Enabling Sustainable Cloud Computing With Low-Carbon Server Design

Jaylen Wang, *Carnegie Mellon University, Pittsburgh, PA, 15213, USA*

Daniel S. Berger, *Microsoft, Redmond, WA, 98052, USA and University of Washington, Seattle, WA, 98195, USA*

Fiodar Kazhamiaka, Celine Irvene, Chaojie Zhang, Esha Choukse, Kali Frost, Rodrigo Fonseca, Brijesh Warrier, Chetan Bansal, Jonathan Stern, and Ricardo Bianchini, *Microsoft, Redmond, WA, 98052, USA*

Akshitha Sriraman, *Carnegie Mellon University, Pittsburgh, PA, 15213, USA*

To combat climate change, we must reduce carbon emissions from hyperscale cloud computing. Compute servers cause the majority of a general-purpose cloud's emissions. Thus, we are motivated to design carbon-efficient compute server stock keeping units (SKUs), or GreenSKUs, using recently available low-carbon components. We built three GreenSKU prototypes, integrating energy-efficient CPUs, reusing old dynamic RAM via compute express link, and reusing old solid-state drives. We reveal challenges that limit GreenSKUs' carbon savings at scale and may prevent their adoption by cloud providers. To address these challenges, we developed a novel framework, GSF (GreenSKU Framework), that enables cloud providers to systematically evaluate GreenSKUs' carbon savings at scale. By implementing GSF within Microsoft Azure's production constraints, we demonstrate that GreenSKUs reduce net cloud emissions by 8%, which is globally significant. This work is the first to demonstrate and quantify how carbon-efficient server designs translate to measurable cloud-scale emissions reductions, enabling meaningful contributions to cloud sustainability goals.

Hyperscale cloud computing has grown rapidly, enabling transformative technologies, from life-saving telemedicine to universal educational resources. However, this expansion has simultaneously made cloud computing a rapidly growing source of global carbon emissions.[1] By 2030, digital infrastructure alone could account for 20% of global emissions, with cloud infrastructure playing a central role.[2] Thus, major cloud providers have pledged aggressive decarbonization targets by 2030.[3] This dual reality of cloud computing's growing demand and rising emissions presents a grand challenge to the computing research community: How can we sustain cloud computing's growth while reducing its environmental impact?

When considering the cloud's environmental impact, the prevailing narrative is about the power consumption of server hardware in data centers; however, ensuing *operational emissions* are only half of the emissions story. Cloud providers are increasingly turning their attention to *embodied emissions*, which include emissions from manufacturing chips in semiconductor fabs and transporting them to data centers that house them.[4] While operational emissions have historically dominated due to cloud servers' high power consumption, energy optimizations and renewable energy deployments have now caused embodied emissions to constitute 50%–82% of cloud emissions.[4] Thus, we ask: Which key cloud server components cause the most embodied and operational emissions, and how can we redesign them to minimize a general-purpose cloud's carbon footprint?

0272-1732 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies.
Digital Object Identifier 10.1109/MM.2025.3572955
Date of publication 23 May 2025; date of current version 12 September 2025.

Authorized licensed use limited to: Carnegie Mellon Libraries. Downloaded on October 27,2026 at 23:15:03 UTC from IEEE Xplore. Restrictions apply.

To reduce cloud computing's emissions, we identified designing carbon-efficient cloud compute server stock keeping units (SKUs) as a promising solution. Server SKU design is the process by which existing hardware components are selected and composed into servers. Typically, cloud providers design compute server SKUs to meet performance and cost goals. To reduce emissions, we introduce a new way of designing carbon-efficient compute server SKUs, or *GreenSKUs*, that trade off performance for lower carbon.

We find that designing and deploying GreenSKUs is promising for three reasons. First, we show that compute servers cause most of the cloud's emissions, and their design directly impacts both embodied and operational emissions. Indeed, with an approximately six-year lifetime for cloud servers,[5] design choices made in the next one to two years will directly affect the industry's 2030 decarbonization targets. Second, cloud servers are often underutilized,[3] motivating designing servers that suitably scale performance to reduce emissions. Third, GreenSKU design and deployment is more feasible today due to the availability of carbon-efficient commodity server components, e.g., energy-efficient cores.[6]

Due to GreenSKUs' promise, we designed and built three GreenSKUs using low-carbon components that mitigate cloud compute servers' key sources of operational and embodied emissions. Our GreenSKUs incrementally incorporate three low-carbon components: energy-efficient high thread-count CPUs,[6] reused old dynamic RAM (DRAM) with compute express link (CXL),[7] and reused old solid state drives (SSDs).

While GreenSKUs promise carbon savings, we demonstrate several challenges that can limit their deployment at hyperscale. For example, a GreenSKU may not achieve the same performance as currently deployed servers, requiring effectively trading off carbon versus performance to enable its deployment (e.g., a GreenSKU with energy-efficient cores and lower single-thread performance would trade per core performance for lower carbon). Thus, we ask: How do we systematically account for these challenges to evaluate a GreenSKU's realistic benefits at scale to enable its deployment in a production cloud?

