Analyzing the Total Cost of Ownership of Carbon in Data Centers

Mitchell Elliott, Andrew Quinn, Heiner Litz
Center for Research in Systems and Storage
University of California, Santa Cruz

UC SANTA CRUZ
BaskinEngineering

Center for Research in Systems and Storage

NSF

Introduction

- ❖ **Context: Data centers are major contributors to global carbon emissions due to high energy consumption and embodied carbon.**
  - ➢ Estimated to be ~1-1.3% of the world's electricity consumption. [1]
  - ➢ Expected growth: 240–340 TWh → 290–600 TWh by 2030. [2]
  - ➢ Conservative Estimate: ~2× growth within 6 years.
- ❖ **Environmental Impact:**
  - ➢ Responsible for ~1% of energy-related GHG emissions. [1]
  - ➢ Comparable to the emissions of the aviation industry.
- ❖ **Rising Scrutiny:**
  - ➢ Increasing demand for digital services raises questions about sustainability.
  - ➢ Need to address both operational and embodied carbon to mitigate environmental impact.

Agenda

CRSS

- Background
- *CarbonStream*
- Evaluation
- Conclusion
- Future Work

Problem Statement and Goals

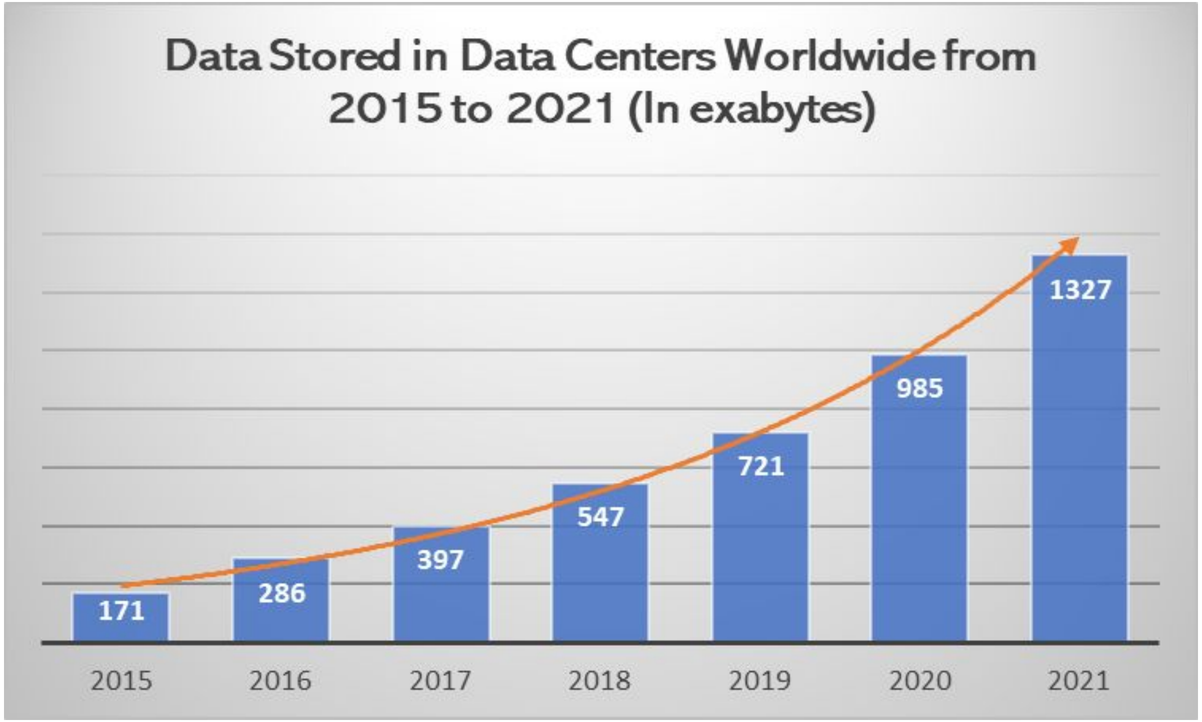
CRSS

- Problem Statement: Storage media contribute significantly to the embodied carbon in data centers, especially for storage-heavy workloads.
- Research Goal: Compare the Total Cost of Ownership (TCO) of carbon across SSDs, HDDs, tape, and glass-based storage, exploring if glass-based storage can reduce total carbon impact.

Background

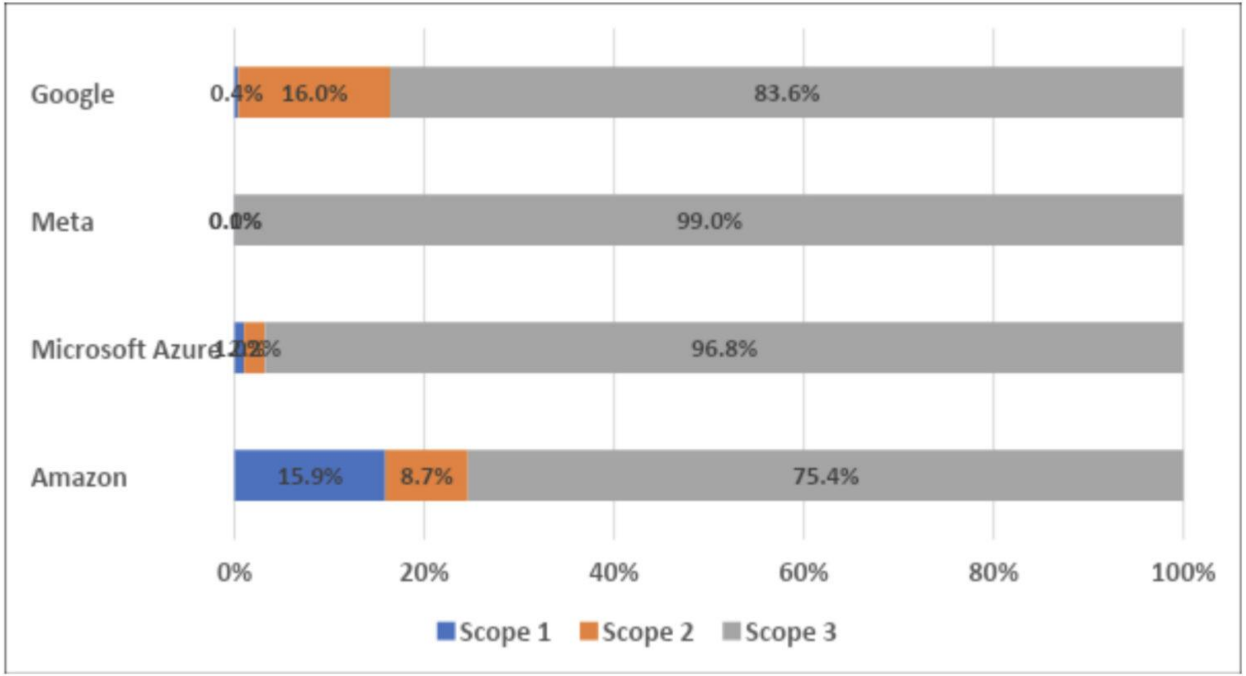
- **Data Centers and Carbon Emissions:**
  - Significant contributors to global greenhouse gas (GHG) emissions.
  - Demand for digital services continues to grow, amplifying energy use and carbon footprint.
- **Greenhouse Gas (GHG) Emissions:**
  - Scope 1: Direct emissions from owned or controlled sources.
  - Scope 2: Indirect emissions from purchased electricity.
  - Scope 3: Indirect emissions from the value chain, including embodied carbon in infrastructure.
- **Total Cost of Ownership (TCO) of Carbon = Operational Carbon (Scope 2) + Embodied Carbon (Scope 3)**
- **Environmental Impact: Rising scrutiny on data center sustainability due to energy consumption and carbon footprint.**
- **Focus Shift: Emphasis traditionally focused on Scope 2 emissions, but Scope 3 is increasingly recognized for its impact.**

