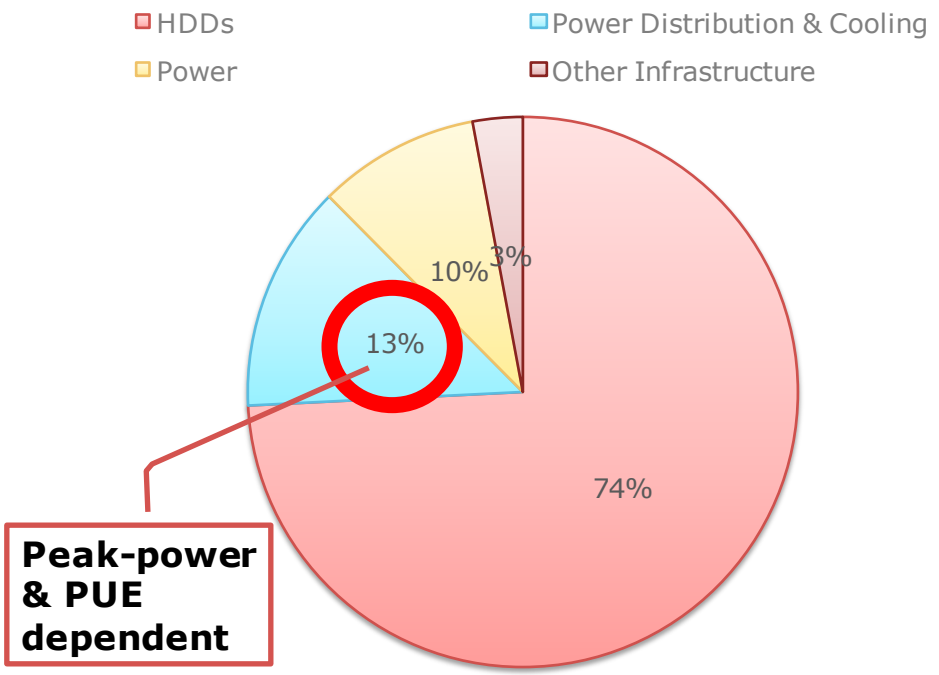
PCAP: Performance-Aware Power Capping for the Disk Drive in the Cloud

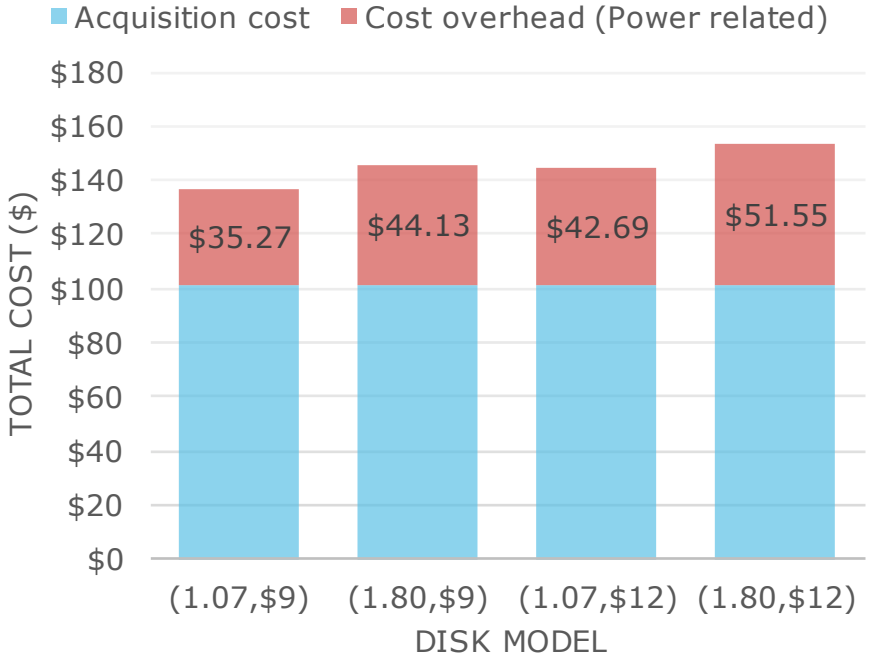
Mohammed G. Khatib & Zvonimir Bandic

WDC Research

HDD's power impact on its cost

3-yr server & 10-yr infrastructure amortization

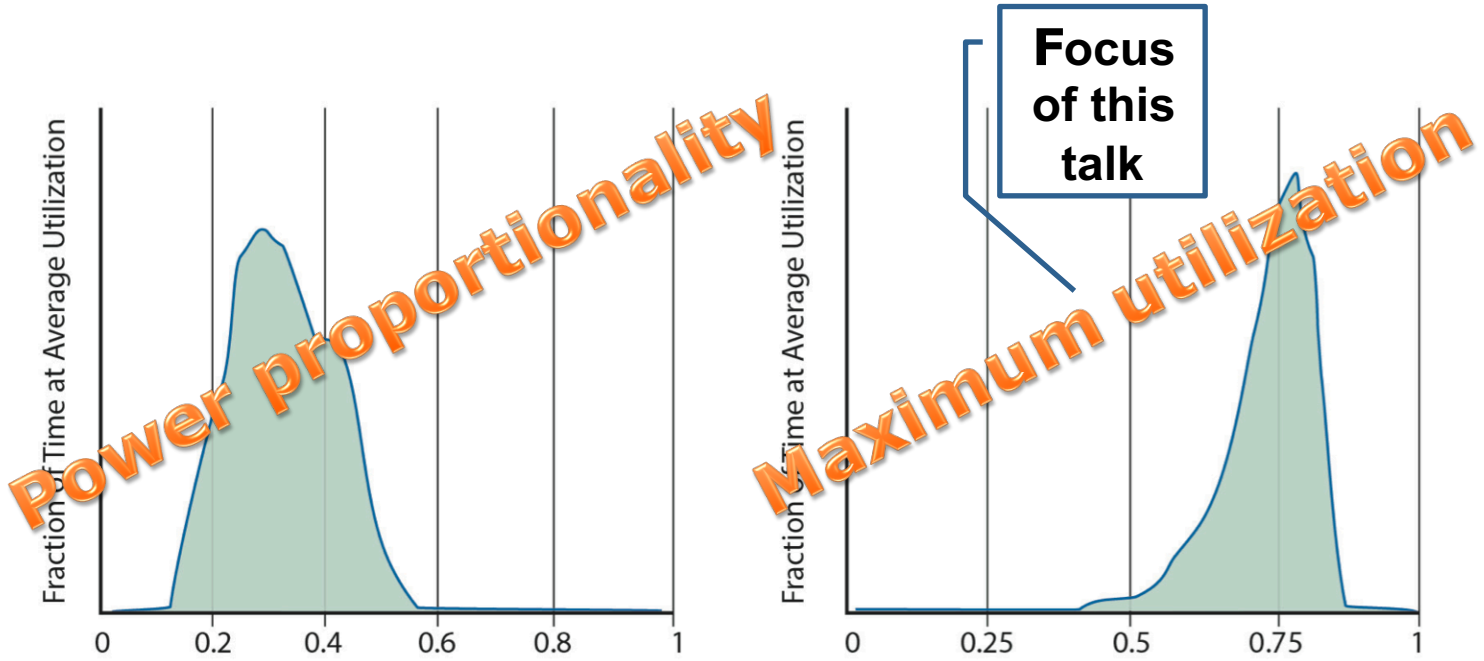




Based on Hamilton's DC cost model [http://perspectives.mvdirona.com/2010/09/overall-data-center-costs/]

Two types of datacenters

The average activity distribution of a sample of 2 Google clusters, each containing over 20,000 servers, over a period of 3 months 2013



Luiz A. Barroso et al. [The Datacenter as a Computer, 2013]

Power over-subscription

- For maximum cost effectiveness, use provisioned power fully
  - If a facility operates at 50% of its peak power capacity, the effective provisioning cost per Watt used is **doubled**!

