Thesios: Synthesizing Accurate Counterfactual I/O Traces from Factual I/O Samples

Phitchaya Mangpo Phothilimthana*,
Saurabh Kadekodi*, Soroush Ghodrati,
Selene Moon, Martin Maas

** equal contribution*

Can we reduce energy consumed by disks in data centers?

Idea: create cold data disks running in low power

Motivation

- Representative I/O traces are critical to the **designs of storage systems**
- Understand **the system** and analyze **proposed changes** on
  - Latency (SLO)
  - Power consumption
- Enable sharing realistic traces from **large-scale data centers** with academia and vendors

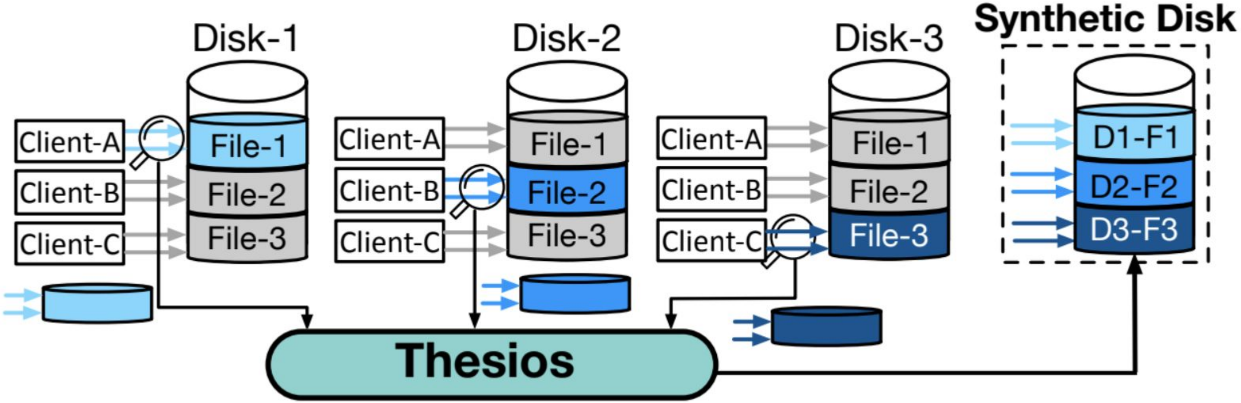
Sampled I/O Disk Traces

- **Distributed storage system**
- **Sampling system** maintains downsampled I/O traces
  - Collect telemetry from **1–in–n RPCs**
  - **n** is between 100 and 10,000

Sampled I/O Disk Traces

- **Distributed storage system**
- **Sampling system** maintains downsampled I/O traces
  - Collect telemetry from **1–in–n RPCs**
  - **n** is between 100 and 10,000
- What **cannot be understood** from **sampled traces**:
  - I/O request interarrival distribution
  - Evaluate latency, utilization, etc. due to placement policy changes
  - Evaluate impact of new hardware such as low RPM, HAMR disks

Thesios: Synthesizing Full I/O traces

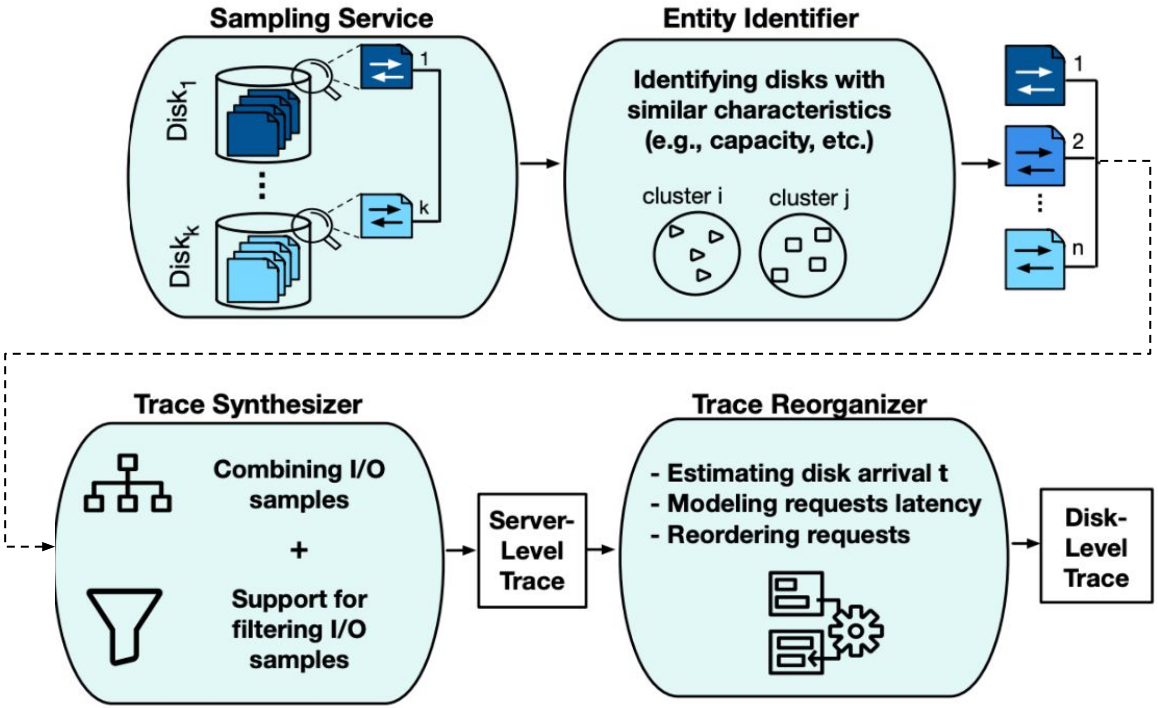


Key idea: combine I/O samples from multiple disks to synthesize full-resolution trace