Global Data Center Storage Capacity



Source: Statista [3]

Carbon Emissions of Cloud Providers



Sources: Schneider Electric 2023 derived from the sustainability reporting of listed organisations. [4]

Embodied Carbon in Data Centers

- Definition: Embodied carbon includes greenhouse gas emissions from the full lifecycle of a product.
- Impact: Storage media are a major source of embodied carbon, particularly in high-demand data centers.
- Storage is not negligible.
- Servers dominate $.

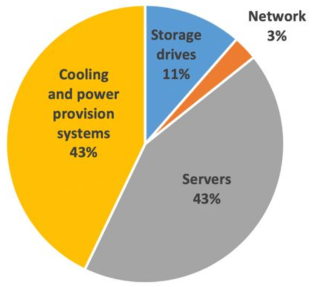
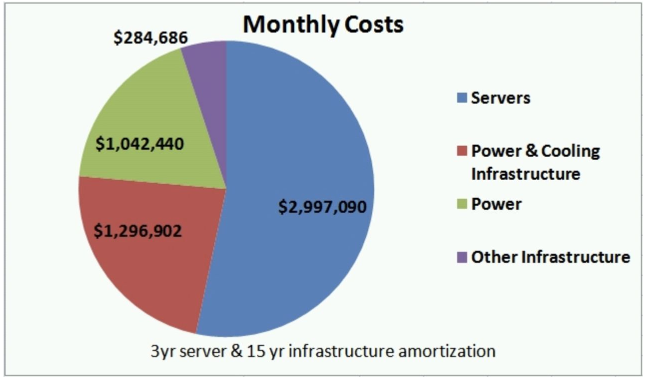


Figure 1. Fraction of U.S. data center electricity use in 2014, by end use. Source: Shehabi 2016.



Cost of Power in Large-Scale Data Centers [5]

Storage Technologies Compared

CRSS

Storage Type	Cost	Performance	Key Characteristics
DRAM	Very expensive	Extremely high speed, low latency	Extremely fast, volatile memory
SSD	Expensive	High speed, low latency	High speed, non-volatile storage
HDD	Cheap	Moderate speed, high latency	High capacity, slower than SSDs
Tape	Cheap	Very low speed, very high latency	Ideal for backups, low operational cost
Glass	?	Very low speed, very high latency	Low carbon cost, highly durable

Previous Research

CRSS

- **Operational Focus (Scope 2): Most prior studies in the last 30 years aimed at reducing operational carbon through:**
  - Dynamic voltage and frequency scaling. [6]
  - Renewable energy adoption. [7]
  - Workload balancing. [8]
- **Emerging Interest in Embodied Carbon (Scope 3):**
  - *Carbon Explorer* [9] and *Chasing Carbon* [10] frameworks highlight embodied carbon as critical for holistic carbon assessment.

Rationale for This Study

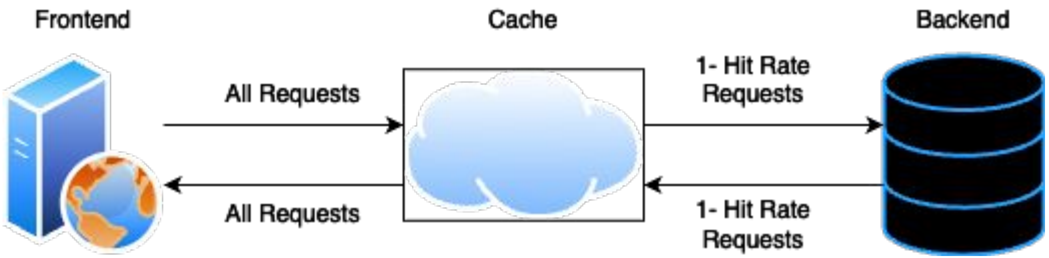
- **Need for Comprehensive Carbon Analysis:**
  - Embodied carbon in data centers is significant yet under-addressed.
  - Lifespan of storage devices directly impacts replacement carbon costs.
- **Focus on Lifespan:**
  - Longer lifespans reduce total replacements, lowering TCO of carbon.
  - Trade-offs exist between durability, performance, and carbon footprint.
- **Contribution:**
  - Provides a comparative carbon analysis of SSDs, HDDs, tape, and glass-based storage, with a focus on device lifespan impacts on sustainability.

CarbonStream

- ❖ Purpose: Custom simulation model to estimate TCO of carbon in data centers.
- ❖ System Architecture: Three-tiered setup (frontend, cache, backend).
- ❖ Metrics: Includes embodied, operational, and replacement carbon costs.
- ❖ Inputs: System configuration and workload parameters
- ❖ Outputs: Graph of calculated costs and optimal storage system setup

System Architecture

- **Components:**
  - Frontend Tier: Processes requests. (DRAM/SSD)
  - Cache Tier: Holds frequently accessed data. (DRAM/SSD)
  - Backend Tier: Primary data storage (SSDs, HDDs, tape, glass).



SLO Parameters

- **Service-Level Objective (SLO): Measurable target that defines the level of service a customer can expect to receive.**
- **Key Metrics:**
  - Latency Requirement: Max end-to-end time for data requests.
  - Throughput Requirement: Data requests per second.

Calculation of Carbon Costs

- **Types of Costs:**
  - Embodied Carbon Cost: Full lifecycle emissions.
    - Embodied Carbon Cost = ∑(# of Devices × Carbon Intensity per Unit)
  - Operational Carbon Cost: Emissions during use.
    - Active Operation Cost = ∑(# of Servers × Active Power Consumption × Carbon Intensity of Energy × Operational Time × Active Time %)
    - Idle Operation Cost = ∑(# of Servers × Idle Power Consumption × Carbon Intensity of Energy × Operational Time × Idle Time %)
  - Replacement Carbon Cost: Emissions from hardware replacements.
    - Replacement Carbon Cost = ∑(# of Replacements × Embodied Carbon Cost per Device)

Calculation of Average Latency

CRSS

- **Includes: Latency from frontend, cache, backend, and network delays.**
- **Formula:**
  - Average Latency = Frontend Latency + Cache Hit Rate × Cache Latency + (1 − Cache Hit Rate) × (Cache Latency + Backend Latency) + Network Latency + Processing Latency

Calculation of Peak Throughput

- **Minimum throughput capacity of frontend, cache, and backend tiers.**
- **Formula:**
  - Peak Throughput = min(Frontend Total Throughput, Cache Total Throughput, Backend Total Throughput)

Calculation of the Number of Servers Needed

- **Objective: Ensure system can meet throughput demands.**
- **Formula:**
  - Number of Servers = ⌈ Desired Throughput / Throughput per Server ⌉
- **Considerations:**
  - Balances required server capacity with carbon costs per server.