    Luiz A. Barroso [The Datacenter as a Computer, 2013]

- *How many servers fit within a given budget?* Hard question!
  - Specs are very conservative → Dell & HP offer online power calculators
  - Actual power consumption varies significantly with load
  - Hard to predict the peak power consumption of a group of servers
    - while any particular server might temporarily run at 100% utilization, the maximum utilization of a group of servers probably isn't 100%.

- Problem: using any power numbers but the specs runs the risk of
  → facility power over-subscription
  → Power capping becomes necessary as a saftey mechanism

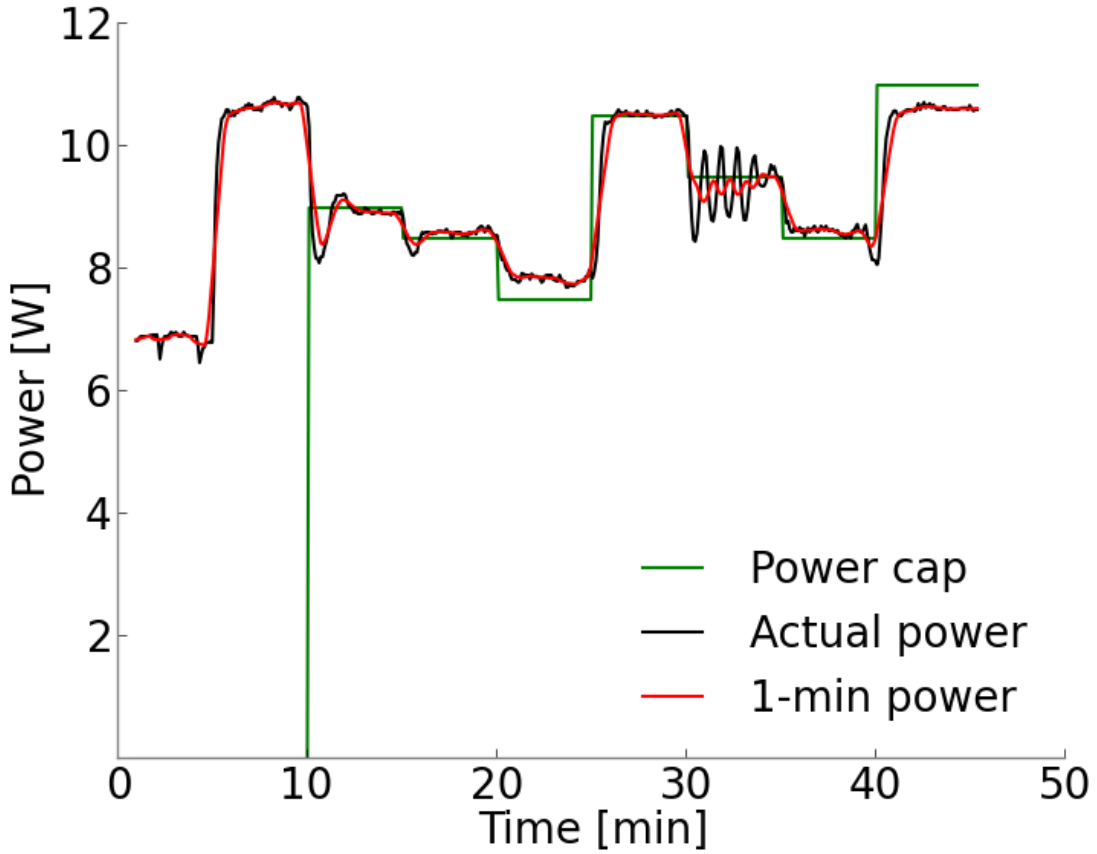
Power capping

- What is power capping?
  - Preventing datacenter's total power usage from violating (crossing) a predefined limit, the power cap (i.e., prevent power overshooting)
- Techniques:
  - Software techniques such as workload re-scheduling
  - Duty cycle adaptation
  - …
- This work:
  - Focuses on the 3.5" enterprise HDD
  - Explores techniques inherently related to the underlying hardware
  - Investigates using the queue size to cap the HDD's power consumption

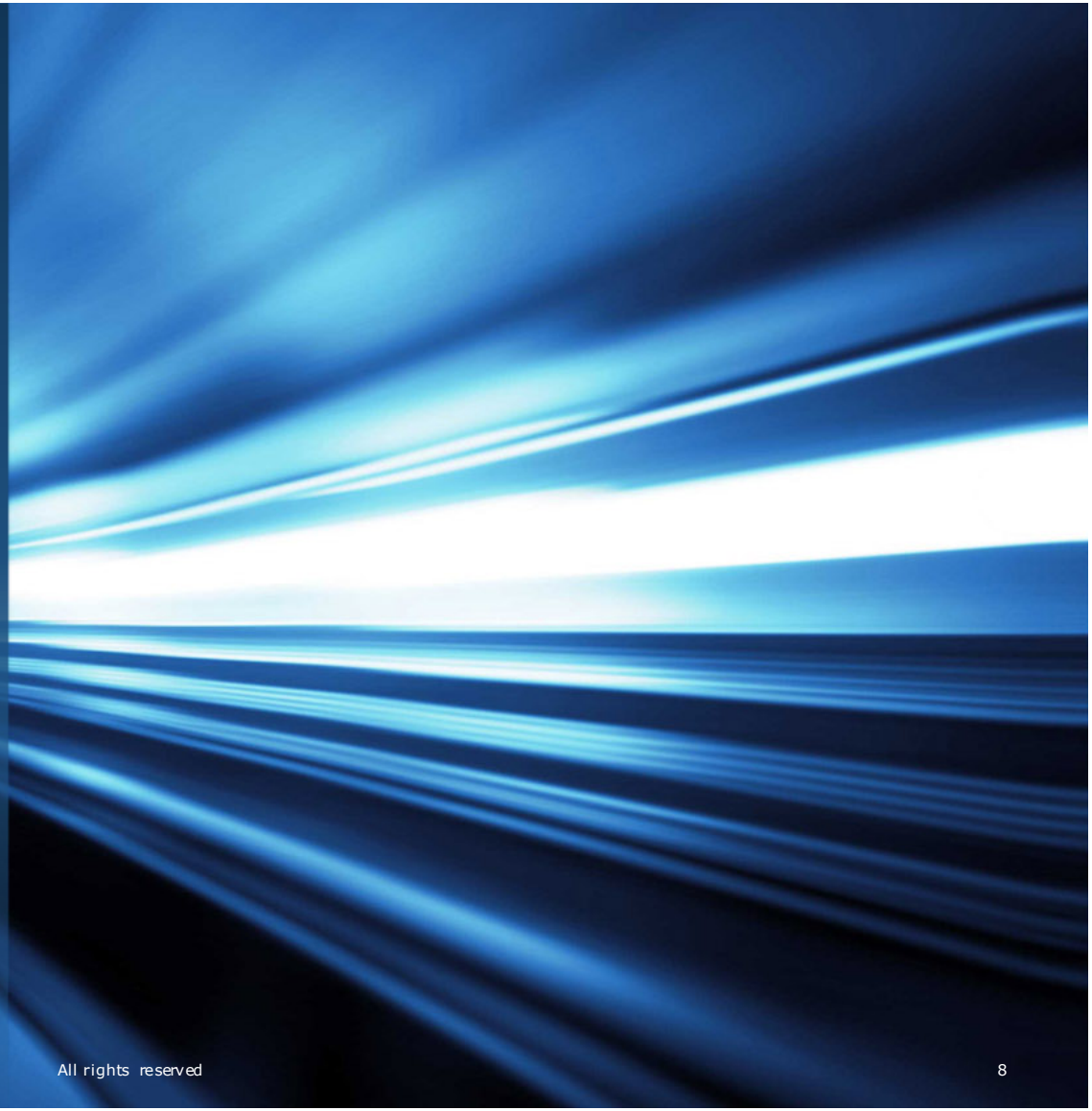
Key contributions

- Investigate throttling HDD's throughput to cap power
  - No strict positive correlation
  - HDD is underutilized
- Investigate resizing HDD's queues and its impact on power
  - Higher HDD utilization
  - Performance differentiation: throughput & tail-latency
  - Limitations under low concurrency and workload
- PCAP system based on queue resizing
  - Make it stable, agile and performance-aware
  - Compare it to throttling
  - Study it for different workloads & settings