- **Representative full-resolution trace** mimics current workload
- **Counterfactual full-resolution trace** for "what-if" scenarios

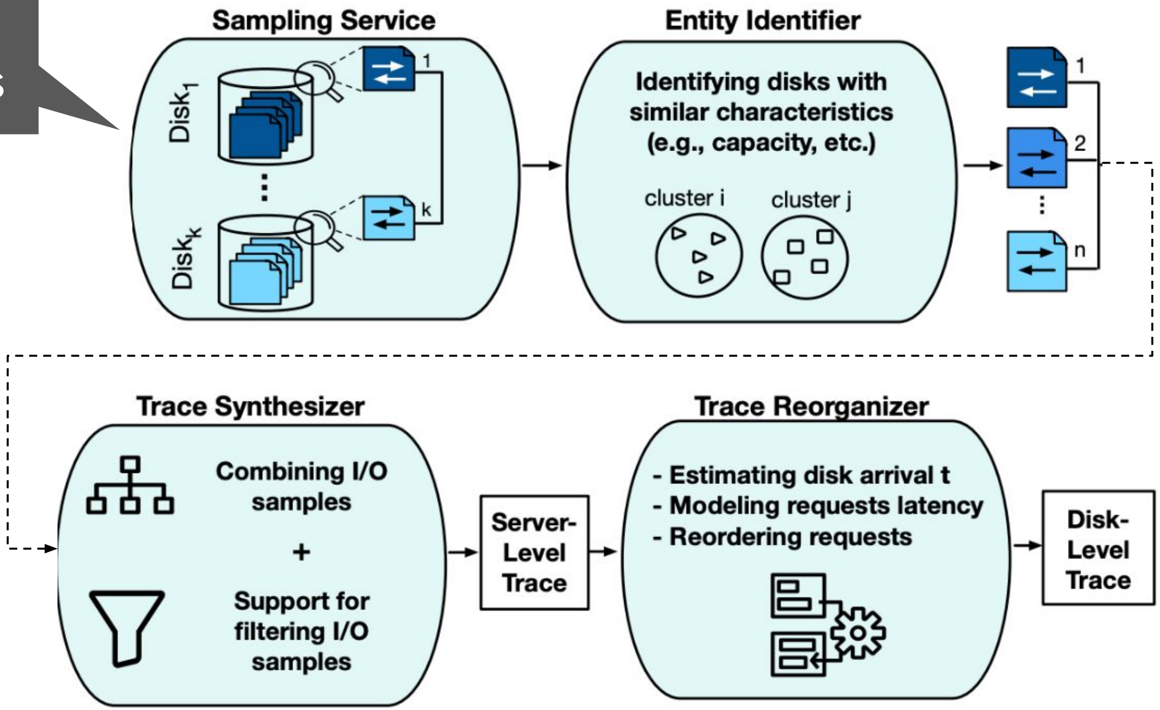
Thesios is a reference to "Ship of Theseus"