Calculation of the Cache Hit Rate

- **Definition: Likelihood that data requests are served from the cache instead of backend storage.**
- **Formula: Cache Hit Rate = Total Cache Size / Total Data Size**
  - Where: Total Cache Size = Number of Cache Servers x Cache Server Size
- **Impact on Performance: Higher cache hit rate lowers backend load, reducing latency and operational carbon costs.**

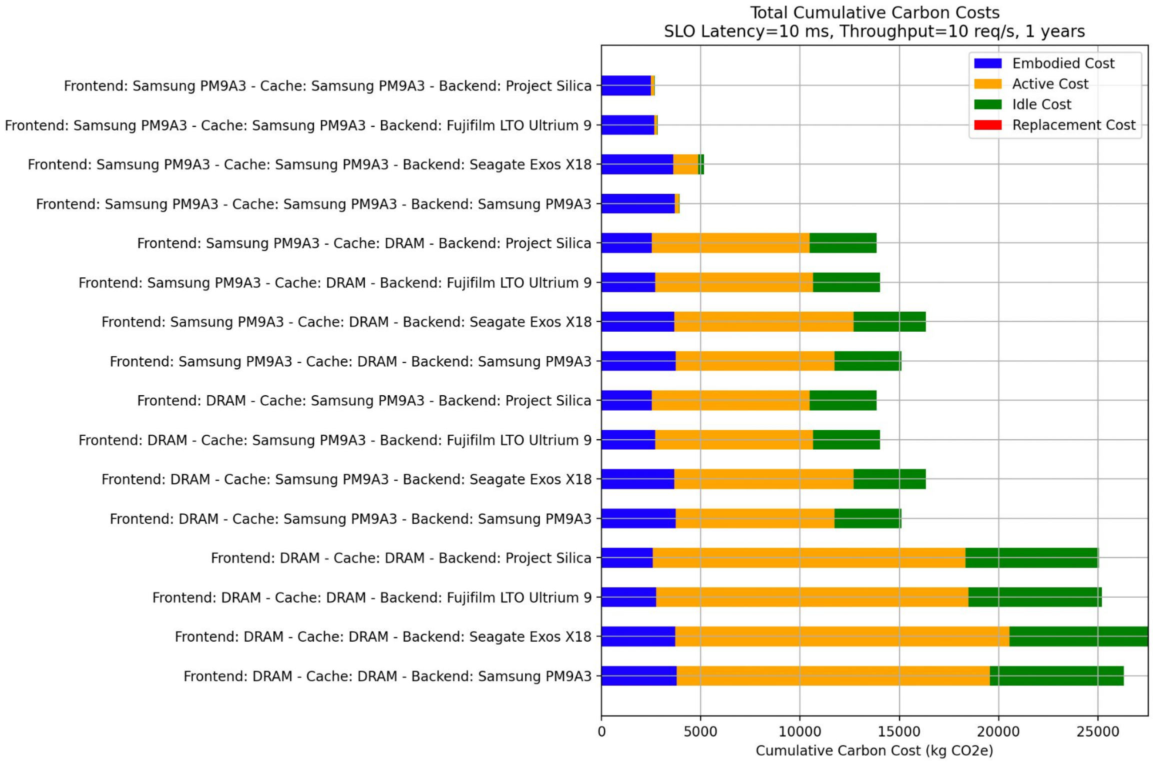
Evaluation - Model Implementation

- Implemented in Python.
- Simulation set to run over a predefined period.
  - 10 years by default.
- Tracks the energy consumption, device replacements, latency, throughput, and overall carbon footprint.
- System is assumed to store a set amount of data.
  - 10 Billion GB (10 EB) by default.

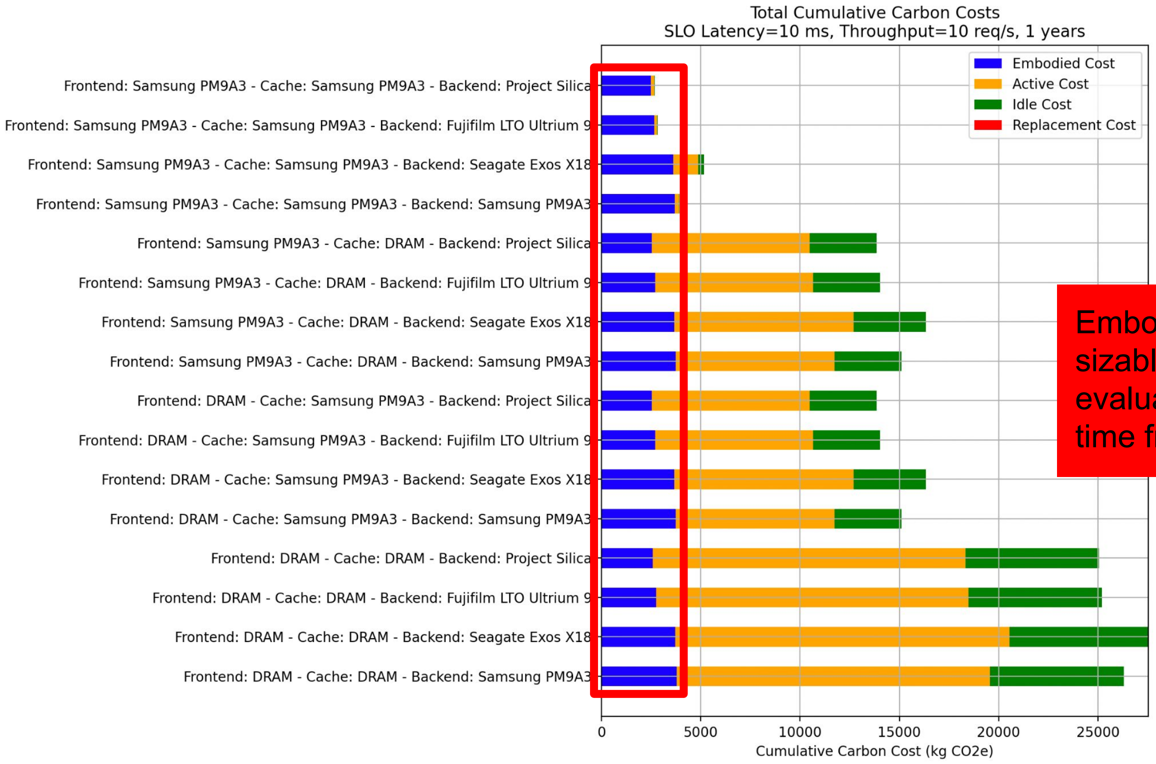
Evaluation - System Hardware

	DRAM	SSD	HDD	Tape	Glass
Device Name	N/A	Samsung PM9A3	Seagate Exos X18	Fujifilm LTO Ultrium 9	Project Silica
Capacity	4 TB	3.84 TB	18 TB	18 TB	7 TB
Latency	10 ns	0.08 ms	4.16 ms	10000 ms	2000 ms
Throughput	20 GB/s	5 GB/s	270 MB/s	400 MB/s	210 MB/s
Embodied Cost	0.31 kg CO2e/GB	0.16 kg CO2e/GB	0.0017 kg CO2e/GB	0.00042 kg CO2e/GB	0.0001 kg CO2e/GB
Power Consumption	2.5 KW	12 W (A), 3.5 W (I)	9.5 W (A), 5.3 W (I)	0.26 W (A), 0 W (I)	0.13 W (A), 0 W (I)
Lifespan	10 years	5 years	5 years	30 years	100 years

