Artificial Intelligence: How Much Will It Cost?

A rapid jaunt through cost models and consequences

Todd Underwood ⧫
@<twitter is gone forever> ⧫
tmu@google.com ⧫
2023-Oct ⧫ SRECon

Photo by Tara Winstead on Pexels

AI Cost

Agenda

AI Costs Are Big….

“AI costs are big. You just won't believe how vastly, hugely, mind-bogglingly big they are. I mean, you may think data centers and servers for search applications cost a lot of money, but that's just peanuts to AI.”

Apologies to Douglas Adams… and you

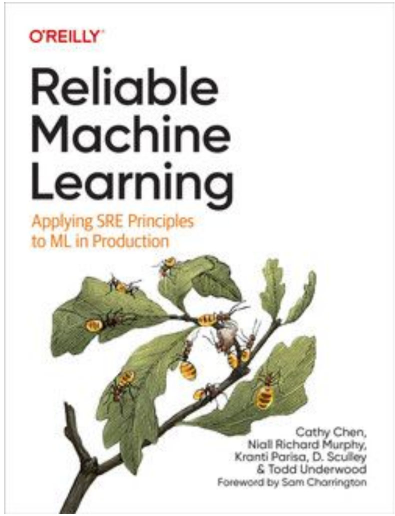
- Intro
- Accounting? Finance? Really
- Generative AI: Why we care
- Large Language Models: understanding costs
- Implications and projections

Photo by Jp Valery on Unsplash

Greetings

AI Cost

Basic Questions (the usual)


O'REILLY
Reliable Machine Learning
Applying SRE Principles to ML in Production
Cathy Chen, Niall Richard Murphy, Kranti Parisa, D. Sculley & Todd Underwood
Foreword by Sam Charrington

Who am I?
Founder of (perhaps the first?) ML SRE team several years ago. Worked on ML systems in SRE at Google for 12+ years. (co)Wrote the “Reliable ML” O’Reilly book. Now work in Capital Engineering, the CFO’s engineering team, managing and understanding ML capital costs.

Who are you?
SREs who get paid for something who are curious about this AI transition. The get paid part is important because that is why (or should be why) you care about how much this stuff all costs.

Why are we here?
To talk about accounting. And technology. And other stuff.

Will AI change everything?
I mean, sure. Doesn’t everything? Anyway, let’s go!

Three Questions

AI Cost

AI Cost

0) How do corporations think about money?
Sigh. I know. We'll be brief.

1) What is Generative AI?
Quick, practical summary of what is happening here. What is new, roughly how it works and what it might do for us.

2) Why does it cost so much?
Review of major costs of these Large models. **There will be accounting words used.** Sorry.

3) What's next?
We'll try to think about where value and cost might intersect in the future.

4) An almost fanatical devotion to the pope?

Photo by Adi Goldstein on Unsplash

Thinking About Money

AI Cost

Important Disclaimers

- I have no business or finance education or credentials (I studied Philosophy in Uni and am happier talking about Wittgenstein's skepticism about meaning).
- I do use finance concepts to provide input to decisions about how much to spend on what kind of ML at my current employer.
- Accounting and finance rules vary (sometimes significantly) among countries, even traditional market economies.
- There are people who know this stuff. Seek them out for real information in your circumstances.

Photo by Nataliya Vaitkevich on Pexels

Accounting: A Wild History

Good accounting tells a (true) story and helps predict the future.

Modern, dual-entry bookkeeping first published in Venice in 1494 by Luca Pacioli:

Summa de arithmetica, geometria, Proportioni et Proportionalita

The same text gave us accounting but also had the math for 2D representations of 3D perspective that helped Leonardo paint *The Last Supper.* **Seriously.**

