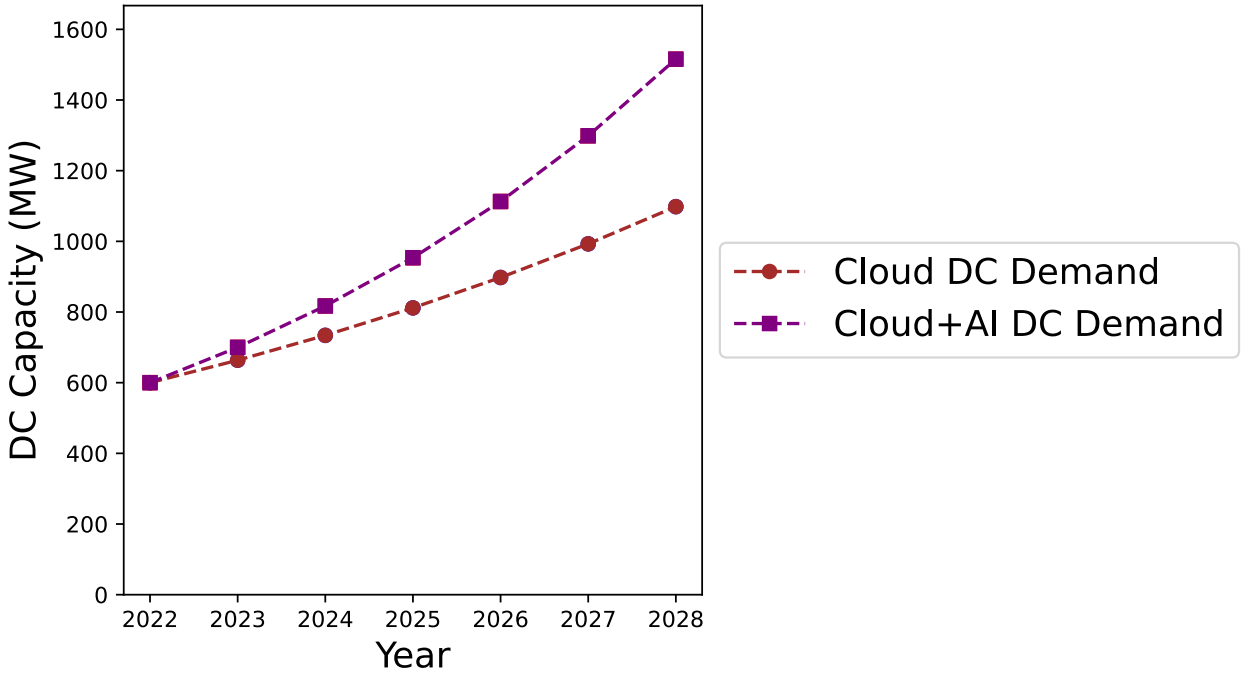
HotStorage 2024

Can Storage Devices be Power Adaptive?

Dedong Xie, Theano Stavrinos, Kan Zhu,
Simon Peter, Baris Kasikci, Thomas Anderson

University of Washington

Power is a major challenge for data centers



DC power consumption is increasing exponentially

- Data centers over-subscribe power
  - > support higher load
- Power availability can vary
  - > Demand-response from the grid
    - > Limit DC power for other users

  - > external: green energy, disruption
  - > internal: power rail failures

Solution: power-adaptivity
Dynamically change power consumption to match available power

[1] Lin et al. " Exploding AI Power Use: an Opportunity to Rethink Grid Planning and Management. " 2024

Power adaptive storage is a crucial building block

- Storage is using a large proportion of power
  - > Accounted for **10%** of total data center power in 2016 [1]
  - > DC storage estimated to use **19.2TWh** of power worldwide in 2021 [2]
- Storage power consumption is **likely to rise**
  - > HDD to SSD transition: SSD higher active power
  - > Higher SSD performance: higher power/disk
  - > ML workloads surge: higher capacity & performance [3]

[1] Shehabi, et al. "United states data center energy usage report." 2016.
[2] IEA. "Global data centre energy demand by end use." 2019.
[3] Zhao, et al. “Tectonic-Shift: A Composite Storage Fabric for Large-Scale ML Training” ATC ‘23

Potential power-adaptive mechanisms

- Different power states in storage devices
  - > Power capping for NVMe SSD
  - > Slumber for SATA SSD
  - > Spin-down for HDD