Simulation 1: Latency = 10 ms, Throughput = 10 req/s, 1 Year

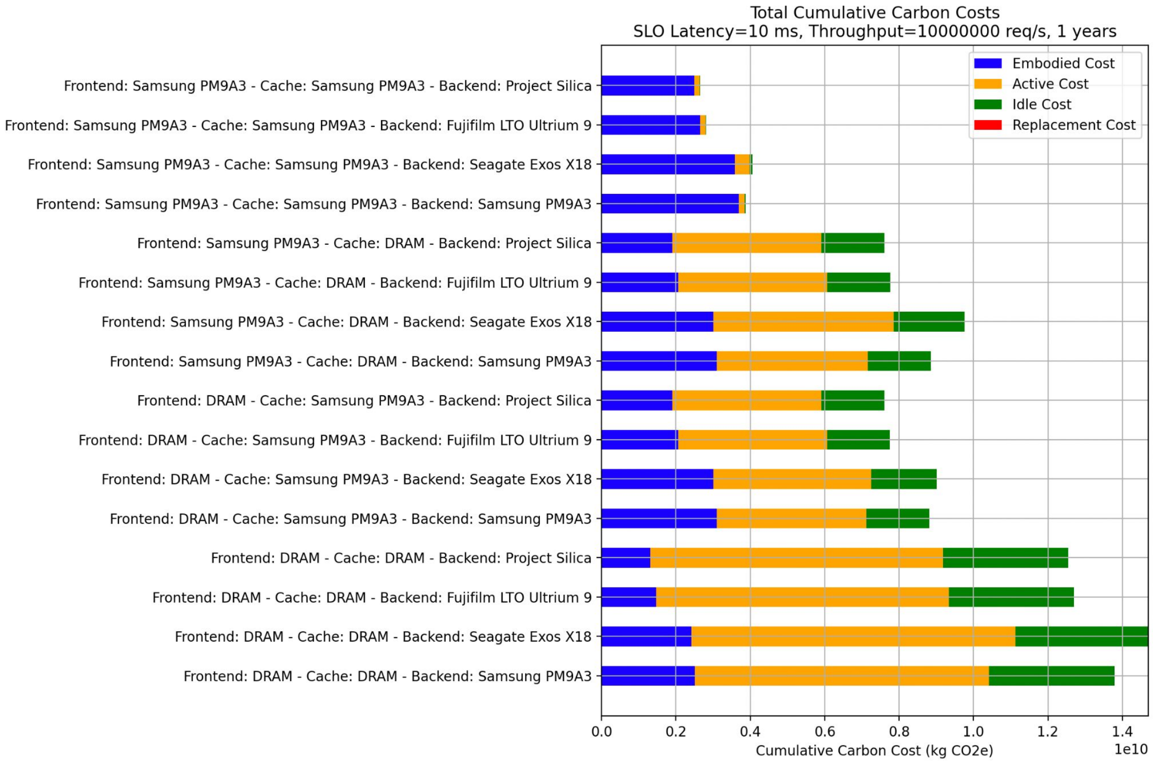


Simulation 1: Latency = 10 ms, Throughput = 10 req/s, 1 Year

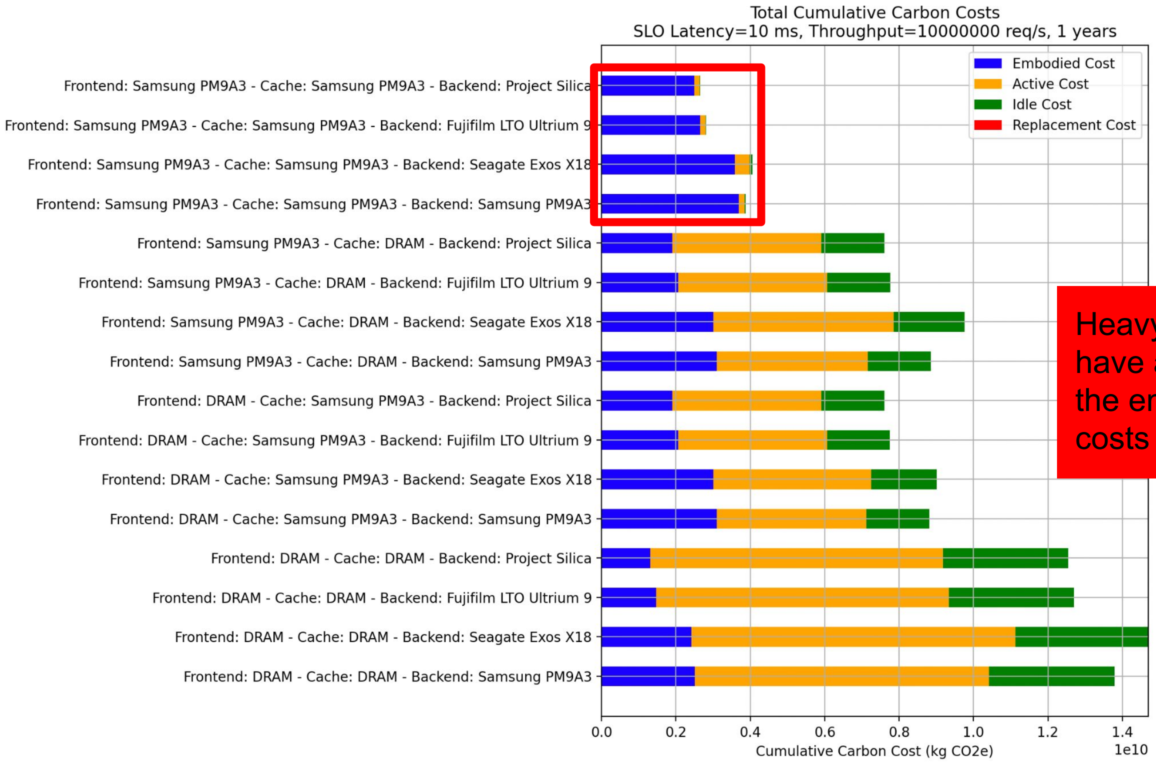


Embodied carbon is a sizable cost when evaluating over shorter time frames.

