Why Professor Barnes Won't Let You Shrink Your Laptop

The thermodynamics behind some of Apple's worst design decisions

Leif Sieben Laptops suck. Humanity has made incredible advances since the invention of the transistor nearly a century ago, and yet one cannot buy a decent new-ish model for less than about a month's worth of rent. A Casio watch calculator from the 1980s has more computing power than the combined efforts of the Apollo 11 space programme but even a used MacBook still easily resells for around CHF 800. Capitalism is a lie and big tech is out to get you!

> "The products suck. There's no sex in them anymore."
>
> Steve Jobs about the Apple product line in 1997

To most people, the incredible technological progress computer hardware has made is nearly invisible. This probably is mostly due to the fact that computational demands of software have scaled with about the same speed as our hardware capabilities. Unfortunately, it is also true that the app on your iPhone that emulates a Casio watch calculator is much more computationally intensive than it was for humans to reach the moon. The simple but unfortunate truth is that **most consumer software today is horrendously inefficient**. Excel won't be fast even when you run it on a supercomputer.

Gordon Moore, the co-founder and CEO of Intel, suggested in 1965 that the number of components per integrated circuit would double every year. **Moore's law predicts that the power of chips grows exponentially over time** which means that to achieve the same amount of computing power we can use ever smaller chips. Surprisingly, Moore's law has held for over 65 years. Even today, many still consider it valid, although recently some naysayers have emerged, such as Jensen Huang, CEO of the chip maker Nvidia.[4]

Nonetheless, chips are now even more in demand than they were 65 years ago. The **recent surge of interest in AI** has seen the stock price of Nvidia more than triple and big tech firms are still scrambling to get their hands on **ever more and ever more powerful processors**. Sam Altman, the founder who brought you ChatGPT, recently made headlines for trying to solve the world's chip shortage by asking for a quixotic seven trillion dollars in venture capital. To put this into context, the en-

tire American federal government is set to spend around 7.2 trillion dollars this year. Back in 2021, Altman posited that with the coming AI revolution, we will live in a world with "Moore's law for everything".[1] Just like with CPU power, technological progress grows strictly exponentially and will hence inevitably reach a point in which nearly all kinds of labour will be automated, or so the thesis goes. As Altman puts it:

> *AI will lower the cost of goods and services, because labour is the driving cost at many levels of the supply chain. If robots can build a house on land you already own from natural resources mined and refined onsite, using solar power, the cost of building that house is close to the cost to rent the robots. And if those robots are made by other robots, the cost of renting them will be much lower than when humans made them.*

If you buy this argument, **seven trillion dollars in exchange for what essentially amounts to an infinite money glitch** still offers a very attractive return on investment (ROI) of $\lim_{t\to\infty}(ROI) \to \infty$. What separates us from this world of incredible abundance, according to Altman, are just a few steps on this logarithmic y-axis. A growth that is currently limited by two things: firstly, by a **lack of sufficiently powerful chips** and secondly, by the **high cost of energy** needed to train today's cutting edge AI models. Logically, Altman's other big investment is in nuclear fusion. Which, if successful, would provide the chips from his seven trillion dollar factory with an essentially endless source of energy.

It seems likely that we will hit some hard limit sooner. Altman requires that nearly all goods and services will be cheaper if produced, planned, and administered autonomously for his argument to work. Instead of everyone becoming richer, **everything will simply become cheaper**. The end effect is the same: we can all afford more stuff, i.e. wealth has increased.

Moore's Law For Law

It is easy to dismiss this argument as utopian. It genuinely is. In some sectors, however, this could very well apply. The reason why consulting a (good) attorney is so expensive comes down to the fact that there are only so many (good) attorneys to go around. It takes years of education for a new attorney to enter practice, and therefore the time of a (good) attorney is limited. As a client, you pay for this limited supply of time. The better the attorney, the more limited their time. An AI model, however, is not time-limited, it has a single (huge) expense in training but then can give out legal advice 24/7 without any additional costs other than running a server farm. Such an AI attorney can have arbitrarily many parallel chats, whereas (good) attorneys generally do not hold two phone calls in parallel. **AI is inherently well scalable.**

