

---

# INFERENCE ECONOMICS OF LANGUAGE MODELS

---

Ege Erdil  
Epoch  
ege@epoch.ai

## ABSTRACT

We develop a theoretical model that addresses the economic trade-off between cost per token versus serial token generation speed when deploying LLMs for inference at scale. Our model takes into account arithmetic, memory bandwidth, network bandwidth and latency constraints; and optimizes over different parallelism setups and batch sizes to find the ones that optimize serial inference speed at a given cost per token. We use the model to compute Pareto frontiers of serial speed versus cost per token for popular language models.

## 1 Introduction

Large language models (LLMs) have significantly influenced the field of artificial intelligence, becoming increasingly prevalent in both research and practical applications. As these models have grown and become used in a wide range of applications, their deployment for real-world tasks has precipitated a notable shift in the economics of AI. The cumulative inference costs—the total computational expenses associated with running these models across all instances—have escalated to levels comparable to, or in some cases exceeding, the initial training costs. Concurrently, a competitive market has emerged, with numerous API providers striving to offer open-source LLMs at optimal price points and processing speeds. Despite these developments, there remains a critical gap in our understanding: a comprehensive, quantitative analysis of the inference costs for LLMs when deployed at scale has yet to be conducted.

This is more than a theoretical problem. In practice, it's crucial for engineers working on serving LLMs for inference cheaply and quickly to have a sense of the limits of what's achievable with their hardware under ideal conditions. For kernel programmers who implement single-GPU kernels in CUDA, the roofline model of taking a minimum over arithmetic and memory read-write times serves this purpose: if a kernel is failing to achieve good performance relative to this baseline, we can often assume its implementation is inefficient and a theoretical best kernel should be able to do better.

Past work on how LLMs can be served efficiently for inference (Pope et al. 2022, Steinhardt 2022) is either "too practical", focusing strictly on kernel performance without comparing it to some theoretical best result, or "too theoretical", abstracting too far from concrete hardware details and therefore not providing useful guidance for real-world inference economics. In short, there's no equivalent to the "roofline model" in the literature about LLM inference right now, though we suspect AI labs and inference providers have their own internal models which they haven't released to the public.

Here are two questions, one very concrete and one more theoretical, that are difficult to answer without such a model:

- Suppose an open-source API provider succeeds in serving 16-bit Llama 3 70B on a DGX H100 machine at short contexts with a serial speed of 70 tokens/second and a hardware utilization rate of 15%. Is this performance good or bad, and is it reasonable to hope we might do better, i.e. increase either serial speed or hardware utilization without decreasing the other?
- Past theoretical work on LLM inference, e.g. Steinhardt 2022, has observed that because individual matrices are two-dimensional while matrix multiplications are three-dimensional, parallelizing a matrix multiplication by a factor of 2 along all its dimensions only increases required memory and network communication for parallelism by a factor of 2. Therefore, even when memory and network reads become binding, we should expect to be able to quadruple the serial speed of matrix multiplications (and thus of LLM inference) for each

halving of hardware utilization by simply parallelizing inference across eight times the number of GPUs. If so, why are LLMs typically not served at much faster speeds?

Our goal is to fill this gap by building a theoretical model that can take an LLM architecture and a GPU as inputs and determine how costly it is to serve this model at a particular context length and a particular rate of tokens per second per request if an optimal inference setup is used. We will then evaluate our model by comparing its predictions against the empirical performance figures for various API providers from using data on AI language models and API providers.

The paper is structured as follows: in Section 2 we first motivate the ingredients that will go into our complete model by analyzing some of the factors determining inference economics in isolation. This section is useful to build intuition for how inference economics works and to motivate why we choose to include some components in the model while excluding others. In Section 3 we give a complete description of our model with all of its parts before using the model to make predictions about well-known models such as DeepSeek-V3 (DeepSeek-AI et al. 2025). Finally, we have a takeaways section in which we list some of the important facts about inference economics revealed by our analysis.

We will assume familiarity with basic features of the decoder-only Transformer architecture and of mixture-of-experts models, but not with the technical details of distributed inference methods.

The code needed to reproduce the results in this paper can be found in [this GitHub repository](#).

## 2 Motivation for the model

### 2.1 Simplest case: single device instances, short context inference, dense model

The simplest case for thinking about inference economics is the case of a dense model where we’re doing short context length inference and each request is routed only to a single device. This removes most of the complications that have to be included in a general model: there’s no possibility of using complex distributed inference methods other than just routing requests to different instances, and the attention computations, along with KV cache reads, are assumed to be negligible because of the short context assumption.

In this case, inference economics of a model are determined by its total number of parameters  $N_{\text{param}}$ , the precision of the model parameters  $p$ , the FLOP per second of the GPU  $C$  and the HBM bandwidth of the GPU  $B$ . If we assume a batch size equal to  $b$  requests, then we will obtain the following:

$$\text{token latency} = \max \left( \frac{p \cdot N_{\text{param}}}{B}, \frac{(2 \text{ FLOP}) \cdot N_{\text{param}} \cdot b}{C} \right) \quad (1)$$

$$\text{GPU seconds per token} = \frac{\text{token latency}}{b} = \max \left( \frac{p \cdot N_{\text{param}}}{b \cdot B}, \frac{(2 \text{ FLOP}) \cdot N_{\text{param}}}{C} \right) \quad (2)$$

where in both equations, the first term in the maximum measures the time taken for parameter reads while the second term measures the time taken for arithmetic. We take a maximum assuming that memory reads and arithmetic are overlapped whenever possible.

In this simple case, the batch size  $b$  is the only variable parameter affecting the inference economics. The optimal batch size, denoted as  $b^*$ , is given by the equation  $b^* = p \cdot C / (B \cdot 2 \text{ FLOP})$ . At this batch size, the time taken for parameter reads and the time taken for arithmetic computations are equal. Consequently,  $b^*$  represents the maximum batch size that allows for the minimum token latency of  $p \cdot N_{\text{param}} / B$  to be achieved.

When not limited by an insufficient number of incoming requests, it is always optimal to set the batch size  $b$  equal to  $b^*$ . If the number of incoming requests exceeds  $b^*$ , it is more efficient to distribute the requests across multiple instances rather than increasing the batch size on a single instance. This is because increasing the batch size beyond  $b^*$  results in higher token latency without improving the cost in terms of GPU seconds per token. Conversely, reducing the batch size below  $b^*$  leads to increased cost without any improvement in token latency.

In situations where the number of incoming requests is insufficient to achieve a batch size of  $b^*$  while maintaining a token latency of  $p \cdot N_{\text{param}} / B$  on a single device, the optimal approach is to set the batch size  $b$  as close to  $b^*$  as possible given the available requests.

### 2.2 Multiple device instances with infinite network bandwidth, short context inference, dense model

In this section, we introduce an additional free parameter, the instance size ( $N_{\text{GPU}}$ ), which represents the number of GPUs utilized in the system. To simplify the analysis, we assume a network with nonzero latency but infinite bandwidth, allowing us to focus on the effects of distributing the workload across multiple GPUs without considering the impact of the number of words being reduced across different GPUs.

We'll assume that we use tensor parallelism across all of the GPUs in one instance: data parallelism is already taken into account because of the batch size being a free parameter, and in most cases pipeline parallelism is going to be inferior to data parallelism and so we can assume that it is not being used. (The only case where this is not true is when memory constraints are binding for otherwise efficient inference setups, which is rarely the case on NVIDIA hardware but is much more relevant when using specialized hardware such as Groq LPUs with very small amounts of memory per chip, as we'll see later.)

Utilizing multiple GPUs enables the distribution of arithmetic operations and parameter reads, potentially reducing the time spent on these tasks. However, this comes at the cost of introducing network latency due to the necessity of performing all-reduce operations in each layer. The token latency now depends on several factors:

- The number of model layers ( $n_{\text{layers}}$ )
- The number of sequential all-reduce operations required per layer ( $n_{\text{reduce}}$ )
- The average latency of performing an all-reduce operation ( $t_{\text{reduce}}$ ).

Considering these variables, a simple model for token latency and GPU seconds per token can be expressed as follows:

$$\text{token latency} = n_{\text{layers}} \cdot n_{\text{reduce}} \cdot t_{\text{reduce}} + \max \left( \frac{p \cdot N_{\text{param}}}{N_{\text{GPU}} \cdot B}, \frac{(2 \text{ FLOP}) \cdot N_{\text{param}} \cdot b}{N_{\text{GPU}} \cdot C} \right) \quad (3)$$

$$\text{GPU seconds per token} = \frac{N_{\text{GPU}} \cdot n_{\text{layers}}}{b} \cdot n_{\text{reduce}} \cdot t_{\text{reduce}} + \max \left( \frac{p \cdot N_{\text{param}}}{b \cdot B}, \frac{(2 \text{ FLOP}) \cdot N_{\text{param}}}{C} \right) \quad (4)$$

Nevertheless, this simple model has several shortcomings:

1. We assume that memory reads cannot be overlapped with network latency: devices are just waiting to receive information from the previous stage in order to start reading information from memory for the current stage. However, it's possible that devices can load some fraction of the weights from HBM into the L2 cache in advance. This fraction is usually negligible for a single device because the L2 cache is usually quite small, but with multiple devices this strategy can actually be quite effective.<sup>1</sup>
2. We assume that  $t_{\text{reduce}}$  is constant, but in general,  $t_{\text{reduce}}$  is an increasing function of the instance size and also depends on which network topology is used to link the GPUs together. For instance, GPUs linked in a ring topology will have poor all-reduce latency at large instance sizes compared to those linked in a tree topology because each all-reduce will need to travel around the ring twice, once for the reduce-scatter and another time for the all-gather operation. The scaling of this nonzero network latency with the instance size  $N_{\text{GPU}}$  turns out to be the key factor that stops us from indefinitely quadrupling speed for each halving of hardware utilization in the  $N_{\text{GPU}} \rightarrow \infty$  limit.

We will ignore point (1) in the rest of this work for simplicity and focus only on addressing point (2), which is far more quantitatively significant.

### 2.2.1 Intra-node and network latency

The latency of an all-reduce operation depends on two factors: the communication latency between GPUs and the number of communications required. For instance, if we have a ring topology of size  $R$  with a maximum communication latency of  $t_{\text{comm}}$  across two adjacent GPUs in the ring, the theoretical ring all-reduce latency will be  $2(R - 1)t_{\text{comm}}$  because we need a total of  $R - 1$  hops in the reduce-scatter phase and another  $R - 1$  hops in the all-gather phase.

Unfortunately,  $t_{\text{comm}}$  is in practice not a simple constant but depends on many details: the network topology, the hardware used for communication and the exact communication protocol, as well as whether communication can be confined to a single node or must occur across node boundaries. For now, we will neglect these details and assume that  $t_{\text{comm}} \approx 1 \mu\text{s}$  on the basis of information from Jeaugey 2019, assuming that we use the lowest latency communication protocol available in NCCL (which is appropriately called "LL"). This is about the best that can be done unless we resort to using the more detailed latency calculations from the NCCL codebase.<sup>2</sup> We leave this work for the next section.

---

<sup>1</sup>For example, Llama 2 70B (Touvron et al. 2023) has 70 billion parameters and 80 layers at 16-bit precision, so each layer is around 1.75 GB in size. With 50 MB of L2 cache per GPU, we only need 35 GPUs to fit the parameters in each layer into L2. An instance with more GPUs than this could therefore overlap most of the network latency with parameter reads from HBM.