Image public domain from wikipedia

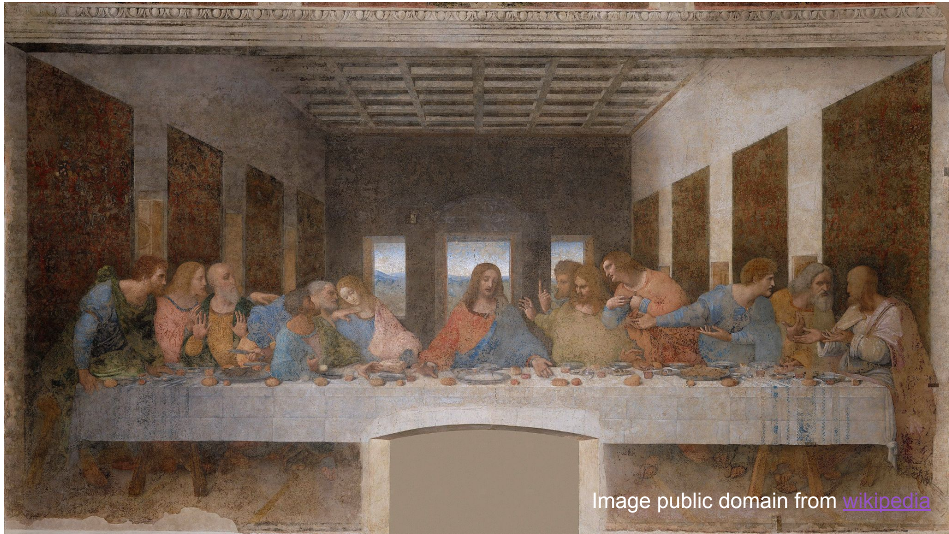


Image public domain from wikipedia

Photo by Savvas Stavrinos on Pexels

Accounting: Stories and Trust

Allows us to trust others

… and understand Ourselves

We need a common way to understand modern organizations. Accounting and finance provide that.

Management: we need ways of tracking how our own organization is (and will be) spending money.

Investment (our own): What should we spend more on? What is a waste of money? What looks profitable but isn't? What will be profitable next year?

Investment (in others): How should I understand the financial circumstances of another organization? Are they healthy? What are the future expenses?

Public Domain Pictures Pexels

AICost

Accounting: The Bare Minimum (for this talk)

Photo by Karolina Grabowska on Pexels

Accounting vs Finance

Accounting systematically tracks money. Finance analyzes and predicts.

Operational Expenses

Costs that we incur where we receive all the value right away (during the current accounting period). Examples: employee salaries, electricity bill.

Capital Expenses

Things (assets) we buy that have (commercial) value over time. Example: an expensive server that we buy for $48k but we think we can rent out for four years (48 months).

Depreciation (related to Amortization)

The periodic reduction in value of an asset (capital expense) over time as its remaining (commercially viable) lifetime is used up. Sad, but nothing lives forever

Historical: SSL Serving circa 2000

Commodity webserver: **$2k**

- ~500 **plaintext** http requests/sec ($0.004/req/s)
- ~5 **SSL/TLS** requests/sec ($40/req/s)

<Context: we figured out we should encrypt stuff>.
Website with peak traffic of ~3000 requests/second.

Plaintext: 8 servers (n+2) or **$16k capex ~= $333 monthly**
SSL: 720 servers(20% reliability) or **$14.4m ~= $30k/month.**

SSL hardware offload: $45k / 500 req/sec or $90/req/s or **$360k** (25% reliability) **~= $7.5k/month**. **Good investment.**

Discussion: faster than planned obsolescence (CPUs do encryption better). Uncertainty about amortization period (48 months used here but it's a guesstimate).