The cost of consulting this AI attorney therefore realistically should tend to zero. But one reason why attorneys' time is in short supply is not just because it takes a long time to attain this knowledge, it is also a very much intentional shortage. Attorneys must be admitted by the bar, the professional organisation governing attorneys, before being allowed to argue a case before court. The supply of attorneys is therefore precisely limited by how many attorneys the bar admits. More likely thus, **legal advice will experience a growth that is more *sigmoidal* than exponential**: If the bar allows an AI attorney to effectively replace a human one, the cost of legal advice will be much lower, else, you still need to consult a human anytime you are in serious legal trouble. Given that the bar itself is run by other lawyers, who presumably would also risk losing their job, we might be stuck on the low end of this sigmoid for quite some time.

No Such Thing As Free Chips For Lunch

So why are laptops so damn expensive and will AI make laptops free? The fact is that **laptops got cheaper early on but have essentially stagnated around the same price** since 2002, see Fig. 7.1 Even when analysing CPU performance per dollar spent, see Fig. 7.2, the same trend emerges. Prior to 2002, the growth of CPU per dollar as well as memory per dollar was indeed strictly exponential. But much sooner than Moore's law itself, the CPU power relative to the cost of the laptop began to stagnate.

The ideal laptop would have an extremely powerful CPU and very large storage (what we may call a large *compute*), it would be very light and very cheap. This however is an impossibility triangle: **You want to have the highest compute at the lowest price with minimal weight. Pick two.**

However, these same constraints do not apply to desktop PCs. In fact, in desktop computing, Moore's law still provides the next-generation chips with ever more compute or the same compute at ever lower prices. Nonetheless, Moore's law will probably soon run out of steam in desktop computing as well: transistors can only get so small. Today's best chips carry around five trillion 2 nm small transistors each. But why did Moore's law hold up so much longer in desktops than in laptops?

Firstly, **laptops are inherently constrained by weight**, which is mostly the same as saying that they are space constrained. If one wants to upgrade the storage in a laptop,

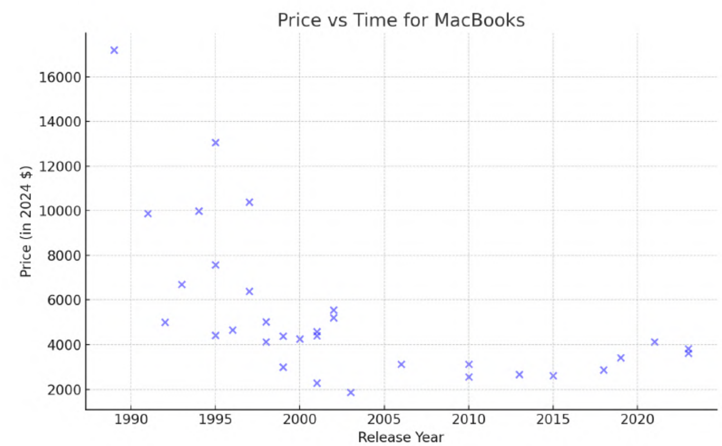

Figure 7.1: Pricing evolution for the top-line MacBook models each year. Note that the prices were inflation adjusted to their 2024 level.

one needs to buy a better memory disk. In a desktop, however, one could simply buy the same memory disk a second time. This is the difference between *scaling up* (buying a larger disk at the same weight which is thus much more expensive) and *scaling out* (buying a second disk to double the storage thus doubling the price tag). **Scaling up is much more difficult to do than scaling out**, the price difference between an 8 TB SSD and a 1 TB SSD is closer to a factor of twelve than a factor of eight. Scaling out just means scaling linearly, the cost of buying eight times the same SSD is simply that price added up eight times.

Given that the weight and the price of laptops have essentially remained constant over time means that at best, laptops can scale with the speed of Moore's law. We can only ever include one chip at a a time, and hence the speed of the laptop will be defined by whatever the fastest chip available at that price is. So why is their growth still so underwhelming?