To address these challenges, we developed a novel methodology and open source framework, *GreenSKU framework or GSF*,[8] that enables cloud providers to systematically evaluate a GreenSKU's carbon savings in the cloud. By applying GSF within Microsoft Azure's production constraints, we demonstrate GSF's utility and GreenSKUs' promise. The results are compelling: Our GreenSKUs reduce emissions per core by 28% compared to currently deployed Azure servers. When configured to meet application performance requirements, they reduce general-purpose cloud compute clusters' emissions by 15%. Critically, when incorporating overall data center overheads, our GreenSKUs reduce net cloud emissions by 8%, a substantial reduction at hyperscale, directly facilitating meeting cloud providers' 2030 decarbonization targets.

In summary, our article,[5] presented at the 2024 International Symposium on Computer Architecture, contributes:

- A quantification of the opportunity to significantly reduce cloud emissions by designing and deploying carbon-efficient servers, i.e., GreenSKUs, at scale.
- The development of three new GreenSKU prototypes with carbon-efficient server components.
- A study of the challenges that limit GreenSKUs' deployment and carbon savings at scale.
- A novel methodology and associated open source framework,[8] *GSF*, that enables cloud providers to systematically evaluate a GreenSKU's carbon savings at scale.
- An evaluation of our GreenSKUs by using GSF within a leading cloud provider's production constraints to demonstrate our GreenSKUs' carbon savings in production.

REDESIGNING SERVERS FOR LOWER CARBON EMISSIONS

We identify key opportunities and design challenges in reducing cloud emissions with GreenSKUs.

Today's Data Center Carbon Landscape

We lack a comprehensive understanding of how emissions arise from data center components today.[5] This knowledge gap makes it challenging to identify which design interventions (e.g., which GreenSKUs) will yield the most environmental benefits. To address this gap, we developed an open source carbon model[8] to break down data center-scale operational and embodied emissions across server types and components.

We used our carbon model to analyze how different server types and their hardware components contribute to a general-purpose cloud's emissions at Azure (Figure 1). Such hyperscale carbon accounting is infeasible with prior component-level models.[4] Using the data center-wide visibility our carbon model provides, we demonstrate that compute servers, which are central to general-purpose cloud computation, are major contributors of operational and embodied emissions at Azure. This finding underscores the urgent need to lower compute servers' emissions.

Authorized licensed use limited to: Carnegie Mellon Libraries. Downloaded on October 27,2026 at 23:15:03 UTC from IEEE Xplore. Restrictions apply. © 2025

Hidden Cost of Performance-First Design

To enable carbon-efficient server design, we reveal the server design process's status quo in the cloud. Traditionally, cloud providers have followed a *faster-at-similar-price* (*FSP*) business model, introducing new virtual machine (VM) generations with enhanced performance per core at comparable price points every few years. This approach has shaped industry practice and academic research, with SKU design primarily optimizing for performance and cost. However, prior work shows that 75% of Azure VMs operate at <25% CPU utilization,[3] suggesting that the FSP model's consistent pursuit of performance may be misaligned with user needs, exacerbating emissions in the process. This mismatch presents a compelling opportunity to design servers that trade off some performance for significant carbon savings.

Recent advances in server component technology make designing low-carbon servers increasingly feasible. For example, we can now design servers with CPUs that contain numerous energy-efficient (i.e., *efficient*) cores,[6] which typically have lower single-thread performance, trading off performance for lower operational emissions. Such emerging carbon-efficient components facilitate effective GreenSKU design.

Design Goals and Constraints for Low-Carbon Servers

We must address three fundamental *design* (*D*) challenges to design and deploy GreenSKUs at scale.

- *(D1) Navigating emissions tradeoffs.* Operational and embodied emissions have complex tradeoffs.[3] For example, while reusing components can reduce embodied emissions,[4] it may increase operational emissions by lowering energy efficiency. Conversely, replacing servers frequently with newer, more energy-efficient hardware can reduce operational emissions while increasing embodied emissions from manufacturing more servers. Successful GreenSKU design requires carefully balancing these tradeoffs.
- *(D2) Accounting for fleet-wide impacts from introducing GreenSKUs.* Introducing new SKUs fleet-wide can unexpectedly affect total emissions. For example, adding new SKU types with diverse resources can affect the efficiency of bin-packing VMs on physical servers. At scale, even small decreases in bin-packing efficiency require deploying many additional servers. Such fleet-wide effects are often overlooked when evaluating carbon-efficient designs at the server or device level, yet they become critical constraints when deploying at hyperscale.
- *(D3) Ensuring performance and adoption at scale.* We must design a server that will be widely adopted in the cloud and in a way that reduces net emissions. This design is challenging as many low-carbon components offer lower performance than servers built under the FSP model. For example, a GreenSKU with energy-efficient cores typically has lower single-thread performance and is less likely to be adopted by applications bound by single-thread performance. Such performance impacts can affect GreenSKU adoption, requiring careful balance between reducing emissions and meeting applications' performance goals.