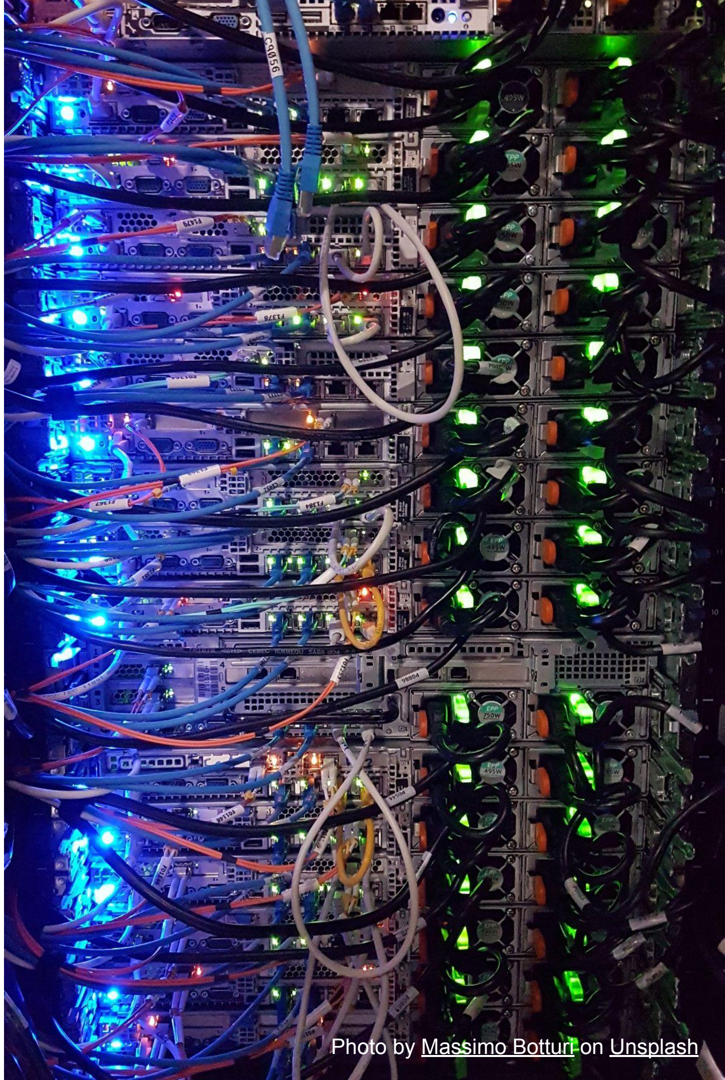


Photo by Massimo Botturi on Unsplash

The Cost of Capex: Predicting the future

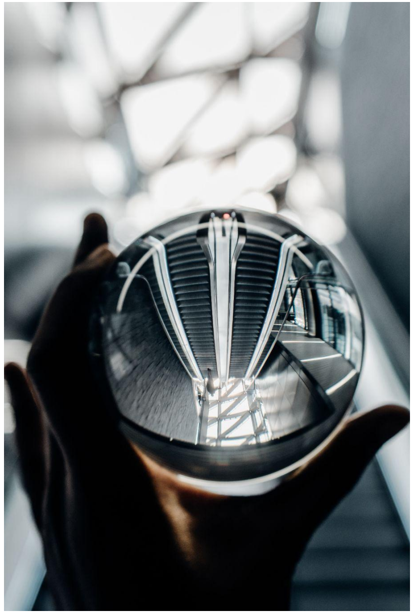


Photo by Nigel Tadyanehondo on Unsplash

The Dark Art of Depreciation
Depreciation is mostly standard. Examples: houses: ~30 years [really 27.5 and only leased houses]; cars: 5 years; Office equipment: 7 years. Generally set by committees. Not perfect, but reasonable.

General rule: the newer something is the harder to predict. So new computer hardware platforms are the worst case scenario. (More on this shortly)

Mistakes
When we get depreciation wrong the consequences are bad (economically).

Equipment lasts longer: incorrectly bad past financial results and now overly good results because of “free” equipment. This can hide successes and hide failures.

Equipment used up faster: underestimate costs and must write off (mark to zero) now-obsolete equipment. This sudden loss is really a bad decision in the past.

Hold this thought.

Generative AI

AI Cost

Generative AI: Why it matters

Photo by Andrea Piacquadio on Pexels

Background: very large models (often trained on language samples) that learn complex probabilities of relationships among terms (or other tokens). LLM == Large Language Model.

“Understanding” Language

LLMs allow for novel interfaces to data, potentially replacing structured languages such as SQL.

Generating Language (and much more)

LLMs can use their “understanding” of language to generate novel strings of contextually correct speech. They can do this with other media (images, video, music, audio) as well.

Sample applications

Write *somewhat* correct software very quickly. Summarize large amounts of text. Write reasonable first drafts of letters and essays. Generate custom artwork.