In fact, laptops suffer from an even worse problem: **unlike in desktops, there is no space for proper cooling.** Performing computations requires letting current flow through a circuit with resistance, which releases heat. Surprisingly, having more transistors does not automatically mean releasing more heat, which is the result of an

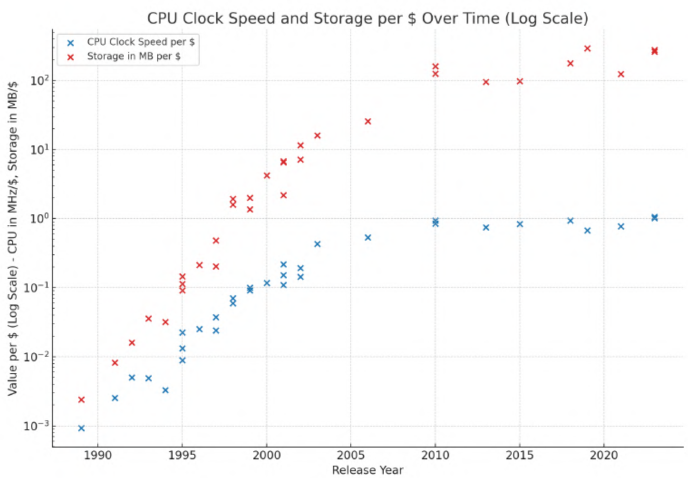

Figure 7.2: CPU speed and storage per dollar spent for a cross section of MacBook models over time. Note that the dollar values are inflation adjusted to their 2024 level.

equally important, but much less known law called *Dennard scaling*. The argument runs like this:

- As we reduce the size of the chip by a factor x, the distance electricity has to travel scales by x and the area of the chip by x^2.
- We know that the power of a chip is CV^2f.
- The capacitance C scales as the area over distance, or with $x^2/x = x$.
- The voltage scales as field times distance, and thus with x.
- The frequency scales with the inverse distance, or with $1/x$.
- Hence, the active power scales with x^2, and thus the power per area actually remains roughly constant.

Birth Of The Cool

Unfortunately, Dennard's law is long dead. It broke down around 2006, meaning that more powerful chips today also create increasing amounts of heat. In fact, the performance data for CPU's today are given at the thermal power limit where the cooling system is just barely able to keep the CPU below melting point. This has become

such a problem that chip designers now implement *dark silicon* or parts of the chip that cannot be turned on because the chip needs to remain below a certain thermal power threshold.

In desktop computers, as well as in server farms, this is compensated by more extensive cooling. It is no coincidence that Europe's fastest supercomputer is situated in Finland, a place with copious amounts of cold air. Laptops, however, simply require better chips with fewer thermal losses and must **compromise between a heavier laptop with better cooling and a less powerful but lighter laptop without it**. Apple has certainly opted for the latter option, where most MacBooks today do not have any cooling system other than the natural thermal conduction of the metal casing. Without Dennard's law, miniaturising the chip further, however, will always lead to more thermal losses, which either requires a larger metal surface for conduction or more cooling systems, in both cases the chip might have gotten smaller but not the laptop overall.

The current most powerful MacBook actually shrinks in comparison to any, much cheaper, desktop PC. Both CPU power and storage per dollar are more than twice as attractive in the desktop PC than in the MacBook. But this comes at the cost of portability: in terms of compute per kilogram, the MacBook easily beats out its desktop rival. The bottom line is that **laptops and desktops are designed with fundamentally different use-cases** (which thus have fundamentally different constraints) in mind: a very light laptop for all things portable, a very powerful desktop for all things computationally intensive and a very expensive server (accessed through a very cheap laptop) for all things truly computationally intensive. Such is a user which truly understands the laws of scaling.

Model	Type	Price [$]	CPU [MHz]	GPU [GFLOPS]	Storage [TB]	Weight [kg]
Macbook Pro M3 Max	Laptop	3823	4050	13600	1000	1.62
Mac Pro M2 Ultra	Desktop	8499	3680	4600	1000	4.6
Assembled PC	Desktop	1836	4700	29150	4000	11.7

Model	CPU/$	CPU/$	Storage/$	CPU/kg	GPU/kg	Storage/kg
Macbook Pro M3 Max	1.06	3.557	0.262	2500	8395	617
Mac Pro M2 Ultra	0.433	0.541	0.118	800	1000	217
Assembled PC	2.56	15.9	2.179	402	2491	342

Most people, however, treat the choice between a laptop and a desktop as mostly one of personal convenience or taste. An impression that is probably due to the ridiculous pricing of Apple's current desktop computer, which seems to have taken a look at our impossibility triangle and said "Could I have none of the three please?" You get less CPU per dollar than even in the MacBook and less CPU per kilogram than even in the PC. Truly the worst of both worlds.