Simulation 2: Latency = 10 ms, Throughput = 10000000 req/s, 1 Year

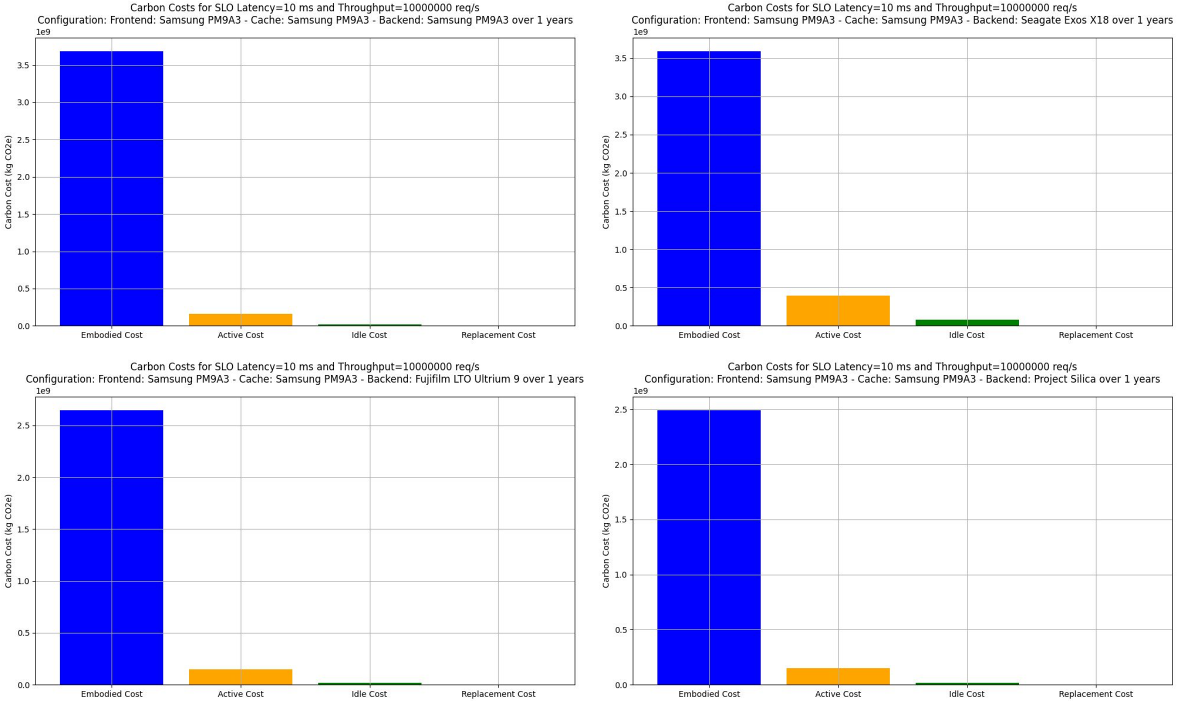


Simulation 2: Latency = 10 ms, Throughput = 10000000 req/s, 1 Year

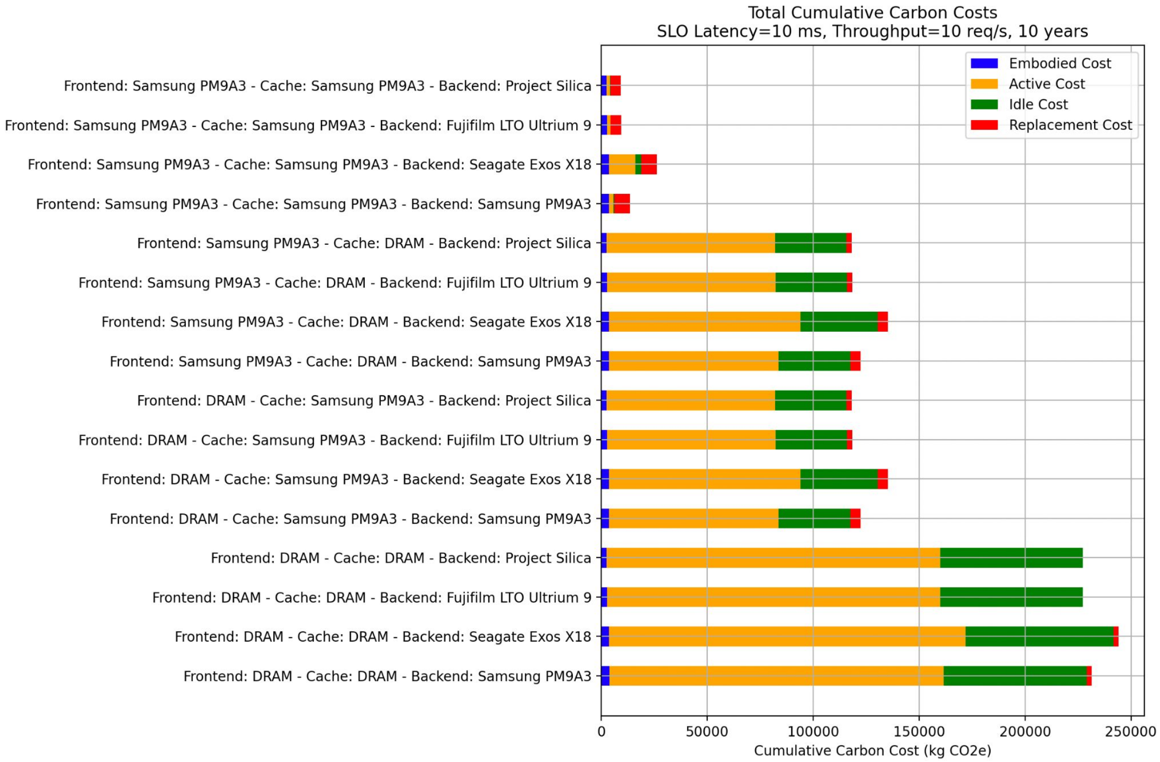


Heavy workloads can have a sizeable impact on the embodied carbon total costs of data centers.

