Something Old, Something New: Extending the Life of CPUs in Datacenters

Amanda Tomlinson
UC San Diego
actomlin@ucsd.edu

George Porter
UC San Diego
gmporter@cs.ucsd.edu

Abstract

Datacenter operators are increasingly powering their operations with low-carbon energy sources such as wind and solar. As a result, the overall carbon footprint of datacenters increasingly comes from the manufacturing phase of server lifecycles. The most direct way to reduce the environmental impact of datacenters is to amortize the manufacturing emissions of hardware over longer timespans by keeping hardware in production beyond today's relatively short refresh cycles.

In this paper we analyze the feasibility of server lifespan extension in datacenters using a series of microbenchmarks. We find that while newer processors outperform older ones, the difference is workload dependent, with some workloads showing promise for older CPUs. We also analyze the potential of incorporating overclocking to further extend server lifespans.

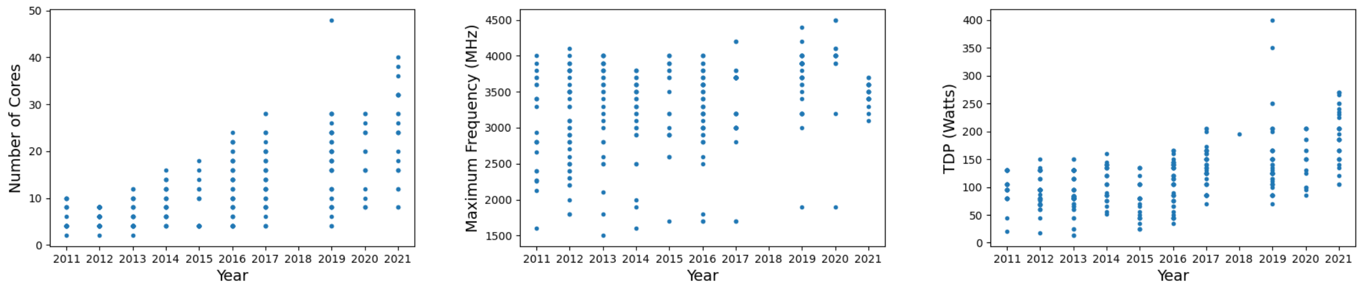
1 Introduction

Demand for cloud computing resources hosted in datacenters has risen considerably since 2010. This growth has fueled an increase in datacenters' operational energy demands. By 2030, datacenters are expected to use 3-13% of the global energy supply [3]. Despite these rising demands, many datacenter operators are pledging to reduce their carbon emissions. For example, Google [8], Microsoft [14], and Facebook [7] have pledged to decarbonize the energy powering their datacenters by 2030, and Amazon has pledged to do so by 2040 [2]. As datacenter operators reduce or eliminate the carbon from their ongoing operations, it becomes the manufacturing step which dominates overall carbon footprints of these deployments [9].

Certainly, working to decarbonize the manufacture of semiconductors and server components is important, however with existing technology progress in this direction has been modest [9]. Thus extending the time over which these manufacturing effects are amortized provides clear progress in lowering the carbon footprint of datacenters. The high footprint of datacenter hardware manufacturing is driven in part by their relatively short hardware refresh cycles, as short as 3-5 years depending on the component and deployment environment [5]. There are several motivations for why datacenter operators might replace hardware, such as increased performance, increased parallelism, new instruction sets, bigger caches, improved security features, and improved energy-efficiency. But indeed, *failure* is generally not the main driver for this relatively rapid hardware replacement cycle. Google found that hardware failures accounted for less than 10% of service disruption events in their datacenters [5].

We argue that it is the right time to reevaluate extending the lifespan of components, especially CPUs, in datacenters to lower their amortized manufacturing emissions. Indeed, it's been over ten years since developers could simply rely on Moore's Law scaling to meet increased application demands [6], and as a result datacenter systems have relied on scale and parallelism. In this work, we evaluate the feasibility of keeping older server hardware in datacenters, putting aside desire for new server features or improved security mechanisms and looking directly at application performance as our metric of choice. We compare the performance of a set of microbenchmark workloads on different generations of server-class CPUs spanning a decade from 2011 to 2021. We seek to understand the tradeoffs in performance between "older" and "newer" processors, especially across different CPU-driven workloads. We find that old processors from 2011 perform surprisingly well compared to processors a decade newer, and that the performance difference in our microbenchmark study is highly workload dependent. This gives hope that for at least some classes of workloads, there is a useful role for older CPUs to play in modern datacenters.