OUR GreenSKU PROTOTYPES

The design space for GreenSKUs is large. We show the practicality and effectiveness of building GreenSKUs that target the top three carbon contributors in

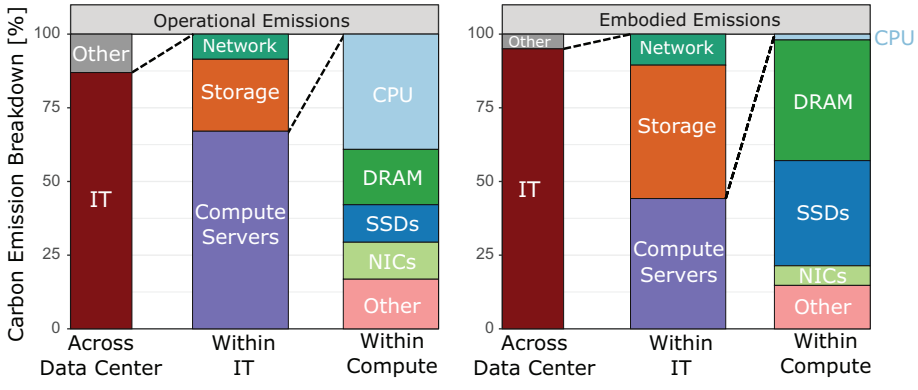


FIGURE 1. Carbon breakdown of a general-purpose data center at Azure.

Authorized licensed use limited to: Carnegie Mellon Libraries. Downloaded on October 27,2026 at 23:15:03 UTC from IEEE Xplore. Restrictions apply.

compute servers: CPUs, DRAM, and SSDs (Figure 1), which cause 67% of a server's net emissions.

Low-Carbon Components

First, we incorporated energy-efficient CPUs with high core counts, which are now widely available.[6] Since these CPUs have 40%–60% more cores at comparable power consumption to mainstream CPUs, they significantly reduce operational emissions per core. We chose AMD Bergamo in our GreenSKU prototype, as it provides full support for type 3 CXL devices (CXL.mem).[6]

Next, we found that it is practical to reuse old DRAM in new servers. In an internal study of decommissioned hardware, we found that DDR4 DRAM dual inline memory modules (DIMMs) from end-of-deployment servers show no sign of increasing failure rates.[5] While DDR generations are traditionally not backward compatible, modern CXL controllers enable connecting older DDR4 memory to newer DDR5 systems. Reusing old DRAM can significantly reduce embodied emissions at the cost of higher operational emissions, due to more power consumed by CXL and old DIMMs' lower density compared to new DIMMs.

Third, we integrated old SSDs from decommissioned Azure servers. Similar to DRAM, we found that most decommissioned SSDs have enough lifetime left for reuse. Reusing SSDs can reduce embodied emissions at the cost of higher operational emissions.

Prototype SKUs

We built three GreenSKU prototypes by incrementally adding each of the three carbon-efficient components: 1) *GreenSKU-Efficient* uses AMD's efficient Bergamo CPU to reduce operational emissions, 2) *GreenSKU-CXL* adds CXL-attached reused old DRAM to reduce embodied emissions, and 3) *GreenSKU-Full* adds reused old SSDs to further reduce embodied emissions. Figure 2 shows our fully built GreenSKU-Full prototype, demonstrating that our GreenSKUs can be readily deployed today.

Performance Characteristics

While our GreenSKUs offer significant carbon advantages, they may offer lower performance than *baseline SKUs*, which are currently-deployed SKUs at Azure following the FSP model. For example, AMD Bergamo operates at lower frequencies with less last-level cache than Azure's baseline AMD SKUs. Similarly, reused DRAM and SSDs may increase latency and reduce bandwidth. These performance considerations could limit adopting GreenSKUs for performance-sensitive applications.

To accurately estimate GreenSKUs' carbon savings at hyperscale, we must systematically account for these performance versus carbon tradeoffs within real-world deployment scenarios.

GSF: THE GREENSKU FRAMEWORK

To evaluate cloud emissions from deploying a GreenSKU, we must 1) account for each factor that affects a GreenSKU's carbon savings and 2) model each factor based on a cloud provider's deployment environment.

To address the first goal of identifying relevant factors, we introduce GSF, a framework that estimates a data center's emissions from deploying a GreenSKU at scale. GSF systematically considers seven major factors that influence a GreenSKU's carbon savings at scale. GSF considers each factor using distinct *components*, as shown in Figure 3. GSF succinctly captures data center complexity by formalizing how components' outputs feed into other components' inputs. GSF can be flexibly used by other cloud providers, as it abstracts components' relationships from cloud provider-specific component implementation details.

To address the second goal of implementing GSF's components at scale, we detailed our implementation of each component under Azure's production constraints to evaluate our GreenSKUs.

We now describe GSF's seven components and how we implement each within Azure.

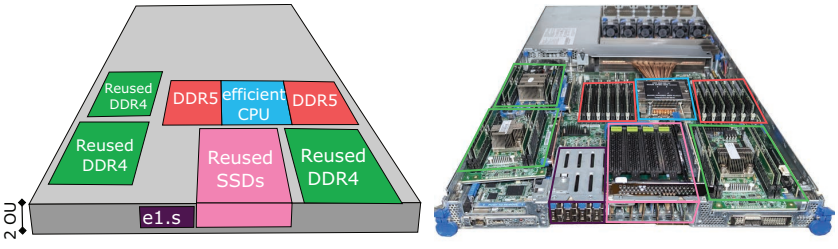


FIGURE 2. Our GreenSKU-Full design with AMD's efficient CPU, reused DDR4 DRAM (via CXL), and reused m.2 SSDs (via e1.s and PCIe adapters).

