MBI Deep Dives

BY **MBI DEEP DIVES** — OCT 15, 2025

Why I don't worry (as much) about big tech's depreciation schedule

In **early 2025**, I wrote about big tech's deteriorating earnings quality primarily because of their gradual extension of useful lives for their servers in the last five years. Over the course of this year, these concerns have reverberated in many corners of the market. Just last month, The Economist warned that if server useful lives were reduced to one to two years, it could shave off $2 trillion to $4 trillion market cap from big tech's valuation.

Similarly, Rihard Jarc at "Uncovered Alpha" echoed similar concerns about depreciation schedule:

> "The life of the current generation of GPUs is shorter than most think, and what many companies are projecting in their amortization plans...But some might say, well, you still see people renting Nvidia H100, which are chips that Nvidia started selling 3 years ago. Yes, but there are two factors to that. The first one is that you have two clients pushing demands sky high, as they are subsidizing the end users, as the computing to do the services that they offer is much more expensive than the price that they are charging the end users. This works out only to the point where investors are willing to give you the money to continue doing that. And the second, even more important point is that the H100 is still useful despite being 3 years old, because NVDA switched to a 1-year product cycle between H100 and Blackwell, so this is in late 2024. Before that, the cycle was 18-24 months. So, in terms of cycle times, the chip isn't that old from a generation perspective compared to looking at it in years. However, with Nvidia now on a one-year product cycle, this change affects things significantly. In my view, the real amortization of these chips should be in 1-2 years."

If Jarc or The Economist are right about chips useful lives of just 1- [illegible] indeed be quite concerning.

So, why don't I worry about it as much as I used to almost a year ago? Thanks to the "Cunningham's Law", the best way to get the right answer online is not to ask a question, but to post "wrong" answer. Following my piece on big tech's earnings quality, I have received some pushback and over time, I came to the conclusion that the critics to that piece had more cogent arguments than I did.

The argument for a rapid, 1-2 year depreciation cycle for GPUs overlooks the critical distinction between different types of AI workloads. The newest and most powerful chips such as Nvidia's Blackwell series are essential for the computationally immense task of **training** next-generation foundation models. However, once a model is trained, the task of **inference** creates a long and valuable life for older chips, which can be efficiently repurposed for these high-volume inference workloads, as well as a broad spectrum of other "accelerated computing" tasks.

This logic is **especially true for big tech**, whose infrastructure supports an incredibly diverse array of services. Jarc cites Groq founder Jonathan Ross who himself believes the chips should be depreciated over just one year. A specialized AI company like Groq might see its hardware's value tied almost exclusively to a narrow set of inference workloads, making it more susceptible to rapid obsolescence. In contrast, a hyperscaler like Google, Amazon, or Microsoft runs everything from cloud databases and video transcoding to scientific simulations and internal analytics. For them, a three-year-old H100 may not be obsolete, and rather can be redeployed to accelerate countless other tasks, delivering a significant performance uplift over traditional CPUs and generating economic value for years.

Big tech operates on a "value cascade" model for their hardware. A new Blackwell GPU takes the top-tier training jobs. The displaced H100s then cascade down to power high-end inference, model fine-tuning, or less-demanding training runs. The A100s they replace might cascade further to handle standard inference or other non-AI accelerated computing tasks. This **systematic repurposing ensures that the chip continues to generate economic value** long after it has been dethroned as the performance king.

One of the pieces that really made me re-think and re-evaluate my concerns around depreciation schedule is this particular piece from Applied Conjecture. I think they made a compelling argument that the evidence so far doesn't suggest big tech being aggressive about the depreciation schedule. Some excerpts from the piece below:

