SNIA DEVELOPER CONFERENCE

SDC 24

BY Developers FOR Developers

September 16-18, 2024
Santa Clara, CA

The Promise of NVMe FDP in Data Center Sustainability

Roshan R Nair
Arun George

Samsung Semiconductor India Research

Global Open-Ecosystem Team (GOST)

SAMSUNG

Agenda

- Motivation
- A Primer on Storage Sustainability
- Write Amplification vs. SSD Lifetime and SSD Power
- NVMe FDP: Quick Overview
- NVMe FDP: Carbon Emissions – Examples with FIO
- Carbon Emission Reductions: Large Scale System - CacheLib
- Summary
- Q&A

Motivation

Why have this session? What will we cover?

Motivation: Why have this session?

- Industry net-zero carbon emission goals.

- Data Center carbon emission forecast
  - "Global data center industry to emit 2.5 billion tons of CO_2 through 2030, Morgan Stanley says" [1]
  - SSDs have a sizable contribution to the carbon footprint.

- Storage – Capacity growth [2]. Demand for storage is ever increasing.

- Adoption of newer storage technologies
  - QLC adoption. Aimed at eventually replacing HDDs with SSDs.
  - SSDs have a higher carbon footprint than HDDs [3].

- A lot of tech we work on has a correlation to sustainability
  - We need to make that connection and report on sustainability as a KPI going forward.
  - e.g. data placement technologies affect SSD lifetime and power.

[1] Global data center industry to emit 2.5 billion tons of CO2 through 2030, Morgan Stanley says | Reuters
[2] The Digitization of the World From Edge to Core. https://www.seagate.com/files/www-content/our-story/trends/files/dataage-idc-report-final.pdf
[3] Tannu, Swamit, and Prashant J. Nair. "The dirty secret of ssds: Embodied carbon." ACM SIGENERGY Energy Informatics Review 3.3 (2023): 4-9.

SAMSUNG

What will we cover?

- Brief discussion on SSD lifetime, power, sustainability and NVMe FDP.
- NVMe FDP's ability to improve SSD lifetime, improve utilization and reduce power consumption thereby leading to reduced carbon emissions.
- Overall carbon emission reductions by using NVMe FDP in a large scale system: CacheLib.
- NVMe FDP can help optimize carbon emissions by enabling deployment flexibility.

A Primer on Storage Sustainability

Carbon footprint from using SSDs

Storage Sustainability: A Quick Overview

Category of Carbon Emission	Description	Typical Contribution	Comments
Scope 1 (Direct Emissions)	Emissions from direct burn of fuel.	Very Low	We don't address this in the talk.
Scope 2 (Indirect Emissions)	**"Operational Carbon Emissions"** Associated with energy purchase.	**Medium**	Sustainable power sources is pegged as the way to solve this.
Scope 3	**"Embodied Carbon Emissions"** Associated with purchase of products, hardware etc. that is used.	**High**	Largest contributor to Data Center carbon emissions.

Sustainability Metrics:

- CO2e – CO2 Equivalent [4]. We will use CO2e to quantify carbon emissions in this presentation
- Power/Energy expenditure in terms of KWh is converted to CO2e (Kg) using the Greenhouse Gas Equivalencies Calculator [5].
- For SSD embodied carbon emissions we use a value of ~0.16 Kg of CO2e per GB* of SSD capacity [3].

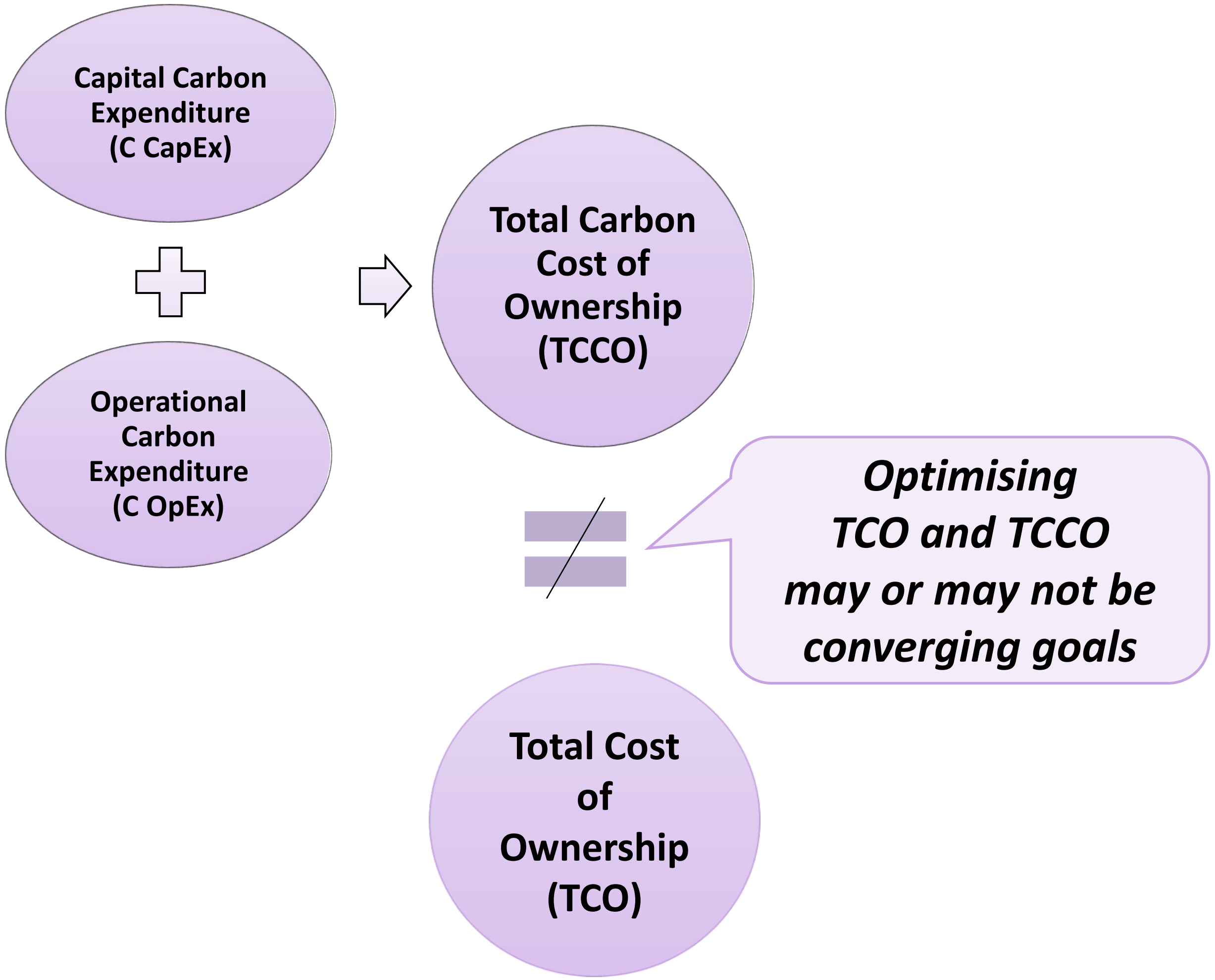
**Disclaimer: Different SSDs might have different embodied CO2e values per GB. This value is used in this talk not for a specific SSD, but to illustrate the methodology to calculate the carbon emissions for systems using SSDs. This value is not to be associated with a specific Samsung SSD product.*

[4] https://www.myclimate.org/en/information/faq/faq-detail/what-are-co2-equivalents/
[5] https://www.epa.gov/energy/greenhouse-gas-equivalencies-calculator
[3] Tannu, Swamit, and Prashant J. Nair. "The dirty secret of ssds: Embodied carbon." ACM SIGENERGY Energy Informatics Review 3.3 (2023): 4-9.

