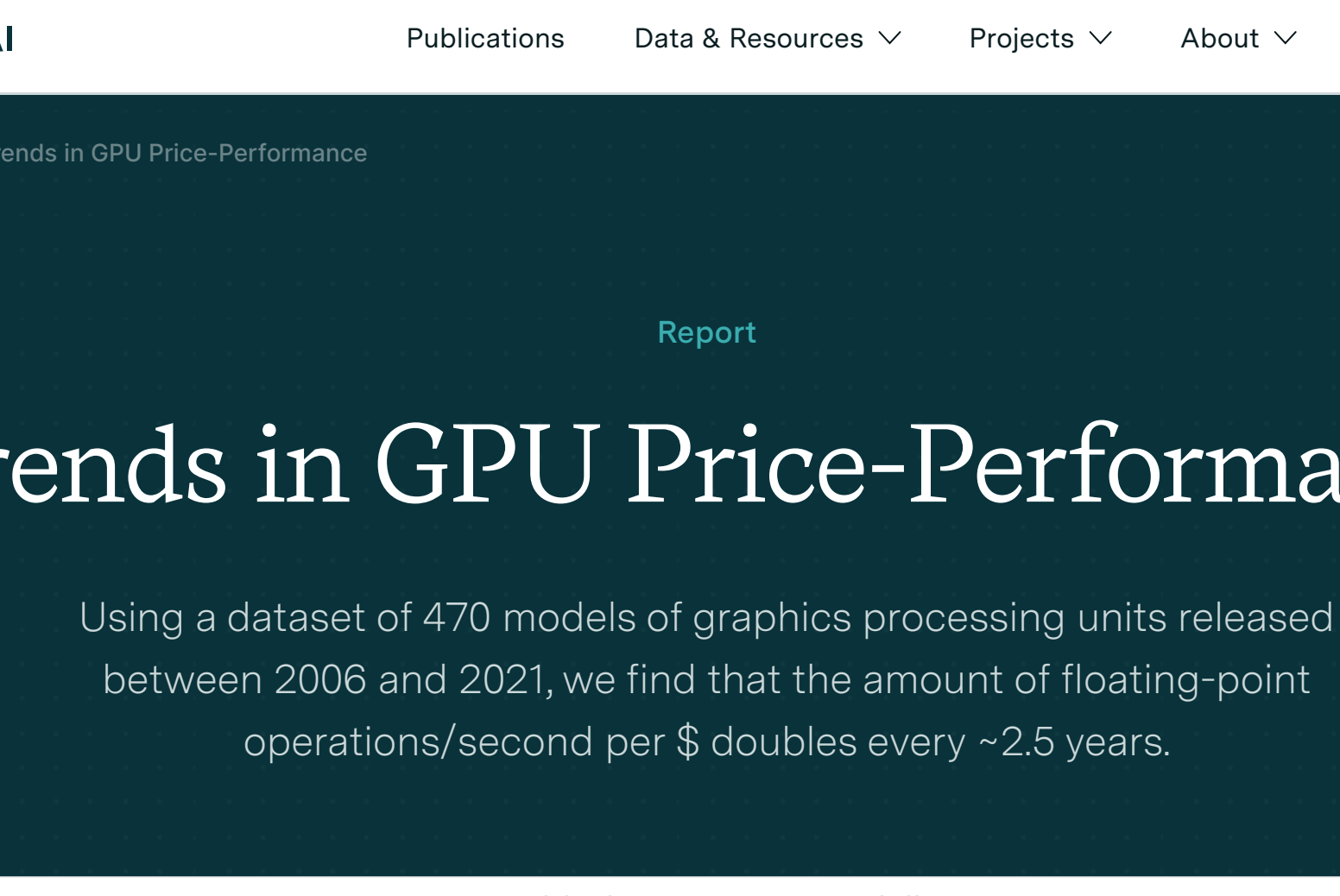
Report

Trends in GPU Price-Performance

Using a dataset of 470 models of graphics processing units released between 2006 and 2021, we find that the amount of floating-point operations/second per $ doubles every ~2.5 years.



Published

Jun 27, 2022

Authors

Marius Hobbhahn
Tamay Besiroglu

Resources

Cite

Update

This report was originally published on Jun 27, 2022. For the latest research and updates on this subject, please see: Data on Machine Learning Hardware.

Contents

We would like to thank Alyssa Vance, Ashwin Acharya, Jessica Taylor and the Epoch AI team for helpful feedback and comments.

Executive Summary

Using a dataset of 470 models of graphics processing units (GPUs) released between 2006 and 2021, we find that the amount of floating-point operations/second per $ (hereafter FLOP/s per $) doubles every ~2.5 years. For top GPUs at any point in time, we find a slower rate of improvement (FLOP/s per $ doubles every 2.95 years), while for models of GPU typically used in ML research, we find a faster rate of improvement (FLOP/s per $ doubles every 2.07 years). GPU price-performance improvements have generally been slightly slower than the 2-year doubling time associated with Moore's law, much slower than what is implied by Huang's law, yet considerably faster than was generally found in prior work on trends in GPU price-performance. We aim to provide a more precise characterization of GPU price-performance trends based on more or higher-quality data, that is more robust to justifiable changes in the analysis than previous investigations.[1]

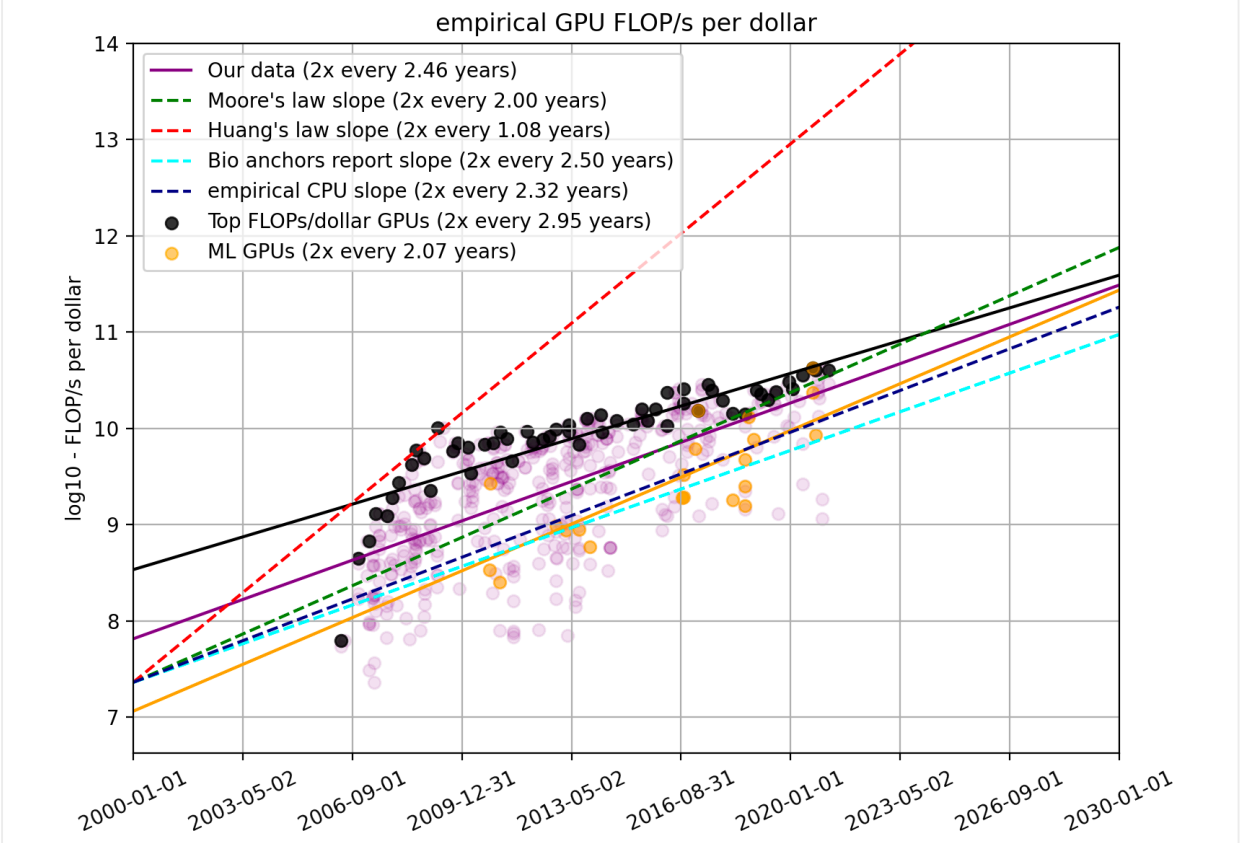


Figure 1. Plots of FLOP/s and FLOP/s per dollar for our dataset and relevant trends from the existing literature

