GPT-4 Architecture, Infrastructure, Training Dataset, Costs, Vision, MoE

Demystifying GPT-4: The engineering tradeoffs that led OpenAI to their architecture.

DYLAN PATEL AND GERALD WONG

JUL 10, 2023 · PAID

189 52 Share

OpenAI is keeping the architecture of GPT-4 closed not because of some existential risk to humanity but because what they've built is replicable. In fact, we expect Google, Meta, Anthropic, Inflection, Character, Tencent, ByteDance, Baidu, and more to all have models as capable as GPT-4 if not more capable in the near term.

Don't get us wrong, OpenAI has amazing engineering, and what they built is incredible, but the solution they arrived at is not magic. It is an elegant solution with many complex tradeoffs. Going big is only a portion of the battle. OpenAI's most durable moat is that they have the most real-world usage, leading engineering talent, and can continue to race ahead of others with future models.

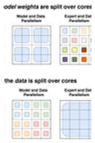
We have gathered a lot of information on GPT-4 from many sources, and today we want to share. This includes model architecture, training infrastructure, inference infrastructure, parameter count, training dataset composition, token count, layer count, parallelism strategies, multi-modal vision adaptation, the thought process behind different engineering tradeoffs, unique implemented techniques, and how they alleviated some of their biggest bottlenecks related to inference of gigantic models.

The most interesting aspect of GPT-4 is understanding why they made certain architectural decisions.

Furthermore, we will be outlining the cost of training and inference for GPT-4 on A100 and how that scales with H100 for the next-generation model architectures.

First off, with the problem statement. From GPT-3 to 4, OpenAI wanted to scale 100x, but the problematic lion in the room is cost. Dense transformers models will not scale further. A dense transformer is the model architecture that OpenAI GPT-3, Google PaLM, Meta LLAMA, TII Falcon, MosaicML MPT, etc use. We can easily name 50 companies training LLMs using this same architecture. It's a good one, but it's flawed for scaling.

See our discussion training cost from before the GPT-4 announcement on the upcoming AI brick wall for dense models from a **training cost** standpoint. There we revealed what OpenAI is doing at a high-level for GPT-4's architecture as well as training cost for a variety of existing models.



The AI Brick Wall – A Practical Limit For Scaling Dense Transformer Models, and How GPT 4 Will Break Past It

DYLAN PATEL · JANUARY 24, 2023

Read full story →

Over the last 6 months we realized that **training cost are irrelevant**.

Sure, it seems nuts on the surface, tens of millions if not hundreds of millions of dollars of compute time to train a model, but that is trivial to spend for these firms. It is effectively a Capex line item where scaling bigger has consistently delivered better results. The only limiting factor is scaling out that compute to a timescale where humans can get feedback and modify the architecture.

Over the next few years, multiple companies such as Google, Meta, and OpenAI/Microsoft will train models on supercomputers **worth over one hundred billion dollars**. Meta is burning over $16 billion a year on the "Metaverse", Google waste's $10 billions a year on a variety of projects that will never come to fruition. Amazon has lost over $50+ billion on Alexa. Cryptocurrencies wasted over $100 billion on nothing of value.

These firms and society in general can and will spend over one hundred billion on creating supercomputers that can train single massive model. These massive models can then be productized in a variety of ways. That effort will be duplicated in multiple counties and companies. It's the new space race. The difference between those prior wastes and now is that with AI there is tangible value that will come from the short term from human assistants and autonomous agents.

The much more important issue with scaling AI, the real AI brick wall, is **inference.** The goal is to decouple training compute from inference compute. This is why it makes sense to train well past Chinchilla optimal for any model that will be deployed. This is why you do sparse model architecture; every parameter is not activated during inference.

The real battle is that scaling out these models to users and agents costs far too much. The costs of inference exceed that of training by multiple folds. This is what OpenAI's innovation targets regarding model architecture and infrastructure.

Inference of large models is a multi-variable problem in which model size kills you for dense models. We have discussed this regarding the edge in detail here, but the problem statement is very similar for datacenter. The quick rundown is that devices can never have enough memory bandwidth for large language models to achieve certain levels of throughput. Even if they have enough bandwidth, utilization of hardware compute resources on the edge will be abysmal.