AI Cost

But Why Accelerators

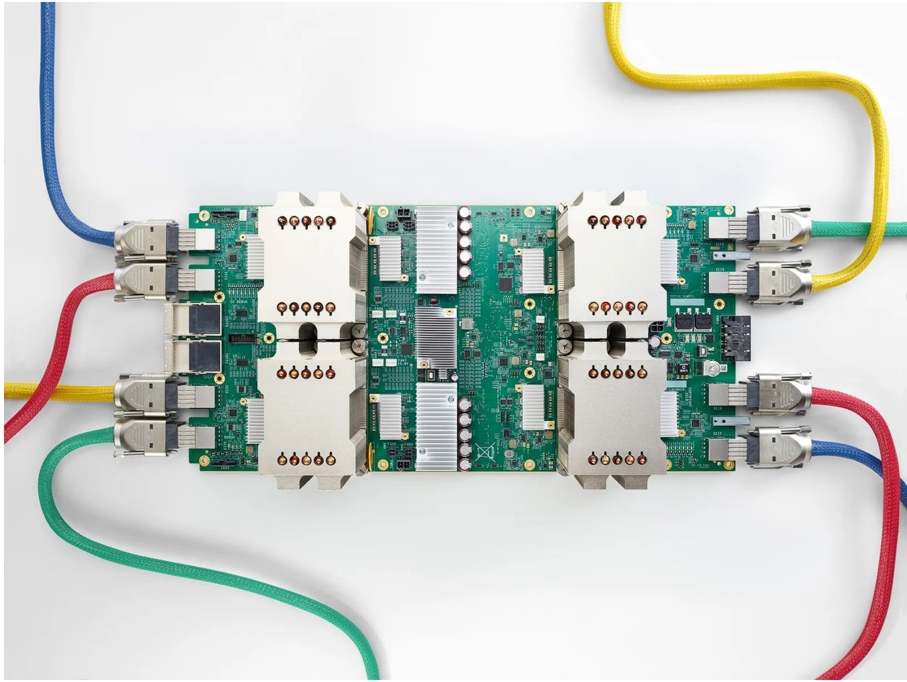
LLM machine cost is mostly compute (not storage/network) (by a lot).

Specific common compute operations: e.g.: low precision matrix operations (common in Deep Neural Networks)

Common ML tasks are faster, more scalable, and cheaper (especially more power-efficient) on accelerators than on general purpose CPUs.

Training/Serving optimization somewhat (not entirely) different. Serving optimized for fast/cheap lookups. Training for updating a large distributed data structure.

Buying requires you amortize hardware costs over lots of computation and and time. Renting better for most.



TPUv2 Image © Google

LLM Costs

AI Cost

AI Cost

LLM Cost Elements

Compute, storage, network

(At most scales **compute dominates** - we will need accelerators here)

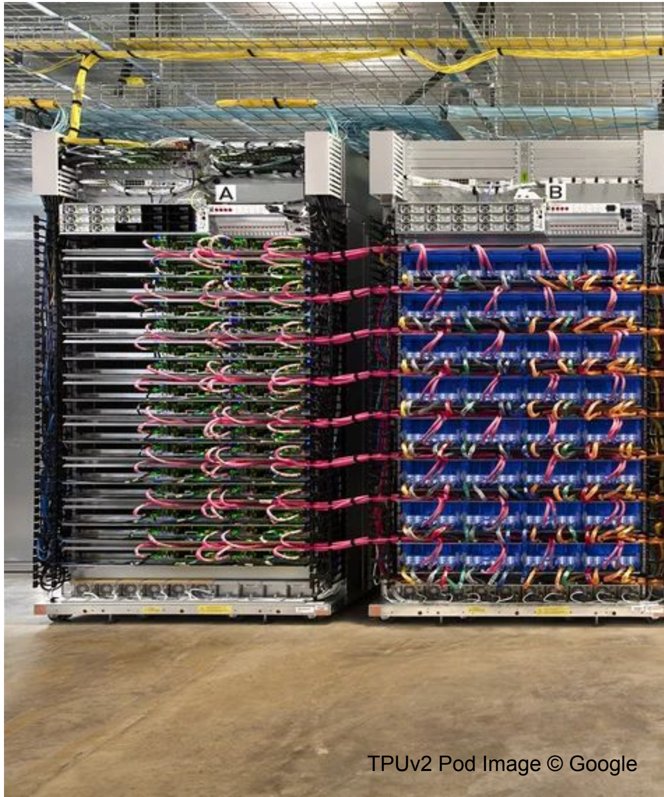
Model cost is computed in tokens and parameters. Tokens are input primitives. Parameters are model features. Typical sizes range from 7-65B parameters (Llama) to a rumored 1.75T parameters (GPT4).