SAMSUNG

Carbon footprint from using SSDs

- Storage System LifeCycle Assessment (LCA)
  - Capital/Embodied Carbon Expenditure (C CapEx)
    - Estimate the number of SSDs (or GBs of SSD capacity) needed during the system's lifecycle.
    - This includes replacement of SSDs due to pre-mature failure.
  - Operational Carbon Expenditure (C OpEx)
    - Estimate the power/energy requirement for the SSDs in your system during your system's lifecycle.
- We use a lifecycle period of 5 years in this talk – different systems can have different periods of operation.



SAMSUNG

Write Amplification vs. SSD Lifetime and Power

WAF vs. SSD Lifetime vs. Embodied Carbon Emissions

- SSD Lifetime is inversely related to the device write amplification factor (WAF).
  - For example, a WAF of 3 would result in the SSD lasting 1/3rd the time it would otherwise.

- A lower SSD lifetime means an increase in SSDs purchased during a system's lifecycle.
  - This contributes to a higher TCO.
  - This also contributes to a higher embodied carbon footprint and TCCO.

- Controlling SSD WAF can therefore help reduce the embodied footprint
  - Host over-provisioning is commonly used to control and manage WAF.
    - It is inefficient and leads to sub-optimal utilization of resources.
  - Data placement is a better way to manage WAF with a reduced need for host over-provisioning.

Note: We use the Samsung PM9D3 NVMe FDP* enabled SSD for all the experiments in this talk.

**DISCLAIMER: The sustainability data points in this presentation are obtained and calculated using example workloads and a value of 0.16 CO2e KG per GB. These data points as such are not to be associated with a specific Samsung SSD product. Using different SSDs might results in varying results.*

SAMSUNG

SDC 24

WAF vs. Power Consumption vs. Operational Carbon Emissions

- SSD WAF >1 is a result of SSD internal operations like garbage collection (we ignore SSD aging related internal operations in this talk).
- Garbage Collection results in additional reads and writes in the SSD
  - This results in an increase in SSD power consumption and an increase in operational carbon emissions.
- Data placement can reduce SSD WAF thereby leading to a more optimal operational carbon footprint.
- The SSD power in Watts is converted to the KWh usage based on the system lifecycle period. This energy usage can be converted to a CO2e Kg value to obtain the operational carbon footprint.

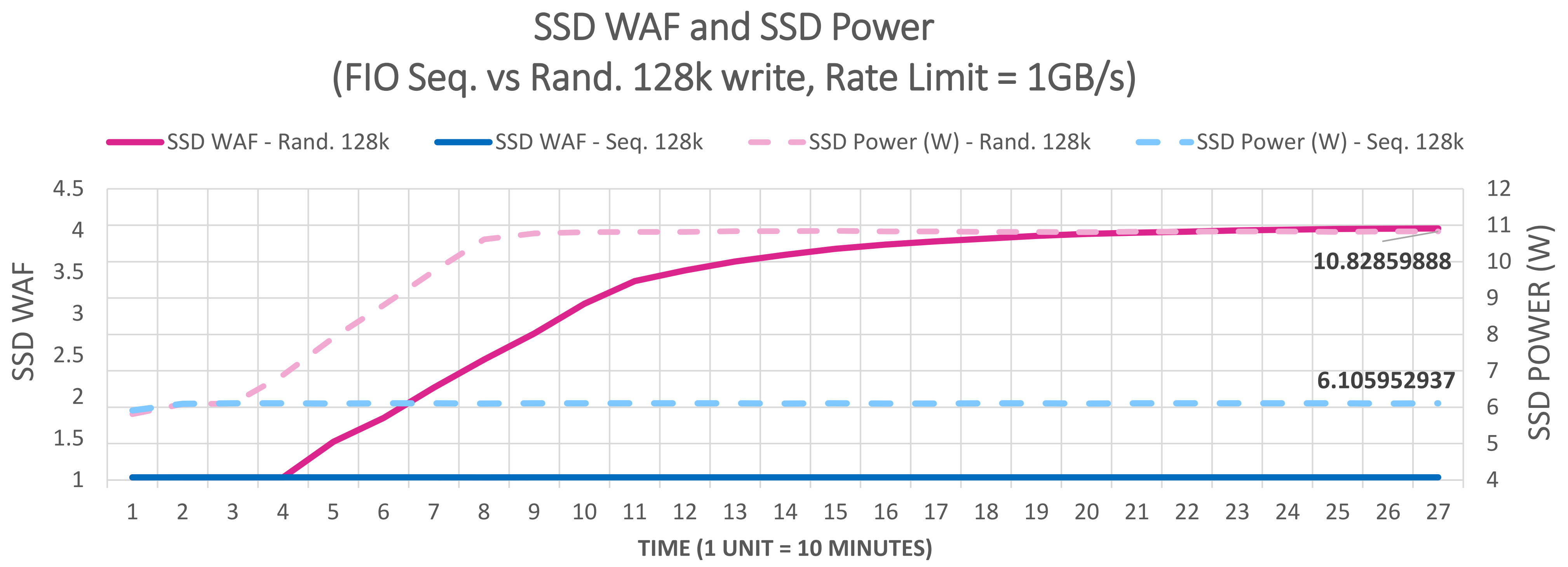
Note: We use the Quarch Power Analysis Module [6] for all the SSD power measurements presented in this talk.

[6] https://quarch.com/products/power-analysis-module/ . Check Appendix for more details.

SAMSUNG

SDC 24

Example: WAF vs. Power Consumption



Main takeaway(s):

- For a fixed host workload, a higher WAF translates to a higher SSD power consumption.

