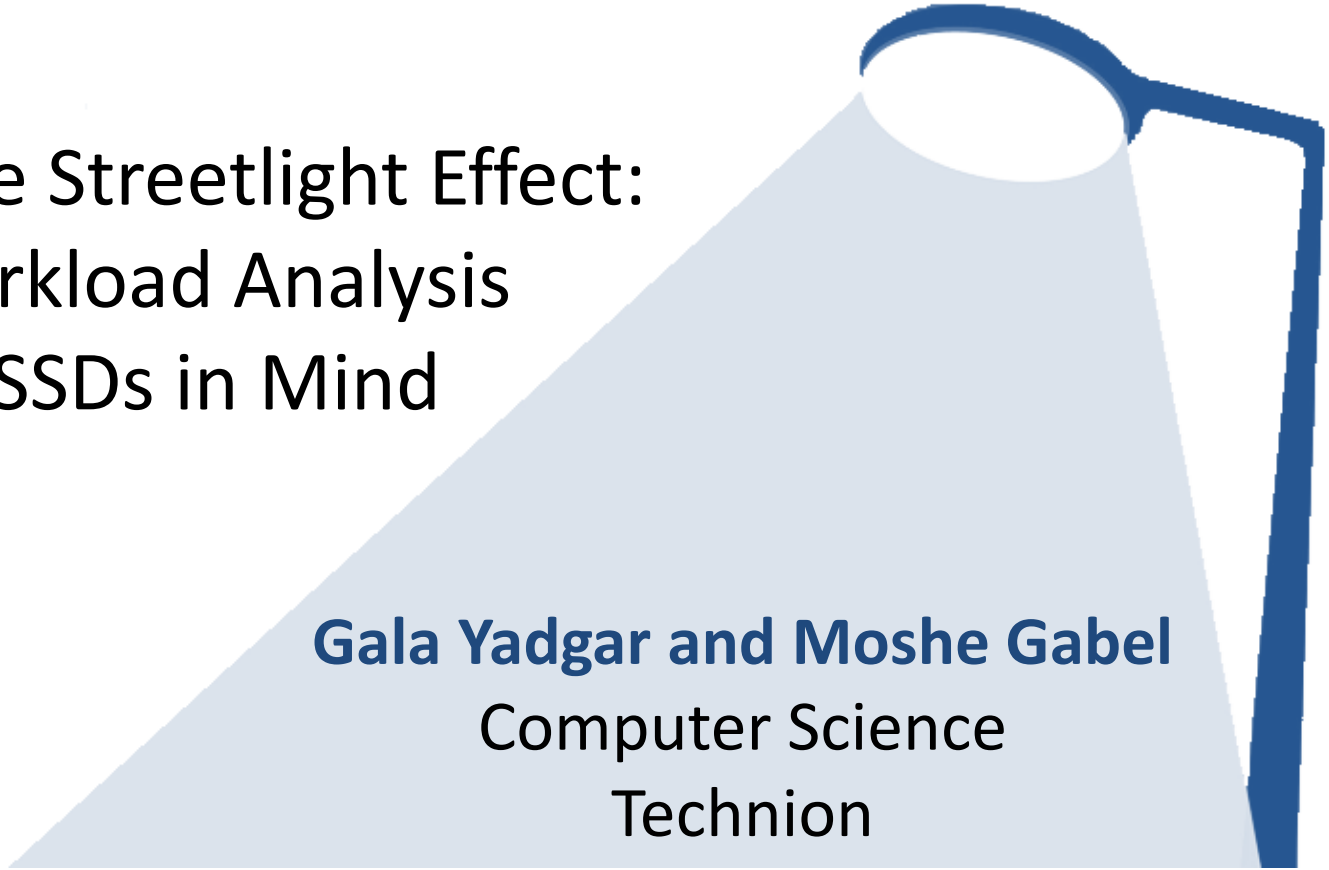
Technion – Israel Institute of Technology