- IO shaping on storage devices
  - > IO chunk sizes
  - > IO queue depth

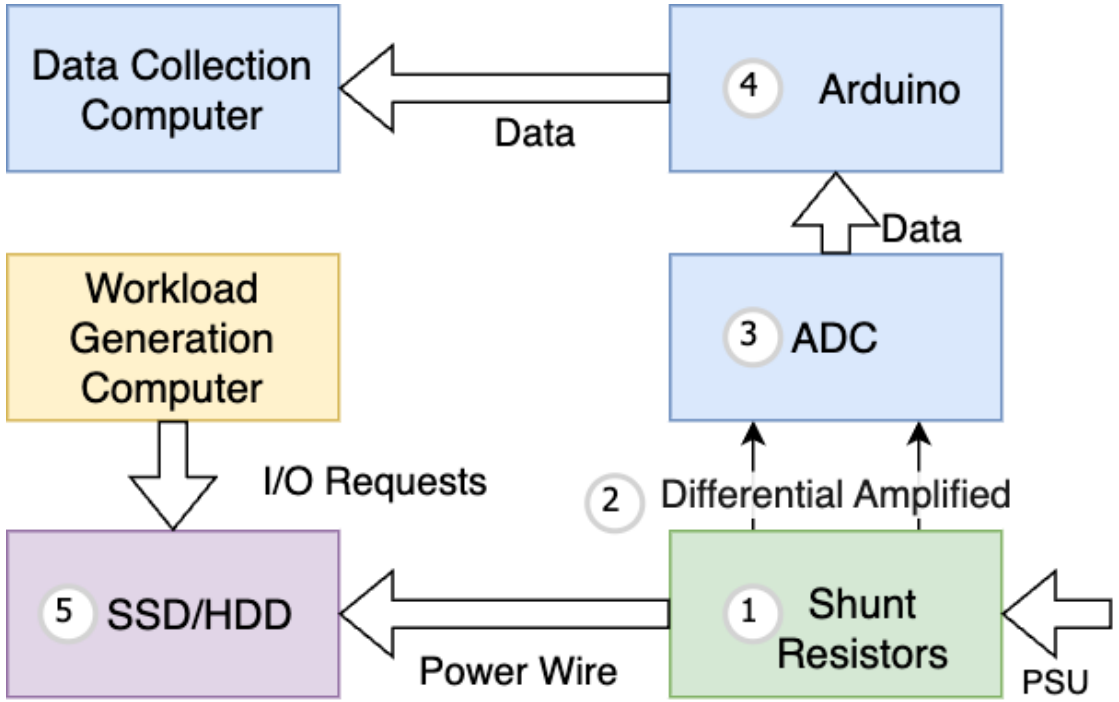
Storage devices studied

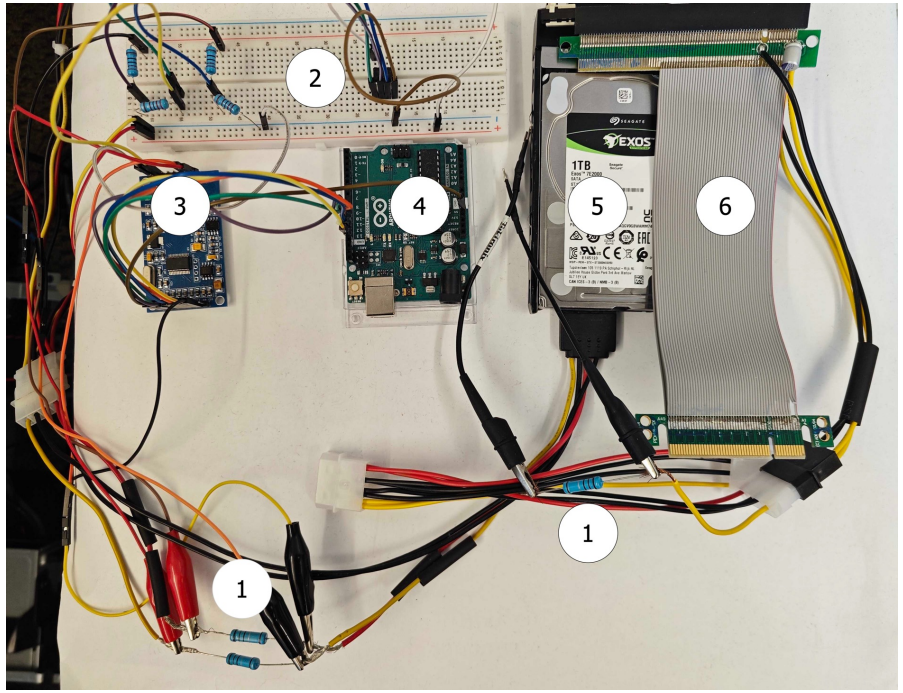
- Datacenter devices

Label	Protocol	Model
Samsung NVMe SSD	NVMe	Samsung PM9A3
Intel NVMe SSD	NVMe	Intel D7-P5510
Intel SATA SSD	SATA	Intel D3-P4510
Seagate HDD	SATA	Seagate Exos 7E2000