Framework Overview



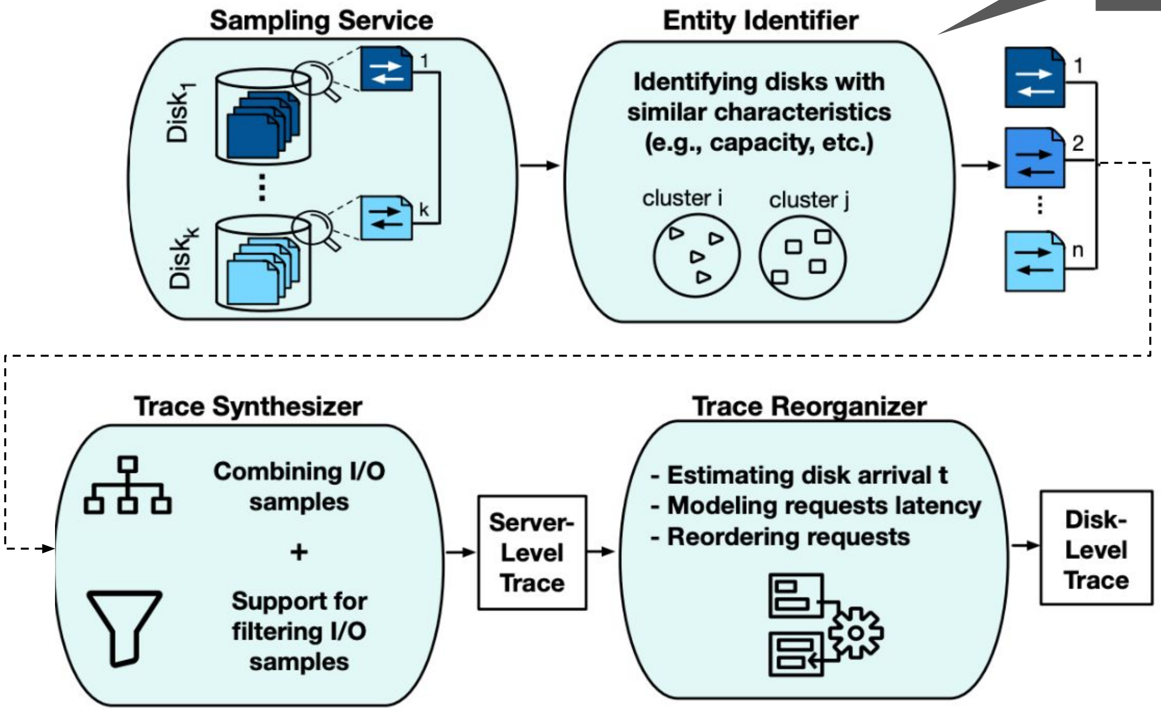
Framework Overview

Sample
1 in n files



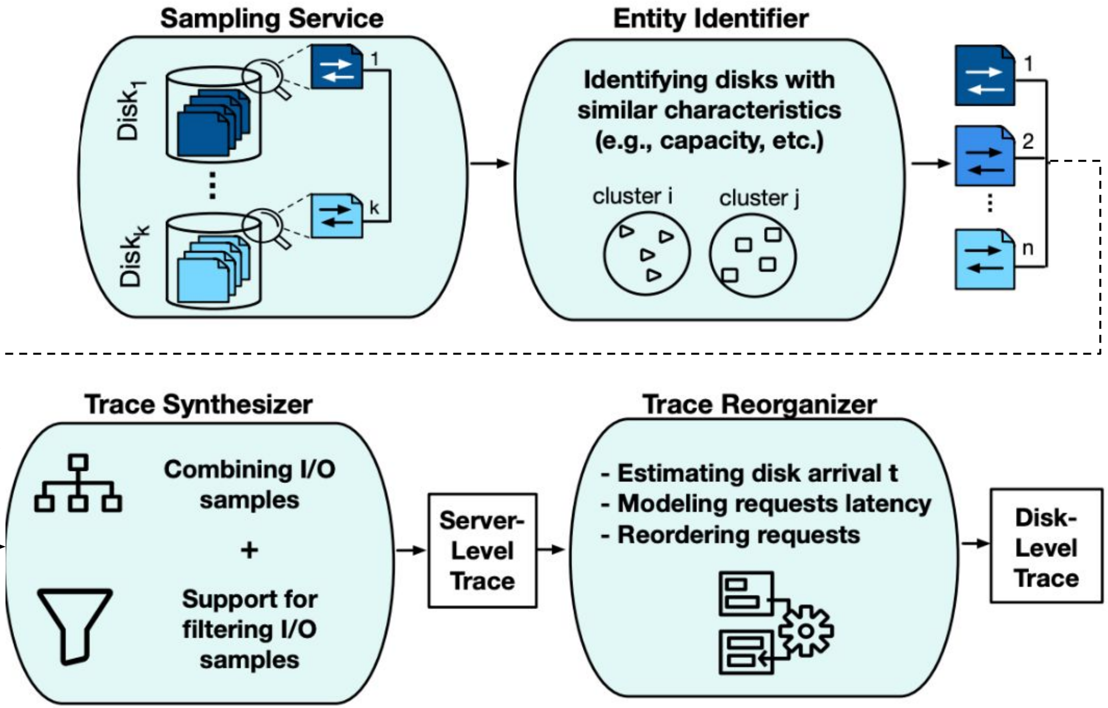
Framework Overview

Samples from which disks?