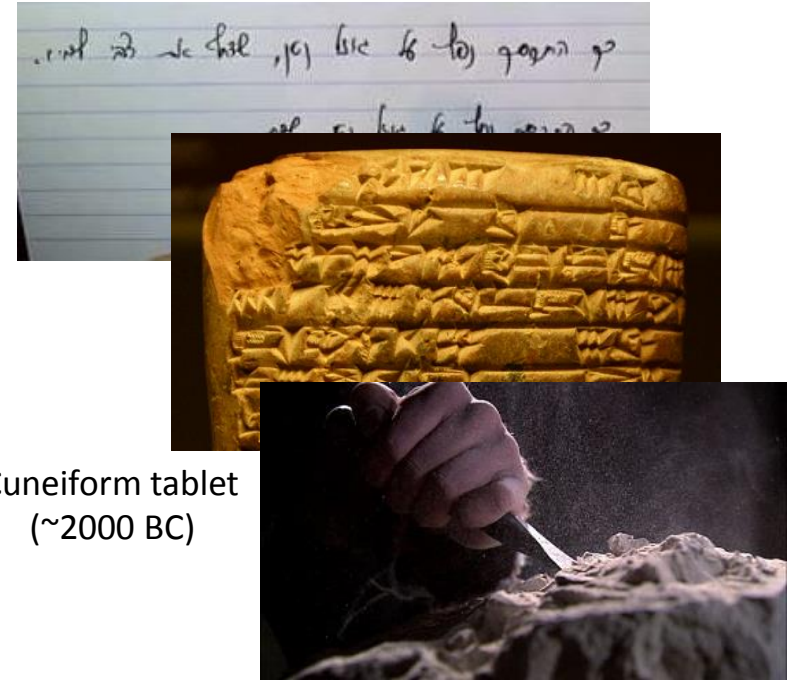
# Avoiding the Streetlight Effect: I/O Workload Analysis with SSDs in Mind



**Gala Yadgar and Moshe Gabel**
Computer Science
Technion

# “Why is Hebrew written backwards?”


Codex Argenteus
(~500 AC)

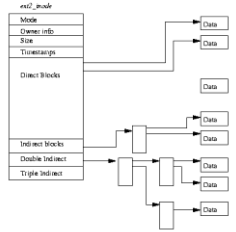


Cuneiform tablet
(~2000 BC)

It’s the storage media!

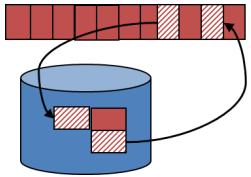
# Workloads Inspire Optimizations

- File size
- File age and functional lifetime
- Directory structure
- Read/write ratio
- Inter-reference gaps
- Working set size
- Access skew
- Request sizes
- Idle times
- Inter-arrival times
- Sequentiality




Metadata

Hard disk



Cache

# Storage Media is Changing

- File size
- File age and functional lifetime
- Directory structure
- Read/write ratio
- Inter-reference gaps
- Working set size
- Access skew
- Request sizes
- Idle times
- Inter-arrival times
- Sequentiality