- Consumer SSD (to study standby state)
  - > Samsung 860 evo

Power study measurement setup





Workloads

- fio 3.28 for workload generation

- Random-only and sequential-only workloads
  > get range of power dynamic control
  > get range of performance trade-offs

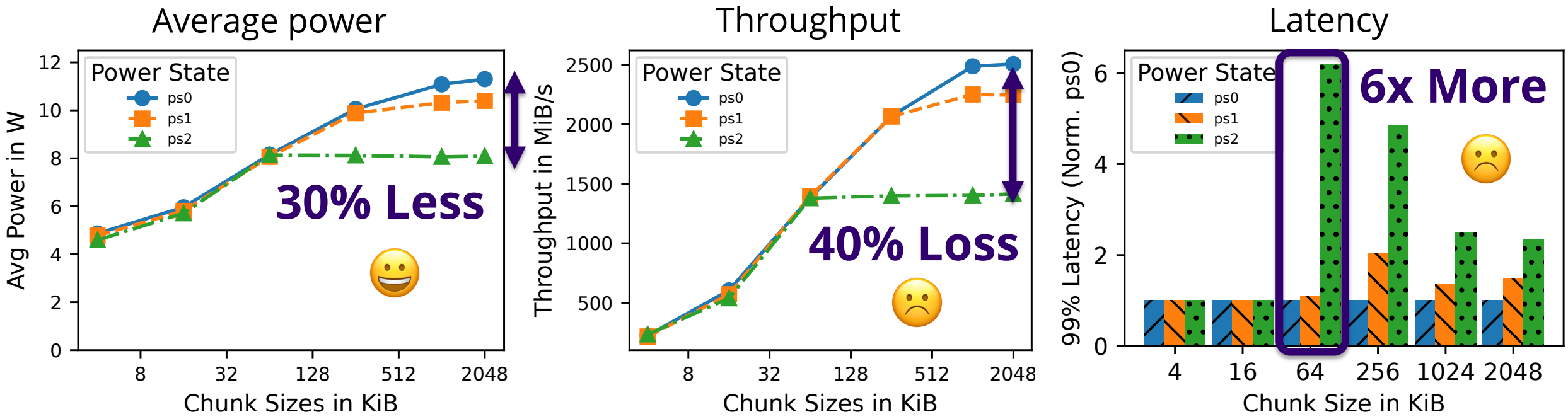
- Runtime: 4GiB of data read/written or 1 minute
  > get stable state readings

What are the power-performance trade-offs?

- When applying power-adaptivity mechanisms
  - > NVMe SSD power capping
  - > Low-power standby: HDD & SATA SSD
  - > IO shaping: IO chunk sizes & IO queue depths

SSD power capping: random write, queue depth 1

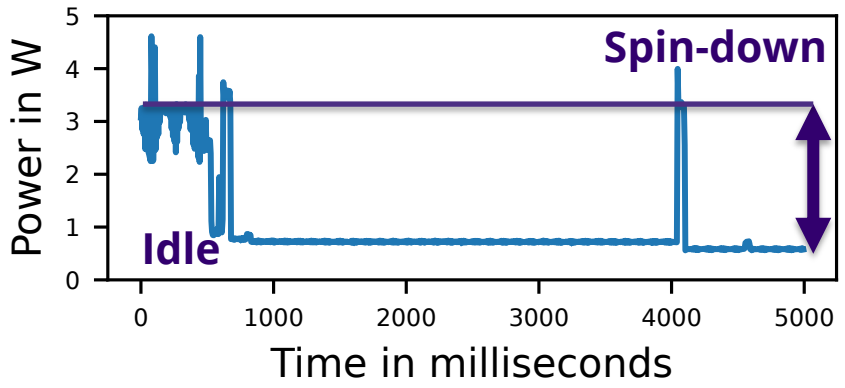
Intel NVMe SSD



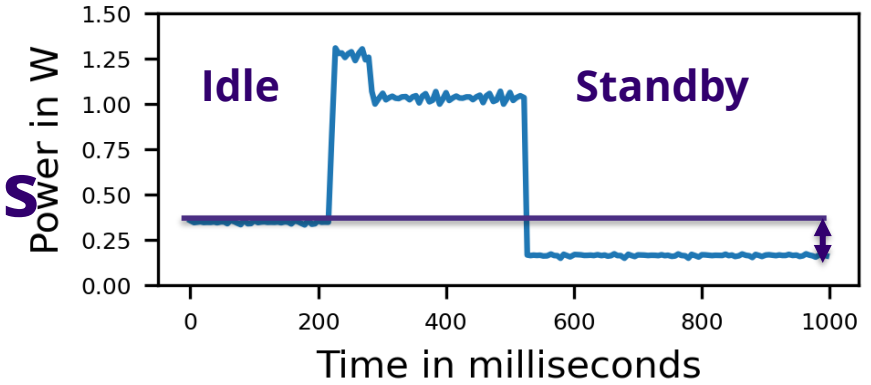
Other workloads:

- Sequential/read: less power difference, less throughput drop, negligible latency change
- Queue depth: similar power difference, similar throughput drop, negligible latency change

Low-power standby

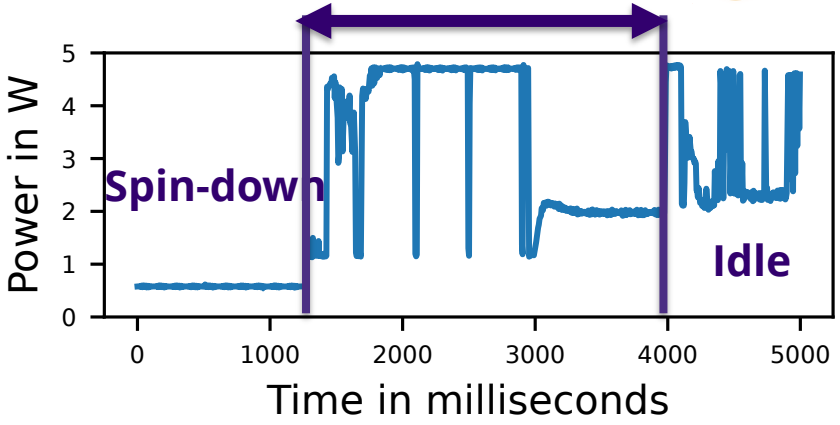


85% Less 😀



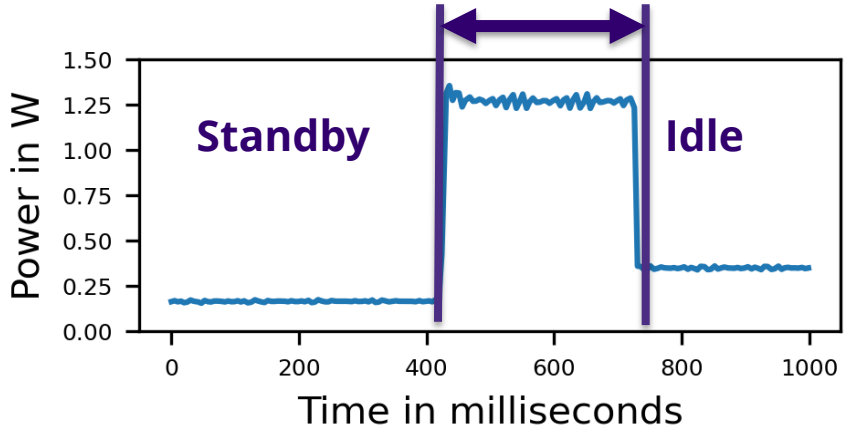
50% Less 😀