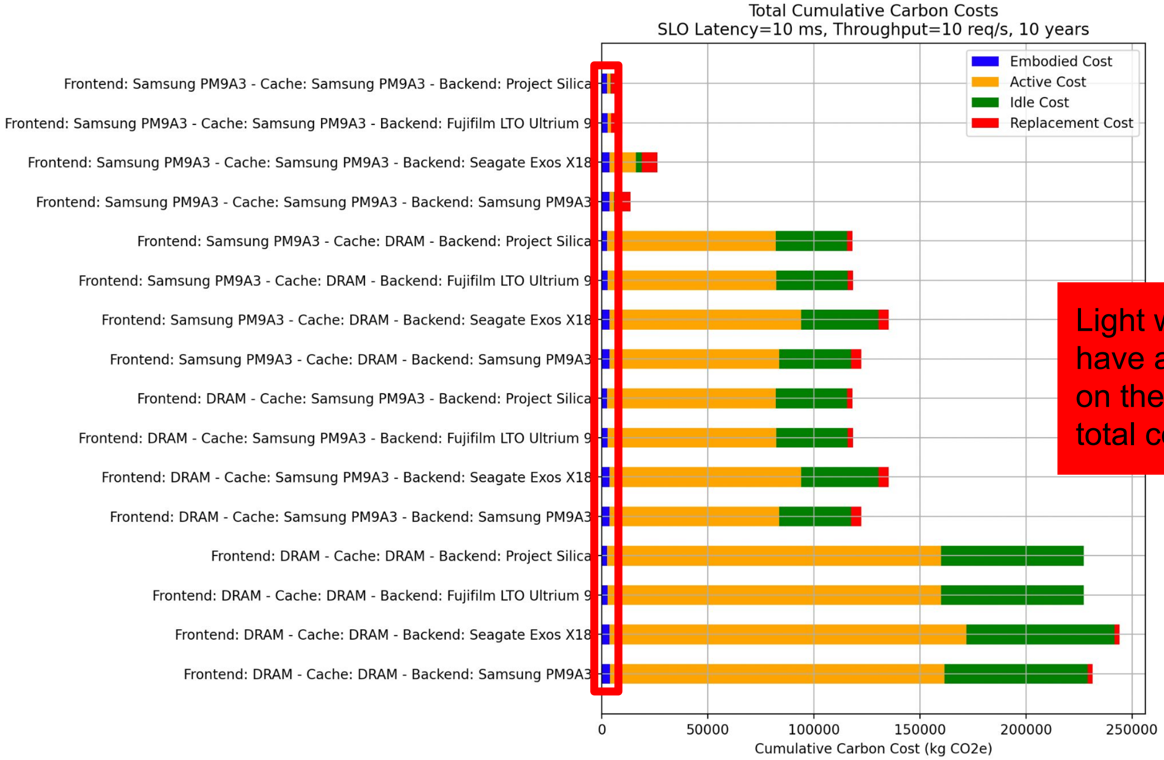
Simulation 2 Backend Configuration Comparison



Simulation 3: Latency = 10 ms, Throughput = 10 req/s, 10 Years

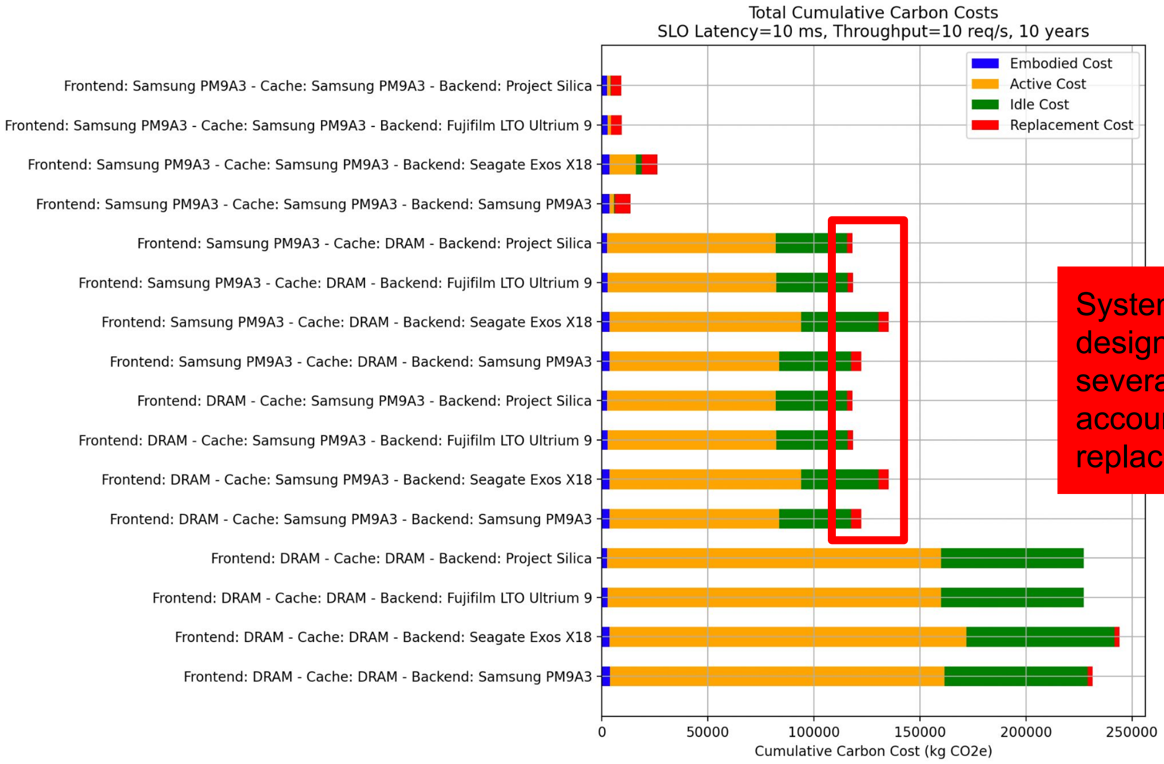


Simulation 3: Latency = 10 ms, Throughput = 10 req/s, 10 Years



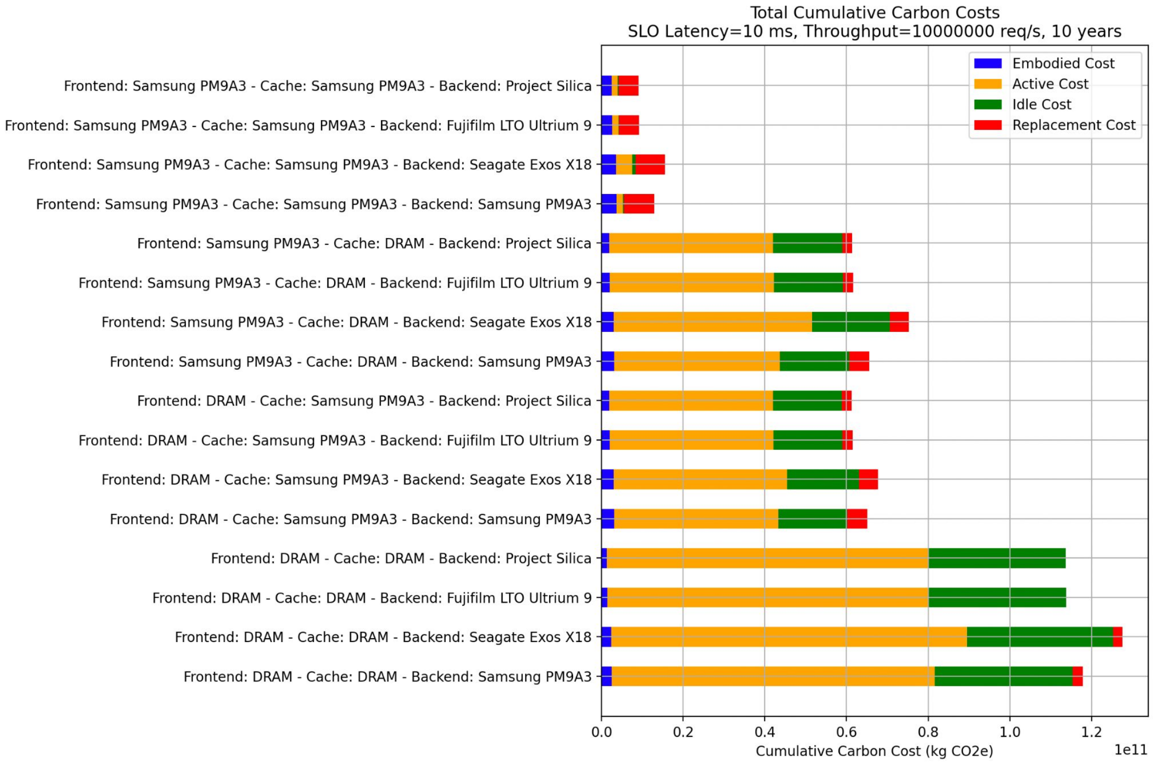
Light workloads often have a negligible impact on the embodied carbon total costs of data centers.