PCAP in the works

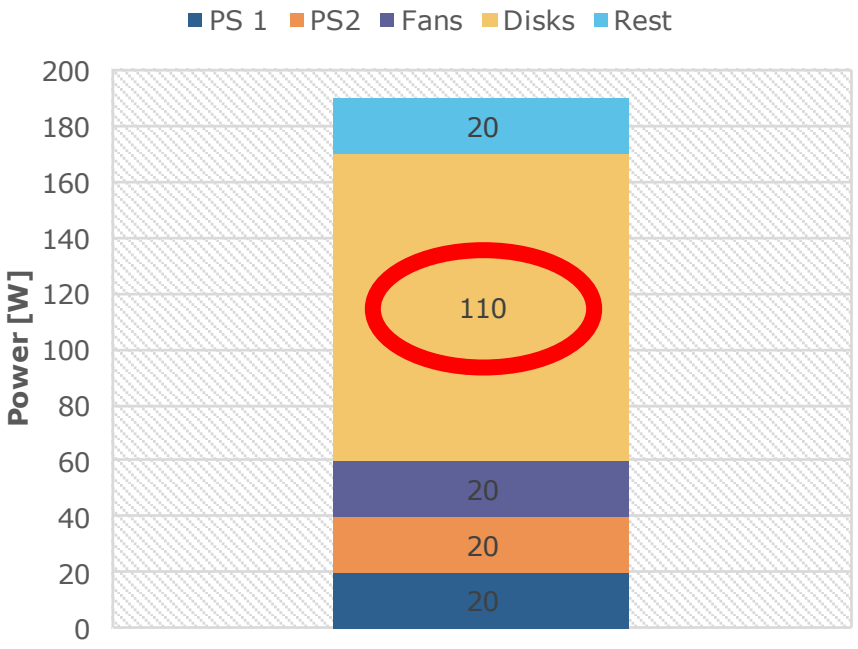


How we do it?

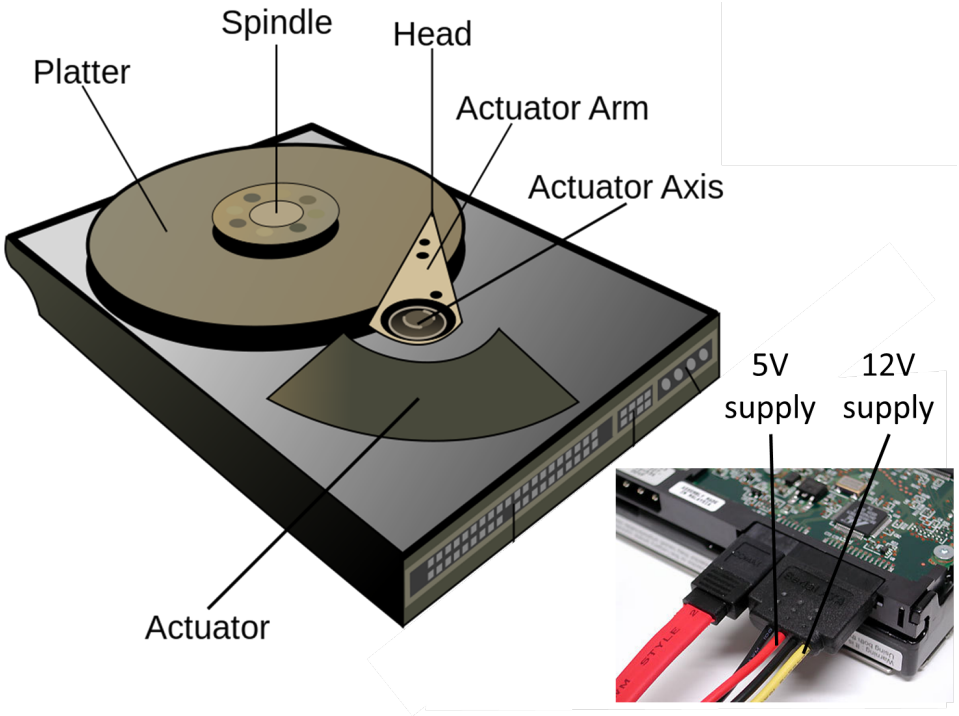


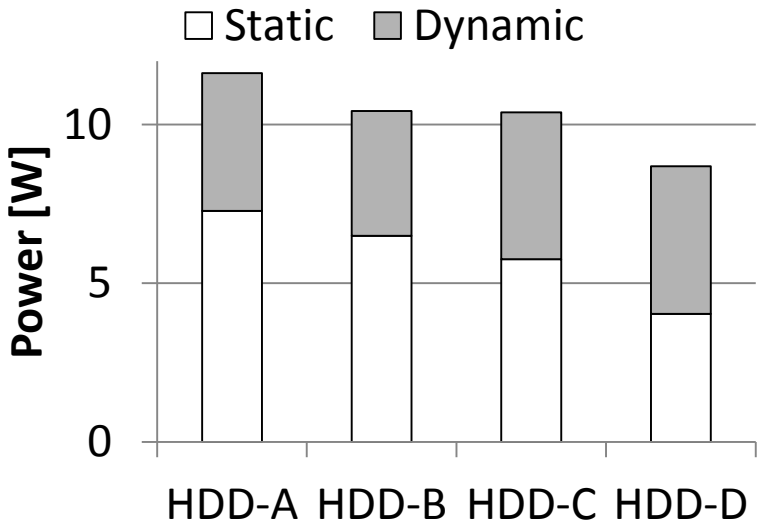
Setup

- A JBOD with 16x 4TB HDDs
- Exercising a single HDD only
- Workload generators:
  - FIO
  - YCSB & MongoDB
- Design space exploration:
  - Reads/writes/mixed
  - 4kB – 2MB
  - Threads: 10-256
  - Varying queue depth (QD)
    - HDD: 1-32
    - IO stack: 4-128
  - Deadline scheduler (default) is used