Trend	2x time	10x time	Growth rate	Metric
Our dataset (n=470)	2.46 years [2.24, 2.72]	8.17 years [7.45, 9.04]	0.122 OOMs/year [0.134, 0.111]	FLOP/s per dollar

Trend	2x time	10x time	Growth rate	Metric
ML GPUs (n=26)	2.07 years [1.54, 3.13]	6.86 years [5.12, 10.39]	0.146 OOMs/year [0.195, 0.096]	FLOP/s per dollar
Top GPUs (n=57)	2.95 years [2.54, 3.52]	9.81 years [8.45, 11.71]	0.102 OOMs/year [0.118, 0.085]	FLOP/s per dollar
Our data FP16 (n=91)	2.30 years [1.69, 3.62]	7.64 years [5.60, 12.03]	0.131 OOMs/year [0.179, 0.083]	FLOP/s per dollar
Moore's law	2 years	6.64 years	0.151 OOMs/year	FLOP/s
Huang's law	1.08 years	3.58 years	0.279 OOMs/year	FLOP/s
CPU historical (AI Impacts, 2019)	2.32 years	7.7 years	0.130 OOMs/year	FLOP/s per dollar
Bergal, 2019	4.4 years	14.7 years	0.068 OOMs/year	FLOP/s per dollar

Table 1. Summary of our findings on GPU price-performance trends and relevant trends in the existing literature with the 95% confidence intervals in square brackets.

Introduction

GPUs are the dominant computing platform for accelerating machine learning (ML) workloads, and most (if not all) of the biggest models over the last five years have been trained on GPUs or other special-purpose hardware like tensor processing units (TPUs). Price-performance improvements in underlying hardware has resulted in a rapid growth of the size of ML training runs (Sevilla et al., 2022), and has thereby centrally contributed to the recent progress in AI.

The rate at which GPUs have been improving has been analyzed previously. For example, Su et al., 2017 finds a 2.4-year doubling rate for GPU FLOP/s from 2006 to 2017. Sun et al., 2019 analyses over 4,000 GPU models and finds that FLOP/s per watt doubles around every three to four years. By contrast, some have speculated that GPU performance improvements are more rapid than the exponential improvements associated with other microprocessors like CPUs (which typically see a 2 to 3-year doubling time, see AI Impacts, 2019). Notable amongst these is the so-called Huang's Law proposed by NVIDIA CEO, Jensen Huang, according to whom GPUs see a "25x improvement every 5 years" (Mims, 2020), which would be equivalent to a ~1.1-year doubling time in performance.

There is previous work that specifically analyzes price-performance across CPUs and GPUs (summarized in Table 1). Prior estimates of the rate

of improvement vary widely (e.g. the time it takes for price-performance to increase by 10-fold ranges from ~6 to ~15 years, depending on the computing precision—see Table 2.). Due to the high variance of previous approaches and their usage of smaller datasets, we are not confident in existing estimates.[2]

Reference	Processor type	Metric	2x time	10x time	Growth rate
Bergal, 2019	GPU	FLOP/s per $ in FP32, FP16, and FP16 fused multiply-add	4.4 years (FP32) 3.0 years (FP16) 1.8 years (FP16 fused)	14.7 years (FP32) 10.0 years (FP16) 6.1 years (FP16 fused)	0.068 OOMs/year (FP32) 0.100 OOMs/year (FP16) 0.164 OOMs/year (FP16 fused)
Median Group, 2018	GPU	FLOP/s per $ in FP32	1.5 years	5.0 years	0.200 OOMs/year
Muehlhauser and Rieber, 2014	Various	MIPS/$	1.6 years	5.2 years	0.192 OOMs/year
Sandberg and Bostrom, 2008	CPU-based	MIPS/$ and FLOP/s per $	1.7 years (MIPS) 2.3 (FLOP/s)	5.6 years (MIPS) 7.7 years (FLOP/s)	0.179 OOMs/year (MIPS) 0.130 OOMs/year (FLOP/s)
Nordhaus, 2001	CPU-based	MIPS/$	1.6 years	5.3 years	0.189 OOMs/year

Table 2. Price-performance improvements found in prior work. See also AI Impacts 2015 for a more detailed overview of prior estimates.

We aim to extend the existing work with three main contributions:

1. Using a larger dataset of GPU models than has been analyzed in previous investigations that includes more recent GPU models, we produce more precise estimates of the rate of price-performance improvements for GPUs than currently exists[3]
2. We analyze multiple key subtrends for GPU price-performance improvements, such as the trends in price-performance for top-performing GPU and for GPUs commonly used for machine learning
3. We put the trends into perspective by comparing them against prior estimates, Moore's law, Huang's law, prior analyses, and public predictions on GPU performance

Dataset

We combine two existing datasets on GPU price-performance. One dataset is from the Median Group, which contains data on 223 Nvidia and AMD GPUs (Median Group, 2018). The second dataset is from Sun et al.,

2019, which contains price-performance data on 413 GPUs released by Nvidia, Intel and AMD.

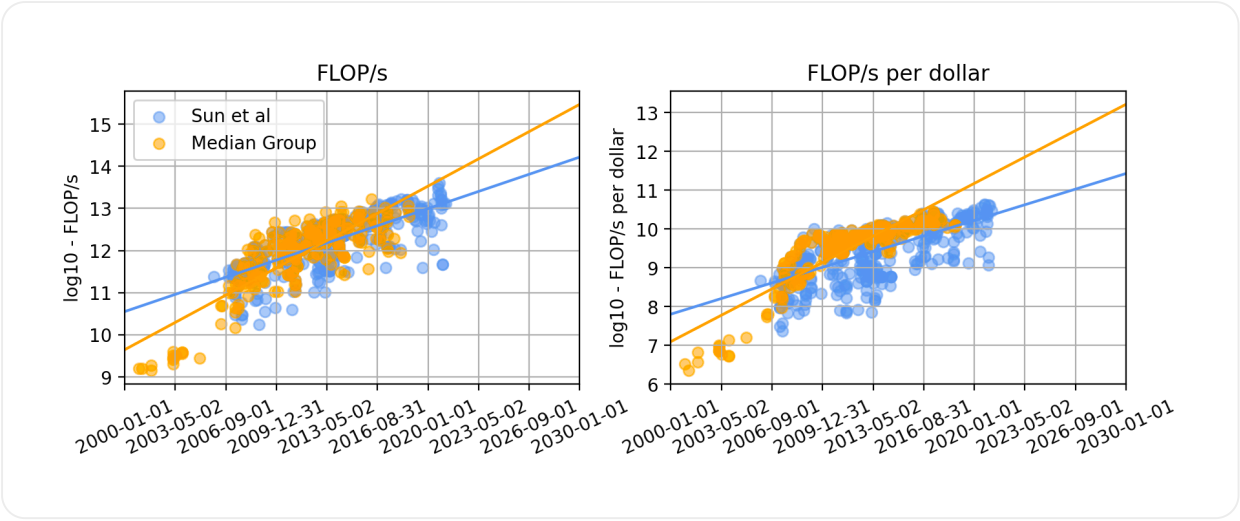


Figure 2. Plots of FLOP/s and FLOP/s per dollar for Median Group's and Sun et al., 2019's datasets.

We merged both datasets and removed duplicate observations, i.e. GPU models that were contained in both datasets. Furthermore, we removed different versions of the same product unless they had different specifications.[4]

We also decided to drop observations prior to 2006 for two main reasons: 1) it is unclear whether the we can meaningfully compare their levels of performance as these models predate innovations that enable general-purpose computing on GPUs, and 2) we were not able to validate the accuracy of the data by looking up the relevant performance details in models' data sheets. For a more detailed discussion see Appendix A.