3 Seconds ☹



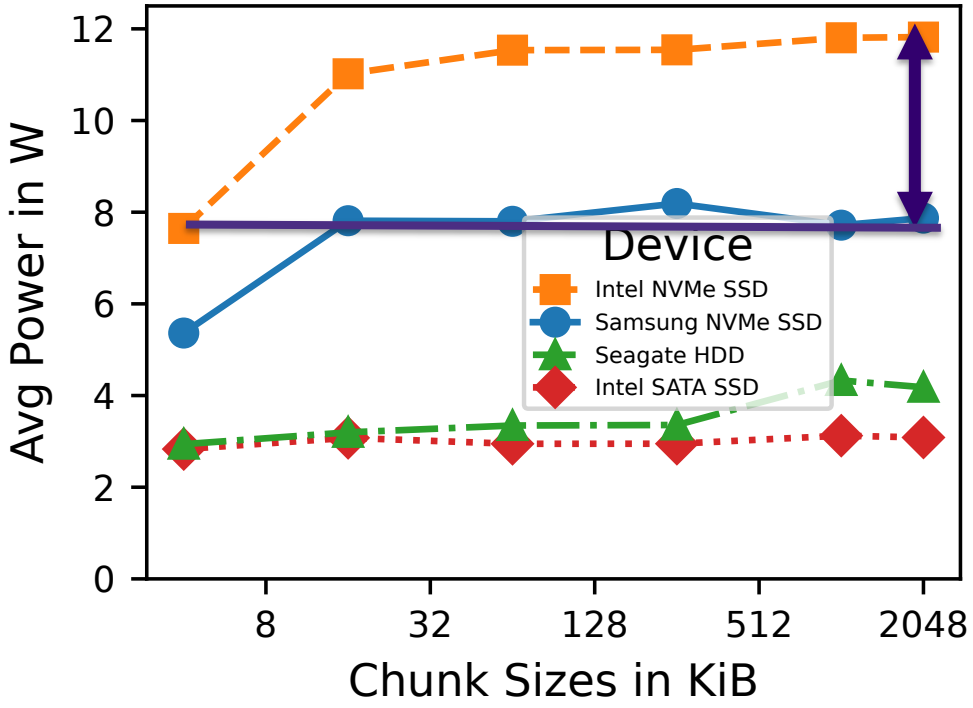
Seagate HDD

0.3 Seconds



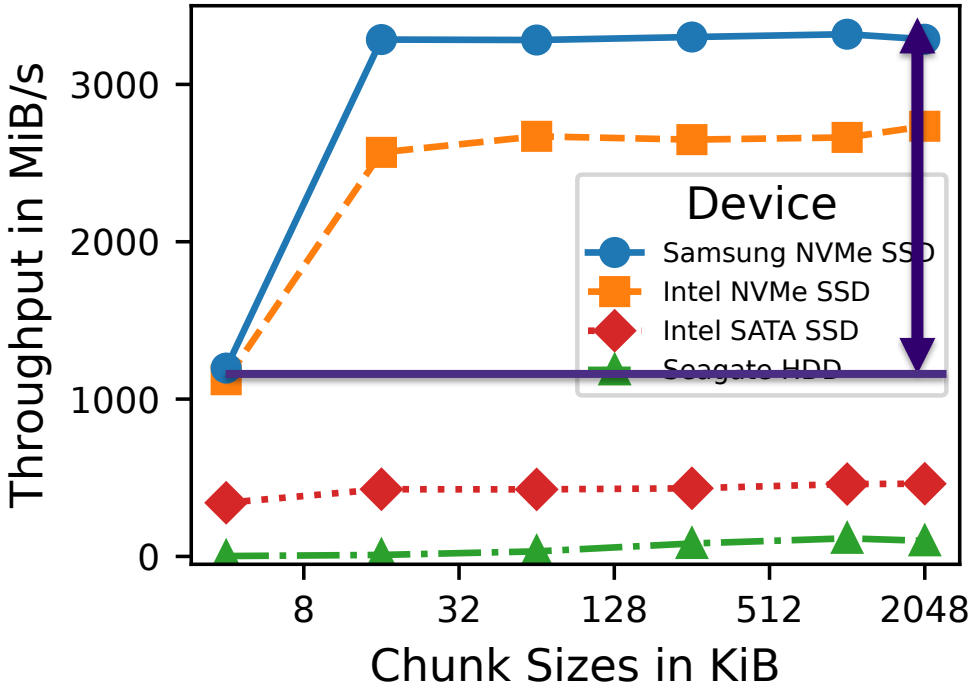
Samsung 860 evo SSD

IO shaping: chunk sizes: random write, queue depth 64



30% Less

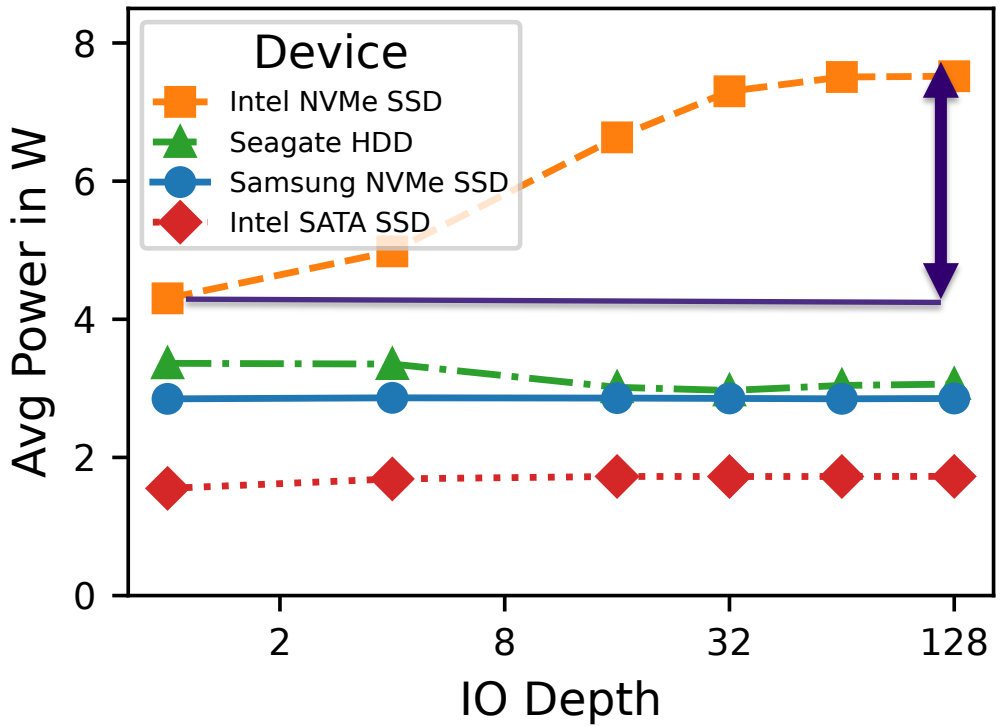
😀



60% Loss

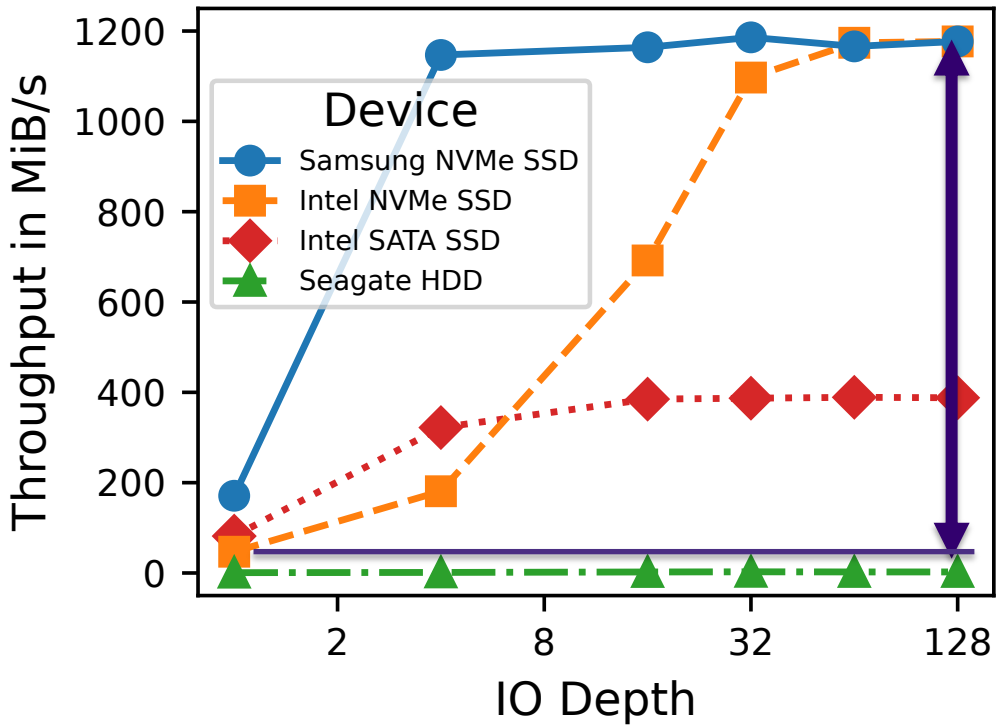
