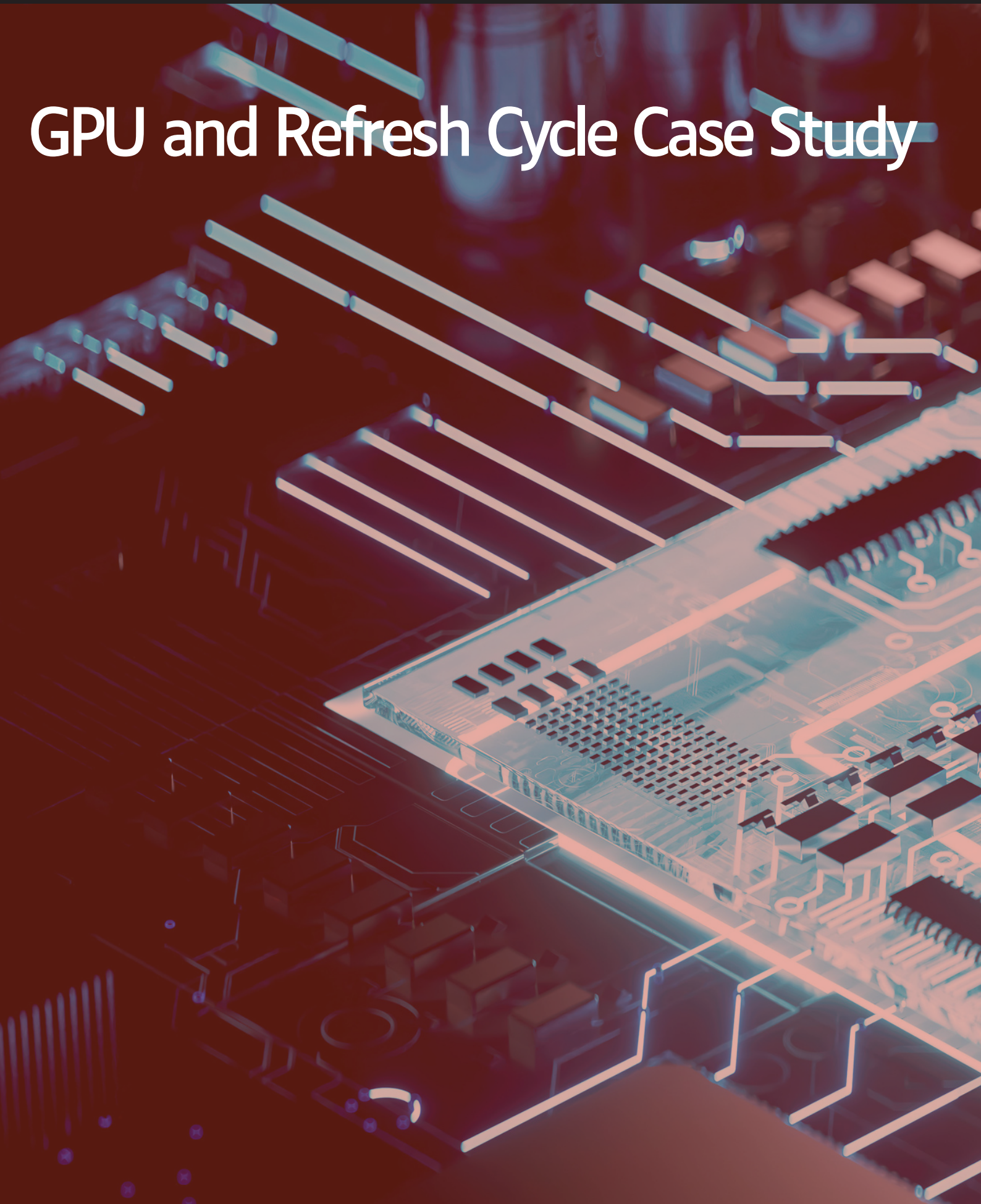


GPU and Refresh Cycle Case Study

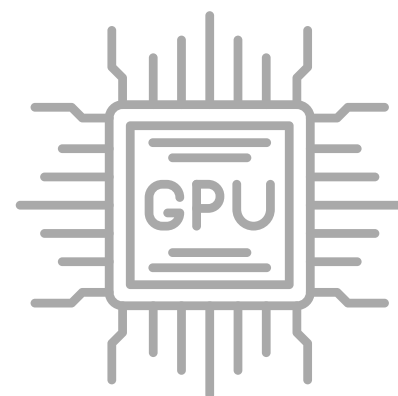


“Extended server refresh cycles hit the ITAD industry, but rising GPU investments in 2024 signal a shift toward AI and may reshape future upgrades.”

AI Will Continue to Lead to Extensions in Refresh Cycles

The ITAD industry has taken a severe hit in the recent year due to the lengthened refresh cycles of many hyperscalers for their data centers. For example, Meta announced in February 2023 they would be extending the expected life span of their servers from 4 to 5 years, following similar decisions from Amazon and Microsoft. The companies extending refresh cycles own at least 20% of the servers in the major public companies alone.

This has negatively impacted cash flows for ITAD companies across the board. For example, multiple main ITAD companies saw a severe decrease in revenue, including Envela, Sims, and CDW Corp.



cumulative % change in revenue	CQ 1 2023	CQ2 2023	CQ3 2023	CQ4 2023	CQ1 2024	YoY	QoQ
Envela (Commercial)	12.62%	-8.34%	-19.32%	-9.49%	-0.28%	-11.45%	10.18%
Sims Lifecycle Services	-7.16%	-7.16%	-10.07%	-10.07%	#N/A	-0.39%	#N/A
CDW Corp ("Other" Segment)	-11.68%	-9.77%	-21.35%	-23.19%	-17.44%	-6.52%	7.49%

Interestingly enough, even though most hyperscalers spent the last two years optimizing their data centers and extending refresh cycles to reduce expenses, all of them have announced they plan on increasing capex spending dramatically throughout 2024, particularly to buy GPUs. GPUs, or Graphic Processing Units, are used within systems to dramatically accelerate calculations.