Finally, we noticed that there is a subset of 20 GPUs for which the 16-bit performance is ~60-fold worse than its performance in 32-bit format, while for all other GPUs the 16-bit performance is at least as good as its 32-bit performance. We dropped these 16-bit performance numbers, which we think might have been erroneous.

The final dataset thus contains 470 GPUs from AMD, Intel, and Nvidia released between 2006 and 2021. We will refer to this merged dataset as "our dataset" for the rest of the report. Throughout, FLOP/s are those in 32-bit (full) precision.

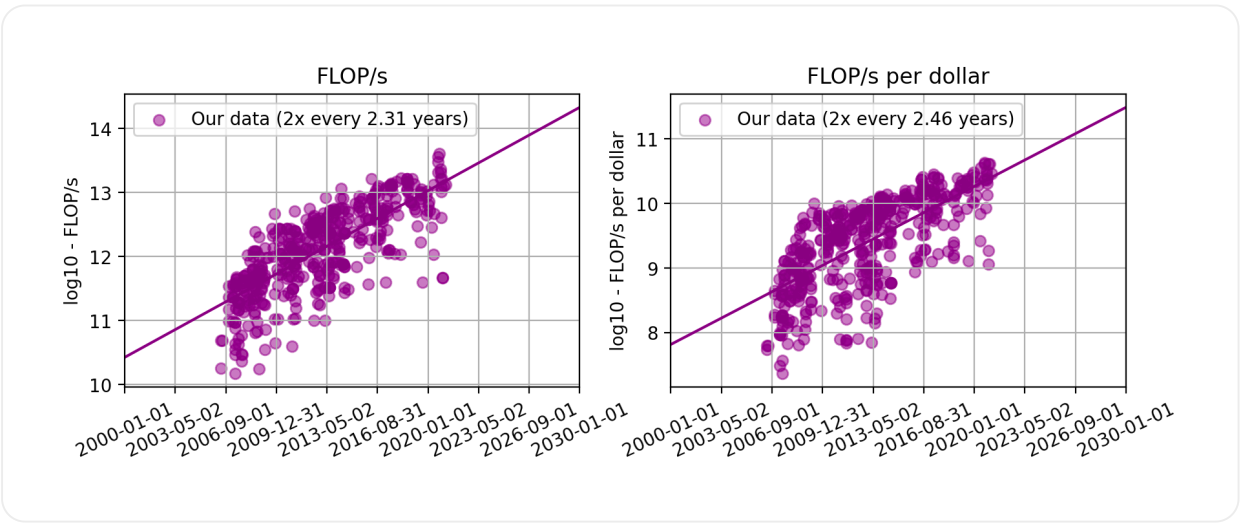


Figure 3. Plots of FLOP/s and FLOP/s per dollar for the dataset used in our analysis

Empirical analysis

In what follows, we analyze trends in price-performance, measured in FLOP/s per dollar as well as raw performance in FLOP/s for GPUs in our dataset. Our analysis considers key subsets, such as GPUs commonly used in machine learning research, as well as top-performing GPUs.[5]

Empirical trend vs. other predictions

To put our findings in context, we compare them with other proposed GPU (price) performance trends found elsewhere. These are

- Moore's law, which states that a transistor density doubled every two years. For the purpose of comparison, we take that to mean that the amount of FLOP/s also doubles every two years
- Huang's law, which describes the rate of performance improvements for GPUs. While there are multiple interpretations of Huang's law, we chose the one that reflects Huang's original wording, namely "25x improvement every 5 years"
- Historical trends in CPU price-performance, which has been found to increase by a factor of 10 every 7.7 years since 1940 (AI Impacts, 2019)
- The prediction made in Cotra 2020 of a 2.5-year doubling time in price-performance of compute relevant to machine learning training runs[6]
- Prior estimates of the rate of GPU price-performance found by Bergal, 2019

We recognize that some of these trends are not quite comparable to FLOP/s per $ (Moore's law relates to the density of circuits, Huang's law relates to theoretical performance improvements, while Cotra 2020's predictions relates to FLOP per dollar[7]). The purpose of these comparisons is just to provide a rough sense of how our estimated trends relate to relevant empirical trends and predictions.

Unless specified otherwise, we will present the results for FLOP/s per dollar. This is because we a) think that FLOP/s per dollar is the more relevant trend as argued previously and b) because there is not that much of a difference between FLOP/s and FLOP/s per dollar trends. For a detailed comparison see Appendix B.

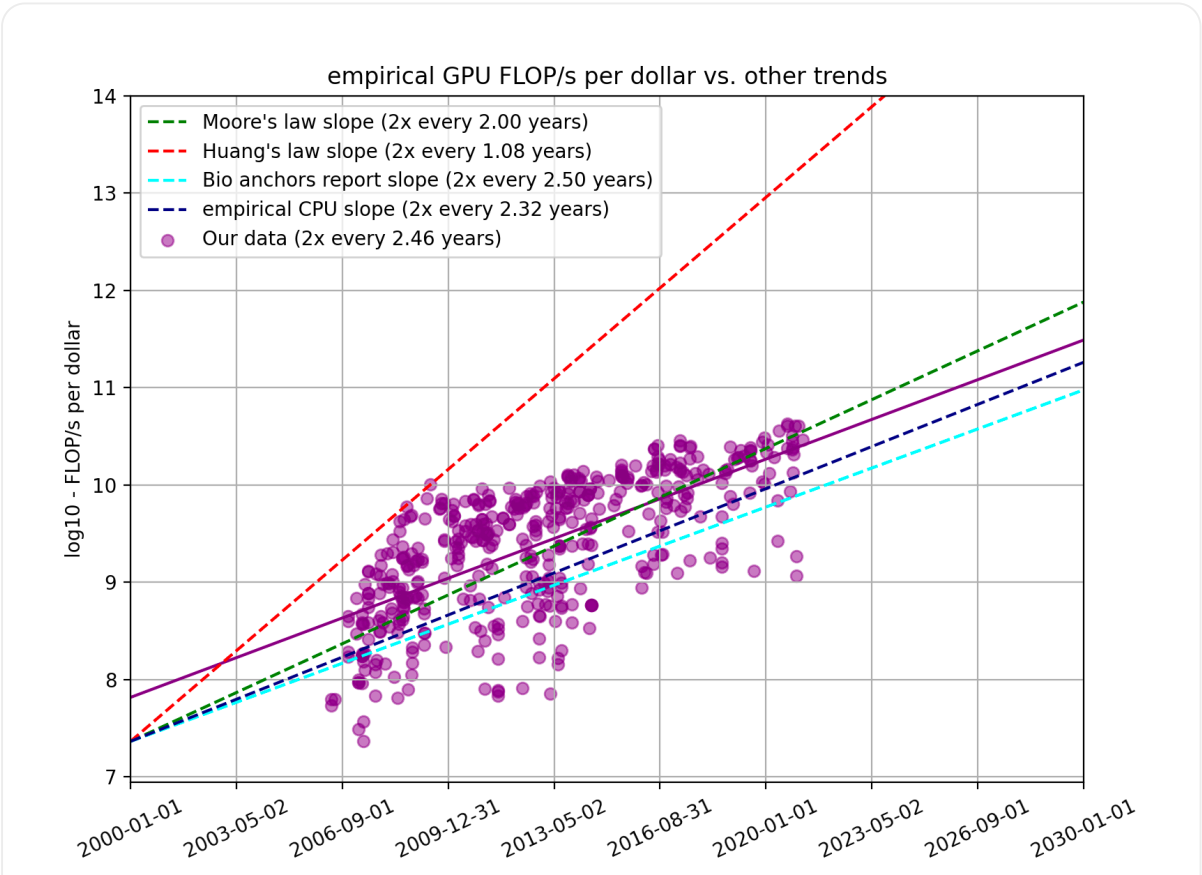


Figure 4. FLOP/s per dollar for our dataset and relevant trends found elsewhere

We find that a linear regression through all of our data shows a doubling time of 2.46 years (95% CI: 2.24 to 2.72 years). This is very well in line with the slope of 2.5 used in Cotra 2020. We can also see that Huang's law is not a good fit for the entire trend and is strongly overstated.