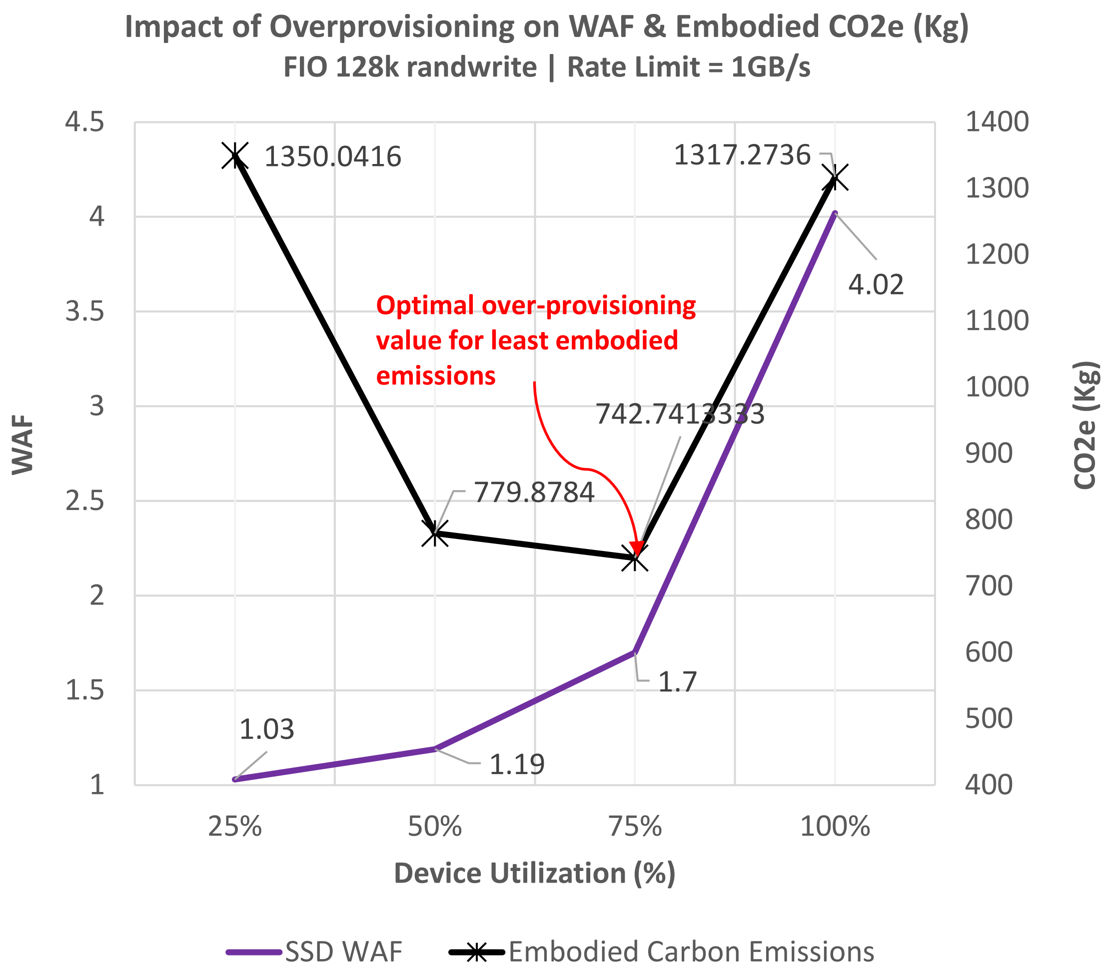
- Controlling WAF helps lower the SSD power and the operational carbon emissions.

SAMSUNG

WAF vs. Over-provisioning vs. Embodied Carbon Emissions

- Over-provisioning helps control WAF
  - A WAF of ~1 is achieved with an OP of 75% (device utilization of 25%). This is neither cost nor carbon effective.
- Over-provisioning has a high impact on embodied carbon emissions due to the extra space used to control the WAF.
- Even with over-provisioning it is important to perform the System LCA and identify what amount of over-provisioning optimizes the carbon footprint of your system.

Main takeaway:
While over-provisioning helps control WAF to some extent, it is not efficient.
- *In this example, 75% device utilization has the lowest capital carbon expenditure (embodied emissions).*



SAMSUNG

SDC 24

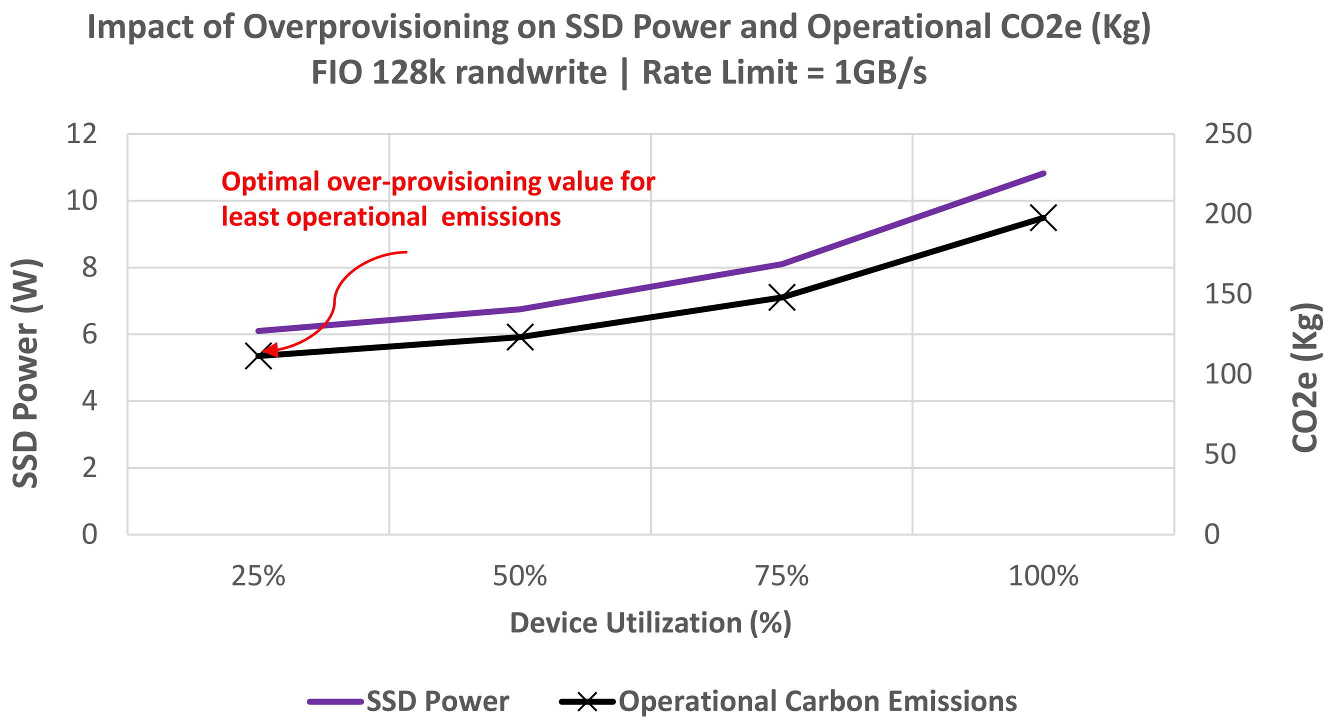
Host Over-provisioning vs. Operational Carbon Emissions

- Over-provisioning helps reduce SSD WAF and leads to lower SSD power consumption.
- The operational carbon footprint with over-provisioning is lower due to the reduced SSD WAF.

Main takeaway:

The operational carbon footprint has a much lower impact than embodied emissions on the overall system's carbon footprint.

- *For example: ~742 Kg CO2e embodied emissions at 75% utilization vs ~148 Kg CO2e operational emissions.*



Device Util. (%)	25%	50%	75%	100%
WAF	1.03	1.19	1.7	4.02

SAMSUNG

SDC 24

Takeaways: WAF vs Carbon Emissions

- Takeaway 1: SSD WAF affects both the operational and embodied carbon footprint of a system.
- Takeaway 2: Reducing SSD WAF helps reduce a system's overall carbon footprint.
- Takeaway 3: Embodied carbon emissions are larger in magnitude than operational carbon emissions.

SAMSUNG

Data Placement and NVMe FDP

Introduction and problems FDP is poised to solve.