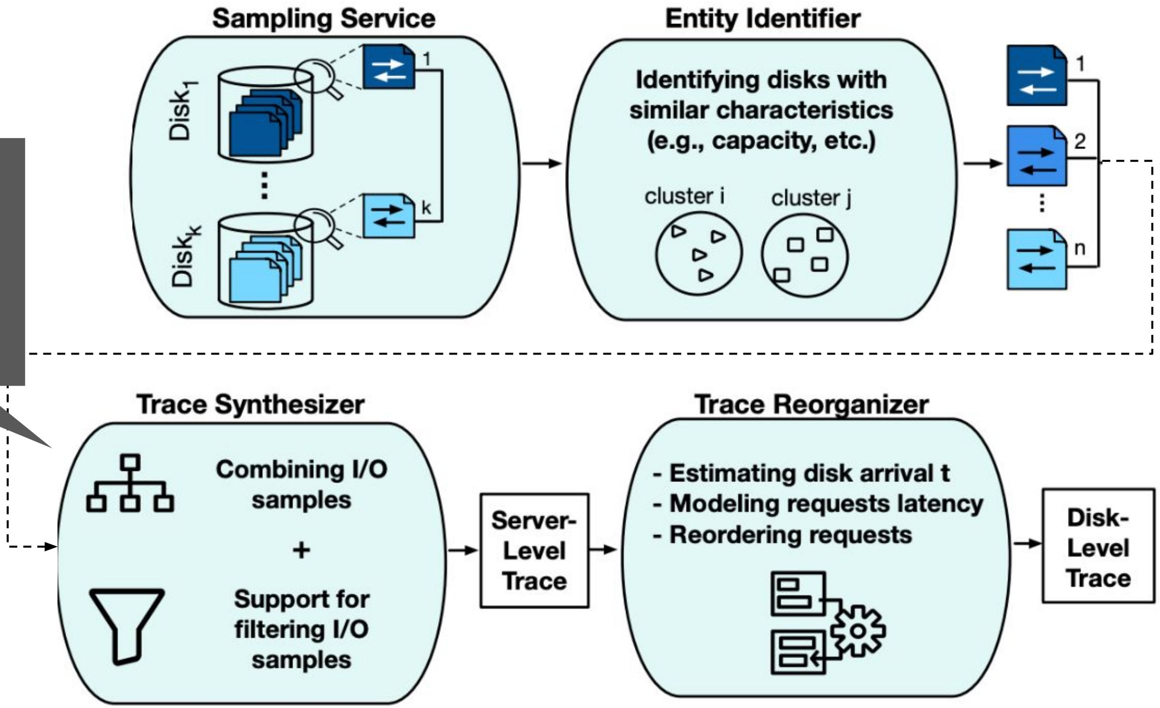


Framework Overview

How many samples to combine?



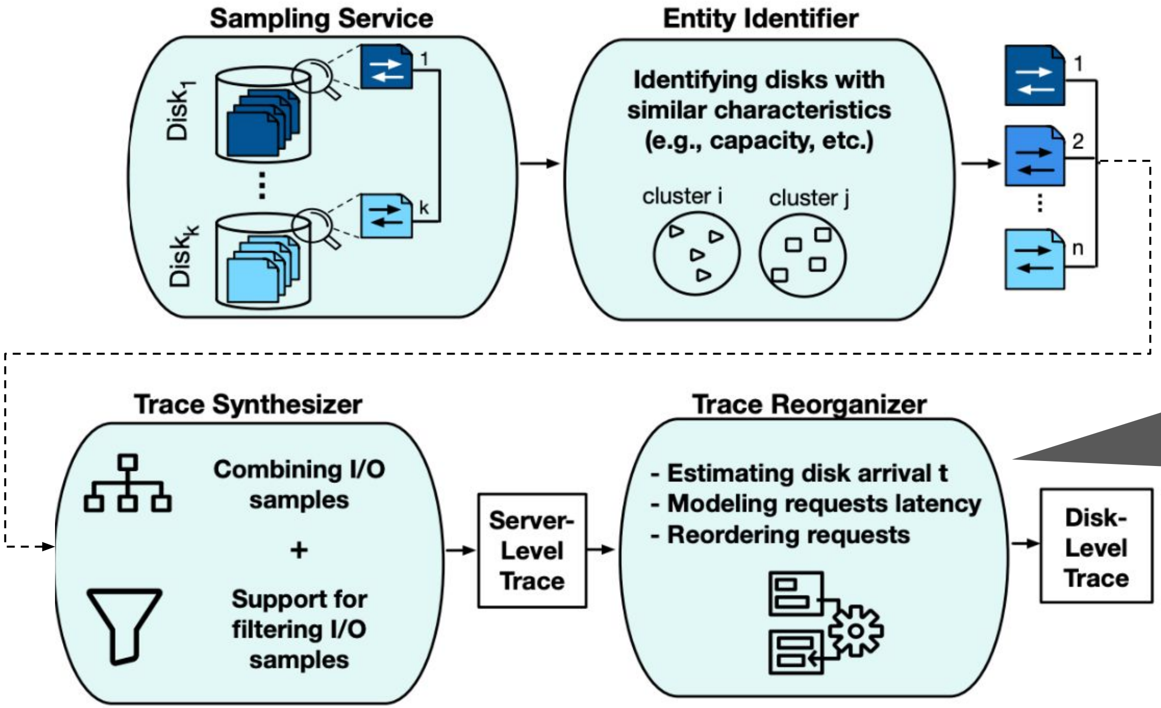
Framework Overview



How many samples to combine?