HDD's dynamic power



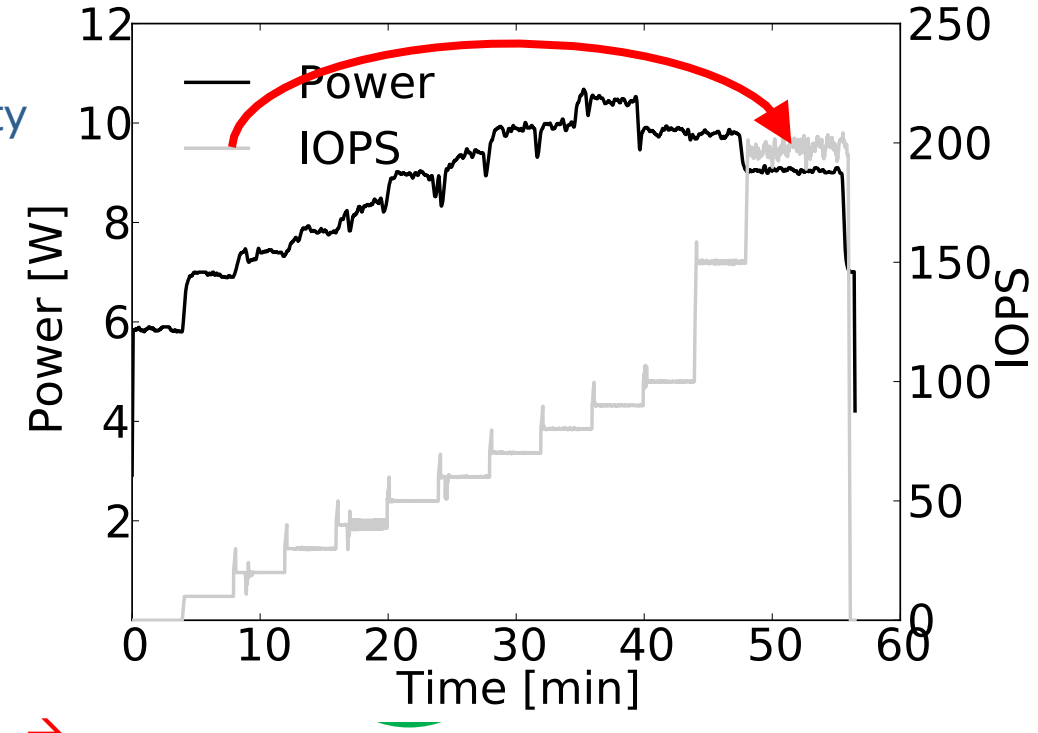


Existing techniques

1. Power off disks
2. Throttling throughput – adapting duty cycle

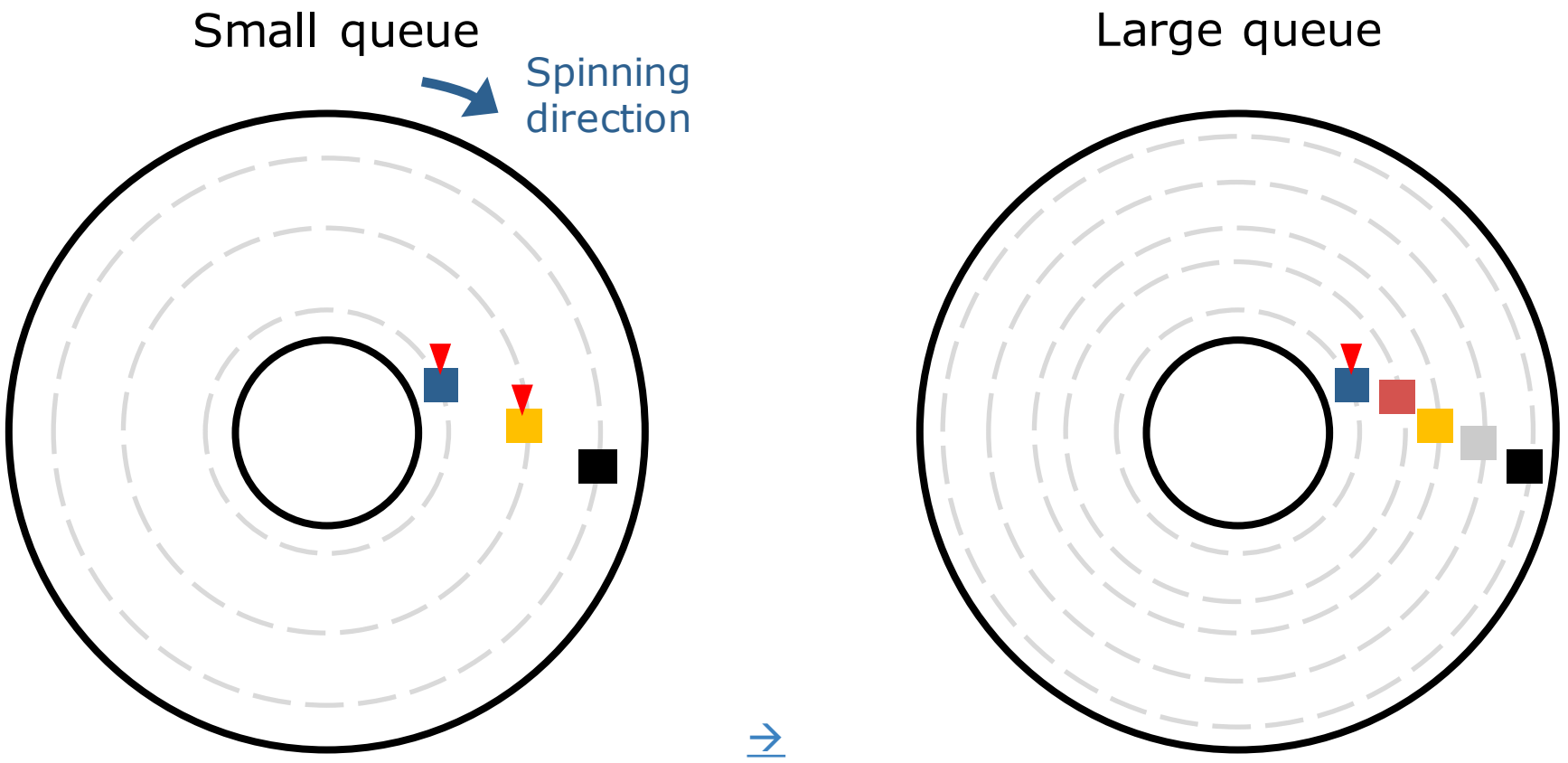
→ ***negatively impact throughput & latency (later)***

→ ***no strict positive correlation between power & throughput***

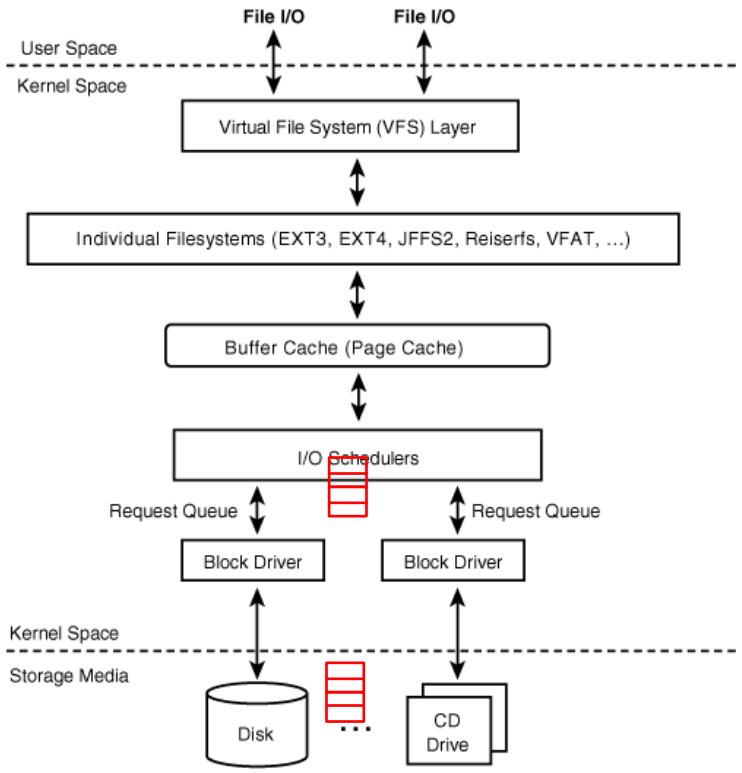


→

outstanding requests matters – queue size



Proposal: Resizing queues



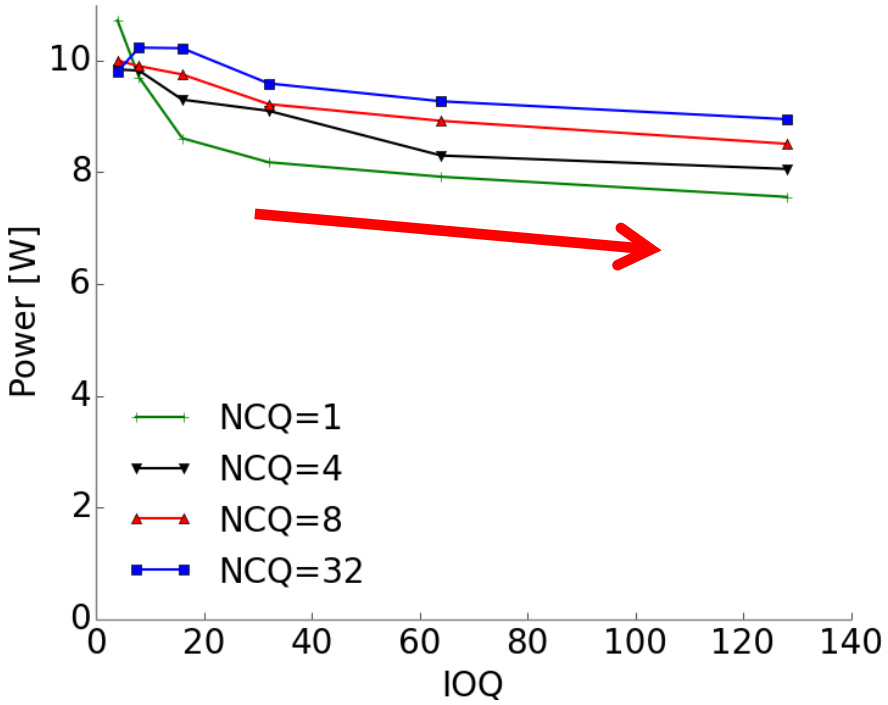
- We propose to resize HDD queues
- Queues:
  - OS I/O scheduler (IOQ)
  - HDD internal queue (NCQ)
- Dynamically resizing as the power cap changes
- Allows to control power
- Minding that queue size influences
  - Throughput
  - Tail-latency