Trends across precision for floating formats

Half-precision computing (FP16) and mixed-precision computing (usually FP16 and FP32) are now commonly used for deep learning. In our dataset, we had 91 GPUs for which we had both price and FP16 performance numbers.[8]

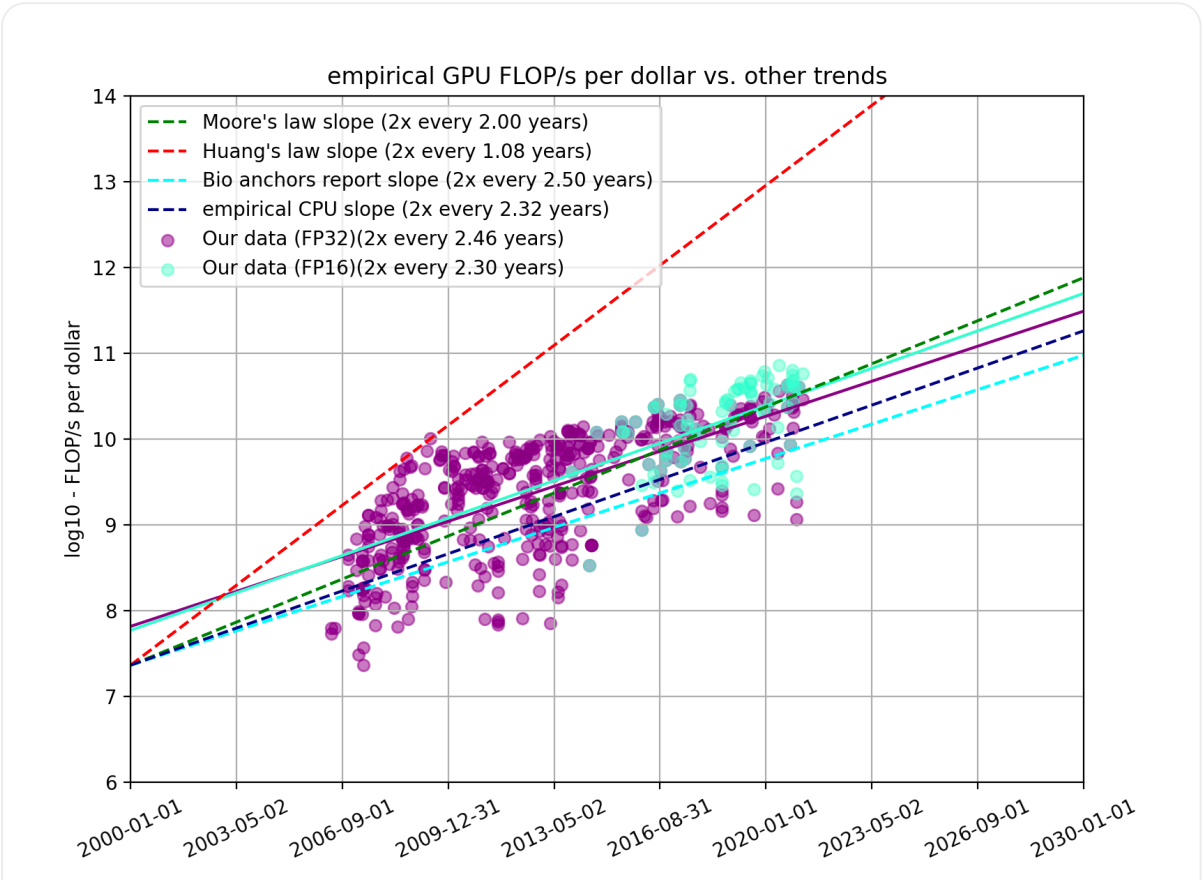


Figure 5. FLOP/s per dollar for FP32 and FP16 performance

We find that the price-performance doubling time in FP16 was 2.32 years (95% CI: 1.69 years, 3.62 years). This was not significantly different from the slope for the doubling time for price-performance in FP32, suggesting that price-performance improvements in FP16 and FP32 are likely to be similar. This stands in contrast to the findings of Bergal, 2019, which finds a 1.8-year doubling time for FP16 FMA.[9] In what follows, we decide to focus on price-performance in FP32 as we do not find a statistically significant difference between the two trends, and we therefore choose to analyze the models for which we have the most data on.

Trends of GPUs used in ML

The vast majority of all ML training is done on a very small number of different models of GPUs. From a previous publication where we looked at 75 papers that present milestone ML models, we collected a total of 42 distinct models of GPUs commonly used to train ML systems. In total, we found 26 of these 42 GPUs in our dataset on GPUs.

We find that the price-performance of GPUs used in ML improves faster than the typical GPU. We find that FLOP/s per dollar for ML GPUs double every 2.07 years (95% CI: 1.54 to 3.13 years) compared to 2.46 years for all

GPUs. This is not significantly different from the slope for the doubling time for price-performance for all GPUs.

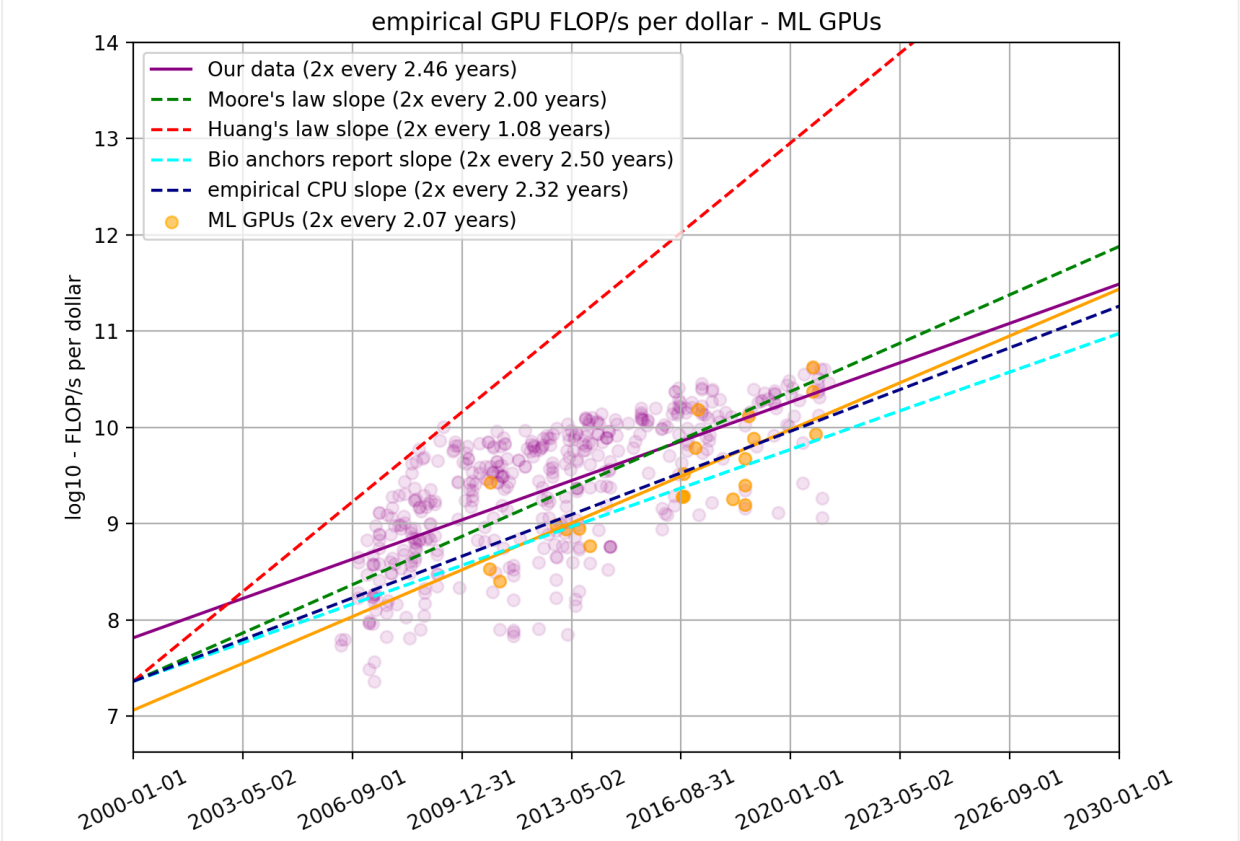


Figure 6. FLOP/s per dollar for our dataset and separately for GPU models commonly used in ML research compared to relevant trends found elsewhere

Furthermore, the latest ML GPUs tend to be among the GPUs with high price-performance, whereas the older ones are more middle of the pack.