☹️

IO shaping: queue depth: random read, chunk size 4KiB



40% Less

😀



90% Loss

☹️

Can storage devices be power adaptive? Yes!

Raises further questions:

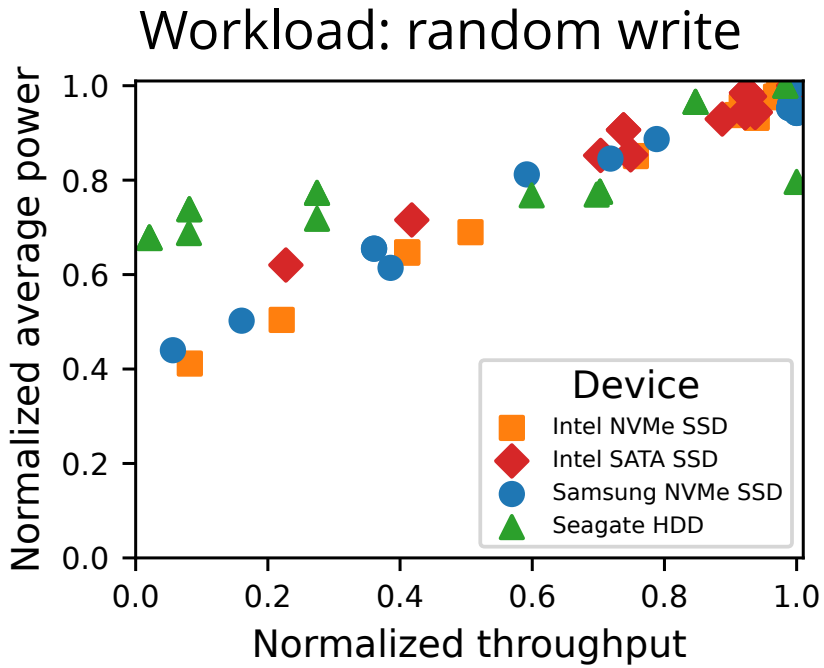
1. **How to find good configuration for a power budget?**
2. **How to make storage stack power-adaptive?**