Authorized licensed use limited to: Carnegie Mellon Libraries. Downloaded on October 27,2026 at 23:15:03 UTC from IEEE Xplore. Restrictions apply. © 2025

- *Carbon model:* GSF requires a carbon model to quantify a SKU's emissions. Our novel, comprehensive carbon model calculates both embodied and operational emissions across server, rack, and data center components. Following established practice,[9] we assigned zero embodied emissions to reused server components that are in their "second life." For complete details, we have open sourced our carbon model.[8]
- *Performance:* Applications running on a GreenSKU may experience reduced performance per core, potentially requiring scaling up/out on GreenSKUs to meet performance goals. To quantify such scaling, we benchmarked 20 open and closed source applications that span the six application classes that run on most VMs in Azure.[10] We measured a GreenSKU's performance by setting a service level objective (SLO) based on a baseline SKU's performance. Our baseline SKUs are three deployed server generations, Gen1, -2, and -3, where Gen3 is a future generation whose deployment scale could be partially offset with a GreenSKU. To achieve comparable performance as the high-performance baseline SKU, we scaled up the number of VM cores on the GreenSKU to eight, 10, and 12 cores and compared the resulting performance against an eight-core VM running on the baseline SKU. We then calculated the *scaling factor*, which defines how many GreenSKU cores per baseline SKU core are needed for a VM to achieve an application's performance goals.
- *Maintenance:* Server failures necessitate standby server capacity, which contributes to a data center's emissions. As GreenSKUs may influence a server's annual failure rate (AFR), we must calculate this maintenance overhead. We estimated this overhead at Azure via our open source maintenance model[8] that uses component AFRs to calculate server maintenance overheads across all SKUs.
- *Adoption:* This component determines which applications can effectively run on a GreenSKU (e.g.,

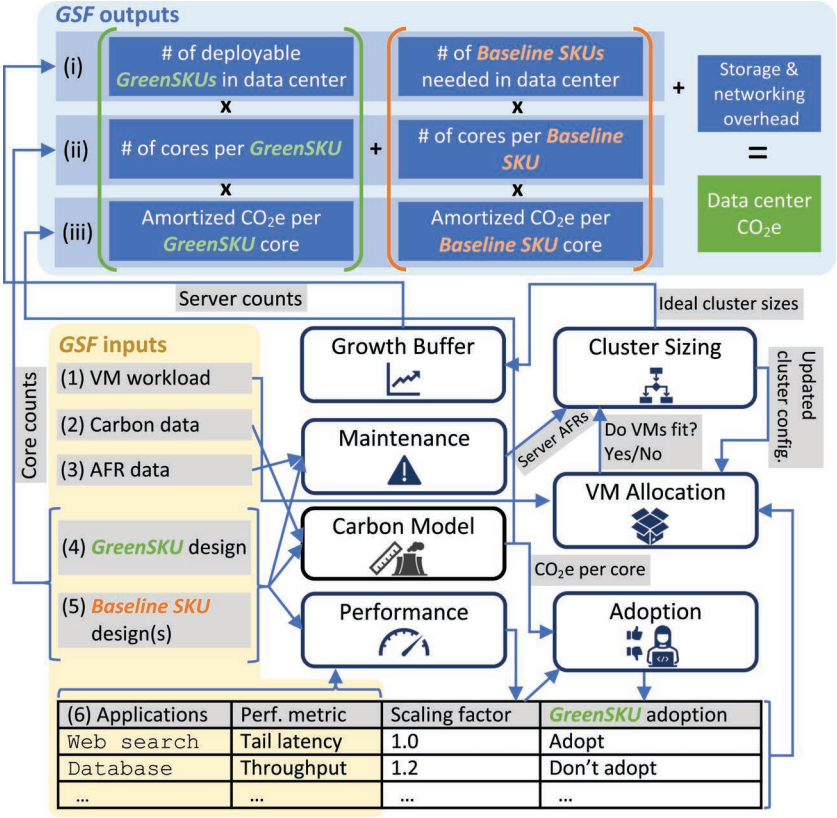


FIGURE 3. Overview of the GSF. Components are solid-line boxes. GSF's inputs are highlighted in yellow. Inputs and outputs between components are in gray boxes. Each component contributes to producing the necessary outputs, as shown in blue, to calculate data center emissions.

Authorized licensed use limited to: Carnegie Mellon Libraries. Downloaded on October 27,2026 at 23:15:03 UTC from IEEE Xplore. Restrictions apply.

while meeting performance goals). Our adoption model assumes that cloud users aim to reduce their applications' emissions while meeting performance goals. Thus, we first calculated the carbon required to service each application on a GreenSKU. To this end, we multiply the scaling factor (determined from GSF's performance component) by the emissions per core determined from the carbon model. We then calculated this value for the baseline SKU, using eight cores and the baseline SKU's emissions per core. We modeled that an application will adopt a GreenSKU if the calculated carbon value to run the application on the GreenSKU is lower than the baseline's, i.e., running on the GreenSKU saves carbon while meeting performance goals. We repeated this step for each representative application.