<sup>2</sup>NVIDIA provides their own model for estimating the latency of GPU primitives at the official GitHub repository for NCCL (NVIDIA 2024) in the file "src/graph/tuning.cc". The variable containing the hardware latency information that we need is called "hwLat" and has latency figures in units of microseconds.

We now turn to characterizing the number of participants  $R$ : in general, when  $N_{\text{GPU}}$  is large the number of participants in each all-reduce operation in a forward pass scales as  $R \propto N_{\text{GPU}}^{1/2}$ . This is because it's optimal for both memory bandwidth and network bandwidth optimization to slice weight matrices close to evenly along both dimensions in the large  $N_{\text{GPU}}$  regime. Because each all-reduce happens across the intermediate dimension  $K$  of a matrix multiplication of type  $M \times K \times N$ , the number of participating ranks is the number of ways the  $K$  dimension is sliced, which will scale proportionally to  $N_{\text{GPU}}^{1/2}$ .

Combining all of this information, we might approximate  $t_{\text{reduce}} \approx 2 \mu\text{s} \cdot (N_{\text{GPU}}^{1/2} - 1)$  for values of  $N_{\text{GPU}}$  that are not too large (because with a sufficiently large instance size we would switch to using e.g. a tree topology, which would have lower latency) and at small enough batch sizes that network bandwidth constraints are not binding even when the LL algorithm is used. This expression is enough to imply an interesting fact: there is an *optimal instance size* to be serving any model at if we want to minimize token latency *even if we do not care about cost*. This is readily apparent from the expression for token latency that we've obtained:

$$\text{token latency} = n_{\text{layers}} \cdot n_{\text{reduce}} \cdot 2 \mu\text{s} \cdot (N_{\text{GPU}}^{1/2} - 1) + \max \left( \frac{p \cdot N_{\text{param}}}{N_{\text{GPU}} \cdot B}, \frac{(2 \text{ FLOP}) \cdot N_{\text{param}} \cdot b}{N_{\text{GPU}} \cdot C} \right) \quad (5)$$

The only parameter we have not characterized so far is the number of sequential all-reduces per layer  $n_{\text{reduce}}$ . When  $N_{\text{GPU}}$  is large, this will be equal to the number of serial matrix multiplications in each layer<sup>3</sup>. In a traditional Transformer operating at short context lengths, the attention mechanism itself is negligible, so we're left with  $n_{\text{reduce}} = 4$  sequential matrix multiplications:

1. The computation of the query, key and value vectors.
2. The post-attention projection.
3. The  $d_{\text{model}} \rightarrow d_{\text{ff}}$  matrix multiplications in the feedforward block. Note that there could be more than one of these in an activation scheme such as SwiGLU (Shazeer 2020), but they can in principle be parallelized and so don't count as multiple serial operations.
4. The  $d_{\text{ff}} \rightarrow d_{\text{model}}$  matrix multiplications in the feedforward block.

While this particular arrangement of a single layer remains common today, it's not the only available option. Some architectures such as the PaLM models from Chowdhery et al. 2022 may opt to use *parallel attention*, which changes the formulation of a Transformer layer (ignoring the layer normalization operations etc.) from Equation 6 to Equation 7.

$$\text{Sequential attention: } x_{\text{output}} = x_{\text{input}} + \text{Feedforward}(\text{Attention}(x_{\text{input}})) \quad (6)$$

$$\text{Parallel attention: } x_{\text{output}} = x_{\text{input}} + \text{Feedforward}(x_{\text{input}}) + \text{Attention}(x_{\text{input}}) \quad (7)$$

The advantage of this change is that it removes the logical dependence of the feedforward layer outputs on the attention layer outputs, allowing the two layers to be executed in parallel. This can potentially reduce  $n_{\text{reduce}}$  from 4 to 2, which is a minor gain during training but a significant gain during inference due to the tighter latency constraints there.

### 2.2.2 Predictions

This model is already rich enough to make useful predictions about the token economics of dense models. To do this, we compute the token latency and the GPU seconds per token across a grid of values for the batch size  $b$  and the instance size  $N_{\text{GPU}}$ , impose an overall throughput constraint

$$b / (\text{token latency}) \leq \text{token throughput}$$

to account for the possibility of limited demand, convert the cost from units of GPU-seconds to units of dollars by multiplying by an hourly rental rate per GPU, and plot the implied Pareto frontiers for the per-request token generation rate on the x-axis versus the cost per million tokens on the y-axis.

---

<sup>3</sup>There are special cases in which this number is less: for example, when  $N_{\text{GPU}}$  is small we can use 1D tensor parallelism in the feedforward block and avoid the all-reduce over the  $d_{\text{ff}}$  dimension, and we might also be able to afford partitioning attention over heads instead of over batch and avoid unnecessary all-to-all communication in that step as well.  $n_{\text{reduce}} = 4$  assumes we're not in the regime where these optimizations would apply.

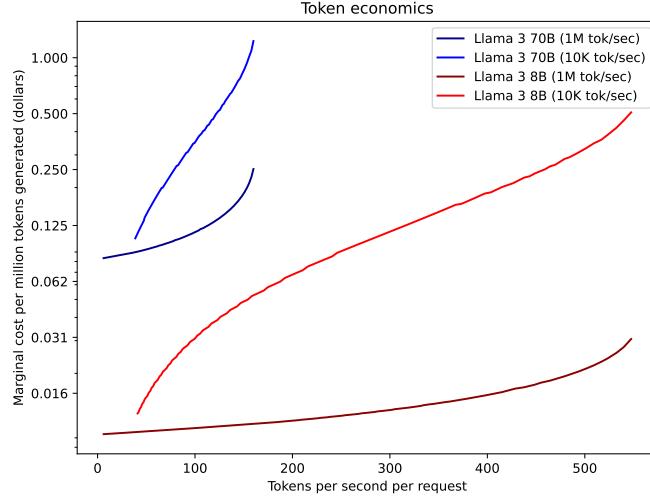


Figure 1: The token economics for Llama 3 8B and Llama 3 70B implied by our current model. We assume the H100 GPU is used at an hourly rental rate of \$2/hour per device. In the legend, "tok/sec" refers to the *total demand* for the model across all users and not the token generation rate seen by a single user. In contrast, the x-axis refers to the rate seen by a single user after they submit a request and not the overall throughput.

The model in this section is crude, neglecting many important details, but as Figure 1 shows it can already roughly match empirical data. If we exclude Groq (since they use their own custom SRAM-based hardware for inference instead of NVIDIA GPUs), [Artificial Analysis 2024](#) reports a maximum inference speed of  $\approx 400$  tokens/second at 20 cents per million tokens for Llama 3 8B and  $\approx 150$  tokens/second at 90 cents per million tokens for Llama 3 70B, both achieved by Fireworks. In both cases, the models appear to be served within a factor of 2 of the maximum tokens per second and markups charged by API providers over just the GPU rental costs appear to be less than a factor of 2.

### 2.2.3 Theoretical analysis

Though more complex than the model from Section 2.1, this model is still simple enough for us to obtain analytic results about it. We repeat the token latency equation with the appropriate substitutions here for completeness:

$$\text{token latency} = n_{\text{layers}} \cdot n_{\text{reduce}} \cdot t_{\text{hop}} \cdot 2(N_{\text{GPU}}^{1/2} - 1) + \max \left( \frac{p \cdot N_{\text{param}}}{N_{\text{GPU}} \cdot B}, \frac{(2 \text{ FLOP}) \cdot N_{\text{param}} \cdot b}{N_{\text{GPU}} \cdot C} \right) \quad (8)$$

The goal of an API provider is to choose  $N_{\text{GPU}}$ ,  $b$  to minimize token latency for a given per-token cost  $c$ , which can be expressed as token latency  $\cdot N_{\text{GPU}}/b = c$ .

We first eliminate one of the variables, namely  $b$ . As in Section 2.1, the efficient unconstrained choice for  $b$  is always  $b^* = p \cdot C / (B \cdot 2 \text{ FLOP})$ . This makes the memory bandwidth and arithmetic terms inside the maximum equal to one another and it's easy to see it achieves both the smallest possible cost and the smallest possible token latency for any fixed value of  $N_{\text{GPU}}$ . If we assume  $b = b^*$ , we can simplify the expression for token latency to

$$\text{token latency} = n_{\text{layers}} \cdot n_{\text{reduce}} \cdot t_{\text{hop}} \cdot 2(N_{\text{GPU}}^{1/2} - 1) + \frac{p \cdot N_{\text{param}}}{N_{\text{GPU}} \cdot B} \quad (9)$$

We now turn our attention to a question which turns out to have a clean analytical answer: *minimum possible token latency*. Because Equation 9 diverges as  $N_{\text{GPU}} \rightarrow \infty$  and is a continuous function of  $N_{\text{GPU}}$ , we know it must have a

global minimum. To find this minimum, we differentiate with respect to  $N_{\text{GPU}}$  and set the result to zero to get

$$n_{\text{layers}} \cdot n_{\text{reduce}} \cdot t_{\text{hop}} \cdot N_{\text{GPU}}^{-1/2} = \frac{p \cdot N_{\text{param}}}{N_{\text{GPU}}^2 \cdot B} \quad (10)$$

$$N_{\text{GPU}}^{3/2} = \frac{p \cdot N_{\text{param}}}{n_{\text{layers}} \cdot n_{\text{reduce}} \cdot t_{\text{hop}} \cdot B} \quad (11)$$

$$N_{\text{GPU}} = \left( \frac{p \cdot N_{\text{param}}}{n_{\text{layers}} \cdot n_{\text{reduce}} \cdot t_{\text{hop}} \cdot B} \right)^{2/3} \quad (12)$$

One caveat is that we might have that this optimal value of  $N_{\text{GPU}}$  is less than 1, which is going to lead to an unrealistic negative value for the latency term. In this case, the optimal solution is always  $N_{\text{GPU}} = 1$ , as using only a fraction of a single GPU would only increase memory read time without having any network latency benefit. Strictly speaking, therefore, the optimal solution is

$$N_{\text{GPU}}^* = \max \left( \frac{p \cdot N_{\text{param}}}{n_{\text{layers}} \cdot n_{\text{reduce}} \cdot t_{\text{hop}} \cdot B}, 1 \right)^{2/3} \quad (13)$$

Plugging this into the token latency expression now gives

$$\text{minimum token latency} = \begin{cases} 3 \cdot (n_{\text{layers}} n_{\text{reduce}} t_{\text{hop}})^{2/3} \cdot (p N_{\text{param}} / B)^{1/3} - 2 n_{\text{layers}} n_{\text{reduce}} t_{\text{hop}} & \text{if } N_{\text{GPU}}^* > 1 \\ p \cdot N_{\text{param}} / B & \text{otherwise} \end{cases} \quad (14)$$

The  $N_{\text{GPU}}^* > 1$  condition is comparing two different timescales: the amount of time it takes to read the parameters of the model from HBM *on a single GPU* and the amount of time spent for each all-gather or reduce-scatter operation per participant in a single forward pass. If the memory read time is greater, then it's optimal to use an instance of multiple GPUs; otherwise it's not.