> "If the latest GPUs become obsolete and uneconomic within a year or two of introduction, ROI on AI CapEx would be hugely negative.
>
> In my view, the consensus is vastly underestimating the useful life of GPUs and thus their lifetime economics
>
> The existence of this large category of throughput-oriented workloads creates a structural demand for older, "good enough" hardware. An older, fully depreciated A100, while slower than a new B200 for a single, latency-sensitive query, can be highly cost-effective for throughput-sensitive workloads. **When running large, batched workloads, the A100 can be driven to high utilization delivering a lower TCO for that workload than a brand new, expensive B200 that might be under-utilize**d.
>
> This creates a situation where hyperscalers and enterprises will deploy their newest, most powerful GPUs for latency-critical tasks, **while repurposing prior-generation GPUs to serve the massive, cost-sensitive market for batch inference**. This dynamic fundamentally alters the traditional IT depreciation curve, giving older hardware an economically valuable and extended useful life.
>
> Essentially, an A100 purchased in 2021 for foundational model training can be strategically repurposed in 2024 for a premium, low-latency inference tier. By 2026, as even faster GPUs (i.e. B100/B200) take over that role, the same A100 can be shifted again to a bulk, low-cost, throughput-oriented inference tier. This deployment model extends the useful economic life of the asset from the oft-cited 2 years to a more favorable 6-7 years.
>
> Real-world evidence supports this model of extended lifecycles. Azure's public hardware retirement policies provide a clear precedent. **For example, Azure announced the retirement of its original NC, NCv2, and ND-series VMs (powered by Nvidia K80, P100, and P40 GPUs) for August/September 2023. Given these GPUs were launched between 2014 and 2016, this implies a useful service life of 7-9 years. More recently, the retirement of the NCv3-series (powered by Nvidia V100 GPUs) was announced for September 2025, approximately 7.5 years after the V100's launch**. This demonstrates the viability of extracting value from GPUs over a much longer period than the consensus implies."

The Chip Letter also shared some historical analogies yesterday to make a similar point:

> There was a famous saying in the 1990s: 'what Andy giveth, Bill taketh away' otherwise known as 'Andy and Bill's law'. Andy was Intel CEO Andy Grove, and Bill, of course was Microsoft's Bill Gates. As Intel delivered performance increases, Microsoft's software updates ate up that performance, making older hardware unusable. This was good news for Andy (and Intel) as users were forced to upgrade to the latest Intel chips.
>
> The AI equivalent of this is driven by the 'bitter lesson'. The latest and best models need a lot more compute. Older hardware just won't do, at least if you need to see the results of your training run this year.
>
> The lifetime of those 'AI chips' is complicated by the fact that they are doing two distinct tasks, training and inference. Both involve lots of matrix multiplications, but being optimal for one doesn't necessarily mean that it's the best for the other. Hyperscalers see Nvidia's chips as the best for training. **Part of Jensen's sales pitch is that these designs can be gently retired to the less onerous challenges of inference, roughly in the same way that you might bequeath your gaming laptop to another family member to do their emails on**.

Given the implications of shorter useful lives and the impact on earnings, it is understandable why many investors are concerned about this. However, many may not appreciate how much talent and effort have always been deployed to keep the infrastructure useful and reliable over the last two decades. The Chip Letter mentioned an interesting paper titled "Extending Silicon Lifetime: A Review of Design Techniques for Reliable Integrated Circuits" which made me appreciate that the estimates of useful life for chips aren't a random number out of a hat and people are constantly trying to figure out new techniques to enhance reliability and useful lives of chips. From the paper:

> ICs have been widely adopted across various sectors and play a vital role in modern electronics. However, they face significant reliability challenges in nearly all of the application domains. These challenges are exacerbated by the

> increasing replacement costs of ICs at advanced technology nodes and the demand for uninterrupted operation, such as during the training of AI models. Aging remains a major threat to the lifetime of IC chips, arising from a combination of device degradation mechanisms, including Bias Temperature Instability (BTI), Hot Carrier Injection (HCI), and Time-Dependent Dielectric Breakdown (TDDB), which affect transistors, as well as electromigration (EM) in on-chip metal interconnects. The confluence of these effects leads to degraded performance and shortened lifespan. The need to understand and address these aging effects has grown significantly. Figure 2 shows **the number of publications addressing each aging mechanism over the past 25 years…The data illustrate that interest in these issues has been steadily increasing, particularly in advanced technology nodes below the 10 nm regime.**