SOPS ‘81

SIGMETRICS ‘99

USENIX ‘00

USENIX ‘06

FAST ‘07

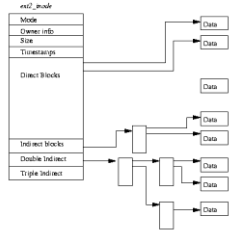
USENIX ‘08

FAST ‘11

HotStorage ‘14

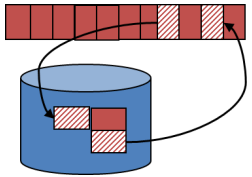
…

Metadata

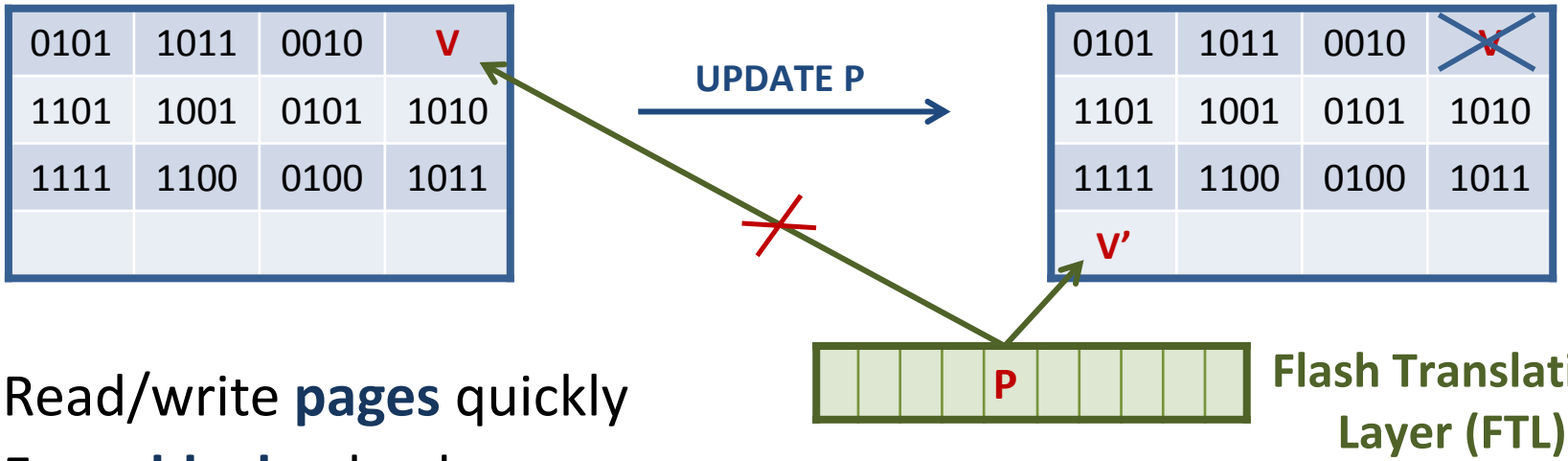



Hard disk

Cache



# Flash Media is Different




- Read/write **pages** quickly
- Erase **blocks** slowly
- **Out of place writes** → Logical page ≠ Physical page
  → Garbage collection
- Limited **lifetime** (≡ number of **erasures**)

# The Streetlight Effect

Let’s look at:

- ? Page sizes
- ? “Temperature” ranges
- ? Logical locality

- ✓ Request sizes
- ✓ Access skew: hot/cold
- ✓ Spatial locality: sequentiality

# Workloads

- University of Massachusetts (36)
- MSR Cambridge (34)
- Microsoft production servers (43)
- Florida International University (9)

UMassTraceRepository

SNIA IOTTA Repository


SyLab Home

- **Duration:** 12 hours – 3 weeks
- **Volume size:** 0.1 – 3200 GB
- **I/O Requests:** 20K – 400M
- **Server categories:** database, development, web, file, mail

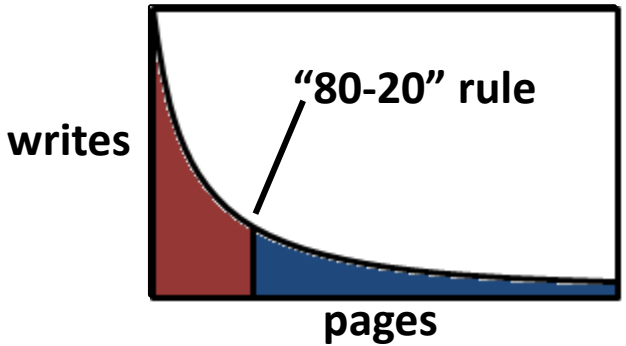
# Access skew

**Previous analyses**

- Working set size
- % of *hot* files/blocks

→ Cache allocation

→ Hot data on outer tracks



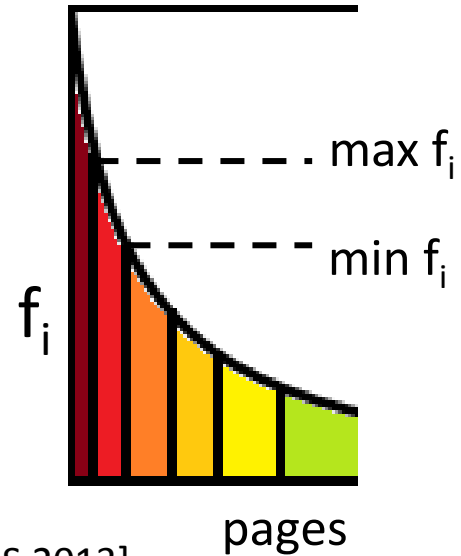

**Flash**

→ Partitions by temperature minimize write amplification

- # temperature ranges
- # required partitions

# Temperature Ranges

- $f_i$ = access frequency of page $i$ (≡ temperature)
- $f_r(p) = \frac{\max f_i}{\min f_i}$ in partition $p$
- $\boldsymbol{f_r = \max f_r(p)}$ in all partitions