We can derive another interesting result in the  $N_{\text{GPU}}^* \gg 1$  regime about the Pareto frontiers of cost versus speed. The cost (in units of GPU seconds) of achieving the minimum possible latency is given by

$$\text{cost per token at minimum token latency} = \frac{N_{\text{GPU}}^*}{b^*} \cdot \text{minimum token latency} \quad (15)$$

$$\approx 3 \cdot \frac{p \cdot N_{\text{param}}}{B \cdot b^*} \text{ (approx. valid because } N_{\text{GPU}}^* \gg 1 \text{ by assumption)} \quad (16)$$

$$= 3 \cdot \frac{(2 \text{ FLOP}) \cdot N_{\text{param}}}{C} \quad (17)$$

In other words, achieving the *best possible token latency* only costs around three times as much as the minimum possible cost per token arising from arithmetic alone. This result is not realistic: in actual inference setups the Pareto frontier of speed versus cost can range across much wider values for cost per token. This toy model, however, is useful because it demonstrates this is entirely because of network bandwidth constraints. The need to economize on it provides an incentive to reduce the batch size beyond  $b^*$  to reduce the sizes of activation tensors, which potentially reduces arithmetic intensity and trades off against the efficient use of memory bandwidth. We will see how this works concretely in later sections when we develop a more complete model taking network bandwidth constraints into account.

Computing the maximum tokens per second for some well-known models running on H100 SXM GPUs using Equation 14 with  $t_{\text{hop}} = 1 \mu\text{s}$  and  $n_{\text{reduce}} = 4$  gives the results presented in Table 1.

It's worth emphasizing that we expect the tokens per second achieved in practice to be below the figures reported in Table 1 for many reasons:

- Even in theory, these are *maximum achievable* figures for tokens per second. In practice, an API provider might opt to provide a model at a higher token latency to be able to serve it more cheaply, especially if the extra serial speed is not useful for the typical use cases of the model being served. This is especially true if the provider faces an overall throughput constraint, as in Figure 1.

Model name	Maximum tokens per second	Optimal instance size (GPUs)
Llama 3 8B	966	11
Llama 3 70B	234	26
GPT-3	148	42
PaLM-540B	86	79
GPT-4 (architecture from Patel and Wong 2023)	56	173

Table 1: Maximum tokens per second and the optimal instance size required to achieve the maximum according to Equation 14 with  $t_{\text{hop}} = 1 \mu\text{s}$ ,  $n_{\text{reduce}} = 4$  and using H100 SXM GPUs.

2. Achieving results that are this good requires being able to eliminate the base latency in GPU collectives such as all-reduce and all-to-all. If this is not feasible, the maximum tokens per second will be less, especially for small models such as Llama 3 8B.
3. Our analysis is based strictly on short-context inference in which the attention mechanism is assumed to be negligible, both for arithmetic and memory reads. We'll see later on how the figures for long-context inference can be significantly worse, especially if multi-query attention (Shazeer 2019) is not used.
4. We assume that  $N_{\text{GPU}}$  takes values in the real numbers, while in practice it not only takes integer values but must satisfy certain divisibility relations to ensure e.g. tensor parallelism can be effectively used during inference.
5. The approximation that  $R \approx N_{\text{GPU}}^{1/2}$  is mildly optimistic: in practice some all-reduces will involve more than  $N_{\text{GPU}}^{1/2}$  ranks while some will involve less, and due to the convexity involved this can slightly raise the token latency we expect.

Table 1 also shows that even for an 8 billion parameter model, we have  $N_{\text{GPU}}^* \gg 1$ , and the optimal instance size increases as the size of the models grow. This is because  $N_{\text{param}} = O(d_{\text{model}}^2 \cdot n_{\text{layers}})$  typically grows much faster than  $n_{\text{layers}}$ , for large models we have  $pN_{\text{param}}/B \gg n_{\text{layers}}n_{\text{reduce}}t_{\text{hop}}$  and the optimal instance size for inference will be  $N_{\text{GPU}} \gg 1$  (this is disregarding the HBM size of a single device, which in practice is also a constraint forcing us to use a bigger instance size when we're working with large models).

In this typical regime when  $N_{\text{param}}$  is large, from equation 14 we know that a good approximation to the minimum token latency is simply

$$\text{minimum token latency} \approx 3 \cdot (n_{\text{layers}}n_{\text{reduce}}t_{\text{hop}})^{2/3} \cdot (p \cdot N_{\text{param}}/B)^{1/3} \quad (18)$$

Equation 18 is quite useful for understanding the factors that slow inference down and their relative importance. The number of layers of a model matters more than its width for how quickly one can inference the model: as  $N_{\text{param}} = O(d_{\text{model}}^2 \cdot n_{\text{layers}})$ , the minimum token latency scales as  $\propto n_{\text{layers}} \cdot d_{\text{model}}^{2/3}$  in the model depth and width, holding hardware features constant. In addition, increasing memory bandwidth also matters less than reducing the latency of GPU collective primitives such as all-reduce and all-to-all. Contrary to the impressions one might get from hobbyist single-device inference, the bigger constraint on inference at scale is not memory bandwidth but intra-node and network latency.

An even simpler approximation can be obtained by assuming that the number of layers  $L$  in a dense model scales as  $L \propto N_{\text{param}}^{1/4}$ , a scaling relationship that holds approximately for the dense models trained by Hoffmann et al. 2022, for example. In this case, the scaling in number of parameters holding all other variables fixed simplifies to

$$\text{token-to-token latency} \propto \sqrt{N_{\text{param}}} \quad (19)$$

This simple rule of thumb agrees with empirical data about dense models surprisingly well. For example, Pope et al. 2022 obtain best latency figures of 4 ms, 13 ms and 32 ms per token for PaLM 8B, 62B and 540B respectively, as can be seen in Figure 2, which agree with square root scaling of token latency in parameter count almost exactly.

We see a similar agreement when looking at more recent data on NVIDIA hardware. Taking Llama 3 8B at 400 tokens/second as a baseline, the square root model predicts 135 tokens/second for a 70B dense model and 56 tokens/second for a 405B dense model, which according to Artificial Analysis 2024 are consistent with the speeds at which leading

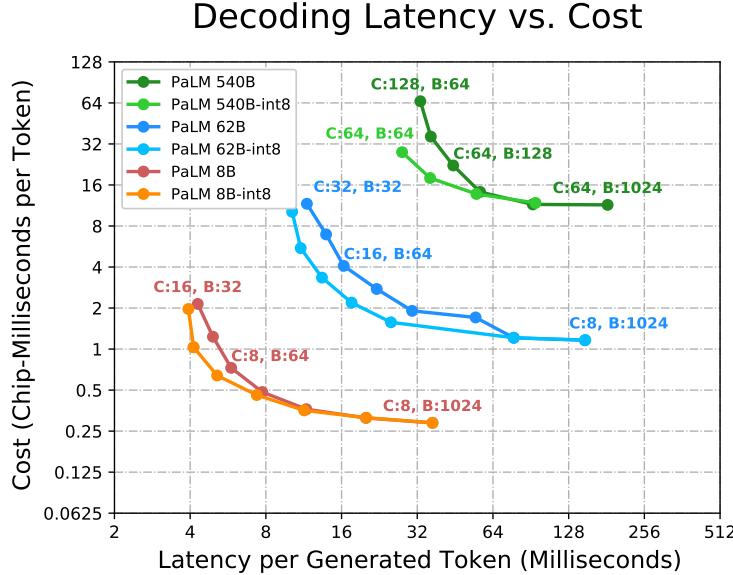


Figure 2: The cost per token versus token latency Pareto frontiers for PaLM 8B, 62B and 540B obtained by Pope et al. 2022 on TPU v4 clusters. The best possible latencies for each model size approximately match the square root scaling predicted by our toy model.

providers on NVIDIA hardware have been able to serve these models. Overall, the toy model performs remarkably well given how simple it is and how many important effects it neglects.

### 3 Model description

We now build up the complete model we'll use in the rest of the paper on top of the more simplified model from Section 2.2. Here are the improvements we'll make:

1. Take attention operations into account so that we can use the model to make useful predictions about long-context inference. This means factoring in both the arithmetic side (the quadratic complexity) and the memory bandwidth side (KV cache reads and storage).
2. Relax the assumption that the model must be a dense model so that we can handle mixture-of-experts models.
3. Use the more precise all-reduce and all-to-all latency approximations from NVIDIA 2024 instead of the rough approximation we used earlier.
4. Drop the assumption that network bandwidth is infinite and take network bandwidth constraints into account alongside network latency.
5. Instead of modeling only tensor and data parallelism, also allow for pipeline and expert parallelism.

We'll do these in order in the next five subsections before presenting the final equations of our model in a self-contained format.

#### 3.1 Handling the attention operations

There are three points we must take into account when it comes to attention operations of a vanilla, decoder-only Transformer:

1. The arithmetic required for taking the inner products of the query and key vectors.
2. The memory bandwidth and memory capacity required for storing the KV cache.
3. The fact that some extra GPU communication might be needed if the attention operations are substantial enough to require being distributed across multiple devices.

The calculations become somewhat more complex for more recent innovations such as the *multi-head latent attention* (MLA) mechanism from DeepSeek-AI et al. 2025. We will address this specific variant of attention after covering the

vanilla case, though it's difficult to give a comprehensive review of all the ways in which the attention mechanism may be modified and all the ways these modifications would alter the calculations in this section.

To handle the arithmetic required for the attention operations, we add  $4n_{\text{layers}}n_{\text{head}}d_{\text{head}}\ell_{\text{input}}b$  FLOP of arithmetic operations that must be completed on top of the weight multiply-accumulates that give  $\approx 2N_{\text{param}}$  FLOP of arithmetic per token. Here,  $d_{\text{head}}$ ,  $n_{\text{head}}$  stand for the dimension of a single attention head and the number of total attention heads,  $\ell_{\text{input}}$  stands for input length, and  $b$  is the batch size.

The size of the KV cache is in general given by  $n_{\text{kv head}}d_{\text{head}}n_{\text{layers}}\ell_{\text{input}}b \cdot p$  where  $n_{\text{kv head}}$  is the number of distinct key and value heads (so double-counting each head to count both key and value vectors). As before,  $p$  is the precision of the model parameters. We add this on top of the parameter reads from HBM that we're already taking into account. If we're willing to accept performance degradation, we can also compress the KV cache using various lossy methods, but for our model we will assume these methods are not used.

For the final point, the exact amount of extra communication needed will in general depend on the partitioning strategy used to divide the attention computation across multiple GPUs. For instance, Pope et al. 2022 notes that partitioning the attention operation across the batch dimension incurs a cost of two additional all-to-all operations. However, the picture is complicated by the fact that when the KV cache is sufficiently small, one can instead partition the attention operations across the head dimension (i.e. assigning different heads to different GPUs) which does not incur the aforementioned extra communication cost. Because an accurate model of this would be rather complicated, in the remainder of this article we will ignore the extra communication cost arising from the attention operations.

Now, let's turn to addressing the case of multi-head latent attention. For our purposes, this variant of the attention mechanism is essentially equivalent to grouped-query attention except with increased arithmetic cost during decoding. Explicitly, the equivalence may be seen as follows:

1. Multi-head latent attention uses a low-rank factorization of the key and value projections and caches the intermediate activations, also called "latent vectors", as a way of achieving KV cache compression. During decoding, one of the low-rank matrices appearing in this factorization are absorbed into the query projection (for the keys) and into the post-attention projection (for the values) respectively.

The net effect is that the effective head dimensionality for computing the arithmetic cost of attention goes up from  $d_{\text{head}}$  to  $d_{\text{latent}}$ , which is typically meaningfully larger. For instance, DeepSeek-V3 (DeepSeek-AI et al. 2025) uses  $d_{\text{head}} = 128$  and  $d_{\text{latent}} = 512$ , raising the arithmetic cost of attention by a factor of 4 relative to a vanilla Transformer.