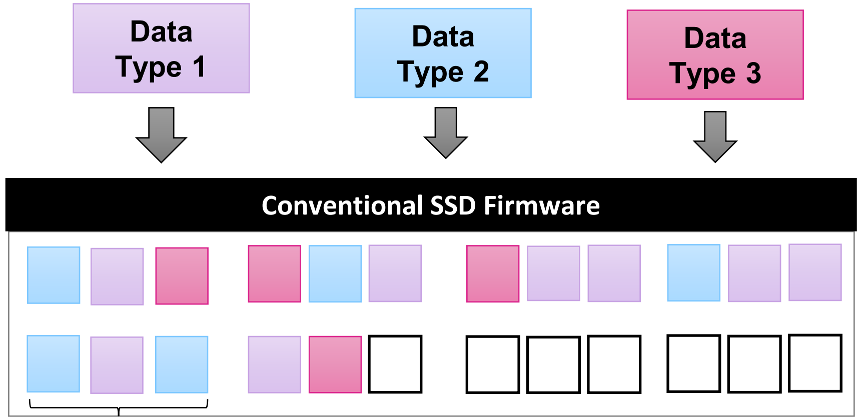
SAMSUNG

SDC 24

NVMe FDP: A Quick Overview

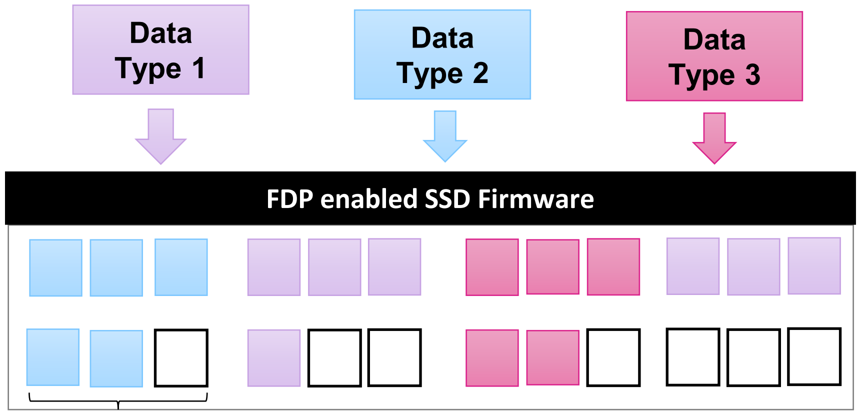
- Data Placement helps control SSD WAF. NVMe FDP is one such data placement method out there.
- NVMe FDP enables the host to segregate it’s data in the SSD. [7,8,9]
  - Host can separate data with different temperatures/patterns on the SSD using “hints” i.e. <Reclaim Group, Reclaim Unit Handle>.
  - Backwards compatible.
  - FDP provides event log pages using which the host can monitor the state of the SSD. Feedback loop mechanism.
- FDP:
  - Helps lower SSD WAF. This results in
    - Improved SSD lifetime and a reduction in SSD power
  - Reduces host over-provisioning
  - No major application design changes needed to use FDP i.e. easy adoption.

Conventional SSD (Host Perspective)



Superblock / Reclaim Unit

NVMe FDP enabled SSD (Host Perspective)



Superblock / Reclaim Unit

[7] TP4146 Flexible Data Placement Ratified Technical Proposal. Available under ratified TPs at https://nvmexpress.org/specification/nvm-express-base-specification

[8] Introduction to Flexible Data Placement: A New Era of Optimized Data Management: https://download.semiconductor.samsung.com/resources/white-paper/FDP_Whitepaper_102423_Final.pdf

[9] Getting Started with Flexible Data Placement (FDP): https://download.semiconductor.samsung.com/resources/white-paper/getting-started-with-fdp-v4.pdf

NVMe FDP: Carbon Emissions

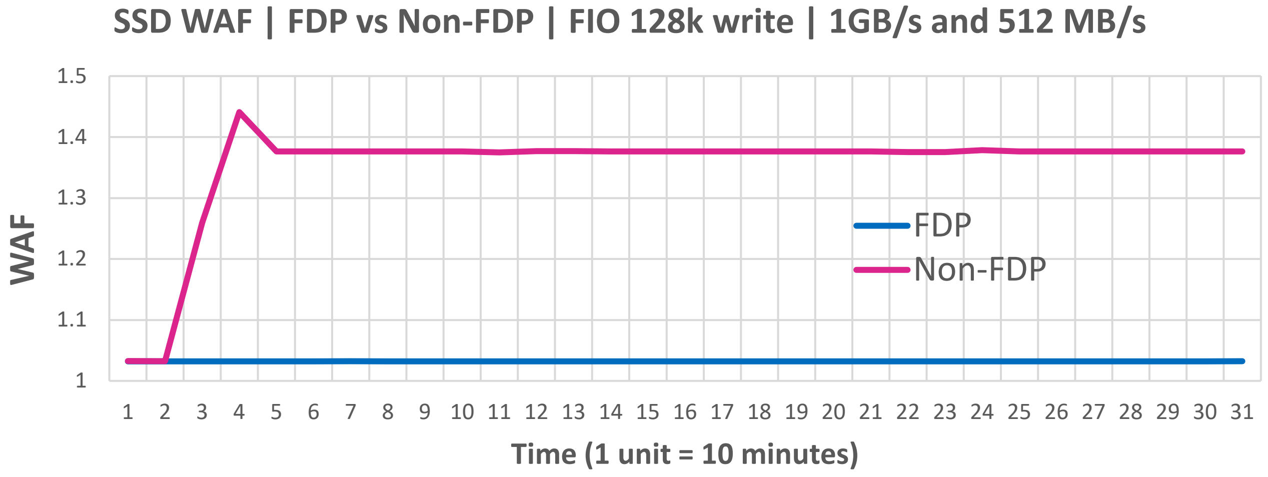
Using NVMe FDP leads to reduced carbon footprint

SAMSUNG

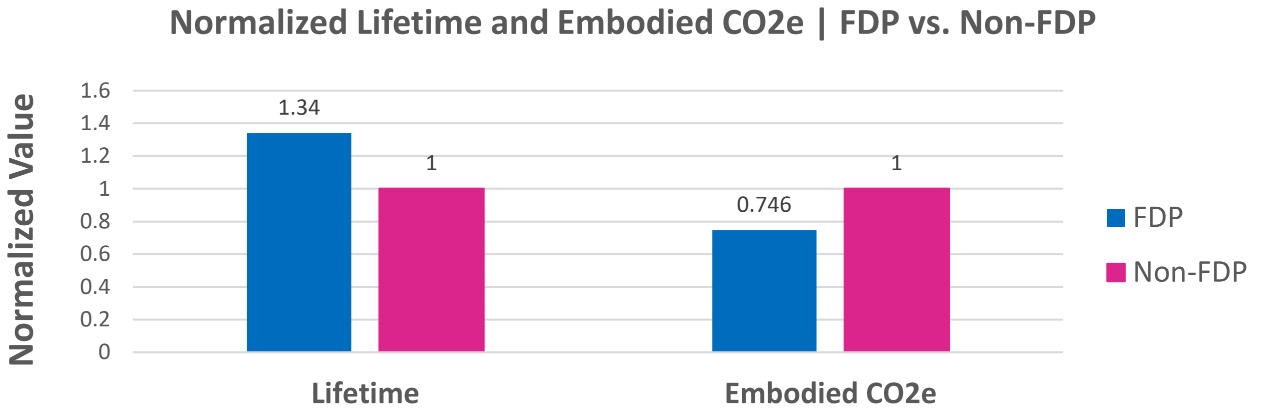
SDC 24

NVMe FDP: Embodied CO2e – Example 1

- FIO synthetic workload run for ~5 hours
  - RUH 1: 128KB Seq. Write. Rate limited to 1 GB/s and using 50% of the LBA space
  - RUH 2: 128 KB Seq. Write. Rate limited to 512 MB/s and using 50% of the LBA space
- SSD WAF:
  - Non-FDP: 1.38 vs. FDP: 1.03
- Carbon CapEx Savings of ~25.3% with FDP.