- *VM allocation/packing:* Server design directly affects VM packing efficiency in a cluster. We accounted for such impacts in this component via an internal VM allocation simulator with Azure VM traces. Since production traces do not specify the applications running in VMs, we mapped each VM to one of our representative benchmark applications. Our adoption model then determines whether each VM can run on a GreenSKU based on its application assignment. For VMs that adopt the GreenSKU, we applied the appropriate scaling factor to adjust resource requirements.
- *Cluster sizing:* This component determines the optimal mix of baseline SKUs and GreenSKUs needed to service a data center's VM workload. At Azure, we employ a search algorithm through our VM allocation simulator to identify the SKU distribution that minimizes emissions while supporting Azure's VM workload.
- *Growth buffer:* Cloud providers maintain a *growth buffer*: extra server capacity to accommodate demand spikes. We account for this capacity through additional baseline SKUs, which simplifies GreenSKU deployment by eliminating multiple buffer SKU types. However, exclusively composing the buffer using carbon-inefficient baseline SKUs marginally increases emissions. We consider these emissions in our carbon savings estimate.

EVALUATING GREENSKUS USING GSF

We applied GSF at Azure to evaluate our GreenSKU prototypes. First, we present results from benchmarking our GreenSKUs using our GSF performance component implementation at Azure. Then, we demonstrate how the complete GSF assesses our GreenSKUs' carbon saving potential.

Evaluating GreenSKUs' Performance

To evaluate our GreenSKUs' performance, we measure the 95th percentile tail latency across varying queries per second (QPS) loads. We define each application's SLO as the 95th percentile tail latency achieved at 90% of the baseline SKU's peak saturation throughput.[5]

Figure 4 shows results for one representative application from three application classes: `Masstree` (Big Data), `Xapian` (Web App), and `Nginx` (Web Proxy). We

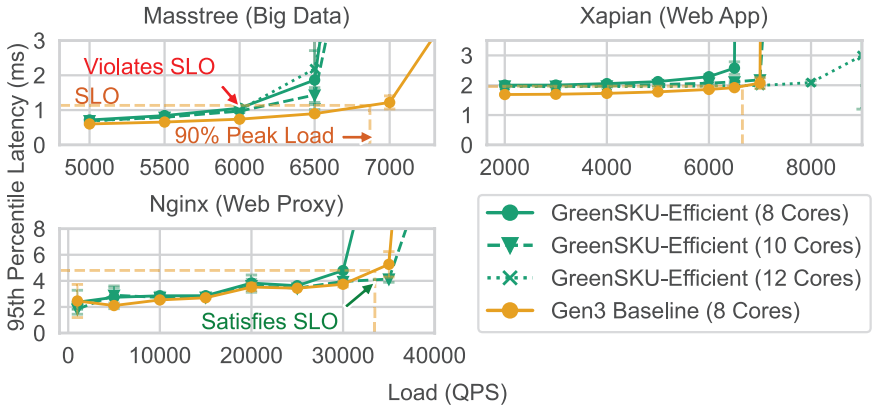


FIGURE 4. Ninety-fifth percentile tail latency versus load (QPS) for applications spanning three of our six application categories. Tail latency is shown for eight-core configurations for the Gen3 baseline SKU (in orange). Results for GreenSKU-Efficient are shown up to the scaling required to achieve comparable performance, if possible, to Gen3. The dotted orange line indicates an SLO set using Gen3's latency at 90% of peak load. For some applications (e.g., `Xapian` and `Nginx`), our GreenSKU-Efficient can achieve the SLO with scaling; for others (e.g., `Masstree`), the scaling required outweighs carbon savings.

Authorized licensed use limited to: Carnegie Mellon Libraries. Downloaded on October 27,2026 at 23:15:03 UTC from IEEE Xplore. Restrictions apply. © 2025 IEEE

show results up to the minimum number of GreenSKU-Efficient cores that achieve a peak saturation throughput closest to our Gen3 baseline SKU. We exclude Gen1/Gen2 results, as Gen3 consistently outperforms them.

We found that GreenSKU-Efficient achieves baseline SKUs' performance, without scaling up the number of cores, for seven of 20 applications, while nine applications require only 25% scaling. With GreenSKU-CXL, 20% of applications meet performance goals running entirely on reused DRAM backed by CXL, allowing other applications to run entirely on local DRAM without performance penalties.

Evaluating GreenSKUs' Carbon Savings

Using the performance results, we calculated applications' scaling factors to determine whether the application can be adopted to run on a GreenSKU. Using these adoption rates, we performed VM allocation simulations across 35 production VM traces to determine cluster and growth buffer sizes. Finally, we took GSF's resulting output, i.e., cluster-wide carbon savings, to evaluate our GreenSKUs' carbon savings at Azure.

Using our carbon model, we show that GreenSKU-Full reduces emissions per core by 28% compared to baseline SKUs at Azure's mean carbon intensity (CI), i.e., emissions produced per unit of energy consumed by a data center. Using this per core carbon savings, we generated GSF's final output of cluster-level carbon savings. Figure 5 shows the cluster-level carbon savings achieved when introducing each of our GreenSKUs compared to an all-baseline-SKU cluster. We evaluated across a spectrum of CI values, showing the values for three Azure data center regions.