2. In return for this increased arithmetic cost, the KV cache is compressed significantly: instead of having to cache key and value vectors each of size  $n_{\text{head}}d_{\text{head}}$  per input position, the vectors that need to be cached are just the latents, which are only of size  $d_{\text{latent}}$  each. Since  $d_{\text{head}} \ll d_{\text{latent}} \ll n_{\text{head}}d_{\text{head}}$  in typical MLA implementations, this represents a huge reduction in the size of the KV cache even though it raises the arithmetic cost of attention.

As we will later see, this memory IO versus arithmetic tradeoff makes MLA models unusual among Transformers in that their inference is prone to be arithmetic bound, rather than memory IO bound, in long contexts where attention dominates the cost of inference.

### 3.2 Generalizing to mixture-of-experts models

To generalize to mixture-of-experts models, we need to make appropriate changes to the arithmetic cost and memory reads necessary for a forward pass. First, to handle the arithmetic cost of a mixture-of-experts forward pass, we decompose the total number of parameters (excluding embedding parameters) of the model as follows:

$$N_{\text{param}} = N_{\text{attention params}} + N_{\text{feedforward params}} + N_{\text{unembedding params}} \quad (20)$$

For a dense model, the arithmetic cost of a forward pass is simply  $2N_{\text{param}}$  FLOP per batch element. However, for a mixture-of-experts model with a sparsity factor of  $s = n_{\text{expert}}/n_{\text{active expert}}$  (where  $n_{\text{expert}}$  is the total number of experts and  $n_{\text{active expert}}$  is the number of experts that are active for any given token), the arithmetic cost is instead  $2N_{\text{active param}}$  FLOP per batch element, where

$$N_{\text{active param}} = N_{\text{attention params}} + \frac{N_{\text{feedforward params}}}{s} + N_{\text{unembedding params}} \quad (21)$$

We only divide the number of feedforward parameters by  $s$  because the attention and unembedding parameters are shared across experts. This gives us the number of FLOP required for a forward pass.

We now move on to handle the memory bandwidth cost. Unlike for a dense model, we might not necessarily read all of the parameters of a mixture-of-experts model from memory in a given forward pass, because the batch might be small enough for some parameters to remain inactive across the entire batch. If we assume that expert activations are independently and uniformly distributed<sup>4</sup>, then the chance that any given expert will be inactive for any specific batch element is  $1 - 1/s$ , and the chance that it will be inactive for *all* batch elements is  $(1 - 1/s)^b$  where  $b$  is the batch size as before.

Therefore, on average we expect the number of active experts to be  $n_{\text{expert}} \cdot (1 - (1 - 1/s)^b)$ , and therefore averaging across many batches, we expect the number of parameters read from memory to be

$$N_{\text{params read}} = N_{\text{attention params}} + \left(1 - \left(1 - \frac{1}{s}\right)^b\right) \cdot N_{\text{feedforward params}} + N_{\text{unembedding params}} \quad (22)$$

When  $b \gg s$ , we'll have  $(1 - 1/s)^b \approx 0$ , so in this case as expected we can approximate without much error that all of the model parameters will be read from memory during a single forward pass.

### 3.3 Improving the latency approximation

To make our latency estimates more rigorous, we use the information from the tuning code in the official NCCL repository at NVIDIA 2024. Based on the information in the file “src/graph/tuning.cc”, we can deduce the following approximate latency expression for an all-reduce when using a tree topology:

$$6.8 \mu\text{s} + (1.2 \mu\text{s}) \cdot \left(\frac{n_{\text{ranks}}}{n_{\text{nodes}}} - 1\right) + (10 \mu\text{s}) \cdot \log_2(n_{\text{nodes}}) \quad (23)$$

Here,  $n_{\text{ranks}}$  is the number of ranks (or GPUs) participating in the all-reduce, while  $n_{\text{nodes}}$  is the number of participating nodes. Of course, while this expression is quite precise, our modeling so far has only used information about the overall number of GPUs participating in an all-reduce, and we've not taken the distribution of those GPUs across different nodes into account. To do this, we'll simply use another approximation: if our inference instance has a total of  $N_{\text{GPU}}$  GPUs distributed across  $N_{\text{nodes}}$  nodes, we'll assume that each all-reduce involves  $\sqrt{N_{\text{GPU}}}$  GPUs across  $\sqrt{N_{\text{nodes}}}$  nodes. While this is still an approximation, it's significantly more accurate than the latency expression we used in Section 2.2.

### 3.4 Taking network communication time into account

Having taken network latency into account, we now move on to consider network bandwidth constraints. This won't affect the order of magnitude of our decoding speed estimates for flagship NVIDIA GPUs, because in practice network bandwidth time only dominates latency for batch sizes larger than the critical batch size on the H100. Specifically, for realistic per-GPU batch sizes of up to 100 and a model dimension of 10,000, network bandwidth limits can add

$$\frac{(2 \text{ bytes}) \cdot 10,000 \cdot 100 \cdot (8 - 1)}{8 \cdot 112 \text{ GB/s}} \approx 16 \mu\text{s} \quad (24)$$

of extra time to each all-reduce operation on a DGX H100 if LL, the lowest latency communication algorithm of NCCL with 112 GB/s of all-reduce algorithm bandwidth, is used. This is not small enough to be negligible, so we have to take it into account in a more realistic model.

To determine how much time is taken up by network communication, we need to answer the following questions:

1. How many bytes do we need to all-reduce?
2. How many nodes and ranks are participating in each all-reduce?
3. To what extent can the network communication be overlapped by memory reads or arithmetic?

---

<sup>4</sup>Note that this assumption would be violated for the "shared experts" of MoE variants such as DeepSeek-V3. We still stick to the independent and uniform routing case because it's the most natural one, but our setup would need modifications before being applied to MoEs for which this assumption is violated.

For the first question, assuming that we have an all-reduce after each matrix multiplication, we just need to add up the number of entries of each output matrix in a single layer and multiply this by the precision of the model’s parameters and the number of layers. This is relatively straightforward: the QKV matrix is of size  $(1 + 2/g)n_{\text{head}}d_{\text{head}}b$ , the post-attention projection result is of size  $d_{\text{model}}b$ , the first and second matrix multiplications in the feedforward block then contribute  $d_{\text{ff}}b$  and  $d_{\text{model}}b$  respectively. Here,  $g$  is the attention group size and  $b$  is the batch size. One caveat to keep in mind is that depending on whether the architecture uses an activation scheme such as SwiGLU, there might be more than one output matrix having the dimensions  $d_{\text{ff}} \times b$ .

Therefore, for a model with a typical Transformer architecture (i.e. no SwiGLU), the total number of bytes we need to all-reduce in a single forward pass using tensor parallelism is

$$\text{bytes reduced} = ((1 + 2/g)n_{\text{head}}d_{\text{head}} + 2d_{\text{model}} + d_{\text{ff}}) \cdot b \cdot n_{\text{layers}} \cdot p \quad (25)$$

where  $b$  is the batch size,  $n_{\text{layers}}$  is the number of layers, and  $p$  is the precision of the model parameters.

For (2), we’ve already answered this question in Section 3.3: the number of participating ranks is  $\sqrt{N_{\text{GPU}}}$  and the number of participating nodes is  $\sqrt{N_{\text{nodes}}}$ . Therefore, we can express the total amount of intra-node and inter-node communication that needs to happen as follows:

$$\text{inter-node network reads} = 2 \cdot (\sqrt{N_{\text{nodes}}} - 1) \cdot \text{bytes reduced} \quad (26)$$

$$\text{intra-node network reads} = 2 \cdot (\sqrt{N_{\text{GPU}}/N_{\text{nodes}}} - 1) \cdot N_{\text{nodes}} \cdot \text{bytes reduced} \quad (27)$$

The expressions here arise from the fact that reducing  $X$  bytes across  $R$  participants requires  $2X(R - 1)$  bytes to be read, in general. Treating each node as a single participant gives the first expression, and then treating each node independently and assuming that an all-reduce is performed across all the GPUs in each node gives the second expression.

Now, to obtain the time taken for the communication, we only need to divide these by the intra-node and inter-node communication bandwidth available, so we obtain the final expression

$$\text{network communication time} = \frac{\text{inter-node network reads}}{N_{\text{GPU}} \cdot \text{inter-node bandwidth}} + \frac{\text{intra-node network reads}}{N_{\text{GPU}} \cdot \text{intra-node bandwidth}} \quad (28)$$

For example, for the H100 the intra-node bandwidth would be 450 GB/s. This is half of the reported NVLink bandwidth of 900 GB/s because the reported bandwidth is not full-duplex: it takes both reads and writes into account, while our network reads expression is counting only network reads. Separately, the inter-node bandwidth for e.g. a cluster consisting of DGX H100 systems would be 50 GB/s, because each DGX H100 system comes together with 8 ConnectX-7 VPI cards, each of which supports 50 GB/s of full-duplex bandwidth. Since the number of GPUs per DGX H100 system is also equal to 8, we have one card per GPU, resulting in a network bandwidth of 50 GB/s per GPU.

For the final question about overlapping, we will conservatively assume that network communication cannot be overlapped with memory reads and arithmetic. This is generally going to be true when using off-the-shelf primitives from NCCL, and overlapping the network communication in an all-reduce with computation using simple methods (e.g. splitting the all-reduce into smaller pieces so that computation can begin on the pieces that are processed first while later pieces are still being reduced) seems to require taking the latency hit from the all-reduce multiple times. This is acceptable during training, but not during inference, because in inference the forward passes are required to be much faster than during training. This inability to split the all-reduce into smaller pieces without multiplying the latency per all-reduce is the reason behind our conservative assumption here.

A significant implication of taking network communication into account is that the critical batch size is no longer the optimal batch size choice in general. This is because unlike parameter reads that can be amortized over a bigger batch, network communication per batch scales linearly with the batch size. Fast inference often requires reducing this communication time by scaling down the batch size beyond the critical batch size as we will see later.

### 3.5 Allowing pipeline and expert parallelism

The model from Section 2.2 only took tensor and data parallelism into account. In this section, we discuss how we can incorporate pipeline and expert parallelism in a simple way into the basic model.

We address pipeline parallelism (PP) first. Ignoring PP is not a major issue in most cases, because if we have a pipeline with  $N_{\text{PP}}$  stages, we’ll generally need  $N_{\text{PP}}$  micro-batches in flight in order to avoid pipeline bubbles. This means

pipeline parallelism slices the batch into smaller pieces just like data parallelism (DP) does. When training a model there's a solid reason to do this: DP requires gradient all-reduces after each forward and backward pass, and using both PP and DP at the same time can economize on this bandwidth cost, essentially because the communication cost of both PP and DP scales linearly with their parallelism degree, while the overall cluster size (and total communication bandwidth available) scales with their product.

However, when we're doing inference, there is no backward pass and subsequent gradient update, which means DP requires no communication whatsoever past the initial routing of tokens to the right DP copies. This eliminates the central reason to use PP over DP in training, and means the only advantage of PP left in inference is that it enables sharding model weights across multiple devices and thereby help the model weights fit into GPU memory at the expense of introducing some communication overhead at pipeline boundaries.

While this can be useful in theory, on NVIDIA hardware efficient inference setups are generally not constrained by availability of memory. An easy way to see this for H100s is that reading the contents of 80 GB of HBM alone would take  $\approx 24$  ms at the theoretical peak HBM bandwidth of 3.3 TB/s, so the Pareto fronts of optimal cost versus speed must be identical for H100s with infinite memory and with 80 GB of memory past a speed of  $(24 \text{ ms/token})^{-1} \approx 41 \text{ tokens/sec}$ , and in practice even sooner once we take into account network bandwidth limits and more realistic sustained HBM bandwidth values.<sup>5</sup> As a result, there's generally no reason to use pipeline parallelism when data parallelism is available as an alternative.