Queue-size & power - causality

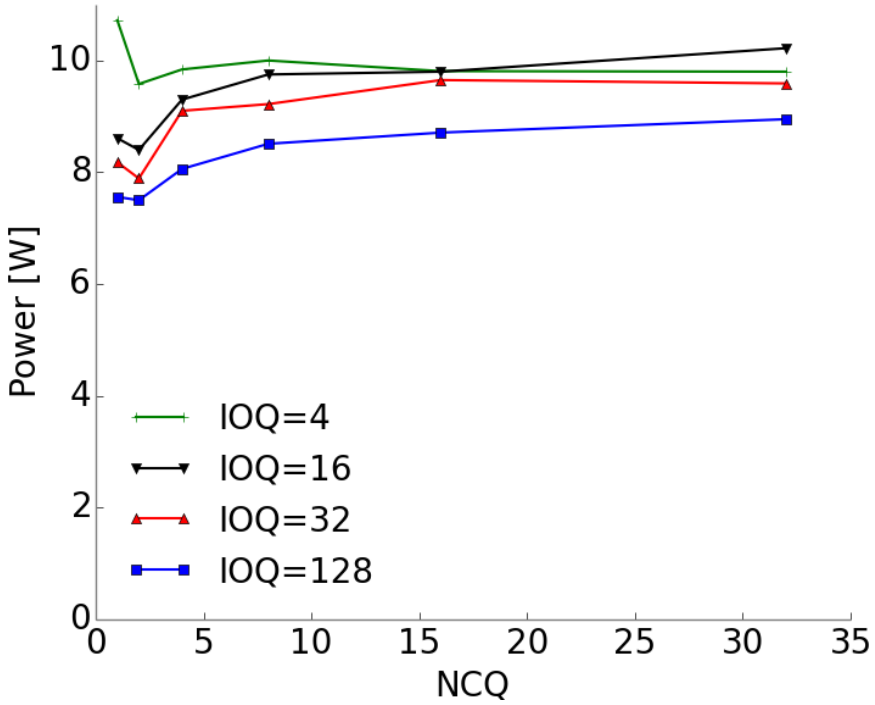
- Queue size influences the seek distance between requests
- A **small** queue results in **long** seek distance due to limited scheduling
- **Long** distances require **acceleration & deceleration**
- Acceleration and deceleration takes relatively **large power**
- And vice versa

Power vs. Queue size

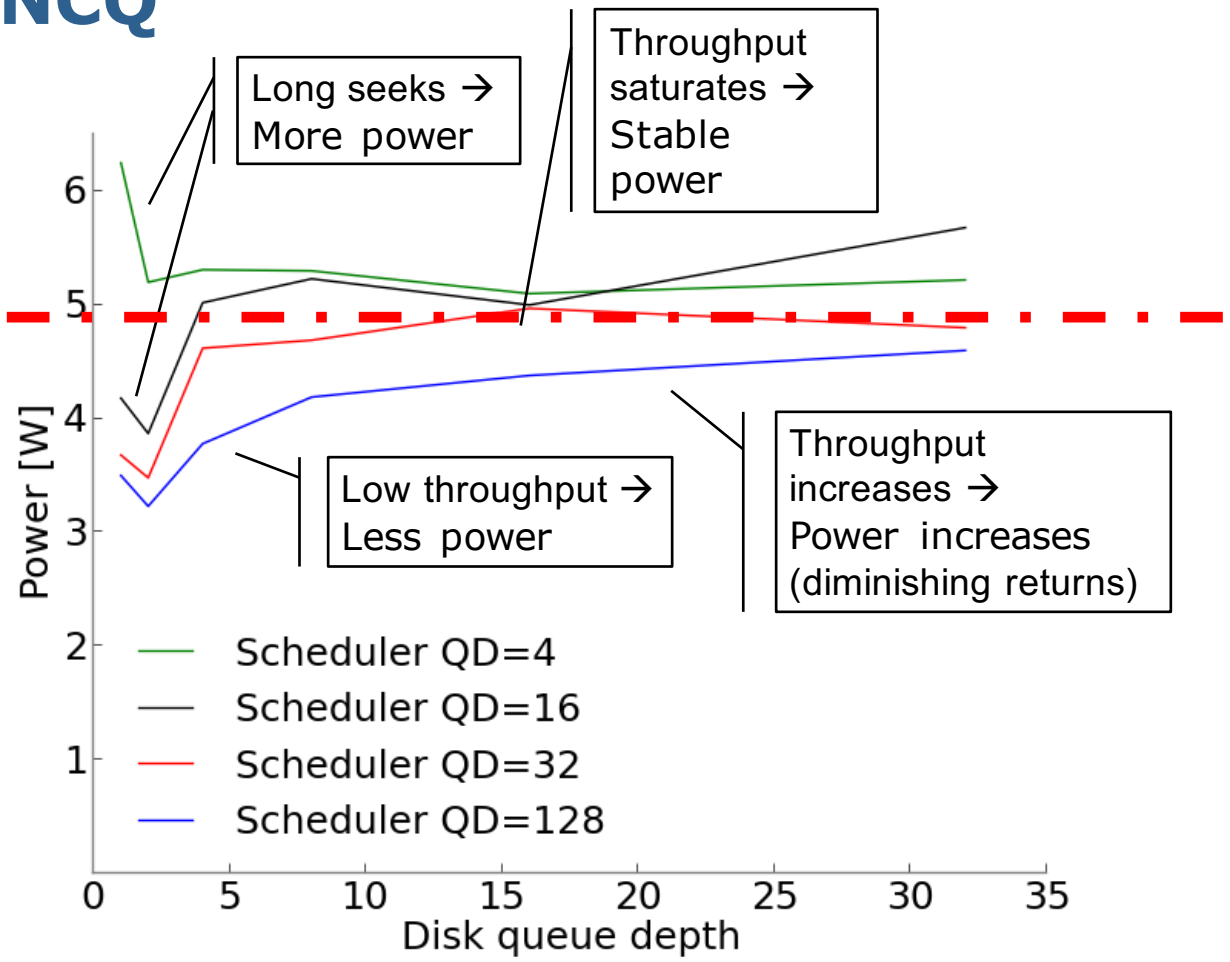
Power vs. IOQ



Power vs. NCQ



Power vs. NCQ



PCAP design

- Reduce HDD's power ***quickly*** to bring it below the power cap
- ***Max out*** HDD's performance when more power is available
- But ***cautiously*** so that power cap is not violated