But unless you really need to perform computationally intensive tasks (which cannot be done on a server) and you need to be able to do them from anywhere – video editing comes to mind – a laptop will always underperform a desktop simply due to thermodynamics. Indeed, **there is a hard lower bound of how much heat any computation will dissipate**, even if we were able to magically resurrect Dennard scaling.

Landau(er) O(ooh)

In 1961, the German physicist Rolf Landauer (1927–1999) derived the minimal energy required to erase one bit of information.[3]

$$E \geq k_B T \ln(2)$$

What we mean by erasing information is that we take a bit, which is either set to 0 or 1, and destroy it physically, for example. Whereas before, the system could be in either a *on* or *off* state, it now does not have a state at all and we cannot recover its original state. Hence, **erasing is a logically and physically irreversible process**. What destroying this bit also means is that the system now has fewer microstates: assume a system with two bits and thus four microstates $\{(1,1),(1,0),(0,1),(0,0)\}$ and an entropy of $S = k_B \ln(4)$. After erasing one bit, the system only has 2 microstates $\{(1),(0)\}$ and an entropy of $S = k_B \ln(2)$. Because $\ln(a \cdot b) = \ln(a) + \ln(b)$, we can also say that we have subtracted an amount of entropy equal to $k_B T \ln(2)$.

But **the entropy of the universe can never decrease**, and so if the entropy is lost in the system, it must increase in its environment by at least the same amount. This will happen through the system dissipating heat. While laptops today actually operate far from this lower bound, it is true that even for a chip made of superconducting transistors, which should have very little other thermal losses, there will be some amount of heat being dissipated simply due to the second law of thermodynamics.

The Rent Is Too Damn High

Upon close enough inspection, nothing in life truly scales limitless for ever. Even in a world in which AI attorneys will be arguing cases before an AI judge, our court system would probably still be hopelessly clogged up. Not because the speed-up was not truly exponential, but because the number of cases grew exponentially with it. Just ask any civil engineer: building a new lane on the highway mostly does not ease traffic jams, it simply attracts more cars.

There is always a bottleneck. Building really great things is also really damn hard and while many things in our current world could (and probably should) be accelerated, this does by no means imply that we live in a universe with infinite resources. Sam Altman notably excludes the price of land, for example, from his exponential scaling for everything: There is only so much real estate to go around and the wealthier people become the more of it they want, which means prices going up. In a world with Moore's law for everything, you could expect the same to be true for water and air pretty soon as well. This presumably is the reason why a group of ultra-wealthy tech investors (excluding Sam Altman) have recently gobbled up a plot of California bigger than the size of San Francisco, in the process turning local farmers into millionaires overnight.[2] Their plan being the construction of a new city to ease Silicon Valley's housing crisis. Perhaps, some bottlenecks are better left in place.

Bibliography

[1] Altman, S. (2021, March 16). *Moore's law for everything*. Sam Altman. https://moores.samaltman.com/

[2] Dougherty, C., & Griffith, E. (2023, August 25). *The Silicon Valley elite who want to build a city from scratch*. The New York Times. https://www.nytimes.com/2023/08/25/business/land-purchases-solano-county.html

[3] Landauer, R. (1961). Irreversibility and heat generation in the computing process. *IBM Journal of Research and Development*, *5*(3), 183–191. https://doi.org/10.1147/rd.53.0183

[4] Witkowski, W. (2022, September 21). *"Moore's law's dead", Nvidia CEO Jensen Huang says in justifying gaming-card price hike*. MarketWatch. https://www.marketwatch.com/story/moores-laws-dead-nvidia-ceo-jensen-says-in-justifying-gaming-card-price-hike-11663798618