While they are not able to handle as diverse of tasks as CPUs, Central Processing Units, the higher speed at which they can calculate makes them extremely enticing for hyperscalers. In fact, GPUs are effectively required for artificial intelligence programs to run within a practical timeframe. As hyperscalers turn their attention to GPUs to fuel their AI creations and data centers, GPUs have begun to play a prominent role within data centers and could even affect their refresh cycles.

HYPOTHESIS:

The race for AI has caused hyperscalers to allocate significant resources to the purchase of GPUs, a vital element of GPU computing. However, due to the expensive nature of GPUs and the rapidly growing space, hyperscalers must find a way to reduce expenditures on traditional servers, specifically by extending refresh cycles. We expect refresh cycles to continue to be lengthened for at least the next five years as the AI boom continues.

Global CPU Market Size: The CPU market, as of 2022, was valued at around USD 116 billion and continues to grow with a moderate CAGR of around 4.3% – 5.8%. However, based on our research, the price of computing parts tends to decrease by approximately 14% each year, using memory as a proxy, meaning that the actual growth rate in units is approximately 22.6%.

Global GPU Market Size: The GPU market was valued at approximately USD 40 billion in 2022 and was expected to grow significantly, reaching USD 400 billion by 2032, with a CAGR of about 25%. Most specifically, the Data Center GPU was estimated at \$13.1-14.3 billion as of 2023 and was expected to grow at a rate of 34.6% up to 2028 or 28.5% up to 2032, to reach about \$120 bn by 2032. However, the boom in AI drove this growth to 140% per quarter in Q1 – Q2 2023 and 27% per quarter or doubling yearly as of Q3 – Q4 2023. This is due to GPUs shifting from a niche entertainment product to an indispensable data center product. The growth should be slower going forward, potentially about 15.6% per quarter next quarter (~62.4% per year) and decreasing each quarter.

Product Outstanding

Total # of CPUs outstanding: As indicated previously, each server has two CPUs, but each of these CPUs has 12 to 32 cores. Our estimate is roughly 180 million CPUs based on the number of servers. **Total # of GPUs outstanding:** there are about 3 million compute GPUs outstanding, as in the chips that are not actually being used for the eponymous graphics, based on prior year sales, pricing, and retention rates.

The present ratio of GPUs to CPUs outstanding in data centers is about 1.67%

Projection: Based on our growth rates, we estimate that by 2030, there will be 10 – 12.5 million data center and other computing GPUs outstanding to 500 million server CPUs outstanding, a ratio of 2.20%.

However, note that GPUs can handle many more Floating-Point Operations per Second (FLOPS, a measure of computation speed) than CPUs. For example, the NVDA H100 processor, the gold standard of parallel computing, can run up to 60 trillion FLOPS. However, the CPU can only run at 50 – 500 billion FLOPS. This difference in computing power means that just one of the GPUs can run many more calculations and generate much more data, and are therefore optimal for AI.

GPU Effects on Refresh Cycles

The effects of GPUs on refresh cycles can be broken down into two main factors - a forecasted increase in capex spending and the rapid energy consumption rates of GPUs.

Increases in Spending - 2024 has proven to be a big year for GPU purchases. Hyperscalers, such as Meta, have forecasted significantly higher capex spending levels for the previous year. For example, Meta has announced that by the end of 2024, they will own about 600,000 NVIDIA H100

For example, Amazon forecasted a 16% increase in capex spending this year, mostly attributed to data center infrastructure and likely GPUs, right after another refresh cycle extension announced in January 2024. It is likely that the increase in spending and extended refresh cycles will follow this pattern going hand-in-hand. Since most of these hyperscalers have huge AI projects planned for the next five years, more refresh cycle extensions may be inevitable.

GPU Energy Consumption - One key drawback of GPUs is their rapid energy consumption. For example, comparing AMD's leading CPU and GPU, their GPU uses almost 5x the energy as the CPU for their maximum theoretical loads. Unsurprisingly, this has driven data center consumption to rise - in fact, the International Energy Agency (IEA) has estimated that the demand for energy globally from data centers will likely double by 2026 and become about equal to the amount of electricity used by all of Japan.

To combat this, hyperscalers have focused on improving operational efficiencies, such as cooling systems. For example, Microsoft recently implemented a new immersion cooling system, instead of their previous air cooling system, which requires significantly less energy. However, these new environmentally-optimized systems don't just improve ESG scores - they also increase efficiency within data centers and reduce the chance for error and hardware damages caused from outdated cooling systems.

Therefore, these new systems allow for servers to last longer and for server refresh cycles to be extended. Since GPUs in data centers are on the rise, it is only natural that hyperscalers will have to keep implementing these new energy-efficient measures, such as different cooling systems, which will in turn likely preserve servers longer. For example, Microsoft attributes its 2022 refresh cycle extension, from 4 to 6 years, mostly to the implementation of their new cooling system.

Conclusion

As spending by hyperscalers is forecasted to continue at least until 2030, due to large upcoming projects such as Microsoft's new supercomputers and Amazon's collaboration with NVIDIA, there is a strong chance hyperscalers will have to continue extending to support this spending.

Additionally, with hyperscalers such as Microsoft already working on new systems to support the energy influx, such as a cooling system that utilizes outside humidity instead of pure liquid and would therefore greatly reduce the risk of water damage to hardware, they will likely have the tools they need to continue extending.