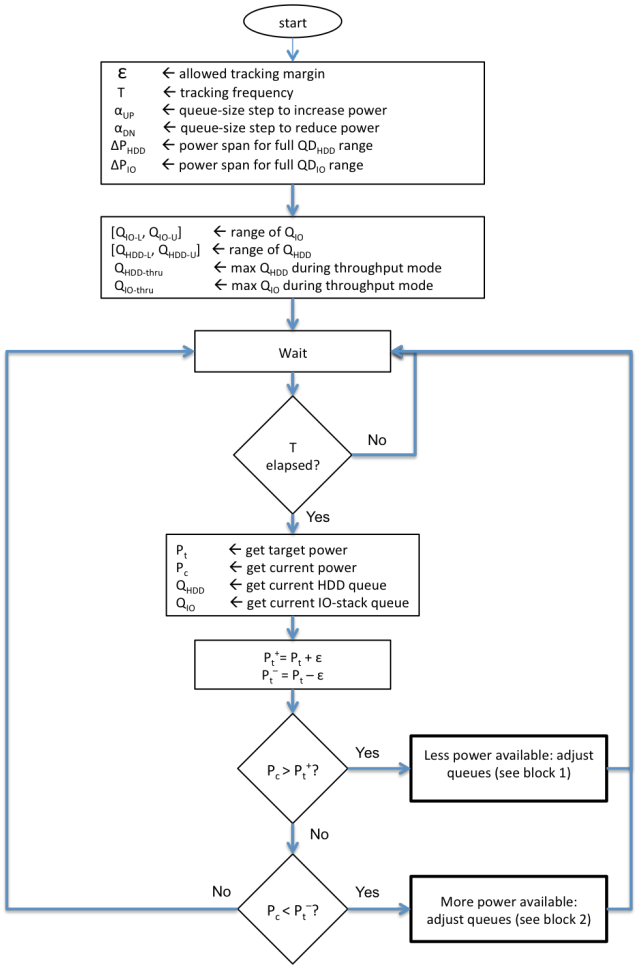
→ Different scaling factors of the queue sizes α_{UP} and α_{DN}

- Reduces oscillations around the target power

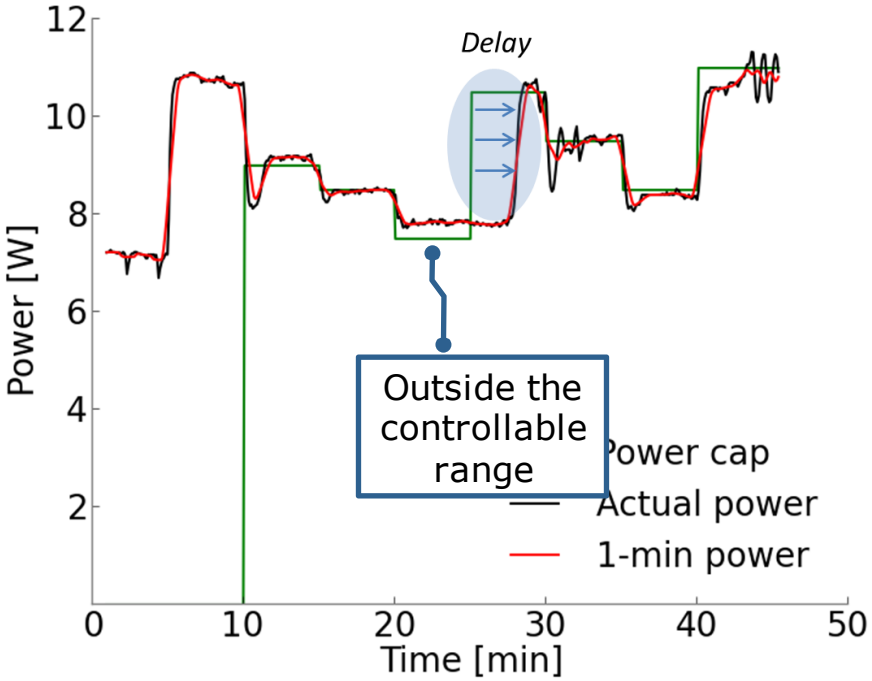
→ Hysteresis with margins [-ε,+ε]

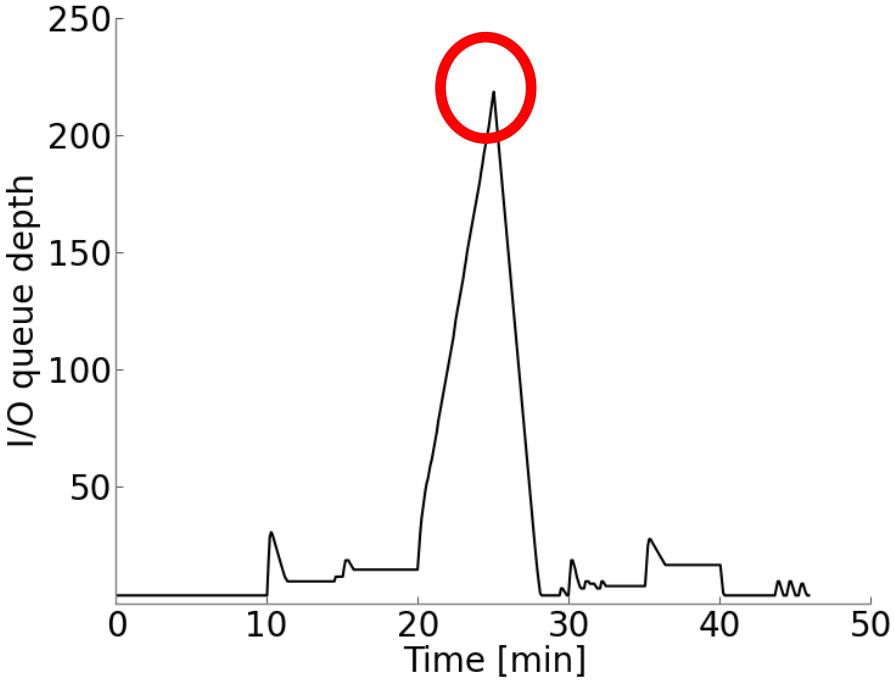
- Periodically adapts queues to ensure cap power

→ Tunable period (T)