Moreover, when looking through the FLOP/s lens, it becomes even more clear that the latest ML experiments use the most powerful GPUs. We think that shows the increased importance of GPUs for modern ML. Once again, the ML GPUs show a steeper slope than the general trend (doubling time of 2.00 years compared to 2.31 years for all GPUs).

Our higher point estimate for the rate of performance improvements amongst GPUs used for ML research could be explained by relevant labs spending more resources on procuring top GPUs over time. If this were the case, this would reflect merely a change in investment decisions by relevant research labs and not a faster-than-usual rate of improvement of the underlying hardware amongst the relevant GPUs suitable for ML workloads. Given this, and because our estimates for the entire GPUs is not statistically significantly different, we expect that the ~2.5 year doubling time to be a more reliable estimate of the underlying rate of hardware price-performance improvements.

Trend of top-performing GPUs

As we saw in the previous section, the latest ML models tend to be trained on state-of-the-art GPUs. Therefore, looking at the trend of top-performing GPUs might be a good indicator for ML capabilities in the future. Note, that this does not imply that we think that GPU performance will grow linearly. We will publish more detailed thoughts on predictions from this data in a second piece.

Here, we select the subset of GPUs that had the highest FLOP/s per dollar values during each month. For this subset of models, we find a doubling

time of 2.95 years (95% CI: 2.54 to 3.52 years), which is statistically significantly longer than the typical doubling time.

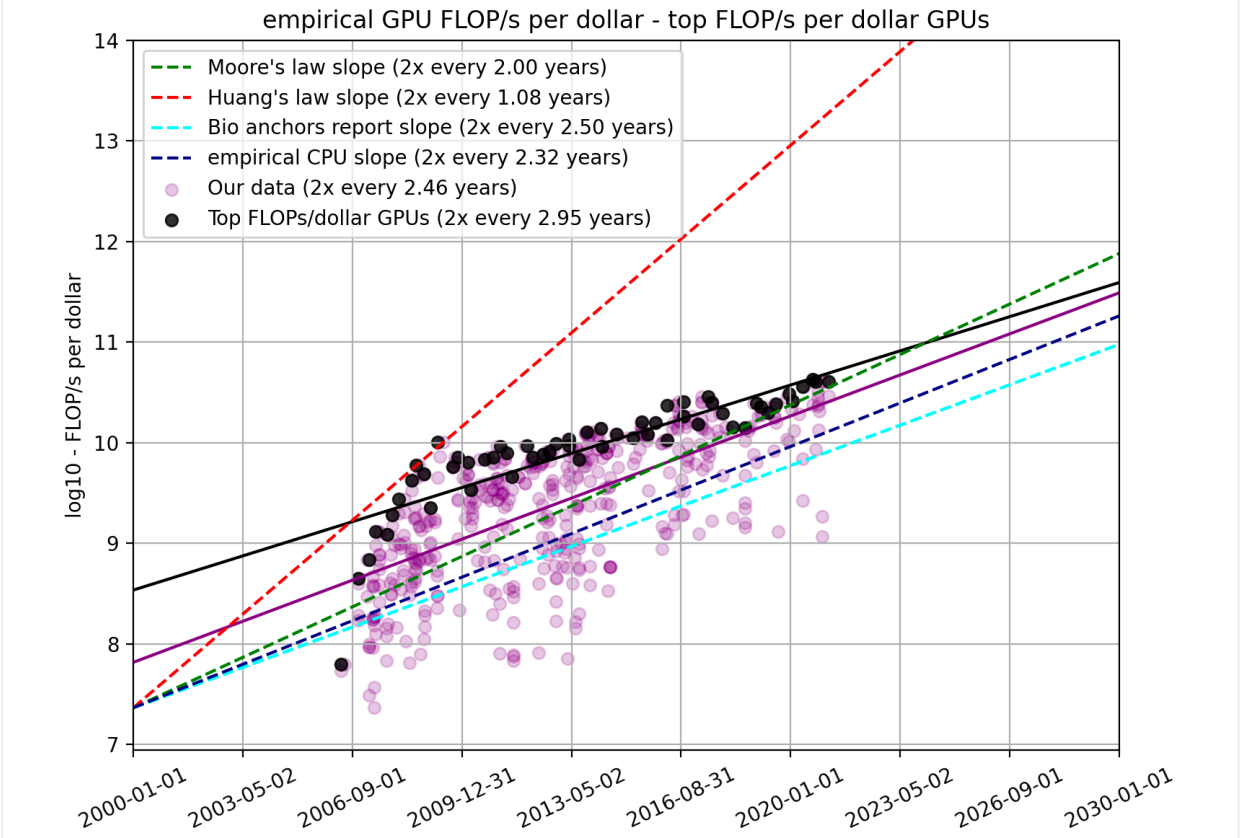


Figure 7. FLOP/s per dollar for our dataset and separately for top-performing GPUs compared to relevant trends found elsewhere

All trends (table & figure)

To compare all the trends we highlighted above and the ones you can find in the appendix, we collected all trends, and for each, report the associated time it takes to increase 2x and 10x.

Trend	Original presentation	2x time	10x time	Growth rate	Reference
Moore's law	2x every 2 years	2 years	6.64 years	0.151 OOMs/year	FLOP/s
Huang's law	25x every 5 years	1.08 years	3.58 years	0.279 OOMs/year	FLOP/s
Biological anchors report (Cotra, 2020)	2x every 2.5 years	2.5 years	8.30 years	0.120 OOMs/year	FLOP/s per dollar
CPU historical (AI Impacts, 2019)	10x every 7.7 years	2.32 years	7.7 years	0.130 OOMs/year	FLOP/s per dollar
Median Group, 2018	2x every 1.5 years	1.5 years	5.0 years	0.200 OOMs/year	FLOP/s per dollar
Our data (n=470)	-	2.46 years [2.24, 2.72]	8.17 years [7.45, 9.04]	0.122 OOMs/year [0.134, 0.111]	FLOP/s per dollar

Trend	Original presentation	2x time	10x time	Growth rate	Reference
Our data (n=470)	-	2.31 years [2.14, 2.51]	7.68 years [7.12, 8.33]	0.130 OOMs/year [0.140, 0.120]	FLOP/s
Our data FP16 (n=91)	-	2.30 years [1.69, 3.62]	7.64 years [5.60, 12.03]	0.131 OOMs/year [0.179, 0.083]	FLOP/s per dollar
Our data FP16 (n=91)	-	2.91 years [1.94, 5.83]	9.68 years [6.45, 19.35]	0.103 OOMs/year [0.155, 0.052]	FLOP/s
ML GPUs (n=26)	-	2.07 years [1.54, 3.13]	6.86 years [5.12, 10.39]	0.146 OOMs/year [0.195, 0.096]	FLOP/s per dollar
ML GPUs (n=26)	-	2.00 years [1.69, 2.43]	6.63 years [5.63, 8.07]	0.151 OOMs/year [0.178, 0.124]	FLOP/s
Top GPUs (n=57)	-	2.95 years [2.54, 3.52]	9.81 years [8.45, 11.71]	0.102 OOMs/year [0.118, 0.085]	FLOP/s per dollar
Top GPUs (n=57)	-	2.69 years [2.40, 3.30]	8.92 years [7.99, 10.95]	0.112 OOMs/year [0.125, 0.091]	FLOP/s

Table 3. Summary of our findings on GPU price-performance trends and relevant trends in the existing literature. 95% confidence intervals are displayed in square brackets.