We also examine whether the performance gap between older and newer processors can be further closed by *overclocking* the older CPUs. While common in the high-end consumer gaming market, overclocking has not traditionally been a major tool of datacenter operators in part due to the high energy and cooling demands of running the CPUs at higher frequencies. The potential abundance of low-carbon energy at certain times of the day coupled with recent advancements in liquid cooling approaches potentially open the door to overclocking as a way to further extend the usable lifespan of older server



(a) Number of Cores per Processor (b) Processor Maximum Frequency (MHz) (c) Processor Maximum Power (Watts)

Figure 1: Intel Xeon processor trends across 447 CPU models. Each point represents an available Xeon CPU introduced in the given year. Despite the wide variation of configurations, core count is generally increasing per year while CPU frequencies and power demands have largely remained stable.

components, and we conclude this work with an analysis of that potential.

2 A Decade of CPU Advancement

Datacenter operators have a wide selection of CPU models by vendors such as Intel, AMD, IBM, Arm, and others. Beyond general-purpose CPUs, an ever increasing array of accelerators such as GPUs, TPUs, and FPGAs tackle the growing diversity of workloads hosted in modern datacenters. In this section we briefly review the processor roadmap spanning the decade 2011 to 2021, focusing our attention on a single type of processor, the Intel Xeon server-class CPU. We use as reference models of this CPU family adopted by Google's Compute Engine and available as a virtual machine selection on that platform [1]. The oldest CPU in that set is from 2014, so from 2011-2013 all Intel Xeon server class processors were included in our analysis (excepting the year 2018, for which there is only one such CPU released without publicly available benchmark results, and so we omit this year from our dataset). We reviewed data sheets for these CPUs provided by Intel (https://ark.intel.com), resulting in 447 data points.

Figure 1 shows the number of cores per processor, and the maximum frequency per processor over this time. In 2011, the maximum number of cores was only 10, where in 2021 the maximum was 40. This trend is driven by increasing virtualization and multi-tenancy. Processor maximum frequency however has been mostly constant, even decreasing slightly year over year in part due to heat dissipation limitations. The relative stability of per-core clock frequencies over this time span is further evidence of the end of Moore's law scaling.

In figure 1c, the Thermal Design Power (TDP) for each processor in the data set is shown. TDP is defined as the power consumption over a theoretical maximum load, and specifies the maximum power used by the processor [10]. Processor TDP is generally rising, due at least in part to the increased number of cores and increased cache sizes.

Although there is an incredible variety of hardware available to datacenter operators, we focus on general purpose CPUs from Intel's Xeon family. We found that from 2011 to 2021, the number of cores per CPU has grown considerably, while processor frequency and energy consumption have stayed relatively constant.

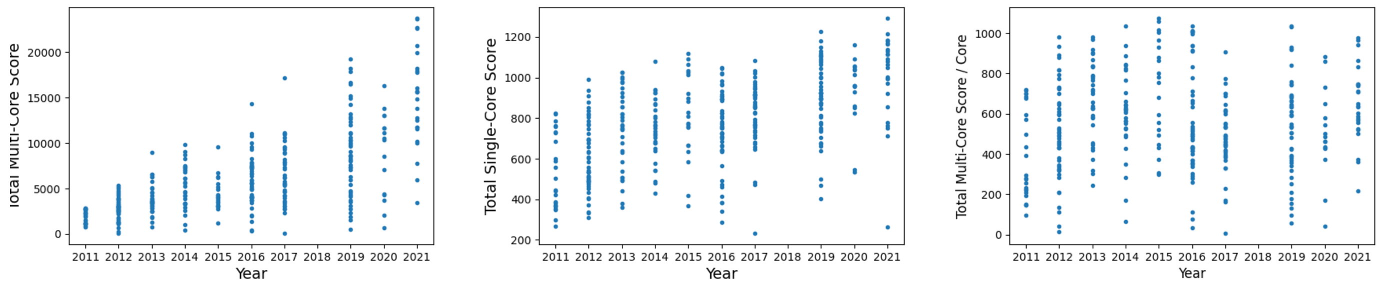
3 Microbenchmark analysis

In this section we study impact of CPU improvements at the hardware level between 2011 and 2021 on application performance and energy efficiency. We then study the potential impact of overclocking on these CPU-bound workloads. Our analysis is based on published performance results across a basket of microbenchmark workloads. Our analysis seeks to answer the following questions:

- How have CPU improvements between 2011 and 2021 affected application-level performance metrics?
- Does the choice of application and input workload affect the improvement of application performance relative to the underlying improvement of the CPU at the hardware level?
- What has been the CPU energy efficiency improvement factor between 2011 and 2021?
- What potential application performance improvements are possible through CPU overclocking?

3.1 Benchmark suite

We selected the Geekbench benchmark suite [13] to evaluate and compare CPUs. Geekbench provides a test suite of 20 tests, and assigns a composite score for single- and multi-core performance. A subset of these 20 tests are detailed in Table 1. The test suite includes three types of workloads: cryptography, integer, and floating point. Each sub-test's score is calibrated against a baseline. Then for each type of workload, a geometric mean is calculated using the sub-tests of that type. Finally, a composite score is calculated with an arithmetic mean of weighted sub-type scores to give the "Single-" and "Multi-" core scores. The single- and multi- core scores are unit-less values that reflect a ratio of performance. For example, if one



(a) Processor Multi-Core Score, by Year

(b) Processor Single-Core Score, by Year

(c) Processor Multi-Core Score per Core, by Year

Figure 2: Intel Xeon CPU performance trends. Each point is a processor in the data set. Multi-core scores have increased greatly between 2011 and 2021, while single-core scores have increased at a much lower rate. The multi-core score relative to the number of cores has remained constant over that time, showing that the increase in multi-core score is due to increased number of cores and cache sizes.

processor has a score of 1 and another has a score of 2, the second processor has twice the performance.

We collected up to five results for each candidate CPU. From the initial results collected, the mean and standard deviation of the multi-core scores were calculated, and any results that were outside one standard deviation from the mean were discarded. The measurements for each benchmark test were averaged among candidate Geekbench reports.

3.2 Overall Performance

To understand how CPU performance has changed between 2011 and 2021, we compare the Geekbench composite multi- and single-core scores of different processors. The Geekbench multi- and single-core scores for all CPUs in our data set are plotted in Figure 2. In Figure 2a, we can see that the composite multi-core score has increased rapidly, with a 5-10x improvement in scores over those 10 years. However, within each generation there is overlap, high scorers from one year may outperform median scorers from the next. Single-core performance (Figure 2b) has also increased year on year, but at a slower rate than multi-core performance. The larger relative gain in multi-core performance is most likely due to the increased number of cores and larger cache sizes on newer processors. When divided by the number of cores, the multi-core score is relatively constant year over year, more in line with the improvement in single core scores.

The single- or multi-core scores are useful to evaluate different classes of applications. A 40-core processor may be useful for virtualization, or for high performance computing where applications may span multiple processors or servers. In another extreme, a simple AWS Lambda function may use as little as one core, and a maximum of six [4]. Although newer CPUs have higher multi-core performance, single-core scores have not improved by the same factor. This makes old processors an especially viable choice for single core applications. The performance ratio for single core applications becomes roughly the ratio of cores, for example a 40 core CPU vs. a 20 core CPU would have a performance ratio of 2.

Overall, there is a large overlap in performance between different generations of CPUs, however in general multi-core applications have higher performance on newer processors, scaling with the number of cores and increasing cache sizes. Single-core application performance, from 2011 to 2021, improves less than mutli-core performance. Although newer processors with more cores can run more instances of single-core workloads, each workload is not drastically more performant.