We can model pipeline parallelism by lowering the effective batch size in our model to  $b_{\text{micro}} = b/N_{\text{PP}}$ , adding  $N_{\text{PP}} - 1$  sequential peer-to-peer communications of size  $d_{\text{model}} \cdot b_{\text{micro}}$  each for the processing of each token that can be parallelized across  $N_{\text{GPU}}/N_{\text{PP}}$  GPUs, and relaxing the memory capacity constraint to allow for weight and KV cache sharding across PP copies. We would then loop over a grid of values for possible PP degrees and see which one gives the lowest token latency for a fixed instance and batch size.

Having handled PP, let's move on to expert parallelism (EP). This method of parallelism can only be used on the feedforward blocks because experts generally share the same attention block, so using EP requires attention and feedforward blocks to be parallelized differently, a detail that's important to handle correctly in an inference model. In general, we'll be using tensor parallelism on the post-attention projection, so we'll end up with a residual stream tensor that's scattered across the  $d_{\text{model}}$  dimension. For expert parallelism, we need to perform the following two communication steps:

1. Each expert parallel GPU needs to receive the batch elements that are routed to it: this is the *all-to-all dispatch*. In the worst case this will require each vector of size  $d_{\text{model}}$  to be transmitted  $r = \min(N_{\text{EP}}, n_{\text{active expert}})$  times, but in practice it could be substantially less if expert activations are correlated and we place experts likely to fire together on the same EP copies. So we get an upper bound on the communication needed:  $r \cdot d_{\text{model}} \cdot b_{\text{micro}}$ , which we can approximately think of as an all-to-all operation across  $r$  ranks of a tensor of size  $d_{\text{model}} \cdot b_{\text{micro}}$ .
2. After each expert has finished its computation, we need to once again scatter the results across the model dimension to proceed: this is the *all-to-all receive*. This operation is simply the reverse flow of the previous operation logically, so we can also model it as an all-to-all with exactly the same type.

This simple estimate shows why expert parallelism is enormously cheaper than tensor parallelism:

1. When  $N_{\text{EP}} \ll n_{\text{active expert}}$ , even if we make the most pessimistic assumption about how active experts distribute over EP copies, the communication volume of expert parallelism is at most identical to that of 1D tensor parallelism: two all-to-alls versus a single all-reduce of tensors of identical size over identically many ranks. If expert activations are suitably correlated, we can often make this communication volume substantially lower by optimizing expert placement across EP ranks.
2. When  $N_{\text{EP}} \gg n_{\text{active expert}}$ , the communication volume of expert parallelism doesn't scale with the number of expert parallel ranks, only with the number of active experts. So parallelizing up to the upper limit of  $N_{\text{EP}} = n_{\text{expert}}$  costs little to no additional communication<sup>6</sup>, though we incur some cost from having to move some of this communication to InfiniBand instead of NVLink for large EP degrees. Still, this is enormously more favorable than what we would have to deal with with large TP degrees in a 2D TP setup.

So an assumption that's generically safe is that whenever the instance size  $N_{\text{GPU}}$  is below the number of experts  $n_{\text{expert}}$  in an MoE model, we would try to maximally use expert parallelism. If  $N_{\text{GPU}} > n_{\text{expert}}$ , then we can split the remaining

---

<sup>5</sup>This calculation assumes we're not using speculative sampling so that each parameter has to be read from HBM once per generated token.

<sup>6</sup>It might cost additional latency depending on the network topology, however.

$N_{\text{GPU}}/n_{\text{expert}}$  degrees of parallelism over TP and PP in whatever way is optimal for latency, holding  $N_{\text{GPU}}$  and batch size constant. This is what we do in our inference economics model.

### 3.6 Equations for the final model

To conclude this section, we present the complete equations of our final model in the 2D tensor parallel case and under the assumption of no pipeline or expert parallelism. This presentation assumes a vanilla decoder-only Transformer architecture that uses ReLU (or similar) activations: small changes need to be made to the memory and network reads/writes if a SwiGLU activation (Shazeer 2020) is used to take the additional weight matrix in the feedforward blocks into account. We also assume that the KV cache and the model weights fit into the HBM of all GPUs combined across the cluster: if this is not true, the model returns an infinite token latency.

The code for the paper contains a more general model which relaxes the 2D TP and no PP or EP assumptions and does casework to handle them according to the discussion from Section 3, but we do not include the equations of the model in this more general case in the paper.

$$\text{token latency} = \text{network and kernel time} + \max \left( \frac{\text{bytes read}}{N_{\text{GPU}} \cdot \text{memory bwd}}, \frac{\text{total FLOP}}{N_{\text{GPU}} \cdot \text{GPU FLOP/s}} \right) \quad (29)$$

$$\text{network and kernel time} = n_{\text{layers}} n_{\text{reduce}} t_{\text{kernel latency}} + \text{network comm time} \quad (30)$$

$$\text{bytes read} = \text{weight precision} \cdot \text{parameters read} + \text{activation precision} \cdot \text{activations read} \quad (31)$$

$$\text{parameters read} = N_{\text{attention params}} + \left( 1 - \left( 1 - \frac{1}{s} \right)^b \right) \cdot N_{\text{feedforward params}} + N_{\text{unembedding params}} \quad (32)$$

$$\text{activations read} = \text{KV cache elements read} + \text{matmul input and outputs read} \quad (33)$$

$$\text{KV cache elements read} = n_{\text{kv head}} d_{\text{head}} n_{\text{layers}} \ell_{\text{input}} b \quad (34)$$

$$\text{matmul input and outputs read} = n_{\text{layers}} \cdot b \cdot (2d_{\text{model}} + (1 + 2/g)n_{\text{head}} d_{\text{head}} + n_{\text{head}} d_{\text{head}} + 2d_{\text{model}} + 2d_{\text{ff}}) \quad (35)$$

$$\text{total FLOP} = b \cdot (\text{pointwise FLOP} + \text{attention FLOP}) \quad (36)$$

$$\text{attention FLOP} = (4 \text{ FLOP}) \cdot n_{\text{layers}} n_{\text{head}} d_{\text{head}} \ell_{\text{input}} \quad (37)$$

$$\text{pointwise FLOP} = (2 \text{ FLOP}) \cdot \left( N_{\text{attention params}} + \frac{N_{\text{feedforward params}}}{s} + N_{\text{unembedding params}} \right) \quad (38)$$

$$\text{network comm time} = n_{\text{layers}} n_{\text{reduce}} t_{\text{reduce}} + \frac{\text{inter-node reads}}{N_{\text{GPU}} \cdot \text{inter-node bandwidth}} + \frac{\text{intra-node reads}}{N_{\text{GPU}} \cdot \text{intra-node bandwidth}} \quad (39)$$

$$t_{\text{reduce}} = 6.8 \mu\text{s} + (1.2 \mu\text{s}) \cdot \left( \sqrt{\frac{N_{\text{GPU}}}{N_{\text{nodes}}}} - 1 \right) + (10 \mu\text{s}) \cdot \log_2(\sqrt{N_{\text{nodes}}}) \quad (40)$$

$$\text{inter-node reads} = 2 \cdot (\sqrt{N_{\text{nodes}}} - 1) \cdot \text{bytes reduced} \quad (41)$$

$$\text{intra-node reads} = 2 \cdot (\sqrt{N_{\text{GPU}}/N_{\text{nodes}}} - 1) \cdot \sqrt{N_{\text{nodes}}} \cdot \text{bytes reduced} \quad (42)$$

$$\text{bytes reduced} = ((1 + 2/g)n_{\text{head}} d_{\text{head}} + 2d_{\text{model}} + d_{\text{ff}}) \cdot b \cdot n_{\text{layers}} \cdot \text{activation precision}^7 \quad (43)$$

$$n_{\text{reduce}} = 4 \quad (44)$$

## 4 Inference economics of example models

Having presented the model, we now move on to discussing the inference economics of some well-known example models running on H100s at a price of two dollars per hour per H100 SXM. The H100 SXM has the following specs which will be relevant for our calculation:

---

<sup>7</sup>This assumes a feedforward block with a vanilla activation function. For more complex activations such as SwiGLU, it might be necessary to add another  $d_{\text{ff}}$  term to the expression inside the parentheses.

- The tensor cores can perform  $10^{15}$  FP16/s,  $2 \cdot 10^{15}$  FP8/s,  $2 \cdot 10^{15}$  INT8/s under ideal conditions. We assume in practice sustained arithmetic throughput caps out at 70% of these claimed numbers due to thermal throttling, unmodeled nonlinearities in the Transformer, et cetera.
- The HBM has a size of 80 GB and a bandwidth of 3.3 TB/s. We assume that in practice sustained bandwidth caps out at 75% of this theoretical maximum.
- Per-GPU bidirectional NVLink bandwidth of 450 GB/s: 450 GB/s of reads and 450 GB/s of writes adds up to the 900 GB/s in official specs. If using the LL (low-latency) protocol for NCCL primitives, which is necessary to get the good latency numbers discussed in Section 3, these bandwidths are further divided by a factor of 2. We also assume a node size of 8 GPUs, though some recent developments might allow this to be increased for Ampere and Hopper GPUs.
- The kernel launch latency (e.g. for CUTLASS kernels doing GEMM operations) is around  $4 \mu\text{s}$ .
- The base latency for launching GPU collectives in NCCL is around  $6.8 \mu\text{s}$  when using the LL protocol, which we assume cannot be eliminated and has to be paid for each GPU collective that's launched sequentially.

Using these numbers for the model in Section 3, we obtain the results in Figures 3 and 4.

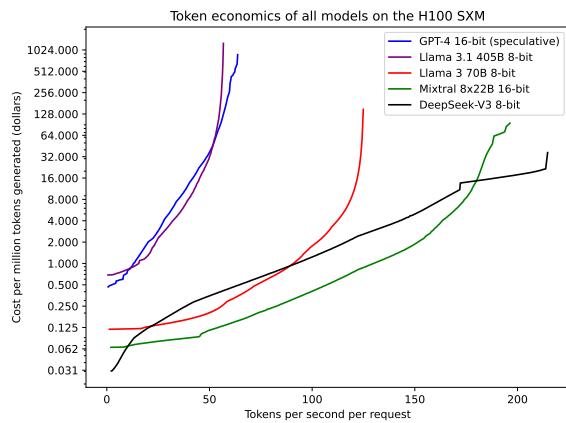


Figure 3: Without speculative decoding.

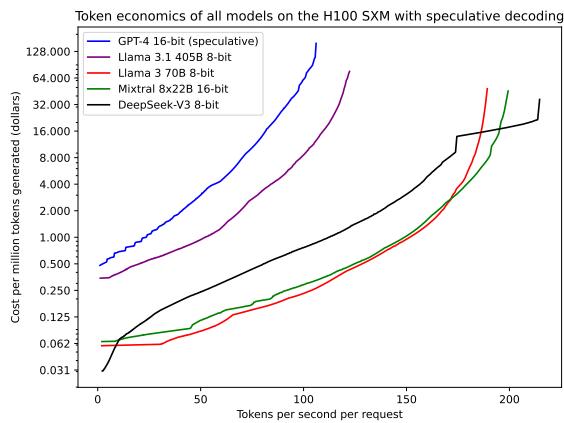


Figure 4: With speculative decoding.

Figure 5: The token economics for GPT-4, Llama 3.1 405B, Llama 3 70B<sup>8</sup>, DeepSeek-V3, and Mixtral 8x22B predicted by our model. In the right plot, we assume speculative decoding by Llama 3 8B (Leviathan, Kalman, and Matias 2023), while the left plot corresponds to the token economics of serving the base models. As before, we assume an effective rental price of 2 USD/hr for each H100 SXM.