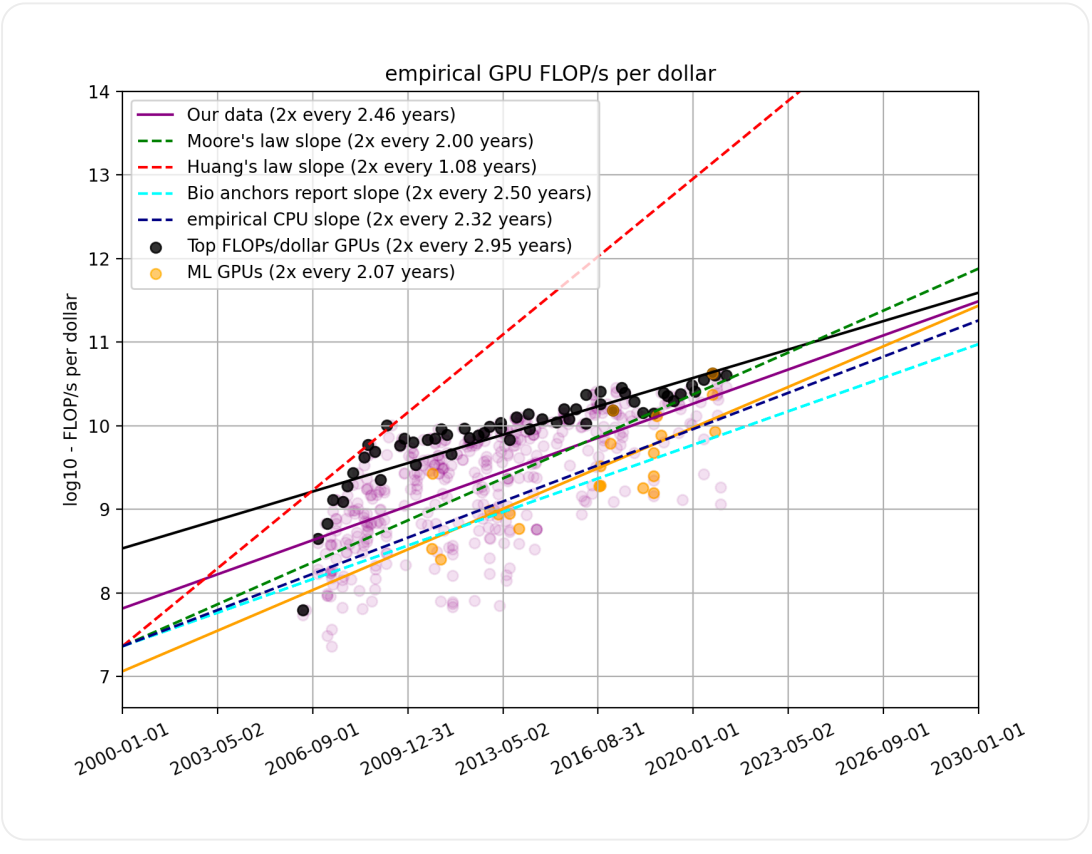


Figure 8. FLOP/s per dollar for our dataset and various subgroups compared to relevant trends found elsewhere

Conclusion

We find that the trend of all data shows a doubling time of 2.46 years, the trend implied by GPUs used in ML shows a doubling time of 2.07 years and the trend implied by every month's top GPU shows a doubling time of 2.95. We think that a doubling time below 2 years or above 3 years is implausible given the data. Furthermore, we think that part of the trend in ML GPUs can be explained by ML prioritizing better GPUs rather than actual hardware advances. We, therefore, think that a doubling time of 2 years is too aggressive and ~2.5 years accurately describes the doubling time of price-performance for GPUs over the past 15 years.

Appendix A - Dropping data before 2006

On balance, we felt like the arguments for keeping the data were weaker than for removing them. In short, it's very unclear whether pre-2006 data are measured in a comparable way and whether pre-2006 GPUs are even really comparable to post-2006 GPUs.

Arguments for including pre-2006 data:

1. The median group provides the data and somehow got a hold of the estimated FLOP/s

Arguments against including pre-2006 data:

1. The data just looks fishy in the graph
2. Alyssa Vance pointed out in a comment that general-purpose GPUs have not been developed until 2005 and cuda didn't exist until 2007. This means we are talking about some very different GPUs that have little to do with the GPUs we are talking about today.
3. The measure of "theoretical performance" is unavailable for most pre-2006 GPUs. It is therefore unclear how this data was initially collected.
    1. List of pre-2006 GPUs:
        1. GeForce 2 GTS Pro (no FLOP/s)
        2. GeForce 3 (no FLOP/s)
        3. GeForce 3 Ti500 (no FLOP/s)
        4. GeForce 3 Ti200 (no FLOP/s)
        5. GeForce FX 5200 (no FLOP/s)
        6. GeForce FX 5200 Ultra (no FLOP/s)
        7. GeForce FX 5600 Ultra (no FLOP/s)
        8. GeForce FX 5800 (no FLOP/s)
        9. GeForce FX 5800 Ultra (no FLOP/s)
        10. GeForce FX 5900 / 5900 XT / 5900 ZT (no FLOP/s)
        11. GeForce FX 5700 Ultra (no FLOP/s)
        12. GeForce FX 5950 Ultra (no FLOP/s)
        13. GeForce FX 5900 Ultra (no FLOP/s)
        14. GeForce 6200 TurboCache 64-TC/256MB (no FLOP/s)
        15. ATI Xbox 360 GPU 90nm (YES FLOP/s)
    2. After 2006 it looks like the GPU model cards contain this FP32 FLOP/s performance. Here are some samples
        1. GeForce 7900 GX2 (2006, no FLOP/s)
        2. GeForce 8800 GTX (2006, YES FLOP/s)
        3. GeForce 8800 GTS 640 (2006, YES FLOP/s)
        4. GeForce 8600 GT (2007, YES FLOP/s)

5. GeForce 8500 GT (2007, YES FLOP/s)
6. Radeon HD 2900 XT (2007, YES FLOP/s)
7. Trend continues afterward

Appendix B - Robustness check for FLOP/s

In our dataset, we only look at GPUs for which we have the FLOP/s and price information since we are interested in performance and price. However, there are many more GPUs that have performance information than ones for which we have both performance and price. We find 1848 data points for which we have FLOP/s data. To make sure that there is no selection effect, we also analyze the trend from “just FLOP/s”.