Main takeaway(s):
- Even two sequential streams of data increases the WAF without FDP.
- With FDP, the WAF can be controlled leading to fewer GB of SSD purchased over the system lifecycle. This reduces the embodied emissions.



SAMSUNG

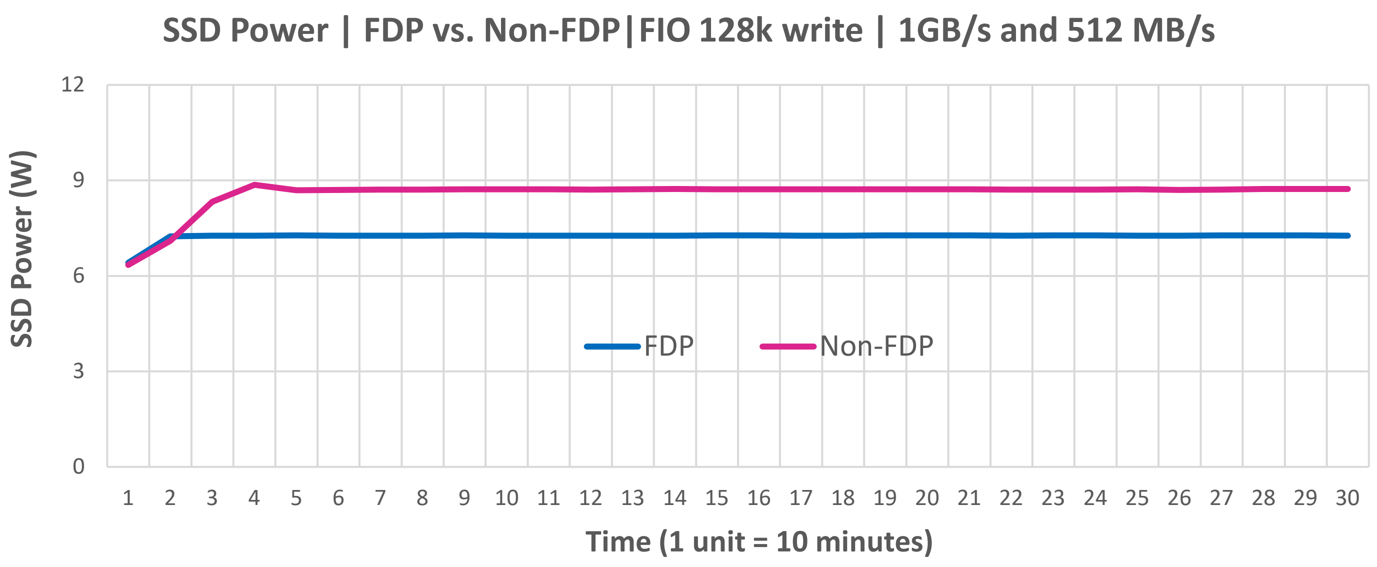
SDC 24

NVMe FDP: Operational CO2e – Example 1

- SSD Power Consumed:
  - Non-FDP: 8.7 W
  - FDP: 7.2 W

  That is a ~17.25% reduction with FDP

Main takeaway:
The WAF gains from using FDP leads to a lower power consumption and lowers the operational carbon footprint.



Configuration	Operational CO2e (normalized)	Number of GC Events
Non-FDP	1.21	8150
FDP	1	2

SAMSUNG

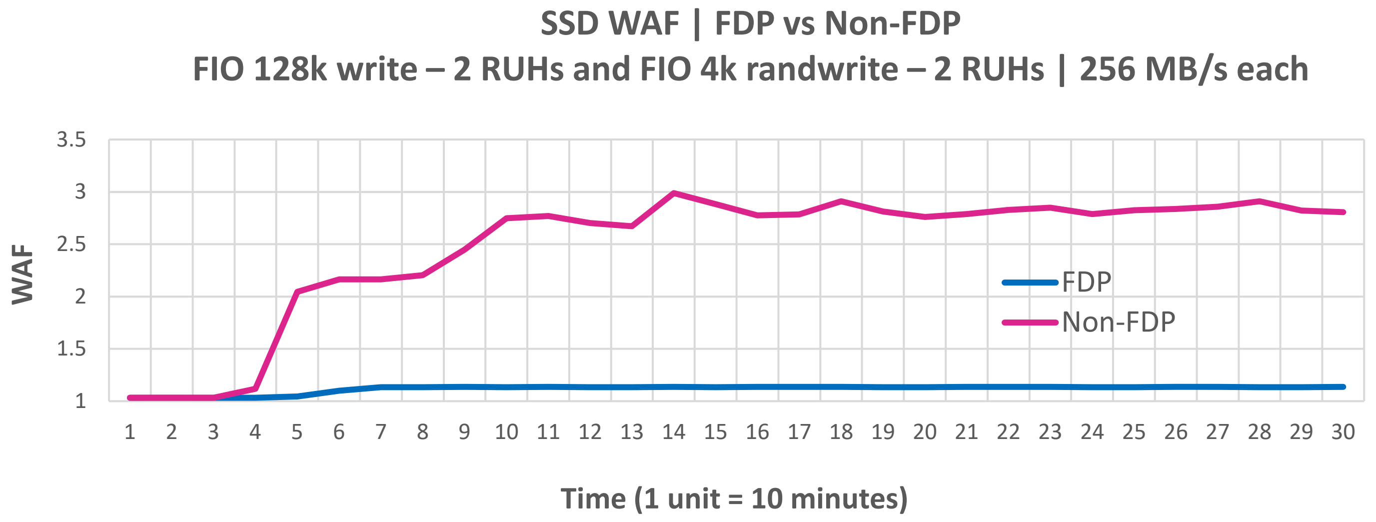
NVMe FDP: Embodied CO2e – Example 2

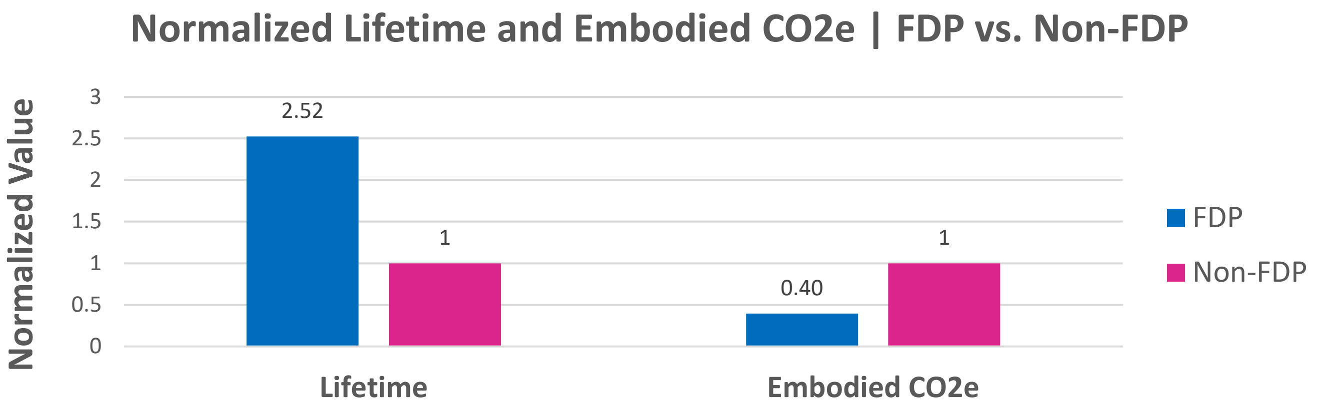
- FIO synthetic workload run for ~5 hours:
  - RUH 1 and 2: 128KB Seq. Write. Rate limited to 256 MB/s and using 45% of the LBA space each.
  - RUH 3 and 4: 4KB Rand. Write. Rate limited to 256 MB/s and using 5% of the LBA space each.
- SSD WAF:
  - Non-FDP: 2.85 vs. FDP: 1.13
- Carbon CapEx Savings of ~60.3% with FDP.