As before, we also provide the maximum achievable tokens per second per request for each model, along with the optimal inference instance size at which this optimal performance is achieved. These results can be found in Table 2.

An important observation to be made in Figure 3 is that even assuming 8-bit quantization for Llama 3 70B is not enough for the observed price-performance point from the leading provider to be to the left of the model's Pareto frontier. This suggests that the observed performance is not achievable when doing naive autoregressive inference using Llama 3 70B alone; it's necessary to use speculative decoding or other similar techniques that reduce the number of calls to the model.<sup>9</sup> We'll later see that the latency gains from speculatively decoding Llama 3 70B using Llama 3 8B as the approximator model are significant: an optimal setup can effectively cut the latency per token by a factor of 2. This is sufficient for our model to be consistent with the data.

Now that we have a method for computing Pareto frontiers, we can ask the question of which point on the Pareto frontier of the API provider makes sense to choose given the preferences of the provider's clients. To perform this analysis in a simplified way, suppose that tokens generated at different speeds are perfect substitutes for each other, and

<sup>8</sup>We assume that Llama 3 70B weights are quantized to an 8-bit format such as FP8 or INT8, while the activations remain at a 16-bit precision.

<sup>9</sup>It's also possible that API providers have more optimized GPU collective kernels than we've assumed thus far, but for the moment we're assuming this is not the case.

Model name	Maximum tokens per second	Optimal instance size (GPUs)
Mixtral 8x22B (16-bit)	199	125
Llama 3 70B (8-bit)	189	24
Llama 3.1 405B (8-bit)	122	48
GPT-4 (16-bit, speculative)	106	460
DeepSeek-V3 (8-bit)	215	14

Table 2: Maximum tokens per second and the optimal instance size required to achieve the maximum according to our final model from Section 3 using H100 SXM GPUs. We assume speculative decoding is used, with the speculative decoder being taken as Llama 3 8B with a token acceptance probability of 80%. Note that these numbers disregard cost and only try to maximize tokens per second per request, so e.g. they assume a batch size of 1 even for mixture-of-experts models to reduce the total amount of memory reads.

the marginal value of a token to a representative customer scales with  $(\text{tokens per second})^\alpha$  for some  $\alpha \geq 0$ : tokens which are generated faster are worth more, all else equal. In this case, the customer will choose to purchase at the speed that maximizes

$$\frac{(\text{tokens per second})^\alpha}{\text{price per million tokens}} \quad (45)$$

Therefore, in equilibrium, we will observe that all API providers serve their models at the point on the cost curves from Figure 3 that maximizes Equation 45. Different values of  $\alpha$  give rise to different outcomes, and empirically we observed that we need to pick  $\alpha \approx 3$  to make the data from [Artificial Analysis 2024](#) consistent with our model. Using this value of  $\alpha$ , we present the implied efficient inference setups in Table 3.

Model name	Tokens per second	Cost per million tokens	Instance size (GPUs)	Batch size
Mixtral 8x22B (16-bit)	128	\$0.54	23	197
Llama 3 70B (8-bit)	107	\$0.27	7	136
Llama 3.1 405B (8-bit)	61	\$1.31	8	58
GPT-4 (16-bit, speculative)	61	\$4.53	100	213
DeepSeek-V3 (8-bit)	116	\$1.10	54	247

Table 3: The efficient inference setups according to our final model from Section 3 that maximize Equation 45, assuming  $\alpha = 4$  and H100 SXM GPUs. We assume speculative decoding is used, with the speculative decoder being taken as Llama 3 8B with a token acceptance probability of 80%.

Instead of fixing a GPU and looking at what happens when we serve different models on it, we can also compare different GPUs with each other to see how much cost savings we get from using newer GPUs such as the H100 compared to older ones such as the V100. For this comparison, we need to assume some rental prices, and we do this by taking the lowest rates we can find on [nebious.ai](#) (which correspond to orders renting many GPUs at once for a long period) and assuming that a third of the price corresponds to markups charged by the cloud provider (this factor doesn't matter for the comparison since it's the same for all GPUs). This gives the following rental prices:

- **H100 SXM:** 2.1 USD/hr
- **A100 SXM:** 1.5 USD/hr
- **V100 SXM:** 0.42 USD/hr

Choosing inference of Llama 3 70B with 8-bit quantized weights as our benchmark to judge the performance of these GPUs, we get the results presented in Table 4 and Figure 6.

Figure 6 and Table 4 suggest the following rough approximation across a wide range of inference speeds: with each new generation of GPUs from NVIDIA, the tokens per second achieved at a fixed cost per token goes up by around 25%, and the instance size required for inference on the Pareto frontier at a fixed cost per token falls by a factor of 2. It remains to be seen whether new generations of GPUs such as the B100 will continue this trend.

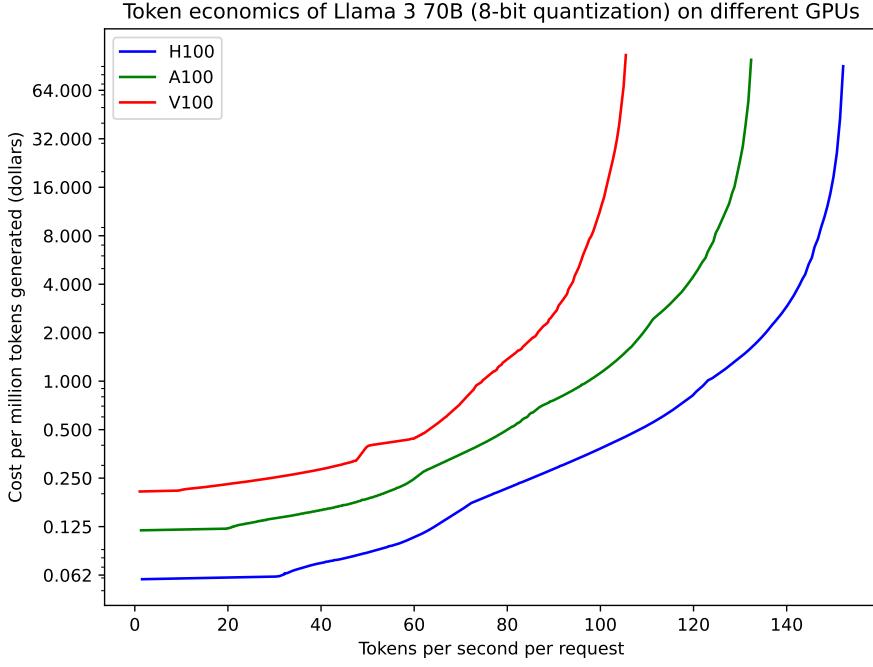


Figure 6: The token economics for Llama 3 70B with weights quantized to 8-bit precision on the V100, the A100 and the H100 (all of them using the SXM form factor). At large batch sizes, the Ampere and Hopper architectures benefit from their tensor cores supporting faster 8-bit precision arithmetic.

GPU name	Maximum tokens per second for Llama 3 70B inference	Optimal instance size (GPUs)
H100 SXM	152	24
A100 SXM	132	32
V100 SXM	105	102

Table 4: Maximum tokens per second and the optimal instance size required to achieve the maximum according to our final model from Section 3 on Llama 3 70B with the weights quantized to 8-bit precision. Different rows correspond to different GPUs.

#### 4.1 The impact of speculative decoding

Our inference economics model so far assumes that each new token generated to respond to a request requires a forward pass of the model that we're serving for inference. This is a reasonable assumption due to the sequential nature of text generation: because responses are generated one token at a time, and because later tokens causally depend on earlier tokens, it might naively seem as if we can't do better than this. However, there are techniques that allow us to circumvent this restriction, and the most popular out of them is speculative decoding (Leviathan, Kalman, and Matias 2023).

The idea of speculative decoding is to use a smaller model which can be inferred more quickly as a "speculator" or "approximator" to the larger model and use a rejection sampling technique to ensure that the tokens thus sampled from the smaller model have the same probability distribution as we would obtain if we had directly sampled from the larger model. More specifically, if  $P$  is the larger model and  $Q$  is the smaller model used as a speculator (both of which are using the same vocabulary and tokenizer), the procedure for sampling tokens looks like the following:

1. Fix an integer  $\gamma \geq 1$ .
2. Sample  $\gamma$  tokens  $x_1, x_2, \dots, x_\gamma$  from  $Q$ . Store the probabilities  $q_i$  of the  $i$ th token under  $Q$  conditional on previous context.

3. Do a single forward pass on all of  $x_1, x_2, \dots, x_\gamma$  with  $P$  to determine their probabilities  $p_1, p_2, \dots, p_\gamma$  of being sampled from  $P$  conditional on previous context.
4. For each  $i = 1, 2, \dots, \gamma$ , accept the  $i$ th token with probability  $\min(1, p_i/q_i)$  and reject the token otherwise.
5. Let  $j$  be the index of the first rejected token. Discard the tokens  $x_j, x_{j+1}, \dots, x_\gamma$  and resample  $x_j$  from the probability distribution whose mass function is proportional to  $\max(0, P - Q)$ .
6. Append  $x_1, x_2, \dots, x_j$  to the previous context and go back to step (2).

It's easy to prove that this procedure is equivalent to sampling the tokens from  $P$  directly. The advantage lies in the fact that if the acceptance rate is high enough for  $j$  to be suitably large in step (5), we can take a large value of  $\gamma$  and ideally only call  $P$  approximately once per  $\gamma$  tokens generated. If forward passes of  $Q$  are much faster than those of  $P$ , the latency advantage we would get from this is apparent. We do incur an additional arithmetic cost from this, as disregarding failed speculative branches each token still has to be seen once by  $P$  and once by  $Q$ , but when  $Q$  is much cheaper than  $P$  this cost is negligible.

If we model token acceptance events as independent and identically distributed, so that each token is accepted with a fixed probability  $\alpha$ , the number of tokens generated in each step follows a truncated geometric distribution whose expected value is given by  $V = (1 - \alpha^\gamma)/(1 - \alpha)$ . If the latency per forward pass of the models is assumed to be constant and given by  $t_P, t_Q$ , the average latency per token under speculative decoding will be given by

$$\text{mean latency per token} = \frac{\text{time per iteration}}{V} = \frac{(1 - \alpha)(t_P + \gamma t_Q)}{1 - \alpha^\gamma} \quad (46)$$

$\gamma$  is an integer-valued parameter we have the freedom to choose, and we can make the choice that minimizes latency for a fixed value of  $\alpha$ . We can estimate  $t_P, t_Q$  given the architectures of the two models using our inference economics model from earlier. Therefore, the only information specific to speculative decoding we need to know is  $\alpha$ , the acceptance probability per token. If we have access to both  $P$  and  $Q$ , we can estimate this quantity for a given prompt  $y$  by the following procedure:

1. Sample a sequence of tokens  $x_1, x_2, \dots, x_n$  from  $P$  as a response to (or continuation of) the prompt  $y$ .
2. For a randomly chosen index  $i$ , sample one token  $z_i$  from  $Q$  conditional on the context  $y; x_1, x_2, \dots, x_{i-1}$ .
3. Compute  $p = P(z_i \mid y; x_1, x_2, \dots, x_{i-1})$  and  $q = Q(z_i \mid y; x_1, x_2, \dots, x_{i-1})$ . Then, we expect  $\alpha = \mathbb{E}[\min(1, p/q)]$ .
4. Using (3), we can obtain more precise estimates of  $\alpha$  by repeating the sampling and computation steps in (2) and (3) for many prompts and many positions in the responses to these prompts, then averaging all of the values  $\min(1, p/q)$ .