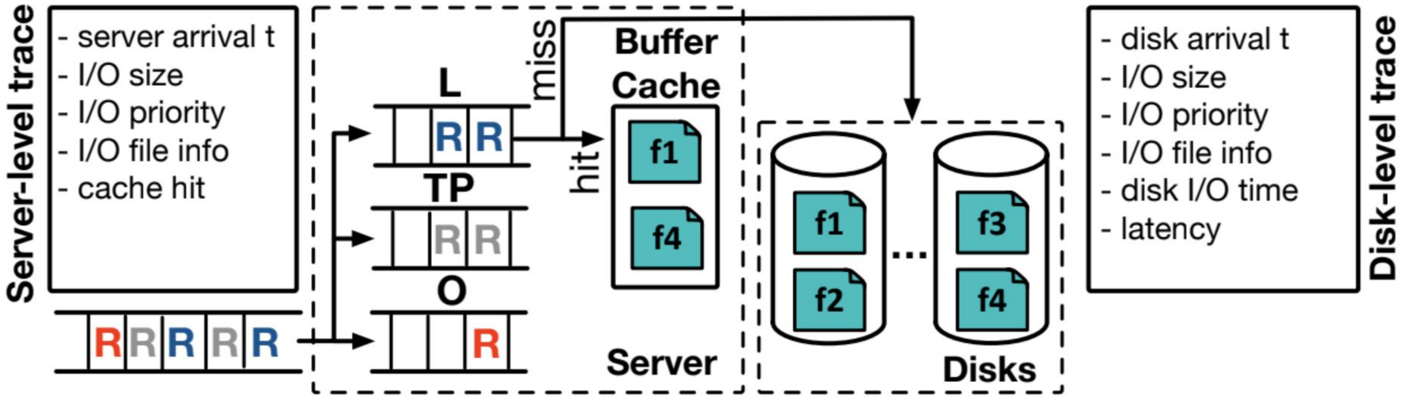
n disks for **1/n** sampling rate

Framework Overview



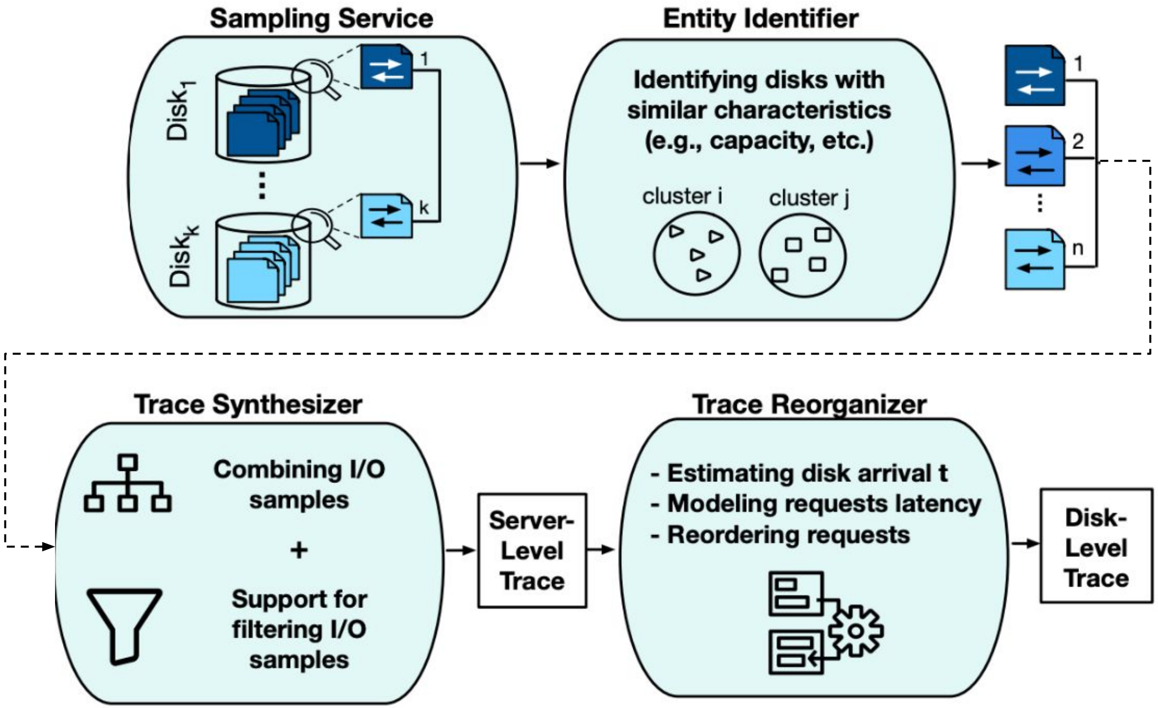
Should requests be **reordered**?
How should **latency** be calculated?

Sever-Level & Disk-Level Traces

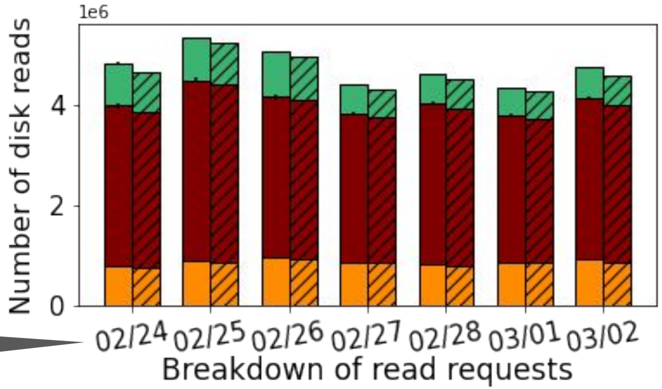


- The server **reorders requests** according to their priorities.
- **Some reads** are served from **buffer cache**.
- **All writes** are written to **buffer cache**, which is a write-back cache.