3.3 Workload Effect on Performance

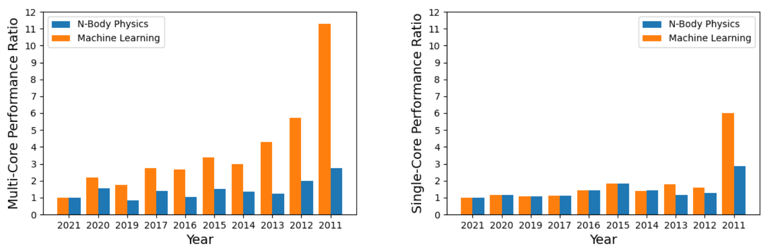
Section 3.2 shows that while single-core performance has only improved slightly from 2011 to 2021, multi-core performance has greatly increased. However, the composite multi-core score is a weighted average of all sub-test scores. We now analyze the sub-test application scores to understand if this performance increase is uniform across all workloads, or if some workloads have increased performance relative to others. For this analysis, we pick the highest performing processors based on multi-core score from each year.

In Table 1, the sub-test results for the highest performing CPU from 2011 are compared to the sub-test results from the highest performing CPU from 2021. The three best and three worst workloads are shown, sorted by multi-core performance ratio. By looking at this varied set of workloads, we can see that some applications show a higher relative improvement. For example, the N-body physics workload has a multi-core performance ratio of 2.7, so a cluster of 2.7N processors from 2011 can match the performance of a cluster of N processors from 2021. In contrast, the machine learning workload shows a relative improvement of over 11x from 2011 to 2021.

Processor performance exhibits large variations year over year, as seen in Figure 2. Comparing processors from different years can lead to different results depending on the sub-test. In Figure 3, the performance ratios for the best (N-Body Physics) and worst (Machine Learning) workloads are shown over the 10 year period from 2011 to 2021. Again, the processor with

Benchmarks	2021 Multi-Core	2021 Single-Core	2011 Multi-Core	2011 Single-Core	**Multi-Core Performance Ratio**	**Single-Core Performance Ratio**
N-Body Physics (Mpairs/sec)	12.9	1.4	4.7	0.5	**2.7**	**2.9**
Ray Tracing (Mpixels/sec)	18.0	0.8	4.6	0.4	**3.9**	**1.9**
HDR (Mpixels/sec)	465.0	14.5	107.7	10.5	**4.3**	**1.4**
Structure from Motion (Kpixels/sec)	255.6	9.7	30.8	2.7	**8.3**	**3.6**
Text Rendering (MB/sec)	7.6	0.2	0.8	0.1	**9.3**	**1.8**
Machine Learning (images/sec)	527.5	42.2	46.8	7.0	**11.3**	**6.0**

Table 1: Performance per Geekbench workload. The performance for the highest performing processors from 2011 and 2021 are compared. Results are sorted by multi-core performance ratio, with the top three and bottom three shown.



(a) Multi-Core Performance Ratio, by Year (b) Single-Core Performance Ratio, by Year

Figure 3: Sub-test performance trends. For multi-core performance, some workloads have experienced year on year improvements in performance while other workloads are relatively constant. The performance of the selected single-core workloads has stayed mostly constant since 2012.

Figure 4: Processor energy efficiency. Each point shows a processor's multi-core score divided by that processor's max power consumption.

the highest composite multi-core score was chosen to compare for each year. There may be a model which has a higher sub-test score in any given year, however we did not carry out an exhaustive search.

Figure 3a shows the multi-core performance ratio for the N-body physics and machine learning workloads, relative to the highest performing CPU from 2021. For the N-body physics workload, there has only been a modest improvement over the measured 10 year period, with the oldest processor having a performance ratio of approximately 3x. In contrast, the machine learning workload has gotten drastically better. However the improvement from 2011 to 2014 is larger than 2014 to 2021. The single-core performance ratios for the N-body physics and machine learning workloads are shown in Figure 3b. Other than a large increase in performance from 2011 to 2012, the single-core performance has remained relatively constant for both workloads over the ten year period. While it may be unlikely that a machine learning or physics simulation would be run as a single-core process, this highlights that there are certain workloads that have not gotten drastically better on newer processors.

While overall multi-core performance is definitively higher in newer generations of processors, the choice of application makes a large difference in the relative performance improvement. In our extreme example of a ten year difference from 2011 to 2021, we see a multi-core performance ratio improvement of as little as 3x to over 11x depending on the workload. Single-core workloads show even smaller differences, with processors from 2012 having almost the same performance as processors from 2021 for certain applications.