Full access to the models  $P$  and  $Q$  is not required for this procedure: it's enough to have access to an API which provides the log probabilities of generated tokens. Many API providers offer this information in their responses to requests, e.g. [together.ai](#) and [octo.ai](#). We can therefore estimate the acceptance rate that would be achieved if a big model served by an inference provider was speculatively decoded using a smaller model that they also serve via API. Typical acceptance rates we've obtained when doing this in the Llama model series at sampling temperature equal to 1 range from 0.7 to 0.9, depending on the exact prompts and the pair of models chosen, and these findings agree with the range of acceptance rates reported for different model families where the two models are around an order of magnitude apart in size in Leviathan, Kalman, and Matias [2023](#).

Using  $\alpha = 0.8$  as our reference value for purposes of illustration, we get the results in Figure 7 and Table 5. Overall, the gains from speculative decoding at an 80% token acceptance rate are substantial: a 66% increase in throughput for Llama 3 70B and a doubling of throughput for Llama 3.1 405B at a fixed cost per token. The size of the effect means that it's difficult for API providers not using speculative decoding to compete with ones that do on latency or pricing.

## 4.2 The impact of quantization

Quantization is one of the primary techniques used by API providers to reduce inference costs and token-to-token latency. In general, weight quantization helps because it reduces memory bandwidth usage and accelerates arithmetic by shifting to lower precision formats supported by tensor cores, while activation quantization helps by reducing the size of the KV cache and by reducing network bandwidth usage.

Once again taking Llama 3 70B as our baseline, we provide some results about how much quantization helps with inference. Our simplified model from Equation 18 would predict that each halving of weight precision should give a

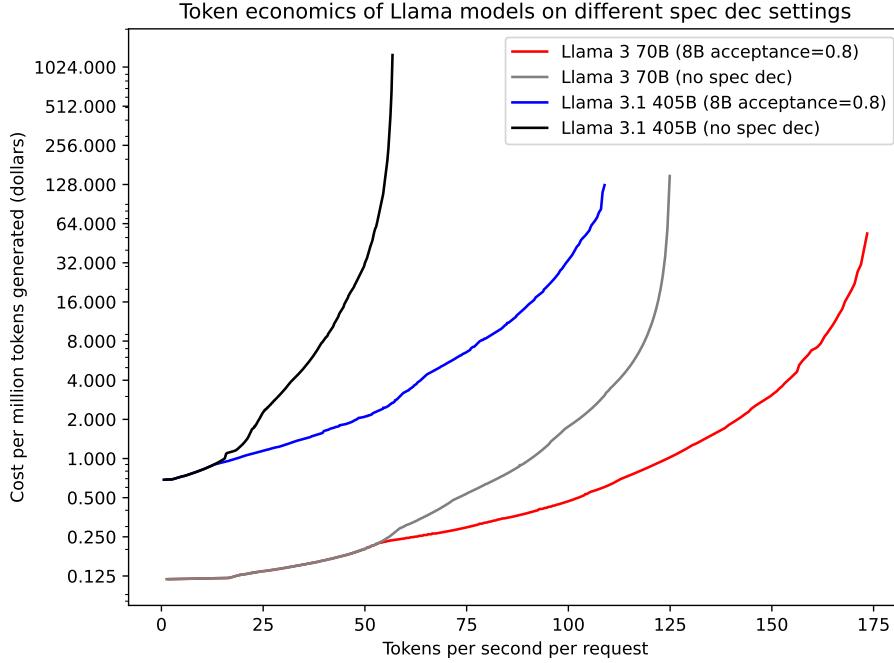


Figure 7: The token economics for Llama 3 70B and Llama 3.1 405B when both models are speculatively decoded with a model having Llama 3 8B’s architecture as the speculator. We provide the curves without speculative decoding for purposes of comparison. Here, all three models involved are served at a 16-bit precision, so no quantization is being employed.

Weight format	Tokens per second	Cost per million tokens	Instance size (GPUs)	Batch size
Llama 3 70B	69	\$0.52	8	127
Llama 3 70B (S)	95	\$0.51	6	73
Llama 3.1 405B	35	\$6.54	31	80
Llama 3.1 405B (S)	71	\$7.33	23	26

Table 5: The efficient inference setups for Llama 3 70B and Llama 3.1 405B with or without speculative decoding with Llama 3 8B as the speculator at a 80% token acceptance rate. (S) following the name of a model means we assume speculative decoding is used for the model in question. As before, we choose the efficient setups by maximizing Equation 45, assuming  $\alpha = 3$  and H100 SXM GPUs. Note that these models are not quantized and remain at 16-bit weight precision.

$2^{1/3} - 1 \approx 26\%$  increase in maximum tokens per second. The results from the more detailed model can be found in Figure 8 and Table 6, and they show that this simple rule of thumb actually works quite well.

#### 4.3 The impact of long context lengths

When the context length is not long enough for KV cache reads or attention arithmetic to make a substantial contribution to the total cost of inference, the main difference between input and output tokens is that input tokens can be processed in parallel and as a result generally have much more forgiving latency constraints. Indeed, in many cases just a ping to a target server might take on the order of 100 milliseconds, so trying to shave off milliseconds from the processing that happens on the server side is usually not worth it. This leads to a much better average arithmetic utilization when processing input tokens compared to output tokens.

For instance, at 100% utilization on an H100, the cost per million tokens of Llama 3 70B at 16-bit precision would be around \$0.11, implying an utilization rate of  $\approx 17\%$  for this precision in Table 6. During prefill (the processing of input tokens before any output tokens are generated) we probably expect this to be 50% or more, suggesting that input tokens

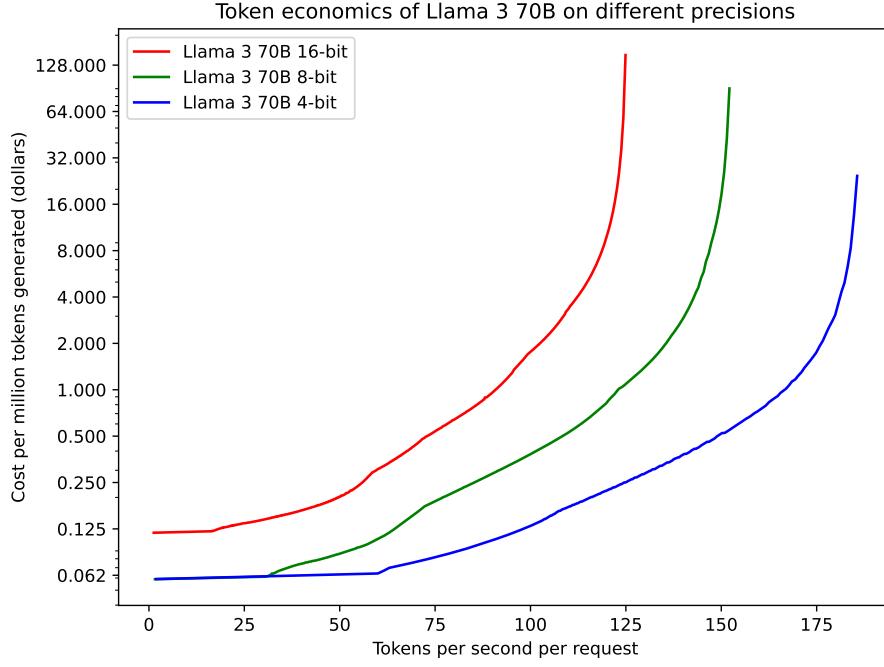


Figure 8: The token economics for Llama 3 70B with weights quantized to different precisions, holding activation precision fixed at 16 bits.

Weight format	Tokens per second	Cost per million tokens	Instance size (GPUs)	Batch size
4-bit weights	122	\$0.23	4	90
8-bit weights	99	\$0.37	7	109
16-bit weights	83	\$0.70	13	136

Table 6: The efficient inference setups for Llama 3 70B with weights quantized to different precisions according to our final model from Section 3 that maximize Equation 45, assuming  $\alpha = 4$  and H100 SXM GPUs.

should be priced 3 times cheaper than output tokens. This ratio matches empirical data about e.g. OpenAI’s pricing of their input tokens versus output tokens for GPT-4o (OpenAI 2024).

In contrast, when the context length is so long that attention dominates the cost of inference, the lower utilization during decoding occurs due to the need to read the KV cache every time a new token is generated. The size of the KV cache per batch element can be expressed as  $2n_{\text{layers}}n_{\text{head}}d_{\text{head}}\ell_{\text{input}}/g$  activations, while the amount of arithmetic per token can be expressed as  $4n_{\text{layers}}n_{\text{head}}d_{\text{head}}\ell_{\text{input}}$  FLOP. Therefore, the arithmetic intensity in this regime is *at most*

$$\frac{4n_{\text{layers}}n_{\text{head}}d_{\text{head}}\ell_{\text{input}} \text{ FLOP}}{\text{precision} \cdot 2n_{\text{layers}}n_{\text{head}}d_{\text{head}}\ell_{\text{input}}/g} = \frac{2g \text{ FLOP}}{\text{precision}}$$

if speculative decoding or other similar methods to decode multiple tokens for one read of the KV cache are not employed. This means that in sufficiently long context lengths, the differential pricing of input and output tokens under an ideal pricing scheme would depend almost entirely on the attention group size  $g$  of the architecture, with all other details being essentially irrelevant. When grouped-query attention is not used, so that  $g = 1$ , the best utilization that can be achieved on NVIDIA hardware will be below 1%. For this reason, GQA is critical for efficient long-context inference regardless of other architectural details.

In practice, because most API providers use fixed prices for input and output tokens, their pricing strategy can only be an approximation to the ideal strategy suitable for the demand pattern that they face for their service, with details such as the ratio of input to output tokens in a typical long-context request being of particular importance.

As an interesting case study, we compare the short and long context token economics of Mistral Large 2, a dense transformer with 123 billion parameters and an attention group size of 12. The predictions of the inference model for these settings can be found in Figure 9.

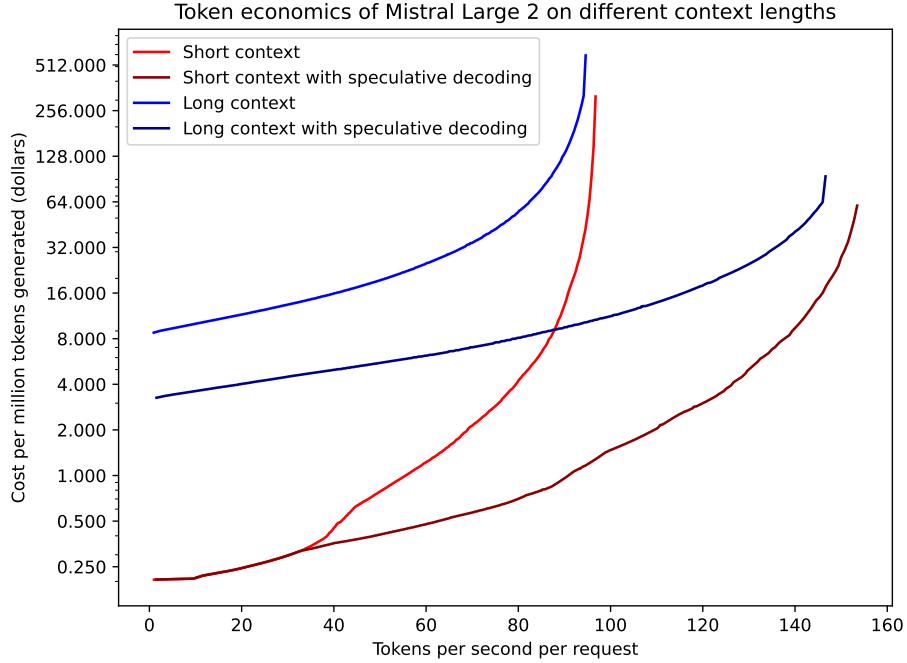


Figure 9: The token economics for Mistral Large 2 at short context lengths and long context lengths (100K). We assume the speculative decoder is a hypothetical model with a variant of Llama 3 8B’s architecture using multi-query attention with an 80% acceptance rate for proposed tokens.