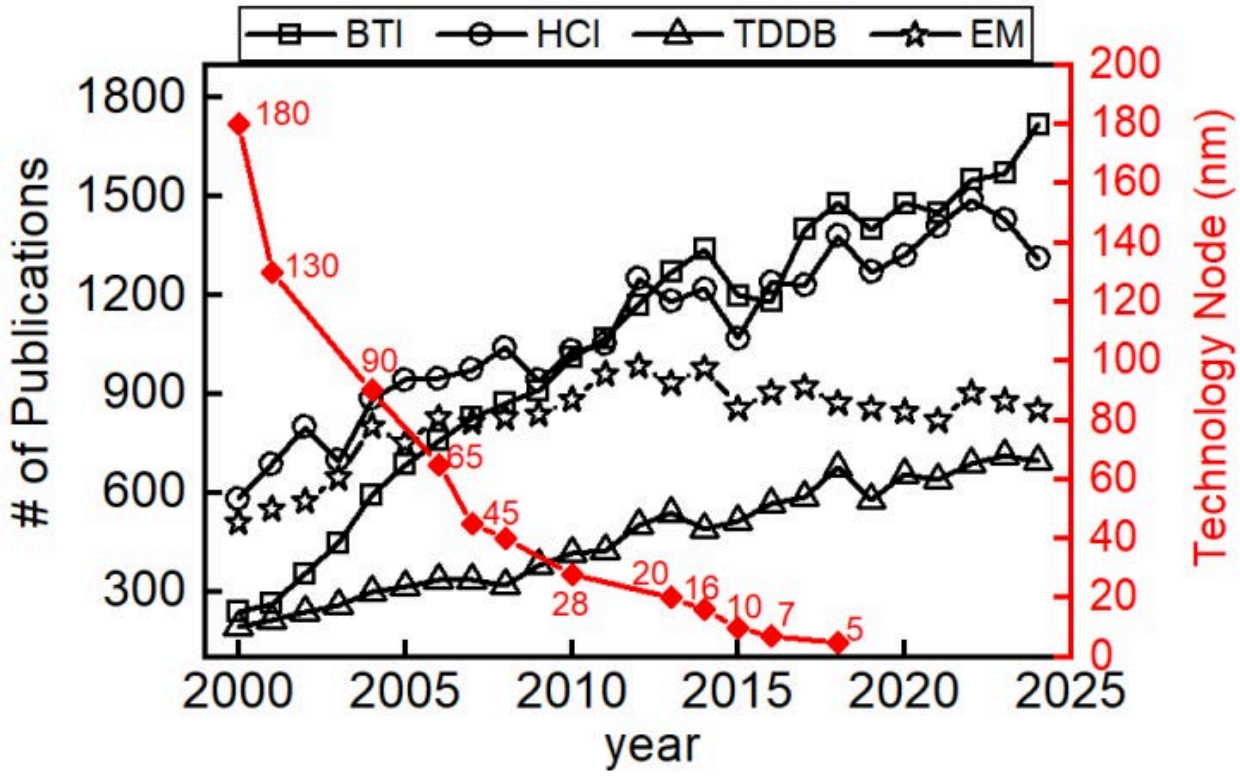

Fig. 2: No. of Publications in device reliability issues and development of technology node [21] over past 25 years (2000-2024).

Given the hundreds of billions of dollars of capex, managing the infrastructure efficiently is going to be a key core part of all big tech's operations. It was always the case, but infrastructure is increasingly at the top of everyone's list. I expect big

tech to deploy a lot of talent in improving useful lives of chips, and the diversity of their workloads will make a lot of old chips useful longer than most bears think. At the end of the day, useful life of chips is an estimate, so the estimate can be off by a little on either side, but I have dialed down my concerns considerably that the big tech may be assuming useful lives of chips ~50-100% longer than they should.

*In addition to "Daily Dose" (yes, **DAILY**) like this, MBI Deep Dives publishes one Deep Dive on a publicly listed company every month. You can find all the 63 Deep Dives here.*

Current Portfolio:

Please note that these are **NOT** my recommendation to buy/sell these securities, but