✓ $f_r = 1 \rightarrow$ minimal garbage collection [Desnoyers, TOS 2013]

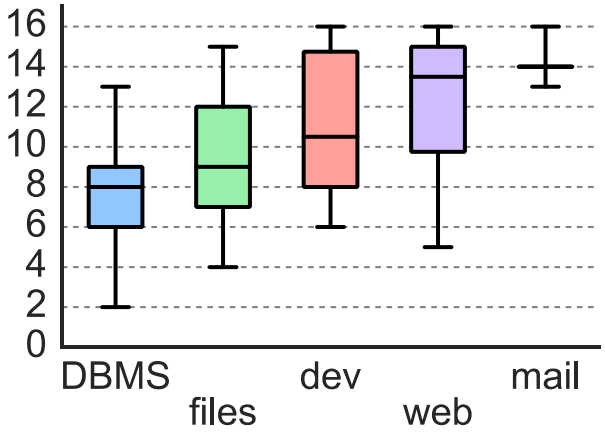
✓ $f_r \leq 2$ is practically sufficient [Stoica and Ailamaki, VLDB 2013]

? How many partitions are needed?

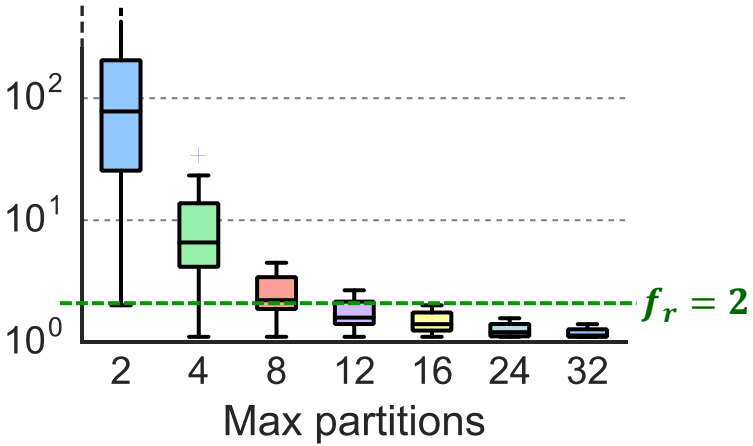
? How bad is it to restrict the number of partitions?

# Findings

# partitions needed for $f_r \leq 2$




$f_r$ when #partitions $\leq N$




→ "Hot" and "cold" are not enough!

? How bad are $f_r \leq 5$ and $f_r \leq 77$?

? How to identify $f_i$ online?