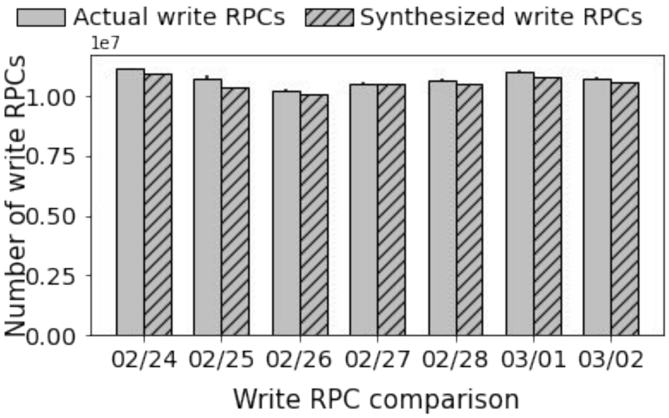
Framework Overview

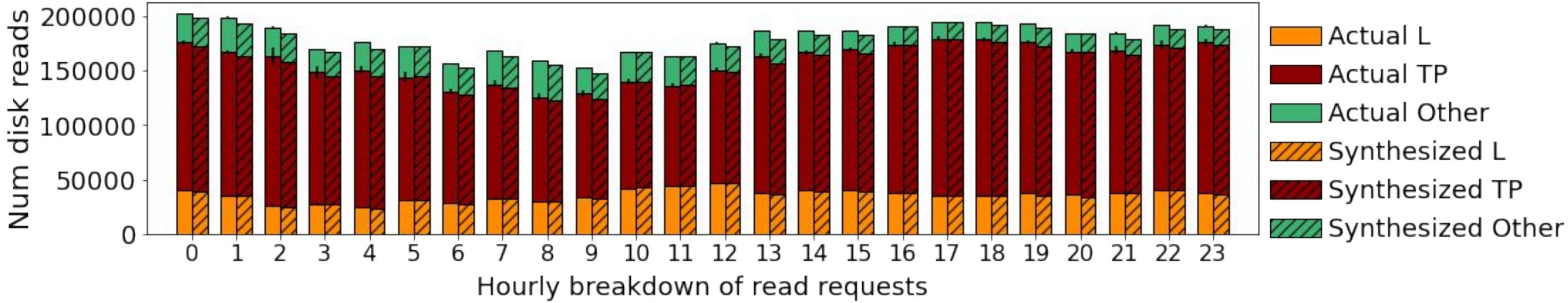


Validation: Number & Breakdown of I/O Requests

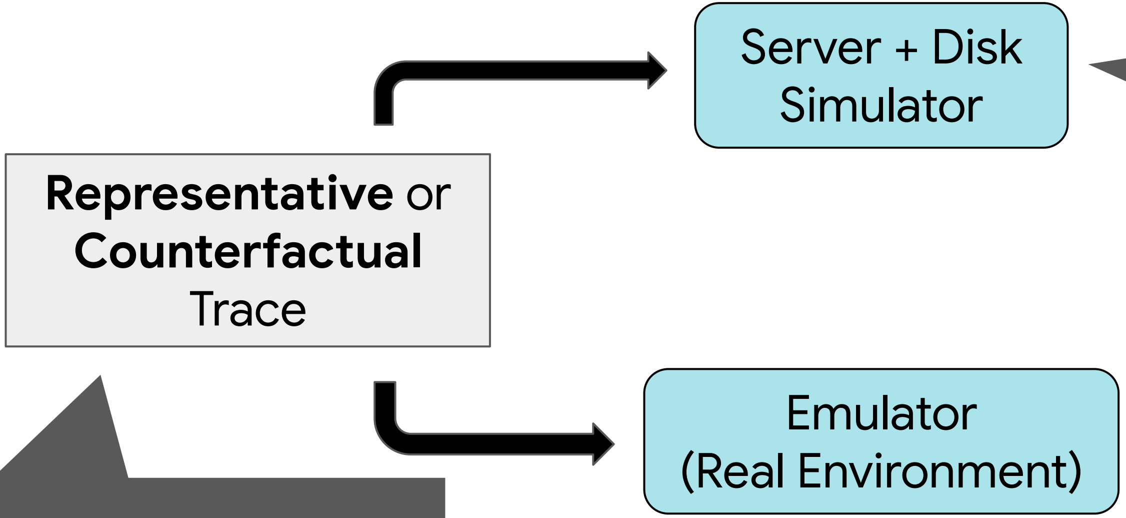


friday





Counterfactual Analysis



Representative or **Counterfactual** Trace

Server + Disk Simulator

Emulator (Real Environment)

Can be extended to model changes, such as:

- New types of disks
- Disk power management

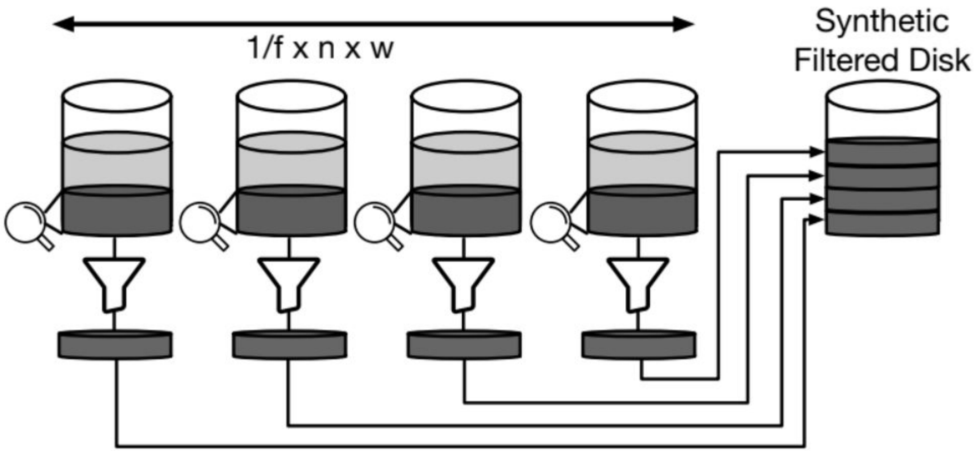
Can be modified directly to experiment with new configurations, such as:

- Cache policy

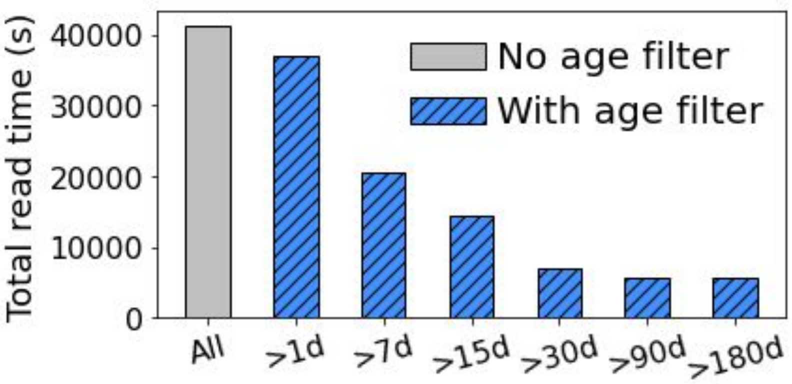
- Synthesize for specific capacity, fullness, and utilization
- Filter by application, age, encoding

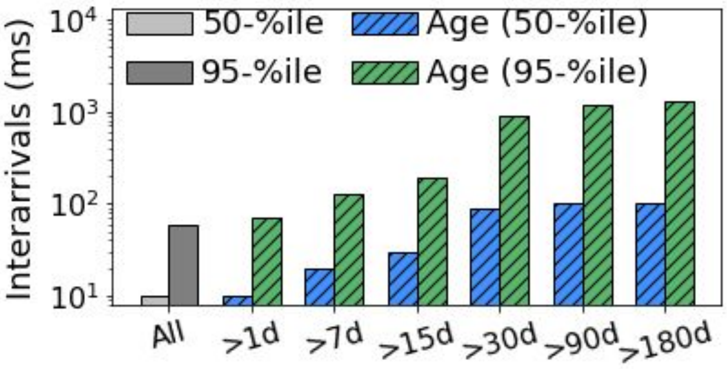
I. Hot & Cold Data Segregation

- Thesios supports **filtering by age, encoding, and application**
- ***f*** = fraction of files that meet the filtering criteria (by size)



I. Hot & Cold Data Segregation

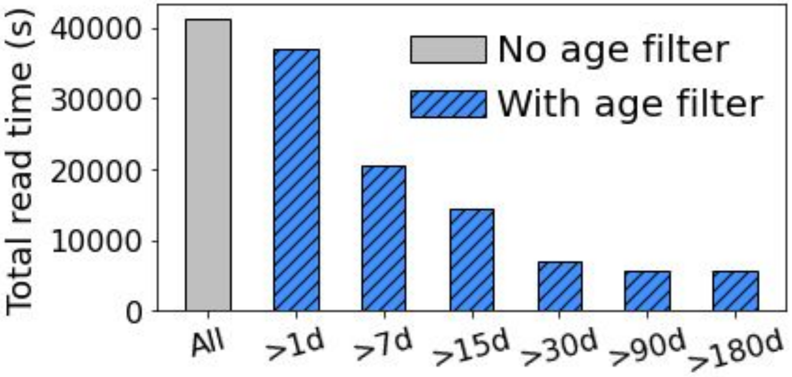


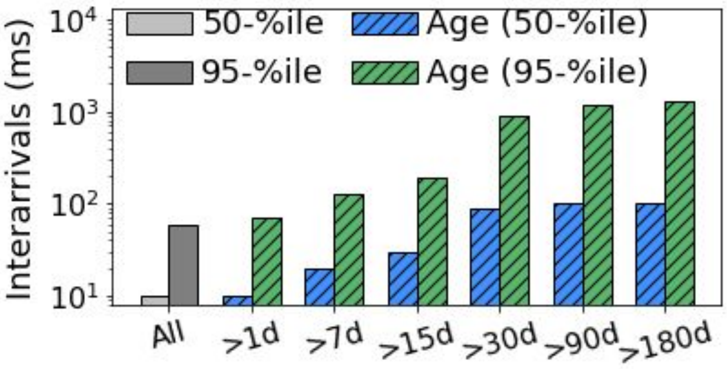


Synthesized trace shows (as expected) **older files → colder**

- **Utilization reduces from 52% to 8%**
- 50th and 95th percentile **request interarrival increases by >10x** for cold trace