3.4 Energy Efficiency

To understand how processor energy efficiency has changed from 2011 to 2021, we calculate each processor's multi-core score, divided by that processor's TDP, to get a ratio of performance per watt. Figure 4 shows each processor's performance per watt. Although there is considerable overlap between generations, this shows that energy efficiency improved roughly 2-3x from 2011 to 2016, but since then has been relatively constant.

The higher energy efficiency of newer processors presents a trade off between the lower operational emissions of newer processor generations and the lower manufacturing emissions of older processor generations. The carbon intensity of the grid, as well as application needs, should also be considered when scheduling a job within a heterogeneous-aged datacenter. The increasing deployment of solar and wind generation has led to a variation in carbon intensity over the course of the day, for example in California the carbon intensity of the grid decreases around 50% in the middle of the day [11]. California is producing so much solar that the California Independent System Operator (CAISO) actually curtails (or disposes of) this excess energy, with 0.6 GWh of curtailment

Year	Model	Cores	Base Speed (MHz)	Top Speed (MHz)	Multi Core Score	**Improvement (%)**
2011	E5606	4	2130	2130	1178	
2011	E3-1290	4	3600	4000	2799	**137.5**
2021	6338	32	2000	3200	18168	
2021	8362	32	2800	3600	22659	**24.7**

Table 2: Potential performance increase from over-clocking. Each pair of processors is equal in everything except clock speed. By comparing the multi-core scores for each pair we can estimate the performance increase from increasing clock speeds.

in April 2022 [12]. We can schedule the activation of older, less energy-efficient hardware with this excess low-carbon energy supply as one avenue for increasing compute without a commensurate increase in carbon footprint.

Newer processors are generally more energy efficient, with a 2-3x improvement in performance per watt from 2011 to 2021. This increased energy efficiency presents a trade off between higher operational emissions of older processors with higher embodied emissions of newer processors.

3.5 Revising the Potential for Overclocking

So far, our analysis has focused on the highest performing processors of each generation. We also want to understand the performance of the other processors in our data set, and see if there is any way to improve their performance. As explained in Section 3.4, low-carbon renewable energy is becoming more common in certain grids at certain times, and shifting demand to meet that excess energy is one way to increase compute while maintaining or even lowering carbon emissions. As mentioned above, we could overclock, or increase the frequency, of CPUs during these times. Recent advancements, potentially open the door to this approach. For example, work on immersion cooling [15] could make overclocking a more viable strategy to increase performance.

Revising our Geekbench data set, we estimate the performance improvement achievable by overclocking in the following way. Intel generally makes many processors in a given year with differing numbers of cores, clock speeds, and cache sizes. For each year, we look for a pair of processors with the same number of cores and the same cache size, but with different clock speeds. By comparing the multi-core scores for these pairs, we can see the effect clock speed has on overall performance. In effect, we're using the potential of overclocking to "convert" a lower-cost, more readily available lower speed processor to the performance of a higher-cost, rarer top-end server-grade processor. In fact, higher speed processors are often those that meet rigid specifications after manufacturing, where those that are not able to meet the specifications are downclocked and sold at a lower frequency.

Table 2 shows the specifications of these pairs and the performance increase with increased clock frequency. Note that this style of analysis underestimates the benefits of overclocking, since it is in theory possible to clock a CPU beyond the highest rated frequency sold under that make and model. The Geekbench scores show an increase of 25% to over 100% improvement based on this simulated overclocking. This analysis shows that even with the more common models produced within a given year, given abundant low-carbon energy, it's possible to get a large performance increase by overclocking.

4 Conclusion and Future Work

In this work we examined "performance ratios" of different generations of Intel server class processors from 2011 to 2021. We analyze these performance ratios to see how many older processors would need to be deployed to match the performance of newer processors. While we found that overall newer processors have higher multi-core performance, the performance improvement is workload dependent. Even in an extreme case of using 10 year old processors, as little as 3x as many processors can match the performance of the top performing model from 2021, for certain workloads. In the worst case, the performance ratio was not more than an order of magnitude. In addition, single-core performance has not increased at nearly the same rate as multi-core performance. This provides an opportunity to schedule single-core applications on older processors for decreased embodied carbon emissions.