1. How to find good configuration for a power budget?

Power-throughput model

- Each point is a power control configuration
  > device power state & IO shaping
- Current model is workload specific

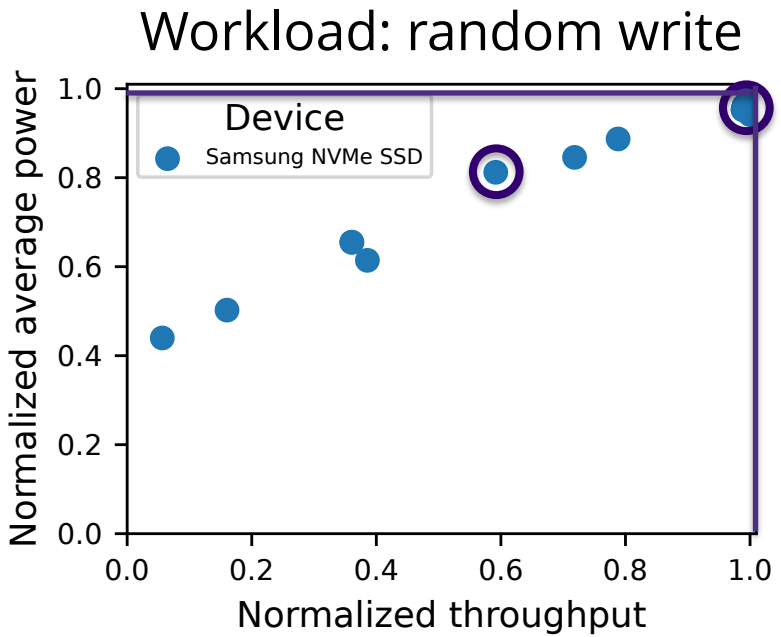
- Use this to
  > get the configuration needed for a specific power budget
  > get the trade-off in throughput



Achieving power adaptivity

Case study: using power-throughput model

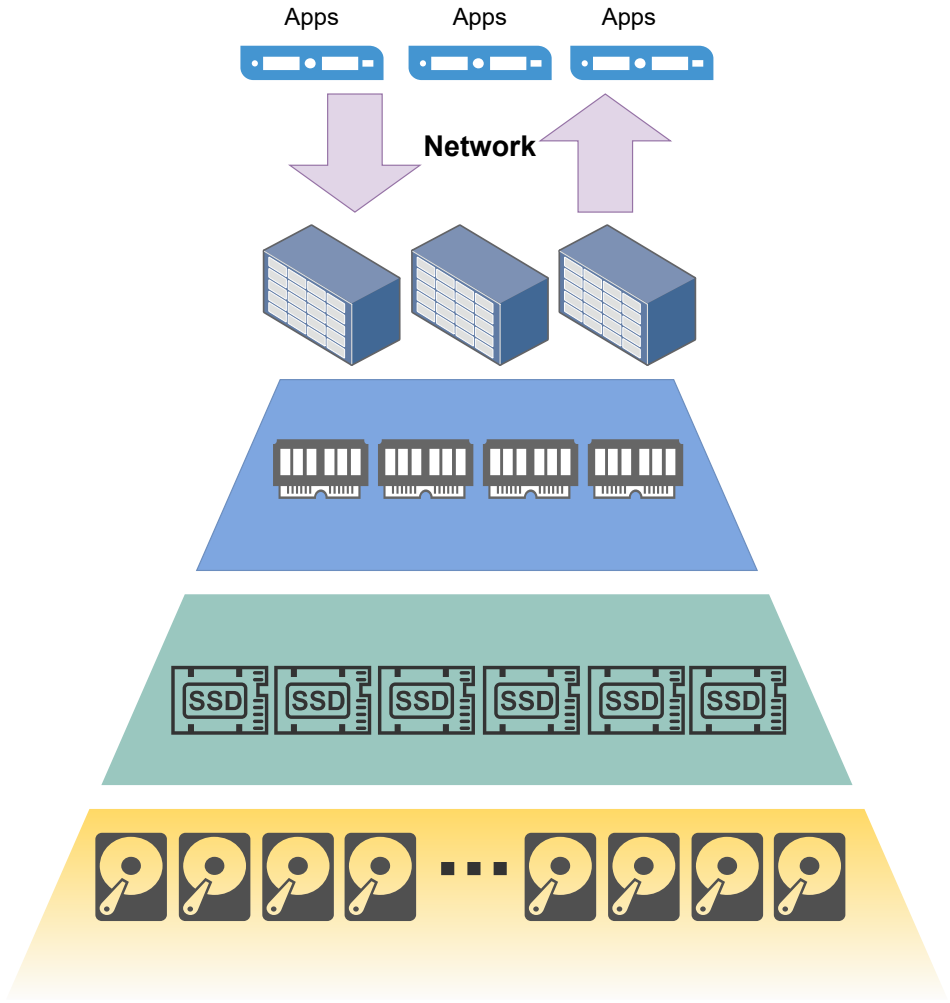
- Configuration:
  - > Samsung NVMe SSD operating at queue depth 64 and chunk size 256 KiB with random write workloads
- Request of reducing 20% of device power
- Target state:
  - > Queue depth 1 and chunk size 256 KiB
- Trade-off:
  - > Model suggests 40% of throughput reduction