Simulation 3: Latency = 10 ms, Throughput = 10 req/s, 10 Years

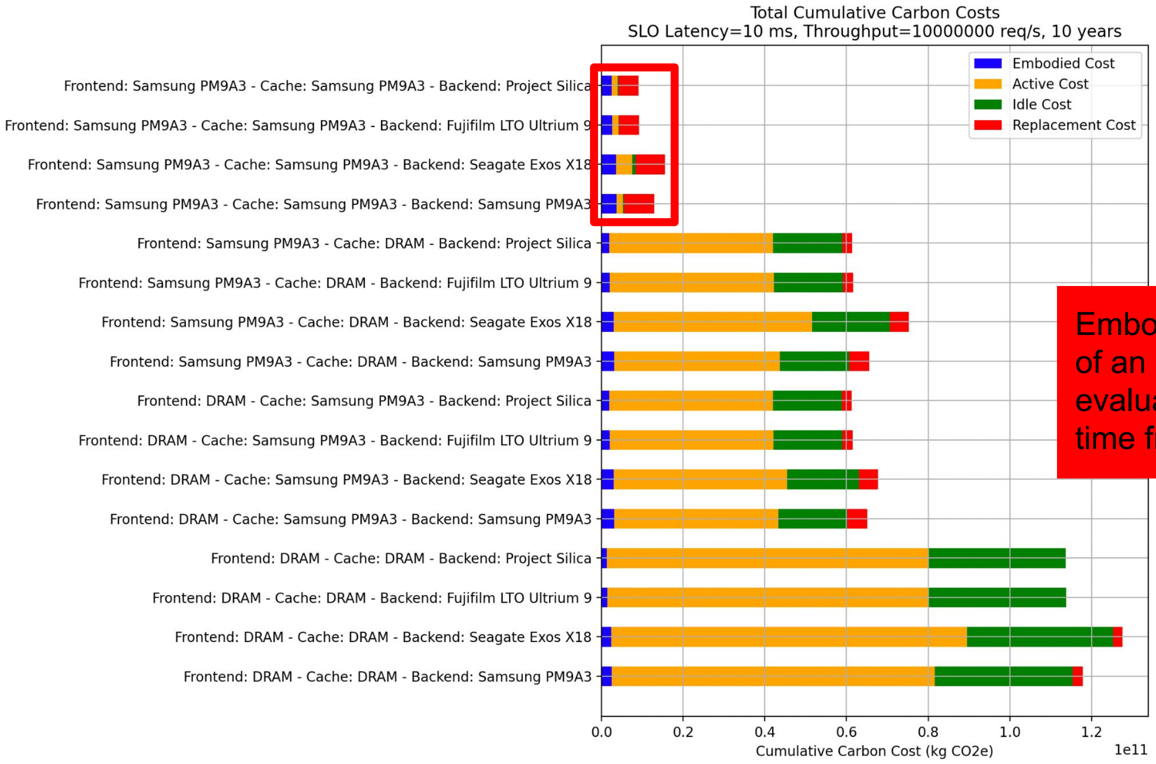


Systems that are designed to last for several years must account for device replacements.

Simulation 4: Latency = 10 ms, Throughput = 10000000 req/s, 10 Years

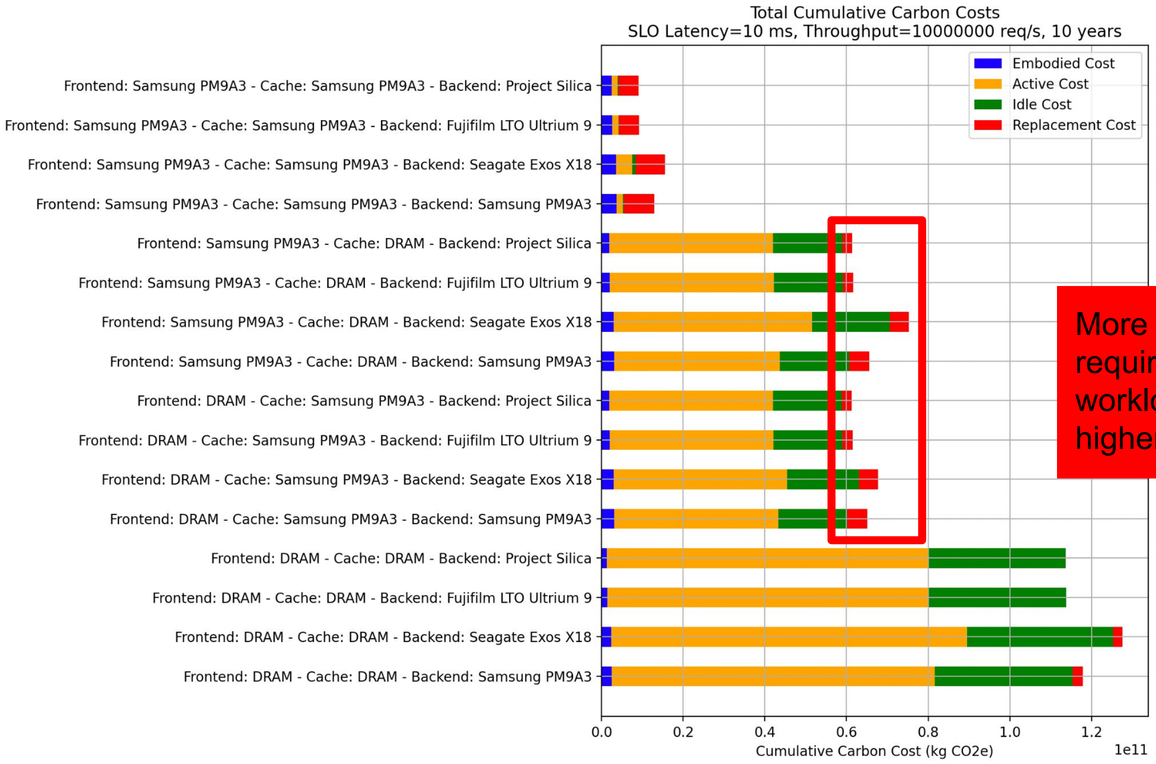


Simulation 4: Latency = 10 ms, Throughput = 10000000 req/s, 10 Years



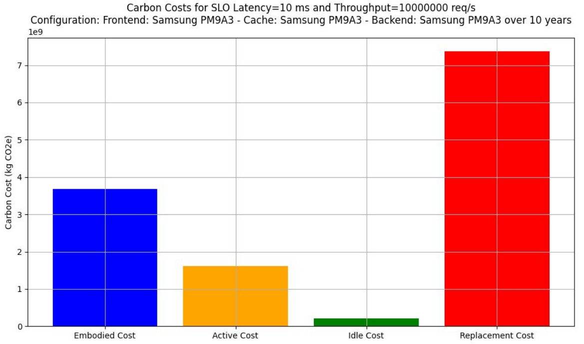
Embodied carbon has less of an impact when evaluating over longer time frames.