Compared to an all-baseline-SKU cluster, GreenSKUs lower cluster-level emissions by up to 25%. Furthermore, CI impacts the lowest-carbon GreenSKU. At high CI, GreenSKU-Efficient is better, due to not having energy-inefficient reused devices. At low CI, embodied emissions dominate, so GreenSKU-Full is better. Such tradeoffs motivate the need for GSF to evaluate GreenSKUs.

Overall, we estimate that GreenSKUs reduce cluster-wide emissions by 15% and Axures net cloud emissions by 8%, a significant reduction at scale.

So, Why Design GreenSKUs?

To put our GreenSKUs' carbon savings into perspective, we detail why deploying GreenSKUs 1) also reduces cost and 2) is more practical compared to other cloud emission reduction approaches. We leverage the flexibility of GSF's carbon model to perform such comparisons.

GreenSKU Cost Analysis

We adapted GSF to analyze GreenSKUs' effect on total cost of ownership (TCO) by replacing the carbon model with a TCO model. Since TCO data are sensitive, we share high-level insights. Our TCO analysis reveals that a cost-efficient SKU is only 5% less costly compared to our carbon-efficient GreenSKU. This modest TCO loss may be acceptable for cloud providers pursuing aggressive decarbonization targets. Moreover, cloud providers can leverage GSF to evaluate other SKUs that achieve the desired carbon versus cost tradeoff.

GreenSKUs Versus Other Carbon-Reduction Strategies

We comment on how designing GreenSKUs can be a more practical and complementary approach to reduce emissions compared to three alternative, commonly proposed carbon reduction strategies. To match GreenSKU-Full's carbon savings, a data center would need to:

- Increase renewable energy use by 2.6%, which is challenging to achieve due to 1) grid decarbonization's slow pace (~1% annually in the United States) and 2) exponentially high costs required to stabilize grid power with more renewables.[3]
- Improve server energy efficiency by 28%, which often 1) takes years to achieve (e.g., AMD Zen 3 to Zen 4 achieves a 25% performance per watt improvement over two years) and 2) requires a complete hardware refresh, increasing embodied emissions.
- Extend server lifetimes from six to 13 years, which can increase 1) maintenance costs prohibitively and 2) operational emissions from energy-inefficient, aging hardware.

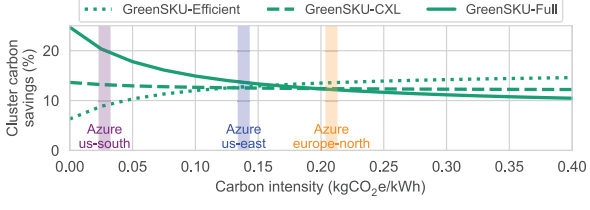


FIGURE 5. End-to-end cluster-level carbon savings relative to an all-baseline-SKU cluster across carbon intensities evaluated for our three GreenSKUs. Vertical lines mark estimated carbon intensities for three Azure data center regions. The best GreenSKU design depends on the data center's operating conditions.

Authorized licensed use limited to: Carnegie Mellon Libraries. Downloaded on October 27,2026 at 23:15:03 UTC from IEEE Xplore. Restrictions apply.

While such strategies are important, by focusing exclusively on operational or embodied emissions, they may fall short under varying carbon intensities. GreenSKUs with GSF can complement them. With GSF, our GreenSKUs navigate the operational versus embodied emissions tradeoff to quickly achieve carbon reductions across diverse carbon intensities.

LONG-TERM IMPLICATIONS

This work has received technical press coverage.[11] We discuss industry and academic impact this work has already had and its long-term potential for catalyzing greener data centers.

GreenSKUs in Production

GreenSKUs have progressed beyond research to early-stage production deployment. Azure is exploring GreenSKUs' adoption for both internal and public cloud customers, with many production teams investigating GreenSKUs' technical, process, and product fit. The business case is compelling: major customers are increasingly interested in adopting GreenSKUs, driven by both carbon and cost reduction incentives. As customer interest grows, Azure is considering GreenSKUs as an important strategy to meet Microsoft's publicly stated 2030 decarbonization targets.

WITH GSF, OUR GREENSKUS NAVIGATE THE OPERATIONAL VERSUS EMBODIED EMISSIONS TRADEOFF TO QUICKLY ACHIEVE CARBON REDUCTIONS ACROSS DIVERSE CARBON INTENSITIES.

Our work has achieved broader industry impact. Beyond Azure, our work influenced other key cloud operators to design GreenSKUs to achieve their decarbonization goals. As examples, our work inspired Google to create a dedicated team of researchers/engineers to design GreenSKUs for their production cloud, using GSF. Meta is exploring DRAM reuse via CXL, directly building on our designs.[12]

Global Carbon Reduction

Under production constraints, GreenSKUs reduce Azure's net emissions by 8%. With interest from other hyperscalers, GreenSKUs' widespread use can cut ~100 million of the 2.5 billion metric tons of CO_2e that the cloud is projected to emit by 2030,[1] equivalent to eliminating annual emissions from entire countries like Qatar/Venezuela. These carbon savings are just the beginning. GSF promises even greater savings by enabling the design and evaluation of future GreenSKUs with more advanced carbon-efficient components.