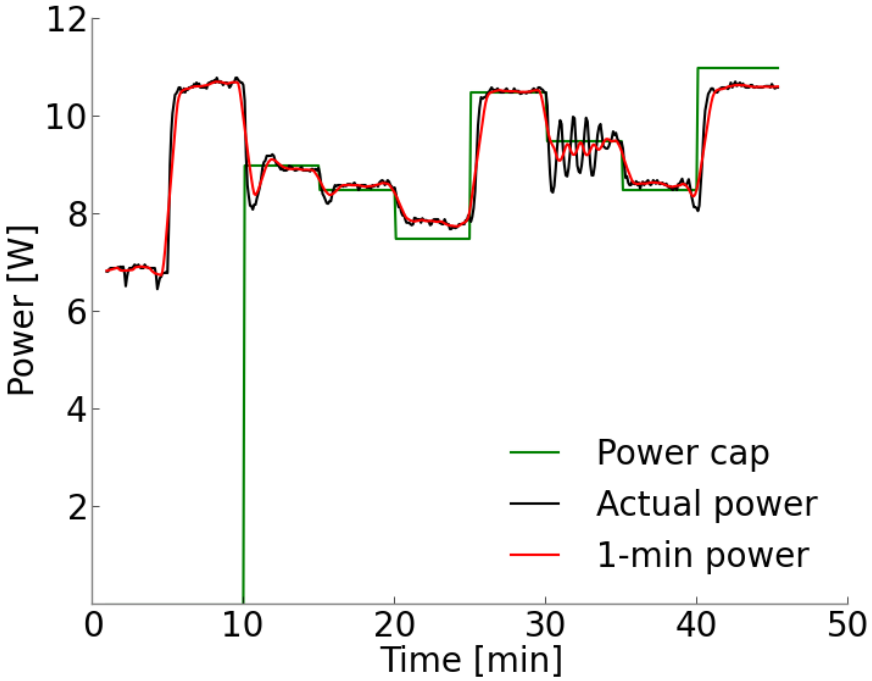


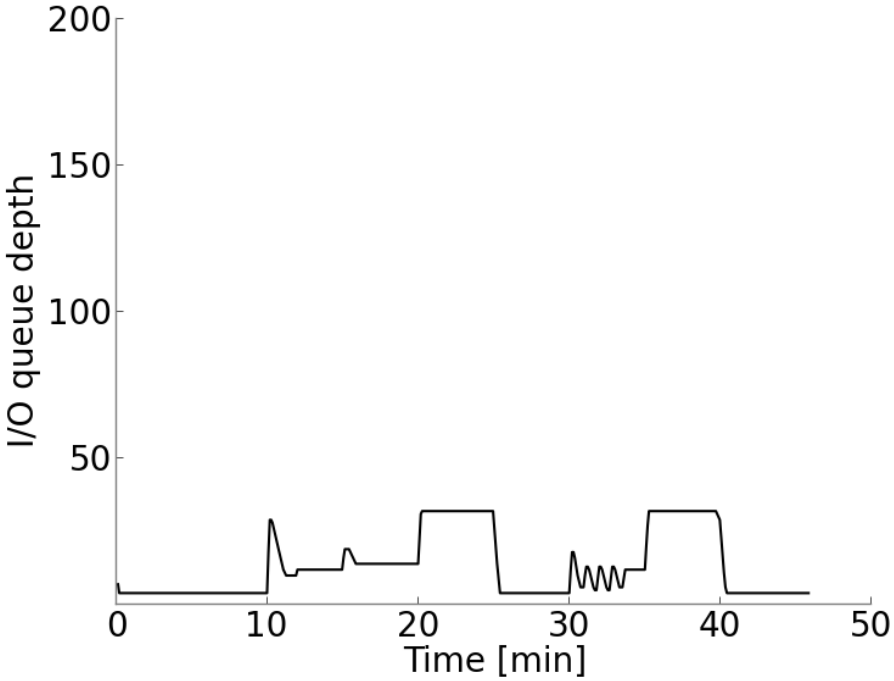
PCAP: basic



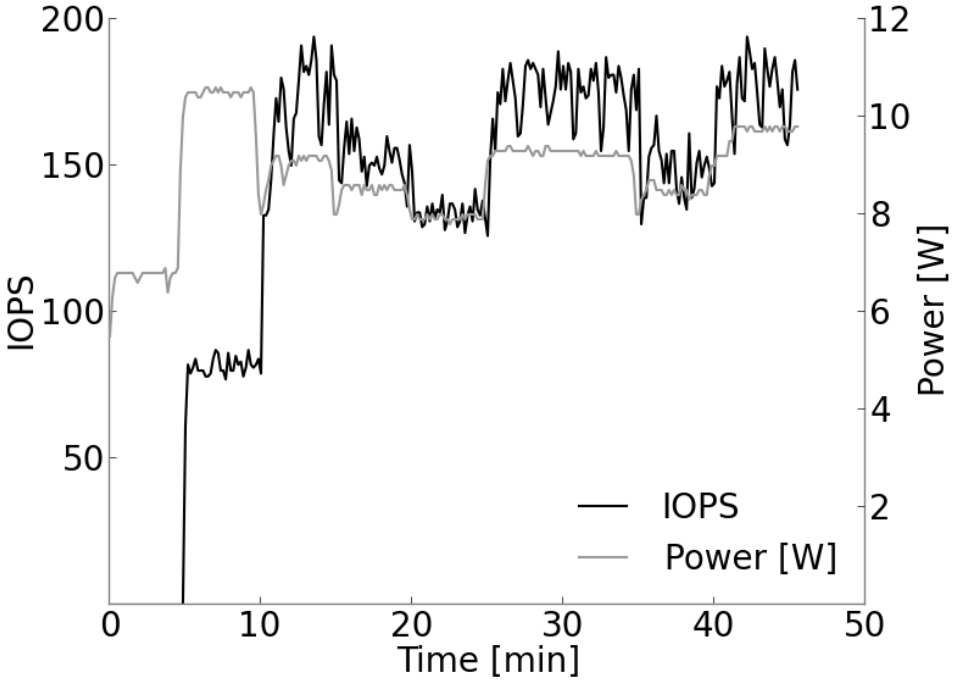


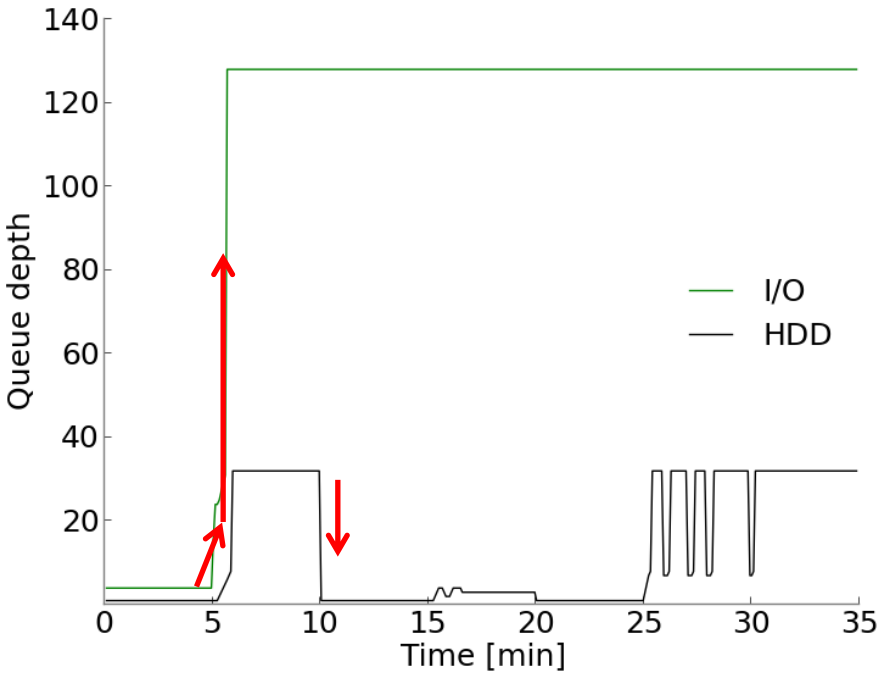
PCAP: Agile (bounded queues)