2. How to make storage stack power-adaptive?

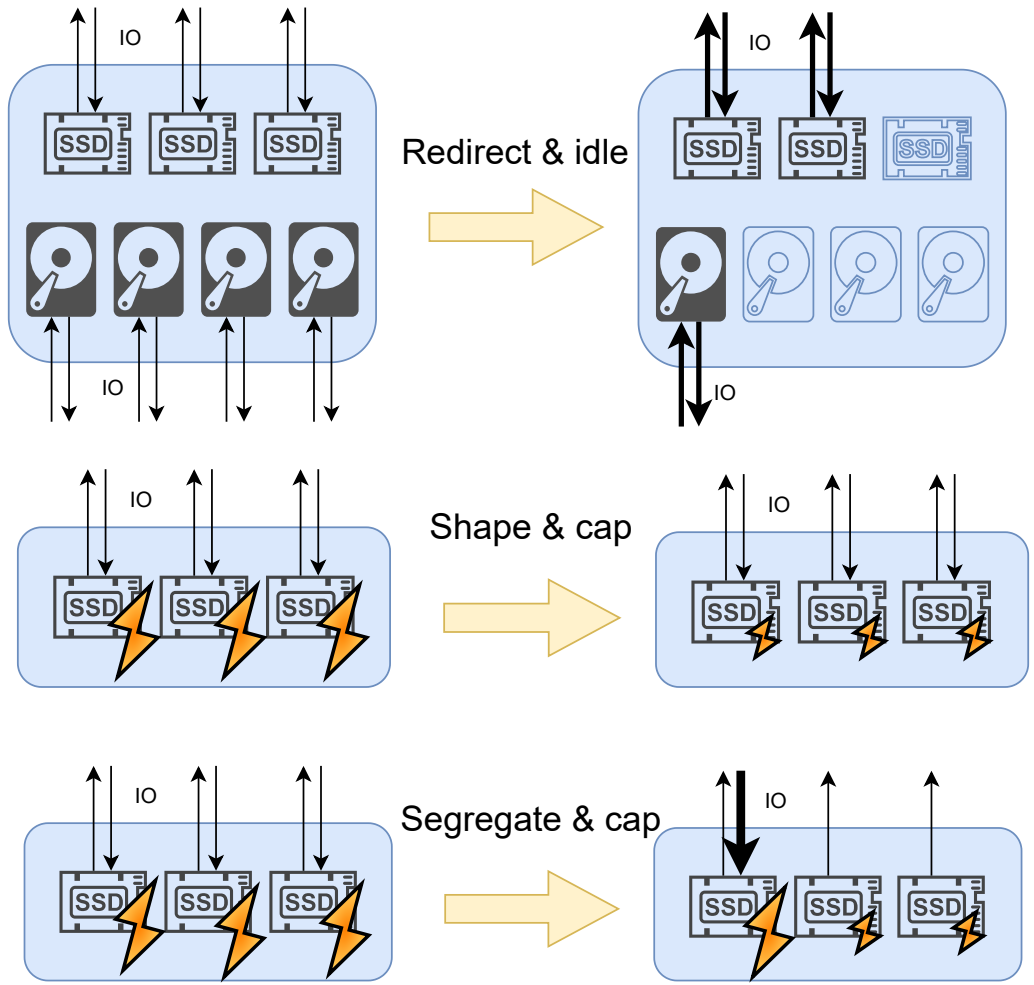
Target: disaggregated storage system

- Modern data center storage system
  - > Disaggregated
  - > Multi-layered
  - > Heterogeneous
    - > SSD & HDD



Implications on power-adaptive storage system design

- Power-aware IO redirection
  - > direct IO to some storage devices to maximize idle period of the rest
  - > spin-down/power-off the idle devices
- Power-capping and IO shaping
  - > match the power-performance models with performance guarantees
- Leveraging asymmetric IO
  - > segregate write traffic to a set of disks
  - > power capping the remainder



Summary

- Conduct a thorough **measurement study**
  - > characterize power control dynamic range of data center storage devices
- Study power control mechanisms
  - > **device power states** and **IO shaping**
  - > potential power control dynamic range of **~60%** (6MW in a 100MW data center)
- Build **power-throughput models** across storage devices
  - > represent feasible power control range and performance **trade-offs**
- Propose designs of power-adaptive storage system

Thank you!