Main takeaway(s):

- Even a small amount of random write leads to a huge increase in the WAF without FDP.

- With FDP, the WAF can be controlled leading to fewer GB of SSD purchased over the system lifecycle. This reduces the embodied emissions.

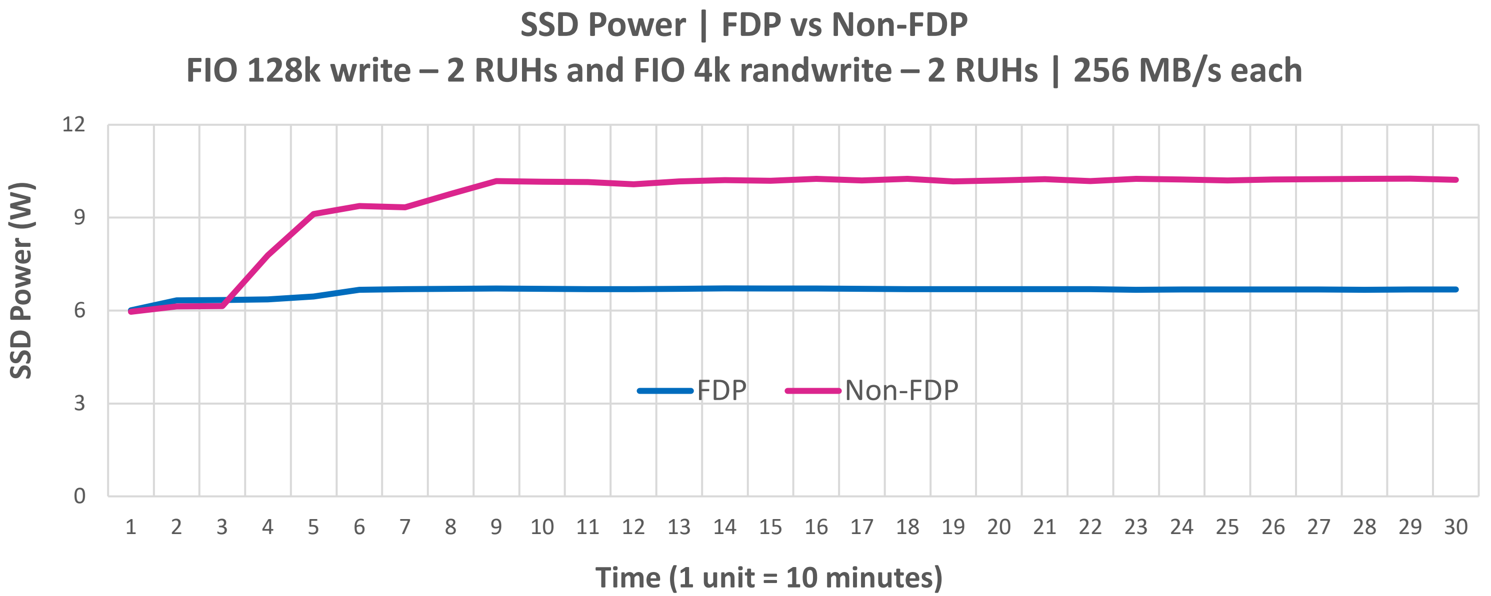




NVMe FDP: Operational CO2e – Example 2

- SSD Power Consumed:
  - Non-FDP: 10.2 W
  - FDP: 6.7 W

  That is a ~34.31% reduction with FDP.



Main takeaway:
The WAF gains from using FDP leads to a lower power consumption and lowers the operational carbon footprint.

Configuration	Operational CO2e (normalized)	Number of GC Events
Non-FDP	1.53	12240
FDP	1	3710

SAMSUNG

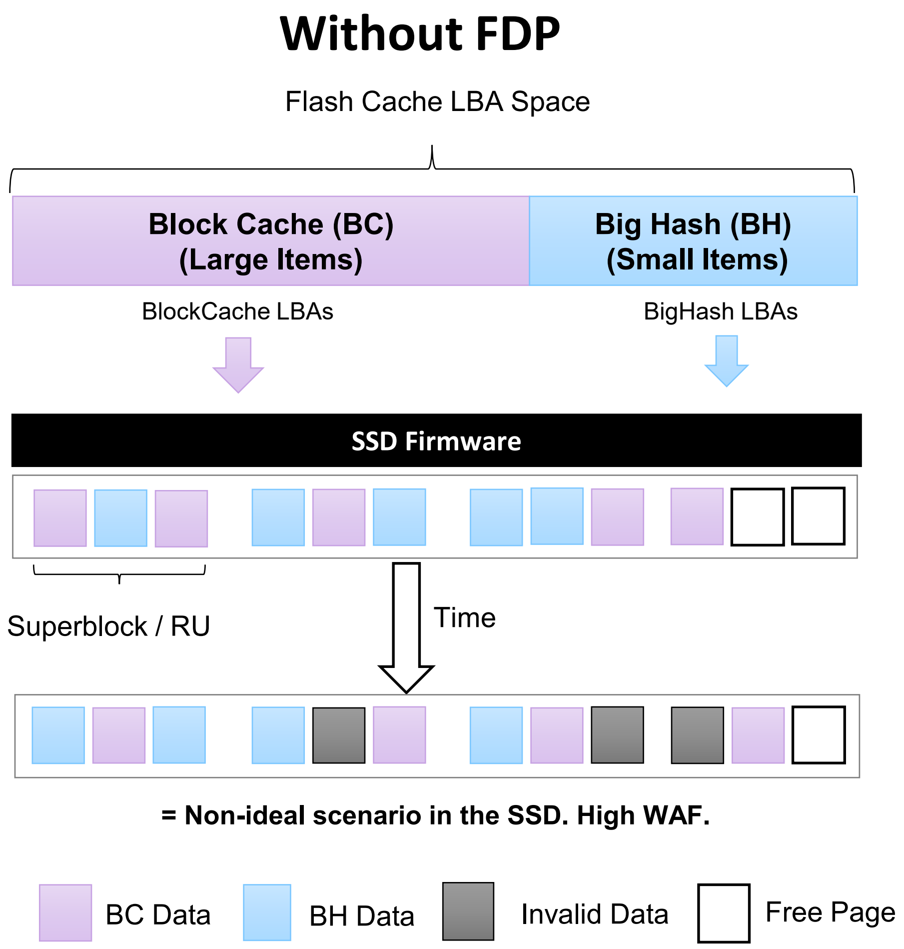
SDC 24

Carbon Emission Reductions: Large Scale System

NVMe FDP + CacheLib

CacheLib

- CacheLib is an open-source Hybrid Cache using both DRAM and Flash [10].
- CacheLib's Flash Cache has two caching engines:
  - Block Cache (Large Items)
    - Log-structured in nature. Relatively cold data.
    - Sequential and less frequent 16 MB (region) sized writes
    - SSD friendly pattern
  - Big Hash (Small Items)
    - Set Associative Cache. Produces hot data.
    - Random and frequent 4KB sized writes
    - SSD unfriendly pattern
  - SSD WAF is a challenge in CacheLib deployments
    - WAF as high as ~3.5
    - Up to 50% host over-provisioning is used to control the WAF (~1.3)



[10] https://cachelib.org/

CacheLib – Deployment and Workload Details

- CacheLib's production deployment typically uses [11]:
  - ~43GB of DRAM Size (varies based on workload and deployment).
  - ~930GB of Flash Size (50% over-provisioned) i.e. a ~1.88TB SSD.
  - 4% of the Flash Size for the Big Hash engine and 96% for the Block Cache engine.
    - Some workloads like CDN don't use Big Hash.