We found that processor energy efficiency has generally increased from 2011 to 2021, with about a 2-3x improvement in performance per watt over that period. Because of this increased efficiency, the carbon intensity of the energy grid where the processors are used needs to be taken into account. One avenue for future work is developing strategies for scheduling workloads to take advantage of green energy while maximizing application performance.

Our work also shows that overclocking has the potential to increase application performance by 25% up to 100%. The barriers to overclocking that exist today, mainly heat dissipation and increased energy demand, are being addressed through liquid cooling and renewable energy generation. As these technologies continue to develop, overclocking will become a viable way to increase performance for applications running on older processors.

Datacenter operators must address manufacturing emissions as they completely decarbonize their operations, and extending server lifespan is a direct and efficient way to amortize those emissions. Although application performance is not the only barrier to extending CPU lifespan, this analysis shows that older CPUs may still be useful in modern datacenters for certain applications. Another area of future work will be to more strictly define which applications are suitable for older processors and schedule those workloads appropriately.

References

[1] CPU platforms: Intel CPU processors. https://cloud.google.com/compute/docs/cpu-platforms#intel_cpu_processors. Accessed: 2022-01-21.

[2] Amazon. The climate pledge. https://www.aboutamazon.com/planet/climate-pledge, 2022.

[3] Anders S. G. Andrae and Tomas Edler. On global electricity usage of communication technology: Trends to 2030. *Challenges*, 6(1):117–157, 2015.

[4] AWS. AWS Lambda now supports up to 10 GB of memory and 6 vCPU cores for Lambda functions. https://aws.amazon.com/about-aws/whats-new/2020/12/aws-lambda-supports-10gb-memory-6-vcpu-cores-lambda-functions/, 2020.

[5] L.A. Barroso, U. Hölzle, and P. Ranganathan. *The Datacenter as a Computer: Designing Warehouse-Scale Machines, Third Edition*. Synthesis Lectures on Computer Architecture. Morgan & Claypool Publishers, 2018.

[6] Hadi Esmaeilzadeh, Emily Blem, Renée St. Amant, Karthikeyan Sankaralingam, and Doug Burger. Dark silicon and the end of multicore scaling. In *2011 38th Annual International Symposium on Computer Architecture (ISCA)*, pages 365–376, 2011.

[7] Facebook. Facebook's net zero commitment. https://sustainability.fb.com/asset/net-zero-commitment/, 2020.

[8] Google. Operating on 24/7 carbon-free energy by 2030. https://sustainability.google/progress/energy/, 2022.

[9] Udit Gupta, Young Geun Kim, Sylvia Lee, Jordan Tse, Hsien-Hsin S Lee, Gu-Yeon Wei, David Brooks, and Carole-Jean Wu. Chasing carbon: The elusive environmental footprint of computing. In *2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pages 854–867. IEEE, 2021.

[10] Intel. Thermal Design Power (TDP) in Intel® processors. https://www.intel.com/content/www/us/en/support/articles/000055611/processors.html, 2019.

[11] California ISO. Emissions, today's outlook. http://www.caiso.com/TodaysOutlook/Pages/emissions.html, 2022.

[12] California ISO. Managing oversupply. http://www.caiso.com/informed/pages/managingoversupply.aspx, 2022.

[13] Primate Labs. Geekbench 5. https://www.geekbench.com/, 2022.

[14] Microsoft. Supporting our customers on the path to net zero: The Microsoft cloud and decarbonization. https://blogs.microsoft.com/blog/2021/10/27/supporting-our-customers-on-the-path-to-net-zero-the-microsoft-cloud-and-decarbonization/, 2021.

[15] Bharath Ramakrishnan, Husam Alissa, Ioannis Manousakis, Robert Lankston, Ricardo Bianchini, Washington Kim, Rich Baca, Pulkit A. Misra, Inigo Goiri, Majid Jalili, Ashish Raniwala, Brijesh Warrier, Mark Monroe, Christian Belady, Mark Shaw, and Marcus Fontoura. CPU overclocking: A performance assessment of air, cold plates, and two-phase immersion cooling. *IEEE Transactions on Components, Packaging and Manufacturing Technology*, 11(10):1703–1715, 2021.