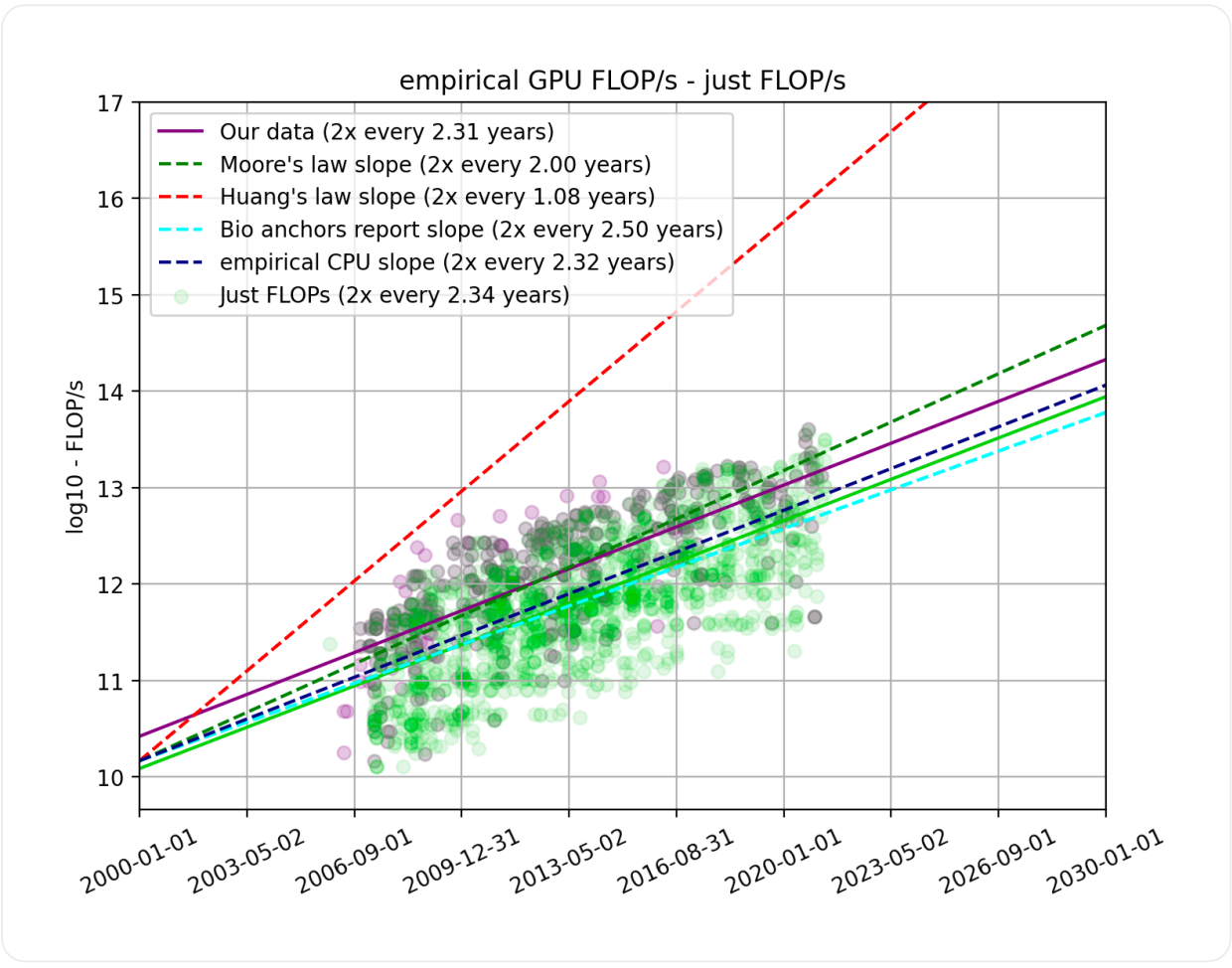


Figure 9. Empirical FLOP/s with all GPUs that we have FLOP/s information for

We find that it pretty much aligns exactly with what we see from our previous selection and therefore preliminary conclude that there is no reason to discard our previous findings. We additionally see that the GPUs for which we have price data tend to be the ones with higher FLOP/s values. We speculate that more powerful GPUs are used more often and therefore have higher availability of price information.

More FLOP/s plots

For all the plots used in the paper, there is also a version in which we only look at FLOP/s information. Note, that this is not the “just-FLOP/s” data from the previous section. Rather it is the same dataset as used in the main text but we didn’t divide it by price.

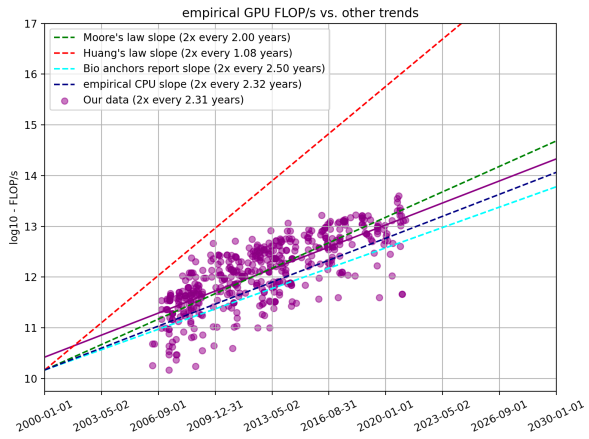


Figure 10. Empirical FLOP/s for our dataset

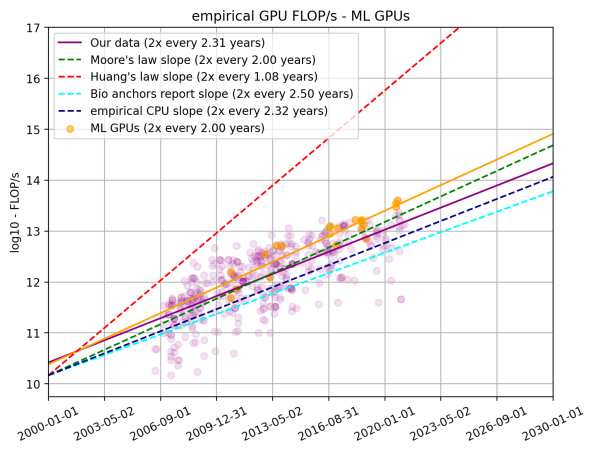


Figure 11. Empirical FLOP/s for our dataset with subset of GPUs used for ML

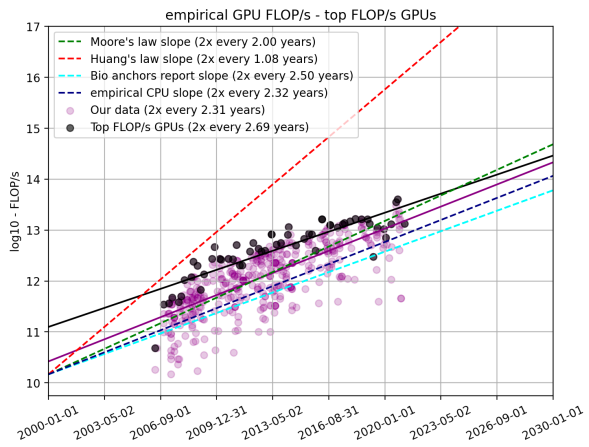


Figure 12. Empirical FLOP/s for the GPUs with the highest FLOP/s value for every month