- KV Cache workload [12]:
  - 80% GETs and 20% SETs.
  - Majority of items are small (< 640 bytes).

- Baseline with KV Cache :
  - With 50% over-provisioning a WAF of ~1.2 to ~1.3 is achieved.
  - Reducing over-provisioning drastically affects WAF i.e. 0% over-provisioning results in a WAF of ~3.5

Reducing over-provisioning while maintaining performance KPIs and getting an acceptable WAF was an open challenge with CacheLib deployments.

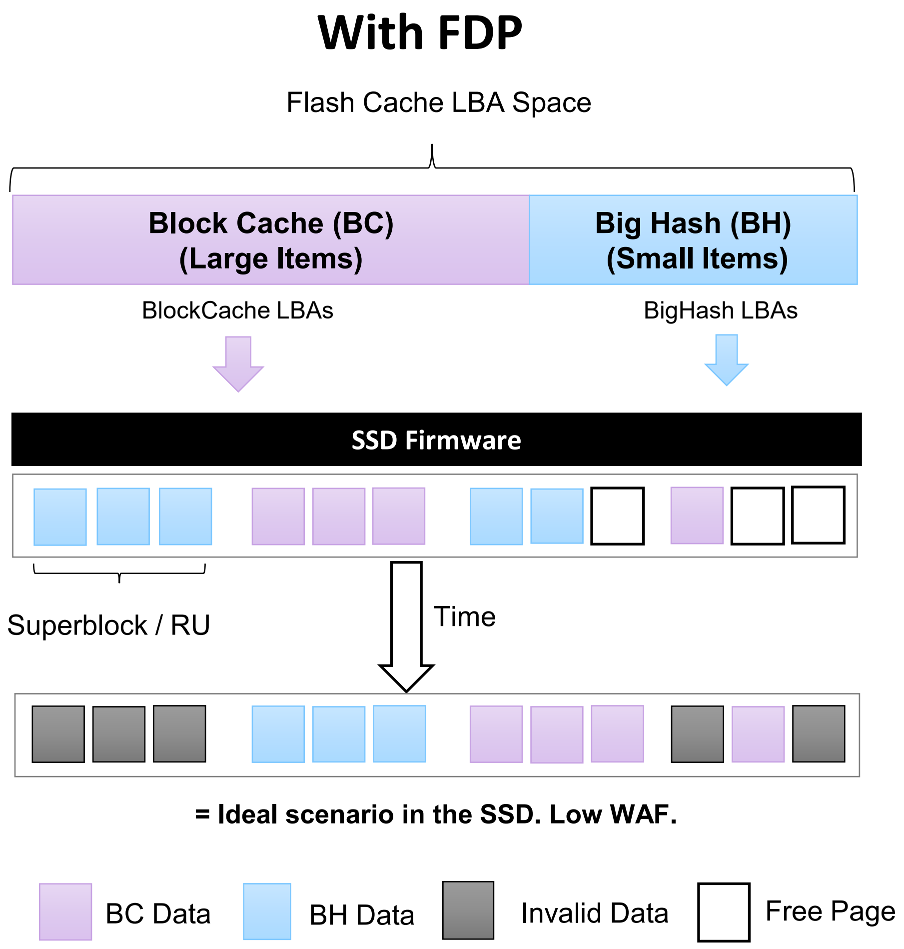
[11] Berg, Benjamin, et al. "The {CacheLib} caching engine: Design and experiences at scale." 14th USENIX Symposium on Operating Systems Design and Implementation (OSDI 20). 2020.
[12] KV Cache workload: https://cachelib.org/docs/Cache_Library_User_Guides/Cachebench_FB_HW_eval/#list-of-traces

SAMSUNG

NVMe FDP + CacheLib

- Segregate the Block Cache and Big Hash data in the SSD using NVMe FDP [13, 14]
  - Lowers the SSD WAF
    - NVMe FDP helps achieve a WAF of ~1
  - Reduces the host over-provisioning used in CacheLib
    - 0% host over-provisioning needed to achieve WAF ~1 with FDP.
  - Improves SSD lifetime
  - Improves SSD power consumption

- Reducing the carbon footprint of CacheLib
  - CacheLib clusters generally have 1000s of nodes, each equipped with an SSD
    - Embodied carbon savings due to improved SSD lifetime
    - Operational carbon savings due to reduced power consumption

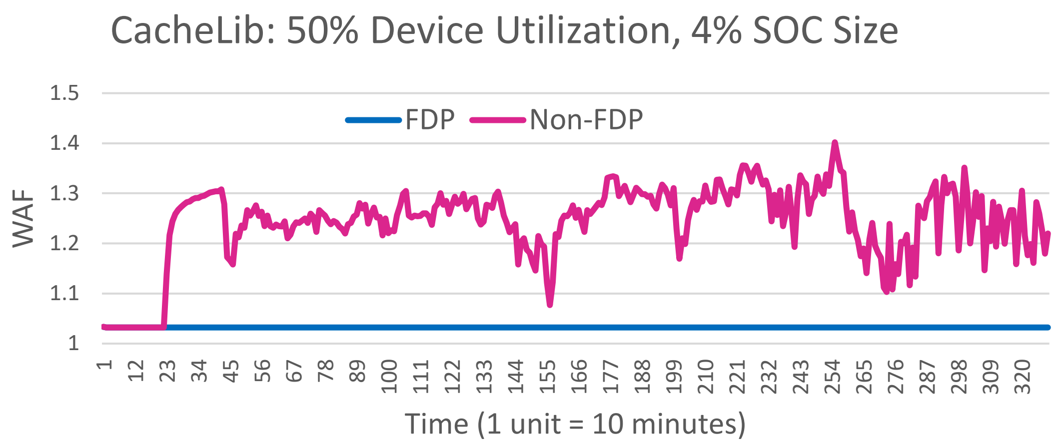


[13] https://cachelib.org/docs/Cache_Library_User_Guides/FDP_enabled_Cache
[14] Towards Efficient Flash Caches with Emerging NVMe Flexible Data Placement SSDs. To Appear in EuroSys '25.