Training (making a new model) vs Serving (using an existing model to perform useful computations)

100% of training costs must be recovered in serving

Other uses: experimentation & development, bulk inference.

Other model types: not everything is an LLM. They **just** showed up! So many other kinds of ML that might benefit from accelerators



TPUv2 Pod Image © Google

AI Cost

LLM Cost Concrete Estimates

Input data size : ballpark ~1.4T tokens in ~5TB.

(Llama from FB) Training cost: 380 tokens/s/GPU
~21 days of 2048 NVIDIA A100s

A100: $15k list
CapEx: $30M (assume 3 year lifecycle?)
OpEx (no markup/margin, for **this** training run):
$580K (power cost <$100k)
Rented on the spot market $15.5M (high estimate)
Real cost in that range.

Assuming:
30% experimentation budget
75% availability during training run
100% utility post-training for 3 years

Estimate: **$1.1M total training cost**

Photo by Dušan veverkolog on Unsplash

LLM Training Size/Cost Guestimates

	Llama 2	GPT-3.5	GPT-4
Parameters	70B	150-175B	1-1.75T
Cost (linear)	$1-2M	$2-5M	$13 - $50M

(sources: Llama paper, various published reports and estimates)

Big error bars.
Biggest error from:

- Cost of experimentation: can't train a model you didn't develop and build.
- Overall utilization during training run
- Estimated useful lifespan of hardware

Photo by Greg Lippert on Unsplash

Photo by Billy Huynh on Unsplash

Cost of Serving

Economic value of ML models is realized in serving.

Principles:

- We need a sufficient volume of sufficiently valuable inferences to justify cost of training **plus** serving.
- More expensive serving == longer/harder payoff
- **Super** model-dependant (big models are more expensive)
- More training can make cheaper/smaller models

Estimates (Llama/GPT3.5; ~5 characters/token):

- Cost/Prompt: ¢0.15 - ¢0.5 / 1k tokens
- Output/Completion: ¢0.2 - ¢1.5 / 1k tokens
- Estimate size of input/output for your use case. Typical chat use case: 1-2k tokens input and 2-4k tokens output. Cost per conversational round-trip: ¢0.55 - ¢7
- **Huge** error bars here

(Sources: Llama estimates plus OpenAI API pricing)

AI Cost

Serving Demand

$4M pretraining
(we need some fine tuning - forgot to mention that)

Using $4m cost to train and $0.035/query:

We need 100M queries before serving cost is >50% (and at that point queries are still $0.07/query, amortizing training costs).

In order to have anything like an economically viable system we either need:

- A huge volume of marginally valuable queries
- A high volume of extremely valuable queries

Photo by Dušan veverkolog on Unsplash

Implications and Projections

AI Cost

AI Cost

What does it all mean?

(Obvious: your mileage may vary and these are just worked examples)

- You should not train your own LLM.
  - Insert many, many caveats here including: specificity of your use case, skill/experience of staff, technical infrastructure, etc.
- You should integrate LLMs into particularly valuable applications (general guideline: things that improve productivity of expensive humans).
- You should rent accelerator hardware where you can.
- “[ML] moves pretty fast. If you don’t stop and look around once in a while, you could miss it”

Photo by Dev Asangbam on Unsplash

Thank You

AI Cost

Todd Underwood ⧫
@<twitter is gone forever> ⧫
tmu@google.com ⧫
2023-Oct ⧫ SRECon