A Series of Firsts

This work is the first to systematically design and evaluate carbon-efficient servers at hyperscale. In fact, due to the challenges in designing low-carbon production servers, this work is a *series of firsts*, establishing a precedent for future carbon-aware hyperscale design. As examples, this work paves the way for future work by being the first to:

- create a carbon-driven cloud business model, while respecting the FSP norm
- show how to systematically study a low-carbon server's potential at scale
- introduce a holistic carbon model incorporating hyperscale parameters, e.g., rack-level power
- reveal key operational versus embodied emissions tradeoffs in low-carbon server design
- break down how data center components contribute to its emissions
- implement hardware reuse at scale while balancing carbon and performance
- record production DRAM failure rates beyond standard lifetime
- apply CXL to reuse DRAM, when CXL is typically used to pool/extend memory
- establish green servers' effectiveness compared to other decarbonization approaches.

Defining Carbon Tradeoffs

Using GSF, we define tradeoffs in GreenSKU design versus deployment capability (e.g., performance) and how that affects net emissions. We show how achieving optimal carbon savings requires carefully balancing operational and embodied emissions. By formulating such tradeoffs concretely, we enable lasting impact on designing greener future computing devices.

An End-to-End Methodology to Design Green Servers

We introduce an end-to-end methodology and framework, GSF, to enable cloud providers to quantify cloud-scale carbon reductions from GreenSKUs. GSF is generic, enabling any cloud provider to plug in their deployment constraints. We also describe Azure's GSF implementation,[5] which can be borrowed.

Authorized licensed use limited to: Carnegie Mellon Libraries. Downloaded on October 27,2026 at 23:15:03 UTC from IEEE Xplore. Restrictions apply. © 2025

We open sourced our carbon model and Azure's GSF implementation,[8] providing valuable data on power consumption, embodied carbon, and component failure rates, greatly expanding carbon data that future work can use.

Long-Term Sustainable Computing

This work fundamentally transforms the hyperscale industry's approach to sustainability. By building real GreenSKU prototypes (unlike the norm of simulating/modeling new servers), it shows that designing/deploying GreenSKUs is not hypothetical, but feasible using commodity devices. Thus, it can drive widespread adoption of low-carbon servers across hyperscalers, greatly reducing global emissions and (potentially) e-waste. Its clear framing of carbon tradeoffs can become a cornerstone principle in sustainable computing. Its open source carbon model and end-to-end methodology for designing green servers can catalyze the application of sustainable design principles to other computing domains. This work can be pivotal in achieving cloud operators' decarbonization targets and setting new standards for environmentally responsible computing at scale.

ACKNOWLEDGMENTS

This work was partially supported by National Science Foundation (NSF) CAREER Award CCF-2340042. Jaylen Wang was supported by an NSF Graduate Research Fellowship, Benjamin Garver Lamme/Westinghouse Fellowship, and Jack and Mildred Bowers Scholarship in Engineering.

REFERENCES

1. I. O'Brien, "Data center emissions probably 662% higher than big tech claims. Can it keep up the ruse?" *The Guardian*, Sep. 15, 2024. [Online]. Available: https://www.theguardian.com/technology/2024/sep/15/data-center-gas-emissions-tech
2. B. Knowles, *ACM TechBrief: Computing and Climate Change*. New York, NY, USA: ACM, 2021.
3. J. Lyu et al., "Myths and misconceptions around reducing carbon embedded in cloud platforms," in *Proc. Workshop Sustain. Comput. Syst.*, 2023, pp. 1–7, doi: 10.1145/3604930.3605717.
4. U. Gupta et al., "ACT: Designing sustainable computer systems with an architectural carbon modeling tool," in *Proc. Int. Symp. Comput. Archit.*, 2022, pp. 784–799.
5. J. Wang et al., "Designing cloud servers for lower carbon," in *Proc. ACM/IEEE 51st Annu. Int. Symp. Comput. Archit. (ISCA)*, 2024, pp. 452–470, doi: 10.1109/ISCA59077.2024.00041.
6. M. Connatser. "AMD adds 128-core Bergamo and 3D V-Cache Genoa CPUs to Zen 4 Epyc lineup." XDA. Accessed: Apr. 26, 2024. [Online]. Available: https://www.xda-developers.com/amd-128-core-bergame-genoa-epyc-cpu/
7. D. D. Sharma, R. Blankenship, and D. S. Berger, "An introduction to the compute express link (CXL) interconnect," *ACM Comput. Surv.*, vol. 56, no. 11, pp. 1–37, 2024.
8. "GreenSKU-model." GitHub. Accessed: Mar. 20, 2025. [Online]. Available: https://github.com/Azure/AzurePublicDataset
9. J. Wang, U. Gupta, and A. Sriraman, "Peeling back the carbon curtain: Carbon optimization challenges in cloud computing," in *Proc. 2nd Workshop Sustain. Comput. Syst.*, 2023, pp. 1–7, doi: 10.1145/3604930.3605718.
10. A. Parayil et al., "Towards workload-aware cloud efficiency: A large-scale empirical study of cloud workload characteristics," in *Proc. 16th ACM/SPEC Int. Conf. Perform. Eng.*, 2025, pp. 136–146.
11. L. Hautala. "Servers get a second life for sustainability." IEEE Spectrum. Accessed: Mar. 20, 2025. [Online]. Available: https://spectrum.ieee.org/amp/server-reuse-2669736974
12. K. Afifi-Sabet. "AMD: Meta are working on revolutionary tech that could recycle petabytes worth of RAM." TechRadar. Accessed: Mar. 20, 2025. [Online]. Available: https://www.techradar.com/pro/amd-meta-are-working-on-revolutionary-tech-that-could-recycle-petabytes-worth-of-ram-cxl-could-help-save-hyperscalers-tens-of-millions-of-dollars-while-improving-performance

JAYLEN WANG is a Ph.D. candidate at Carnegie Mellon University, Pittsburgh, PA, 15213, USA. His research interests include computer architecture and software systems to enable environmentally sustainable data center systems. Wang received his B.S. degree in electrical engineering from Harvard University. Contact him at jaylenw@andrew.cmu.edu.