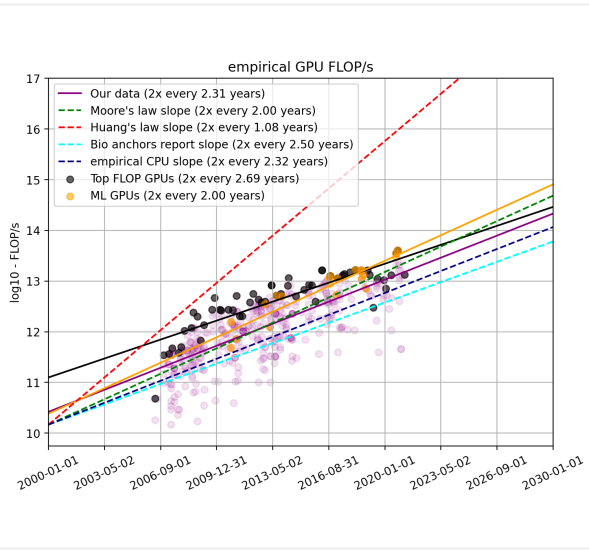


Figure 13. Empirical FLOP/s for top FLOP GPUs and ML GPUs combined

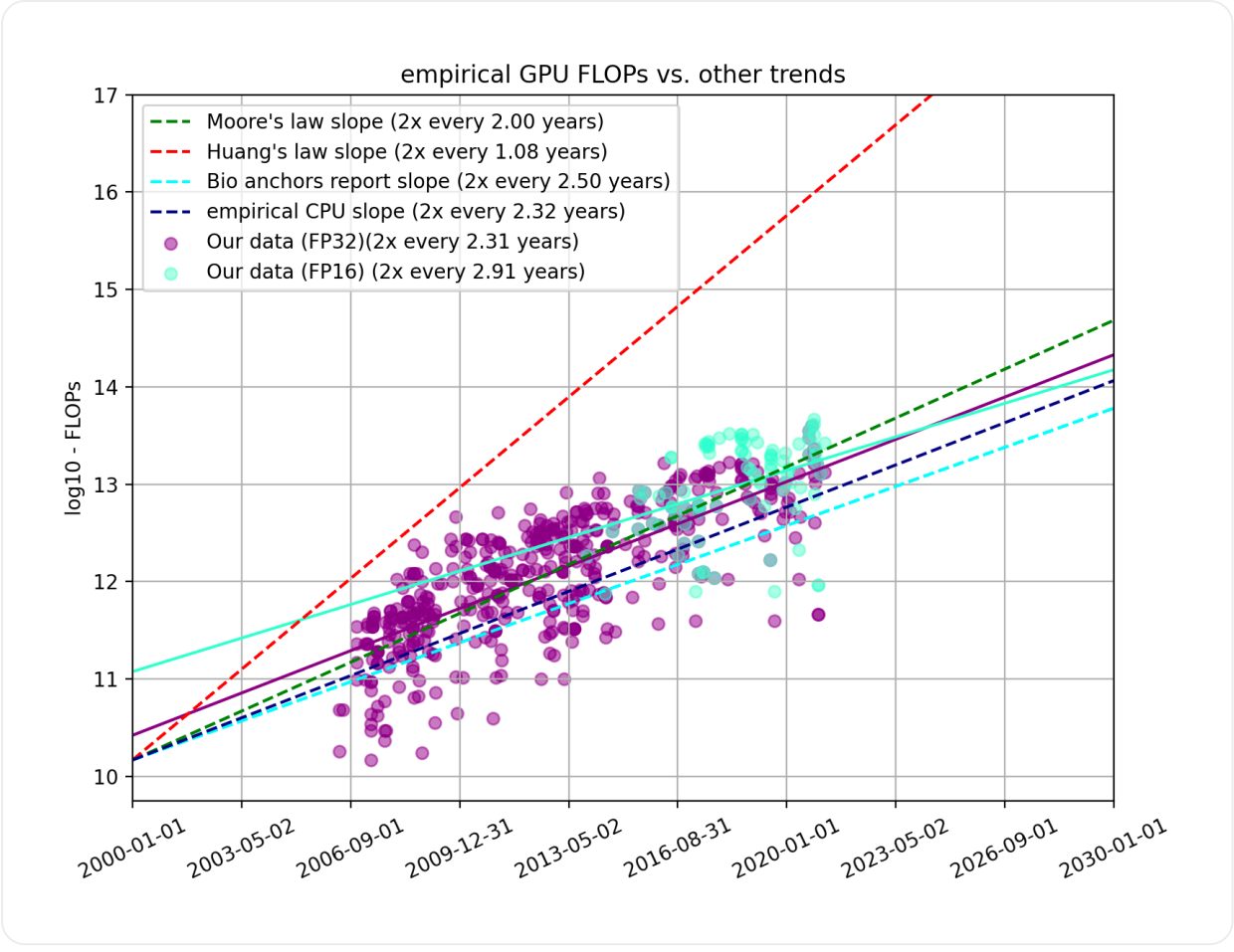


Figure 14. Empirical FLOP/s for FP16 and FP32

We did not include these in the main text because they show very similar slopes as the FLOP/s per dollar and we think FLOP/s per dollar is the more important metric.

---

1. This work *does not* attempt to project future GPU price-performance but merely to take stock of the recent historical trend. In future work, we intend to investigate what GPU price-performance trends informs us about historical growth in hardware spending in machine learning and future large machine learning training runs. ↩

2. Moreover, we had discovered issues in estimates from a prior investigation by the Median Group (2018), which we have since pointed out to them, and which they have since corrected. ↩

3. We unfortunately can't publish our data but the code that generated all the figures can be found here. ↩

4. For example, consider NVIDIA Tesla K40c and NVIDIA Tesla K40d to be the same models, as these have essentially identical specifications. ↩

5. We focus primarily on these metrics as we are mostly interested in questions related to the amount of compute that might be deployed for large AI experiments. While there are other metrics that might be of interest (such as energy efficiency), we do not consider these here as they relate less directly to the questions motivating our work, and because these have been analyzed in prior work, notably in Sun et al., 2019. ↩

6. This is our interpretation of section 4 of her draft report, where she writes "I also assume that effective FLOP/s per dollar is doubling roughly once every 2.5 years around 2025. This is slower than Moore's law (which posits a ~1-2 year doubling time and described growth reasonably well until the mid-2000s) but faster than growth in effective FLOP/s per dollar from ~2008 to 2018 (a doubling time of ~3-4 years)" ↩

7. FLOP per dollar in Cotra 2020 refers to the total amount of computation that can be done per dollar. ↩

8. We also took an initial look at FP64 performance but decided not to include the analysis because FP64 performance seems to be much lower than FP32 performance for newer GPUs. We interpret this as GPU companies deprioritizing FP64 in favor of FP32 and FP16. ↩

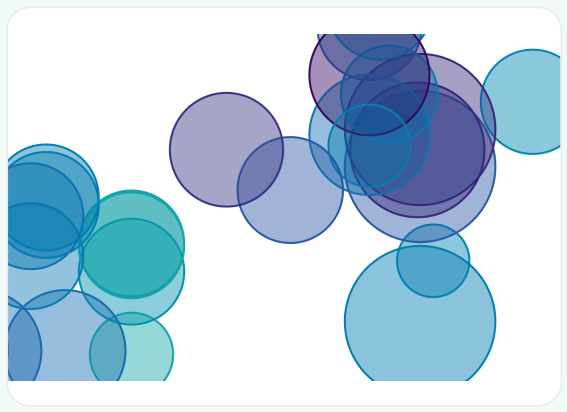
9. This is, we suspect, due to the fact that their approach involves analyzing the moving optima—which, in their case, involves analyzing 9 data points, which we think is insufficient to yield confidence in their point estimate. ↩

About the authors

Former employee
Marius Hobbhahn builds models for AI timelines and takeoff using historical trends and his best understanding of the future.

Tamay Besiroglu is the associate director at Epoch AI. His work focuses on the economics of computing and big-picture trends in machine learning. Previously, he was a researcher at the Future Tech Lab at MIT, led strategy for Metaculus, consulted for the UK Government, and worked at the Future of Humanity Institute.

Related work



DATA

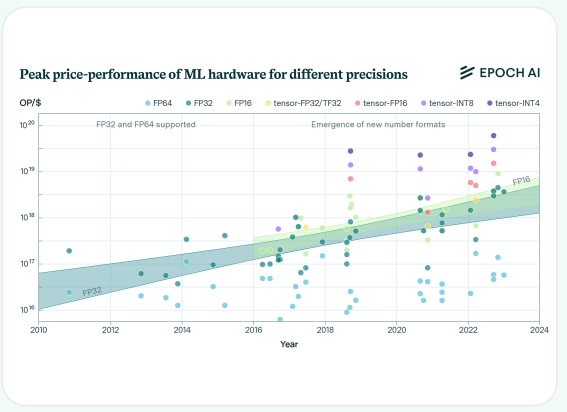
Data on Machine Learning Hardware

We present key data on over 110 AI accelerators, such as graphics processing units (GPUs) and tensor processing units (TPUs), used to develop and deploy machine learning models in the deep learning era.

PAPER · 7 MIN READ

Compute Trends Across Three Eras of Machine Learning

We've compiled a comprehensive dataset of the training compute of AI models, providing key insights into AI development.



REPORT · 28 MIN READ

Trends in Machine Learning Hardware

FLOP/s performance in 47 ML hardware accelerators doubled every 2.3 years. Switching from FP32 to tensor-FP16 led to a further 10x performance increase.

Updated December 23, 2024

Feb 16, 2022 · Updated May 02, 2022 · By Jaime Sevilla, Lennart Heim, Anson Ho, Tamay Besiroglu, Marius Hobbhahn and Pablo Villalobos

Nov 09, 2023 · By Marius Hobbhahn, Lennart Heim and Gökçe Aydos

Excited about our work? Talk to us Support our research

EPOCH AI

Sign up for our newsletters to receive the latest updates on our research.

- [x] The Epoch Brief

Our most recent research and publications.

- [x] Gradient Updates

Weekly analysis of the latest in AI news and developments.

Enter your email

Sign up

RESEARCH

Publications

Data on AI

Key Trends & Insights

Gradient Updates

PROJECTS

FrontierMath

Distributed Training

COMPANY

About Us

About Our Research

Careers

Support Us

Contact

Epoch AI is fiscally sponsored by Rethink Priorities.

Privacy Notice