I. Hot & Cold Data Segregation





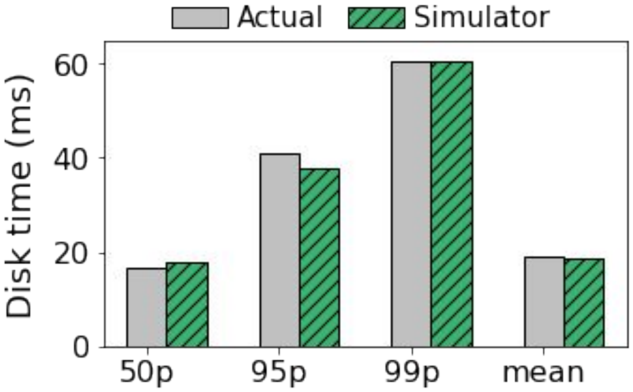
Synthesized trace shows (as expected) **older files → colder**

- **Utilization reduces from 52% to 8%**
- 50th and 95th percentile **request interarrival increases by >10x** for cold trace
- **Still enough to turn off disks**

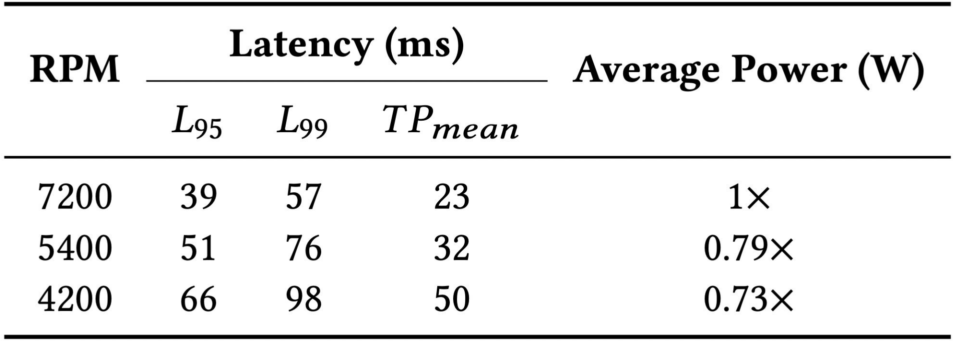
II. Low-RPM Disk

Simulate **impact of low-RPM disks** on individual requests

- **T_d** (disk time) = **T_s** **(**seek time) **+** **T_r** (rotational latency) **+** **T_t** (transfer time)
- Low-RPM increases **T_r** and **T_t** wrt RPM slowdown
- Evaluate 5400 RPM and 4200 RPM against 7200 RPM (current)



Validation with real low-RPM disk

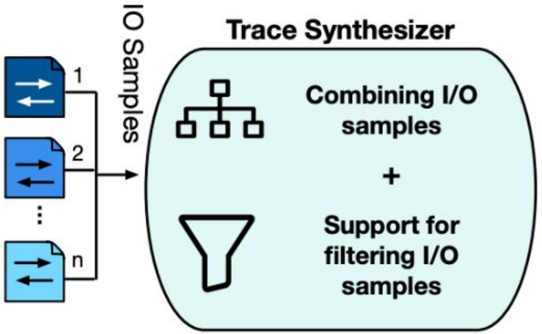


RPM	Latency (ms)			Average Power (W)
	L_{95}	L_{99}	TP_{mean}	
7200	39	57	23	1×
5400	51	76	32	0.79×
4200	66	98	50	0.73×

Summary

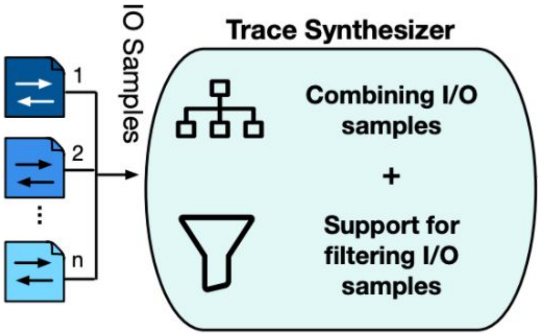
- Thesios synthesizes representative traces with **high accuracy**

- Thesios **enables risk-free "what-if" evaluations** of policy and hardware changes



Summary

- Thesios synthesizes representative traces with **high accuracy**
- Thesios **enables risk-free "what-if" evaluations** of policy and hardware changes
- Release **2-month-long synthesized traces** from Google storage clusters: github.com/google-research-datasets/thesios


1
2
⋮
n
IO Samples
Trace Synthesizer
Combining I/O samples
+
Support for filtering I/O samples