Mistral Large 2 has a per-token KV cache size of around 360 KB. At a context length of 100K, this means each generated token requires  $10^5 \cdot 360 \text{ KB} = 36 \text{ GB}$  of KV cache reads without speculative decoding. Since the model size itself is only 244 GB, KV cache reads become a bigger bottleneck than parameter reads at batch sizes larger than 7. In addition, at 3.3 TB/s of memory bandwidth per H100 GPU with GPU time costs at 2 USD per hour, the price of the memory read time required for KV cache reads per million tokens can be calculated as

$$\frac{36 \text{ PB/million output tokens} \cdot \$2/\text{hour}}{3.3 \text{ TB/s}} = \$6.06 \text{ per million output tokens}$$

Since this price is a lower bound on the actual price per million tokens in most situations<sup>10</sup>, the difference between the short and long context Pareto frontiers ends up being quite large.

It’s instructive to contrast this with the long-context behavior of a highly sparse MLA model such as DeepSeek-V3. With a per-token KV cache size of  $\approx 60 \text{ KB}$ , we expect DeepSeek-V3 to be around 6 times cheaper than Mistral Large 2 if we’re doing inference slowly. This turns out to roughly check out in our model as seen in Figure 10.

<sup>10</sup>It’s possible to get around this to some extent using clever scheduling methods, for example by batching long context requests with short context ones and keeping the tensor cores occupied during the memory reads for the KV cache. However, these methods only yield limited gains, and our model assumes a fixed context length over the entire batch.

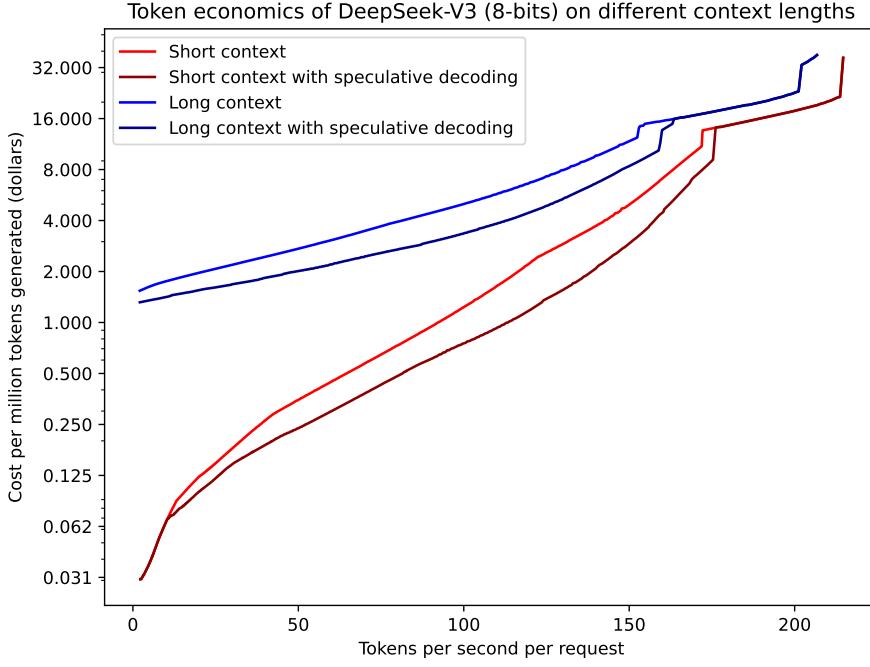


Figure 10: The token economics for DeepSeek-V3 (8-bit weight quantization) at short context lengths and long context lengths (100K). We assume the speculative decoder is a hypothetical model with a variant of Llama 3 8B’s architecture using multi-query attention with an 80% acceptance rate for proposed tokens.

## 5 Conclusion and important takeaways

We list some of the most important takeaways from our analysis below.

### 5.1 Takeaways for short-context inference

1. The most important bottleneck to fast decoding in short contexts is **latency**, with network and NVLink latency being the biggest culprits on NVIDIA hardware. Even small all-reduce operations can cost tens of microseconds of pure latency.

If latency constraints didn’t exist, then we could in principle increase the serial speed of token generation by four times indefinitely for each doubling of cost per token (Steinhardt 2022), because each matrix multiplication can be parallelized along three different dimensions, so increasing our number of GPUs by 8 times with 4 times more 2D tensor parallelism and 2 times more pipeline or data parallelism would only increase both memory and network bandwidth usage by a factor of 2. Any modeling of real-world token economics has to begin by understanding how latency causes this basic logic to break down.

2. A rough rule of thumb is that the speed at which a dense model with  $N$  parameters can be served for inference scales as  $\propto 1/\sqrt{N}$ : there’s an inverse square root law. This can be empirically verified using data from sources such as [Artificial Analysis 2024](#) but it also drops out of the theoretical analysis that we perform on a toy model in Section 2.2.3 if we make some reasonable assumptions about model aspect ratio scaling.<sup>11</sup>
3. Higher memory bandwidth does speed up inference, but the increase is slow in short contexts: a doubling of memory bandwidth alone can only be expected to deliver a reduction of token-to-token latency of around 20% or less. This is implied by toy models of inference such as in Section 2.2.3 and is also confirmed by empirical data: Groq LPUs have up to 80 TB/s of memory bandwidth per chip compared to 3.3 TB/s per H100 SXM,

<sup>11</sup>This agreement shouldn’t be overstated, as in practice there are multiple factors that would lead us to expect latency to increase faster with model size than Section 2.2.3 does not consider. The fact that the scaling law nevertheless empirically holds shows that the factors that this model does not consider appear to roughly cancel out in practice up to an overall constant factor in the scaling law.

and as our model would predict, the latency gap between the fastest NVIDIA provider and Groq on *Artificial Analysis 2024* is generally a factor of 2 to 3.<sup>12</sup>

4. Speculative decoding at plausible token acceptance rates has a large impact on token-to-token latency, possibly reducing it by a factor of 2 at a fixed cost per token in realistic situations. This is in spite of the added arithmetic cost of speculative decoding and the potential for speculated tokens to be rejected.
5. Quantizing models is most useful when quantization allows the model to fit inside a discontinuous hardware boundary such as a single GPU or a single node. In general, reducing the precision of a model by a factor of 2 seems to reduce token latency by around 20%, but the decrease in latency can be significantly more than this in the event that the inference setup is running up against a natural hardware boundary.
6. Assuming that an inference provider is able to get good utilization during the prefill phase, the hardware utilization during short-context decoding can be anywhere from 3 to 5 times less than the one achieved during prefill at practical speeds.

## 5.2 Takeaways for long-context inference

1. In leading open source models today which use a traditional attention mechanism (Llama series of models, Mistral Large 2, *et cetera*), the use of limited grouped-query attention with group sizes ranging between 4 and 12 means that long-context inference is **always memory-bandwidth bound** due to KV cache reads. This does not depend on the batch size used for inference, because the size of the KV cache scales linearly with the batch size. Overcoming this bottleneck requires big group sizes (on the order of  $g = 100$ ), KV cache compression methods, or revisions to the attention mechanism itself.

One such revision is *multi-head latent attention* (MLA), used for example in DeepSeek-V3 (DeepSeek-AI et al. 2025). This variant of attention makes a significant tradeoff, raising arithmetic cost by many times in order to match the efficient KV cache size reduction of GQA with large group sizes without the same performance degradation. The large arithmetic cost of the attention mechanism of MLA models means their inference often becomes attention arithmetic bound at context lengths as short as ten thousand tokens.

2. As a consequence of (1), increases in memory bandwidth are typically more important for long-context inference compared to short-context inference, as in longer contexts even increasing the batch size and slowing down inference cannot improve utilization when KV cache reads become the bottleneck.

Overall, the theoretical model we've developed accurately predicts many empirical facts about LLM inference economics without having been fit to the data or adjusted with "fudge factors" by hand. The framework has the potential to serve as a practical "roofline model" for understanding the fundamental limits of LLM inference performance.

## References

- Artificial Analysis (2024). Accessed: 2024-04-26. URL: <https://artificialanalysis.ai>.
- Chowdhery, Aakanksha et al. (2022). *PaLM: Scaling Language Modeling with Pathways*. arXiv: [2204.02311 \[cs.CL\]](https://arxiv.org/abs/2204.02311).
- DeepSeek-AI et al. (2025). *DeepSeek-V3 Technical Report*. arXiv: [2412.19437 \[cs.CL\]](https://arxiv.org/abs/2412.19437). URL: <https://arxiv.org/abs/2412.19437>.
- Hoffmann, Jordan et al. (2022). "Training compute-optimal large language models". In: *arXiv preprint arXiv:2203.15556*.
- Jeaugey, Sylvain (Feb. 2019). *Massively Scale Your Deep Learning Training with NCCL 2.4*. NVIDIA Developer Blog. Accessed on 16 April 2024. URL: <https://developer.nvidia.com/blog/massively-scale-deep-learning-training-nccl-2-4/>.
- Leviathan, Yaniv, Matan Kalman, and Yossi Matias (2023). *Fast Inference from Transformers via Speculative Decoding*. arXiv: [2211.17192 \[cs.LG\]](https://arxiv.org/abs/2211.17192). URL: <https://arxiv.org/abs/2211.17192>.
- NVIDIA (2024). *NCCL (NVIDIA Collective Communication Library)*. Accessed: 2024-05-23. URL: <https://github.com/NVIDIA/nccl>.
- OpenAI (2024). *OpenAI API Pricing*. <https://openai.com/api/pricing/>. Accessed: 2024-08-01.
- Patel, Dylan and Gerald Wong (July 2023). *GPT-4 Architecture, Infrastructure, Training Dataset, Costs, Vision, MoE*.
- Pope, Reiner et al. (2022). *Efficiently Scaling Transformer Inference*. arXiv: [2211.05102 \[cs.LG\]](https://arxiv.org/abs/2211.05102).
- Shazeer, Noam (2019). *Fast Transformer Decoding: One Write-Head is All You Need*. arXiv: [1911.02150 \[cs.NE\]](https://arxiv.org/abs/1911.02150).
- (2020). *GLU Variants Improve Transformer*. arXiv: [2002.05202 \[cs.LG\]](https://arxiv.org/abs/2002.05202).

<sup>12</sup>Groq is able to overcome their very small amount of memory per chip (around 230 MB) by making extensive use of pipeline parallelism during inference. This allows them to approach the performance that would be attained by the same memory bandwidths at much higher amounts of on-chip memory, which is what the model in Section 2.2.3 estimates.

- Steinhardt, Jacob (2022). *How fast can we perform a forward pass?* Accessed: 2024-07-24. URL: <https://bounded-regret.ghost.io/how-fast-can-we-perform-a-forward-pass/>.
- Touvron, Hugo et al. (2023). *Llama 2: Open Foundation and Fine-Tuned Chat Models*. arXiv: [2307.09288 \[cs.CL\]](https://arxiv.org/abs/2307.09288).