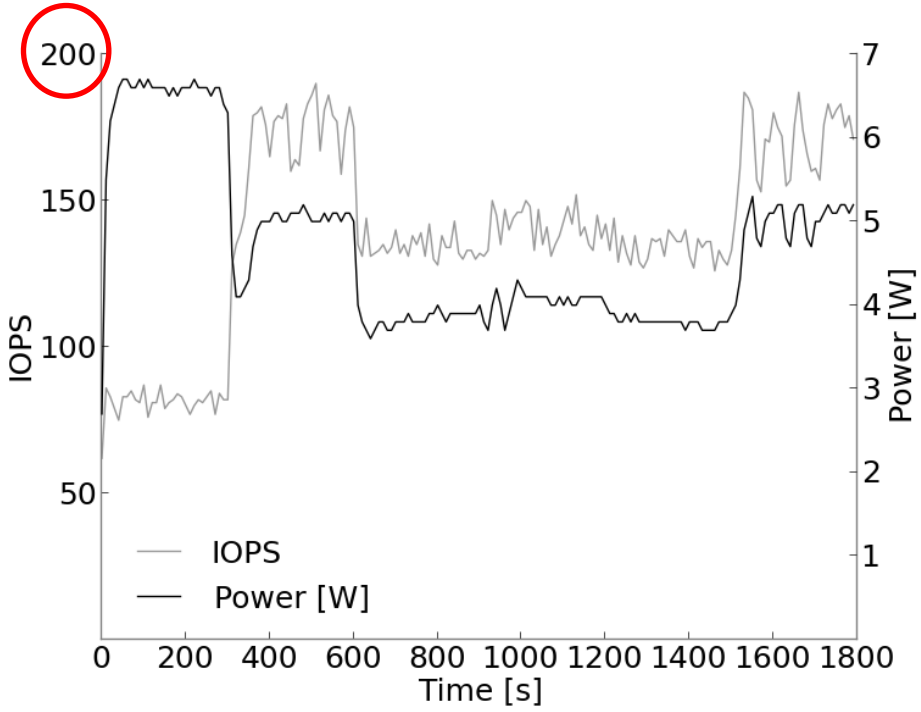
PCAP: Agile w/ improved throughput



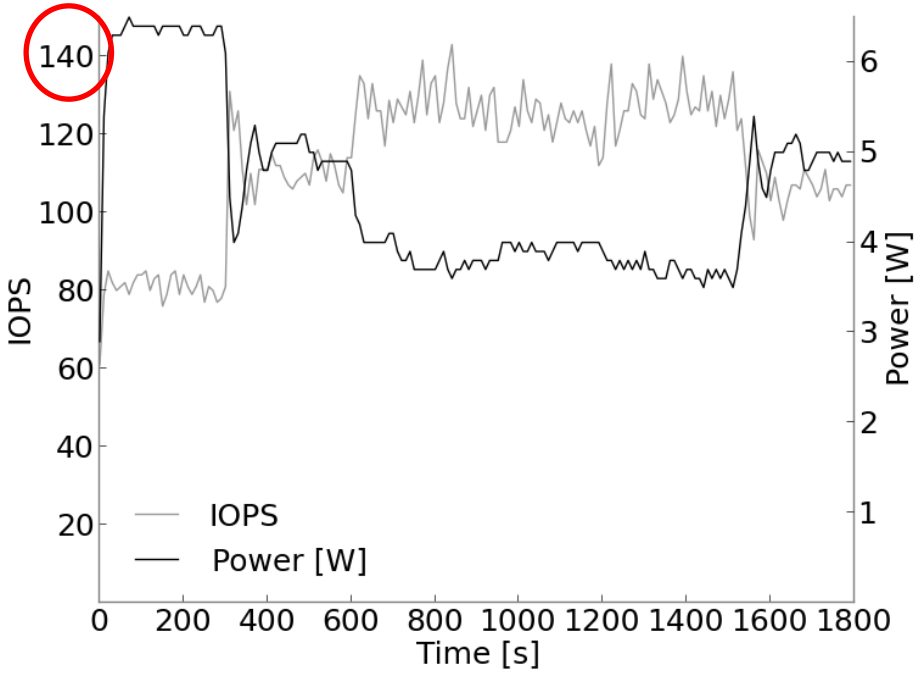


PCAP: Dual-mode

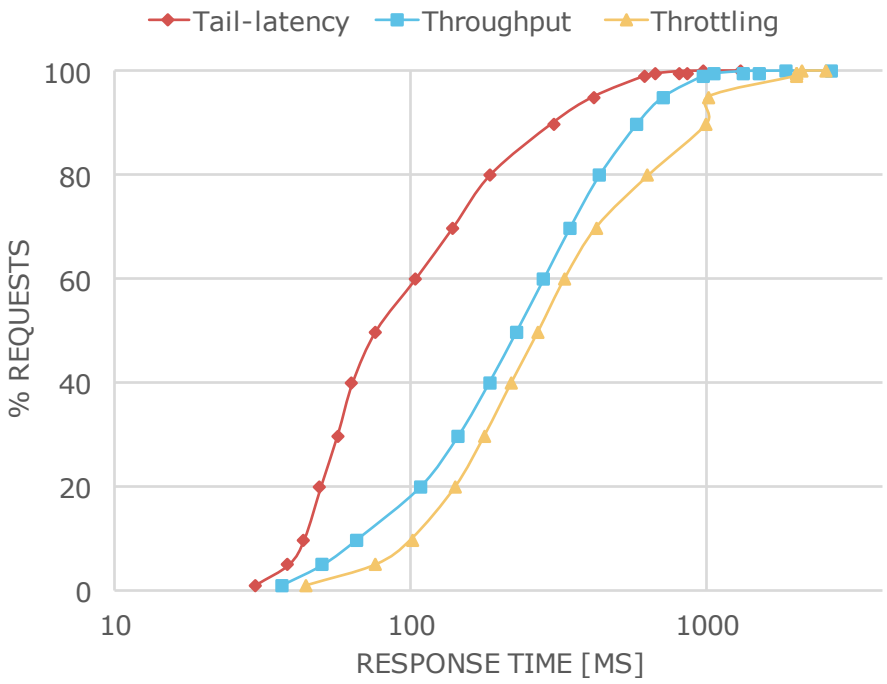
Throughput



Tail-latency

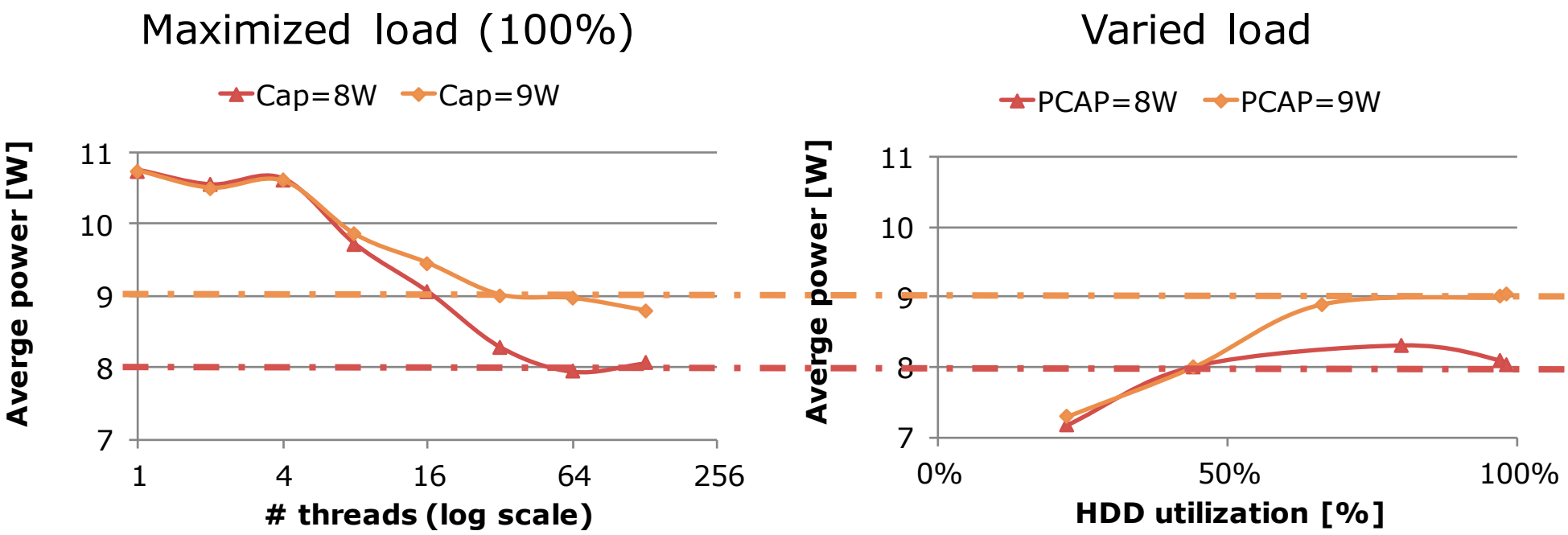


Summary of performance



	Avg. throughput [IOPS]	% requests < 100ms	Max. latency [ms]
Throttling	117	10%	2.5
PCAP - Throughput	154 (32%)	20%	2.7
PCAP – tail latency	102 (-15%)	60%	1.3

PCAP limitations
Effective queue size matters



Conclusions

- Throttling underutilizes HDD's performance
  - Useful under low concurrency and sequential throughput
- PCAP: resizing queues
  - Improves HDD's utilization
  - 32% more throughput
  - 50% more requests < 100ms
  - WC latency reduce by 2x
- PCAP performance is limited under
  - Low concurrency
  - Light workloads
- PCAP works for multiple HDDs
- Please see the paper for more observations and results

Thanks for your attention!

Q & A

Interested in internship in WDC research?

Apply at: http://bit.do/FAST16

Email: mohammed.khatib@hgst.com