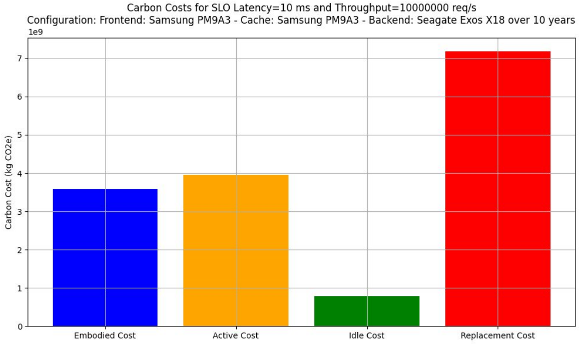
Simulation 4: Latency = 10 ms, Throughput = 10000000 req/s, 10 Years

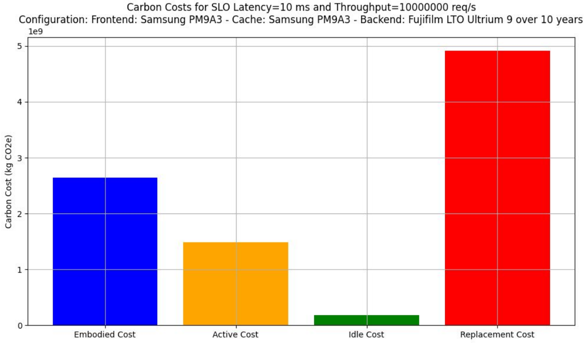


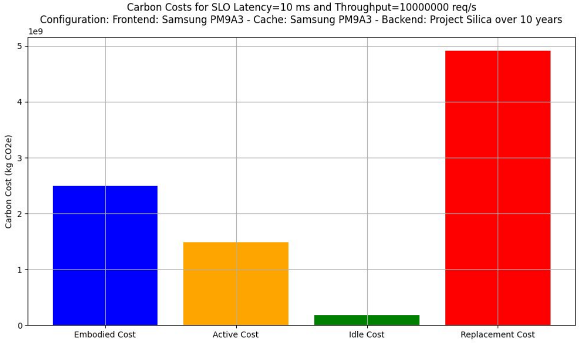
More storage devices are required to support heavy workloads, which leads to higher replacement costs.

Simulation 4 Backend Configuration Comparison









Impact of Frontend and Cache Configurations

- **DRAM**
  - Provides the highest performance and the lowest latency.
  - Higher embodied and operational carbon costs.
- **SSDs**
  - Slower performance.
  - Lower environmental costs.
- Takeaway: Use DRAM when performance is critical and SSDs when carbon cost is a concern.

Total Carbon Cost Analysis

CRSS

- ❖ DRAM: Highest performance but highest embodied carbon; impact lessens over time.
- ❖ SSDs: High performance with lower carbon cost than DRAM; effective for long-term use.
- ❖ HDDs: Lower embodied carbon than SSDs but highest total carbon usage overall.
- ❖ Tape Storage: Low carbon cost, ideal for infrequent archival storage.
- ❖ Glass-Based Storage: Emerging, durable, and environmentally promising, especially with SSDs in hybrid systems.
- ❖ Conclusion: Balancing performance and carbon impact is critical; tape and glass storage excel in sustainability for long-term use.

Conclusion

- Glass-based storage stands out as a viable, low-carbon solution for storing archival data in data centers.
- Tape also presents a strong option for archival storage.
- SSDs, while fast, come with high embodied costs.
- Key Insight: Informed storage choices and cache configurations can significantly impact sustainability.

Future Work

- ❖ Data compression integration.
- ❖ Variable data access patterns.
- ❖ Redundancy management strategies.
- ❖ Carbon intensity fluctuations.
- ❖ Durability testing for glass storage.
- ❖ Workload influence on storage configurations.

References

[1] Data Centres and Data Transmission Networks:

https://www.iea.org/energy-system/buildings/data-centres-and-data-transmission-networks

[2] Climate Progress 2024: REPEAT Project's Annual U.S. Emissions Pathways Update:

https://zenodo.org/records/13345138

[3] Amount of data actually stored in data centers worldwide from 2015 to 2021:

https://www.statista.com/statistics/638613/worldwide-data-center-storage-used-capacity/

References

[4] Scope 3 is the next challenge on the journey towards decarbonisation in the data centre industry. Here are some thoughts on how to meet it: https://w.media/scope-3-is-the-next-challenge-on-the-journey-towards-decarbonisation-in-the-data-centre-industry-here-are-some-thoughts-on-how-to-meet-it/

[5] Cost of Power in Large-Scale Data Centers: https://perspectives.mvdirona.com/2008/11/cost-of-power-in-large-scale-data-centers/

[6] David el al. Memory Power Management via Dynamic Voltage/Frequency Scaling. In ICAC ’11: Proceedings of the 8th ACM international conference on Autonomic computing (June 2011), pp. 31– 40. https://dl.acm.org/doi/pdf/10.1145/1998582.1998590.

References

[7] Souza et al. 2023. Ecovisor: A Virtual Energy System for Carbon-Efficient Applications. In Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS 2023). Association for Computing Machinery, New York, NY, USA, 252–265. https://doi.org/10.1145/3575693.3575709.

[8] Radovanovic el al. 2021. Carbon-Aware Computing for Datacenters. https://arxiv.org/pdf/2106.11750.

[9] Acun et al. 2023. Carbon Explorer: A Holistic Framework for Designing Carbon Aware Datacenters. In Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS 2023). Association for Computing Machinery, New York, NY, USA, 118–132. https://doi.org/10.1145/3575693.3575754.

References

[10] GUPTA et al. Chasing Carbon: The Elusive Environmental Footprint of Computing. In IEEE Micro, Volume 42, Issue 4 (July 2022), pp. 37– 47. https://doi.org/10.1109/MM.2022.3163226.

[11] SAMSUNG. PM9A3 Whitepaper, August 2021. https://download.semiconductor.samsung.com/resources/white-paper/PM9A3_SSD_Whitepaper_230327_2.pdf.

[12] SEAGATE. EXOS X18 Sustainability Report, 2021. https://www.seagate.com/esg/planet/product-sustainability/exos-x18-sustainability-report/.

References

[13] FUJIFILM. LTO Ultrium 9 Sell Sheet, 2021. https://asset.fujifilm.com/www/us/files/2024-01/ce72cf783970aa0d05d4463e4febcb62/FB_LTO9_SellSheet_V1.pdf.

[14] ANDERSON et al. Project Silica: Towards Sustainable Cloud Archival Storage in Glass. In The 29th ACM Symposium on Operating Systems Principles (October 2023), pp. 166–181. https://dl.acm.org/doi/pdf/10.1145/3600006.3613208.

Thank You

Questions?