DANIEL S. BERGER is a principal researcher at Microsoft, Redmond, WA, 98052, USA, in the Azure Research-Systems group, and an affiliate assistant professor in computer science at the University of Washington, Seattle, WA, 98195, USA. His research interests include improving data center efficiency and sustainability. Berger received his Ph.D. degree in computer science from the Technical University Kaiserslautern. Contact him at daberg@microsoft.com.

Authorized licensed use limited to: Carnegie Mellon Libraries. Downloaded on October 27,2026 at 23:15:03 UTC from IEEE Xplore. Restrictions apply.

FIODAR KAZHAMIAKA is a senior researcher at Microsoft, Redmond, WA, 98052, USA, in the Azure Research-Systems group. His research interests include the design and operation of efficient data center systems. Kazhamiaka received his Ph.D. degree in computer science from the University of Waterloo. Contact him at fkazhamiaka@microsoft.com.

CELINE IRVENE is a research software development engineer at Microsoft, Redmond, 98052, WA, USA, in the Azure Research-Systems group. Her research interests include efficient cloud capacity modeling and artificial intelligence for system design and operations. Irvene received her Ph.D. degree in electrical and computer engineering from the Georgia Institute of Technology. Contact her at celineirvene@microsoft.com.

CHAOJIE ZHANG is a research software development engineer at Microsoft, Redmond, WA, 98052, USA, in the Azure Research-Systems group. Her research interests include artificial intelligence efficiency and data center infrastructure. Zhang received her Ph.D. degree in computer science from the University of Chicago. Contact her at chaojiezhang@microsoft.com.

ESHA CHOUKSE is a principal researcher at Microsoft, Redmond, WA, 98052, USA, in the Azure Research-Systems group. Her research interests include efficient artificial intelligence deployment across the stack, crossing the layers of the artificial intelligence platform, hardware, and data center infrastructure. Choukse received her Ph.D. degree in computer architecture from the University of Texas at Austin. Contact her at esha.choukse@microsoft.com.

KALI FROST is a senior technical program manager in the Microsoft Cloud Supply Chain, Redmond, WA, 98052, USA. Her research interests include environmental lifecycle assessment of cloud hardware, including semiconductor impact quantification. Frost received her Ph.D. degree in environmental engineering from Purdue University. Contact her at kalifrost@microsoft.com.

RODRIGO FONSECA is a senior principal research manager at Microsoft, Redmond, WA, 98052, USA, where he leads the Azure Research-Systems group. His research interests include improving the efficiency of cloud infrastructure. Fonseca received his Ph.D. degree in computer science from the University of California, Berkeley. Contact him at fonseca.rodrigo@microsoft.com.

BRIJESH WARRIER is a Partner Group engineering manager at Microsoft, Redmond, WA, 98052, USA. He leads the software-defined data center team that focuses on infrastructure efficiency and is involved in research on datacenter power, cooling, and operations. Warrier received his M.S. degree from the University of Cincinnati. Contact him at brijeshr@microsoft.com.

CHETAN BANSAL is a senior principal research manager at Microsoft, Redmond, WA, 98052, USA, where he leads the M365 Research-AIOps research group. His research interests include the design of data-driven techniques for improving cloud reliability, efficiency, and developer productivity. Bansal received his M.S. degree from the Birla Institute of Technology and Science, Pilani. Contact him at chetanb@microsoft.com.

JONATHAN STERN is a principal hardware product manager at Microsoft Azure, Redmond, WA, 98052, USA. His research interests include system design and definition, including modeling total cost of ownership and lifetime carbon emissions. Stern received his B.S. degree in computer science from Trinity University. Contact him at jonstern@microsoft.com.

RICARDO BIANCHINI is a technical fellow and corporate vice president at Microsoft Azure, Redmond, WA, 98052, USA, where he leads the compute capacity and efficiency organization and drives research in cross-stack cloud datacenter optimization. Bianchini received his Ph.D. degree in computer science from the University of Rochester. Contact him at ricardob@microsoft.com.

AKSHITHA SRIRAMAN is an assistant professor at Carnegie Mellon University, Pittsburgh, PA, 15213, USA. Her research interests include computer architecture and software systems to enable sustainable and impartial web systems that also achieve high performance. Sriraman received her Ph.D. degree in computer science and engineering from the University of Michigan. Contact her at akshitha@cmu.edu.

Authorized licensed use limited to: Carnegie Mellon Libraries. Downloaded on October 27,2026 at 23:15:03 UTC from IEEE Xplore. Restrictions apply.