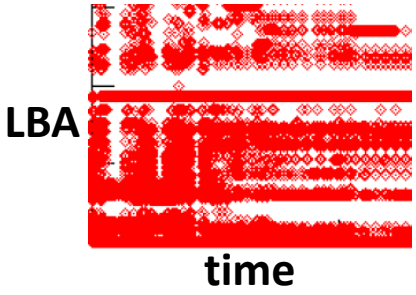
# Access Locality

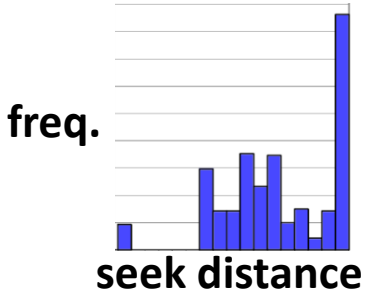
**Previous analyses**

- Sequentiality
- Seek distance

→ Avoid cache pollution

→ Reorder disk I/Os




**Flash**

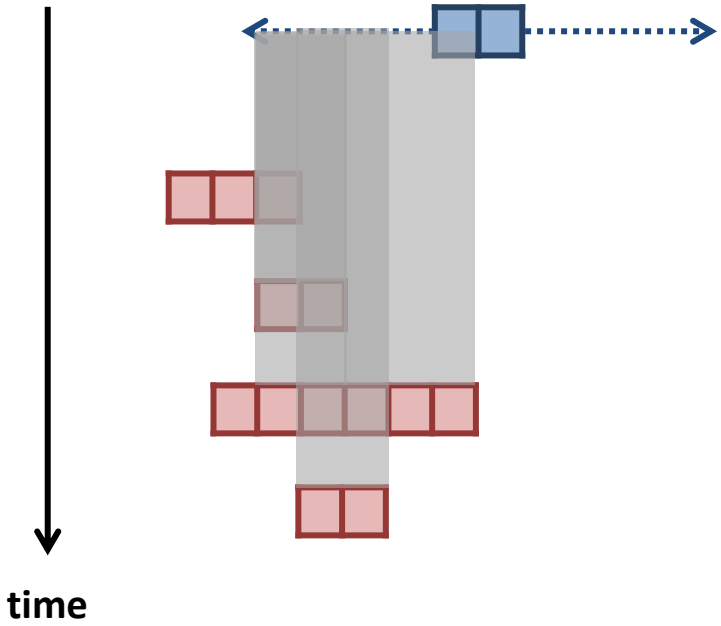
→ Delay RAID parity updates

→ Delay block-merge garbage collections

? What is the probability that **a nearby page** will be written **soon**?

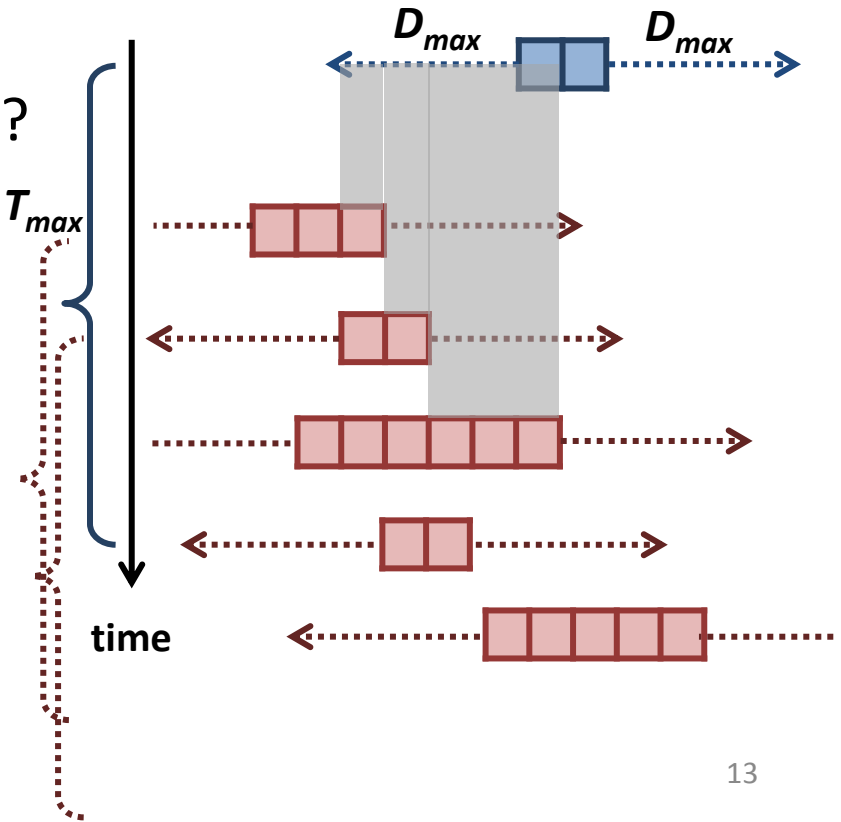
# Spatial Logical Locality

- What is the probability that **a nearby page** will be written **soon**?
- $P_{D,T}$: a page within distance $D$ will be written within time $T$