On Device AI – Double-Edged Sword

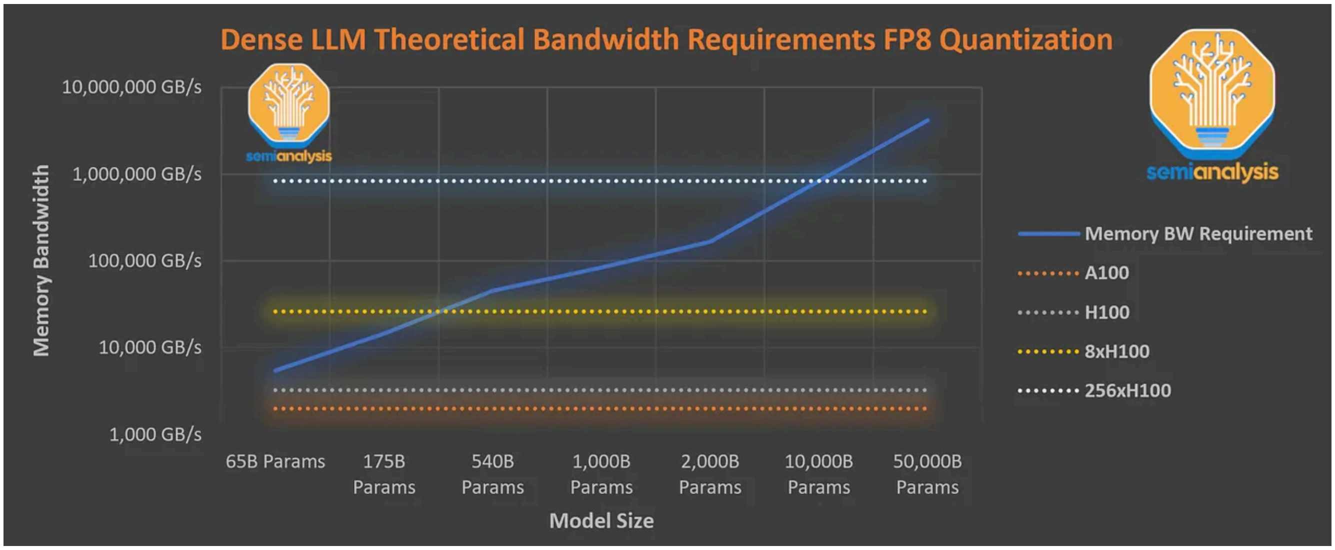
DYLAN PATEL AND SOPHIA WISDOM · MAY 13, 2023

Read full story →

In the datacenter, in the cloud, utilization rates are everything. Half the reason Nvidia is lauded for software excellence is because over a GPU's generations lifespan, Nvidia is constantly updating low level software that pushes FLOPS utilization rates up with smarter movement of data around a chip, between chips, and memory.

LLM inference in most current use cases is to operate as a live assistant, meaning it must achieve throughput that is high enough that users can actually use it. Humans on average read at ~250 words per minute but some reach as high as ~1,000 words per minute. This means you need to output at least 8.33 tokens per second, but more like 33.33 tokens per second to cover all corner cases.

A trillion-parameter dense model mathematically cannot achieve this throughput on even the newest Nvidia H100 GPU servers due to memory bandwidth requirements. Every generated token requires every parameter to be loaded onto the chip from memory. That generated token is then fed into the prompt and the next token is generated. Furthermore, additional bandwidth is required for streaming in the KV cache for the attention mechanism.



This chart assumes inefficiencies from not being able to fuse every op, memory bandwidth required for the attention mechanism, and hardware overhead are equivalent to parameter reads. In reality, even with "optimized" libraries such as Nvidia's FasterTransformer library, the total overhead is even larger.

The chart above demonstrates the memory bandwidth required to inference an LLM at high enough throughput to serve an individual user. It shows that even 8x H100 cannot serve a 1 trillion parameter dense model at 33.33 tokens per second. Furthermore, the FLOPS utilization rate of the 8xH100's at 20 tokens per second would still be under 5%, resulting is horribly high inference costs. Effectively there is an inference constraint around ~300 billion feed-forward parameters for an 8-way tensor parallel H100 system today.

Yet OpenAI is achieving human reading speed, with A100s, with a model larger than 1 trillion parameters, and they are offering it broadly at a low price of only $0.06 per 1,000 tokens. That's because it is sparse, IE not every parameter is used.

Enough waffling about, let’s talk about GPT-4 model architecture, training infrastructure, inference infrastructure, parameter count, training dataset composition, token count, layer count, parallelism strategies, multi-modal vision encoder, the thought process behind different engineering tradeoffs, unique implemented techniques, and how they alleviated some of their biggest bottlenecks related to inference of gigantic models.

Model Architecture

← Previous Next →

A guest post by

Gerald Wong

Call me Howie