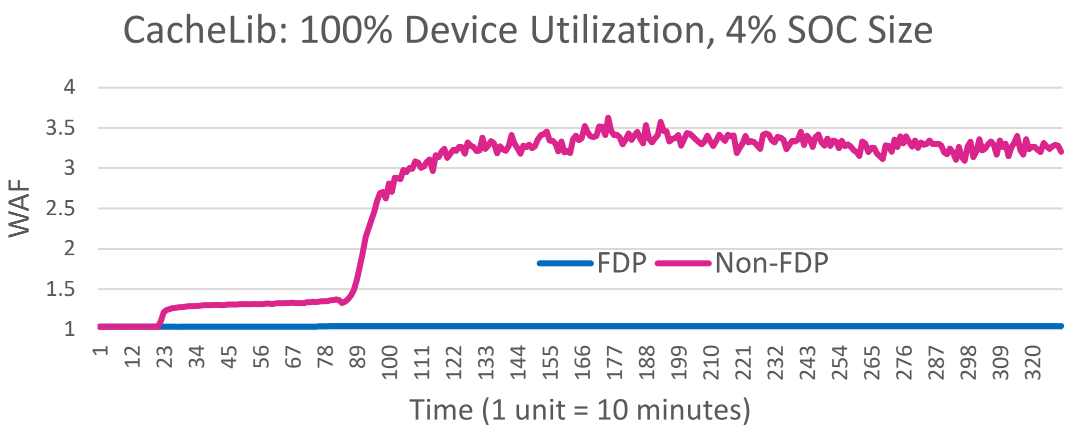
SAMSUNG

Embodied Carbon Emission Reduction in CacheLib

CacheLib with 50% Host Over-provisioning i.e. ~930 GB Flash Cache Size:



CacheLib with 0% Host Over-provisioning i.e. ~1.88 TB Flash Cache Size:



Main takeaway(s):

- With FDP, a WAF of ~1 is achievable in CacheLib with 0% host over-provisioning i.e. using the entire SSD.

- FDP reduces the embodied carbon emissions of CacheLib by reducing WAF and increasing utilization.

Configuration	Effective Lifetime	Effective Embodied CO2e
50% OP: Non-FDP	1	1
50% OP: FDP	1.27	0.79
100% OP: Non-FDP	1	1
100% OP: FDP	3.44	0.29

SAMSUNG

Operational Carbon Emission Reduction in CacheLib

The WAF reductions achieved with FDP results in fewer GC events leading to lower SSD power consumption*:

- Using FDP with the KV Cache workload results in SSD power reduction of ~16.52%.
- With a write only KV Cache workload, the SSD power with FDP power reduces by ~30%.

Workload	Configuration	Power (W)	Takeaway
KV Cache Workload (80% GETs, 20% SETs)	100% device utilization Non-FDP	5.69 ± 0.037	**FDP helps reduce Carbon OpEx by ~16.52%**
	100% device utilization FDP	4.75 ± 0.007	
Write-only KV Cache Workload (100% SETs)	100% device utilization Non-FDP	6.77 ± 0.154	**FDP helps reduce Carbon OpEx by ~30%**
	100% device utilization FDP	4.74 ± 0.051	

* The power measurements here use a DRAM size of 4GB in CacheLib due to setup limitations with the system equipped with the power analysis tools.

SAMSUNG

SDC 24

CacheLib: Deployment Flexibility to optimize carbon emissions

- FDP enables the usage of the entire SSD space while still maintaining a WAF of ~1.
- This enables CacheLib deployments with reduced DRAM.
  - The reduction in DRAM is compensated for by the increased Flash Cache size.
  - This is more carbon efficient as DRAM has much higher carbon footprint than SSDs.
  - This is also more cost effective as DRAM is much more expensive than SSDs.
- Comparing Non-FDP with 43GB of DRAM and FDP with 4GB of DRAM for the trade-offs:
  - ~ 30% drop in throughput.
  - ~70% reduction in carbon footprint.

DRAM Configuration	Hit Ratio (%)	Flash Hit Ratio (%)	Throughput (KGETs/s)	Embodied CO2e (Kg)*
FDP 4GB	*86.3*	*37.74*	*303.1*	*347.2*
Non-FDP 4GB	86.11	37.34	298.8	1081.1
FDP 20GB	91.9	31.37	412.2	372.8
Non-FDP 20GB	92.1	33	399.1	1106.8
FDP 43 GB	90.32	17.51	445.9	409.6
Non-FDP 43GB	*90.22*	*17.34*	*434.4*	*1143.6*

*We calculate the embodied carbon emissions of both the SSD and DRAM components together.

Main takeaway:
Without FDP it was inconceivable to reduce the DRAM used in CacheLib. For a trade-off in performance, FDP enables more carbon efficient deployments.

SAMSUNG

SDC 24

Summary

Key Takeaways

SAMSUNG

SDC 24

Key Takeaways

- NVMe FDP helps reduce SSD WAF thereby leading to reductions in embodied and operational carbon emissions.
- NVMe FDP helps reduce host over-provisioning which optimizes carbon emissions.
- The reduced need for over-provisioning with NVMe FDP allows greater deployment flexibility and helps optimize your system's overall carbon footprint.

SAMSUNG

SDC 24

Questions?

Please take a moment to rate this session.

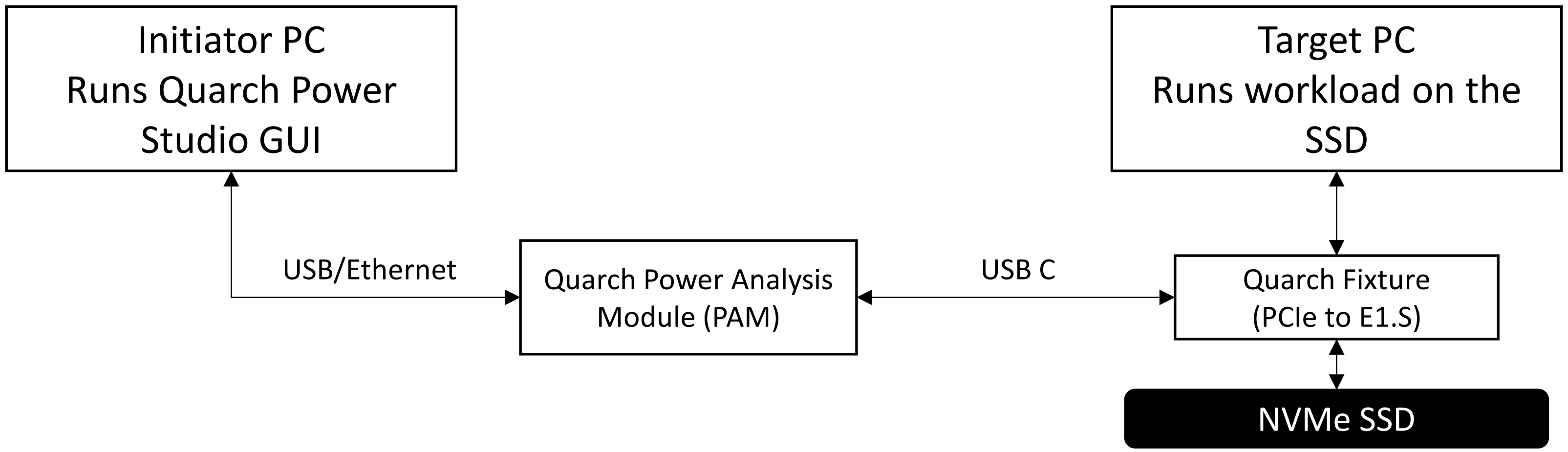
Your feedback is important to us.

SAMSUNG

SDC 24

Appendix

Quarch Power Analysis Module – Setup Block Diagram



SAMSUNG