponto flutuante que são entregues.

	v4	v5e	v5p
Chips por unidade	4.096	256	8.960
Largura de Banda Interconectores	2.400 Gbps	1.600 Gbps	4.800 Gbps
BF16 TFLOPS	275	197	459
Memória HBM	32 GB	16 GB	95 GB
Largura de Banda HBM	1.228 GB/s	820 GB/s	2.795 GB/s

Figure 1: Comparativo entre TPUs Google para cargas de trabalho em IA e LLM

7 Google Colab e distribuição

References

TPU e Arquitetura
Comparativos oficiais em modelos de IA (TPU vs GPU)
TPU v5p
Conceitos básicos
Wikipedia, tem tudo