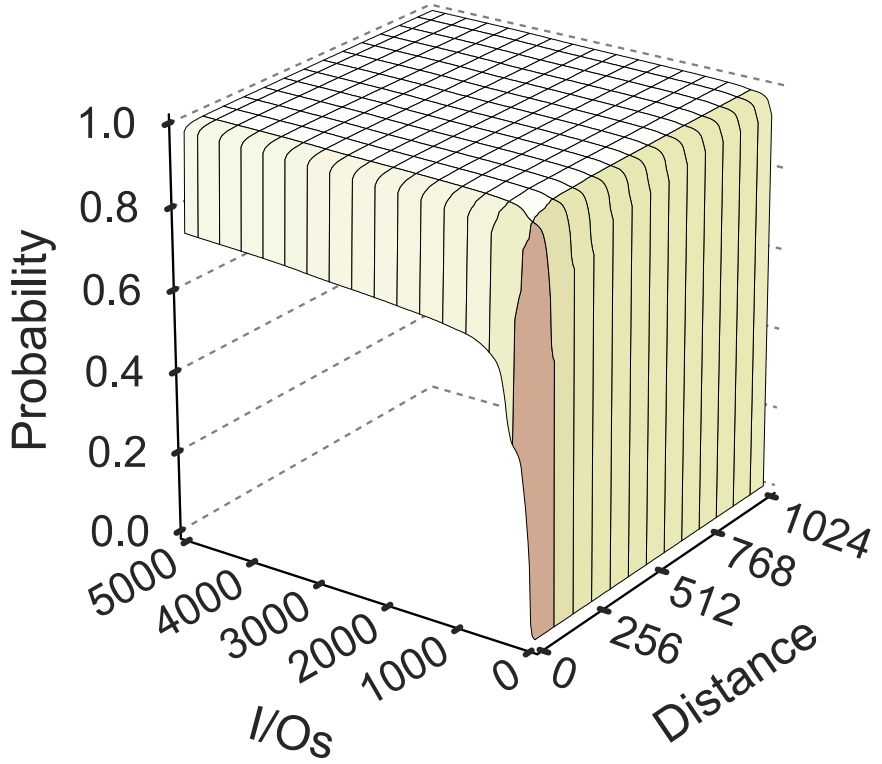


# Spatial Logical Locality

- What is the probability that **a nearby page** will be written **soon**?
- $P_{D,T}$: a page within distance $D$ will be written within time $T$
- CDF calculated with sliding window of $D_{max} = 1024\,, T_{max} = 5000$

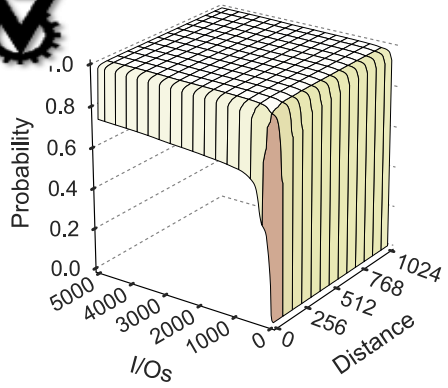


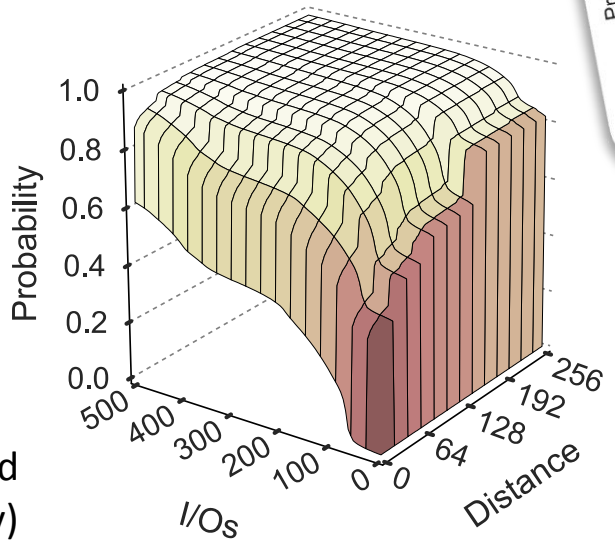
# Findings

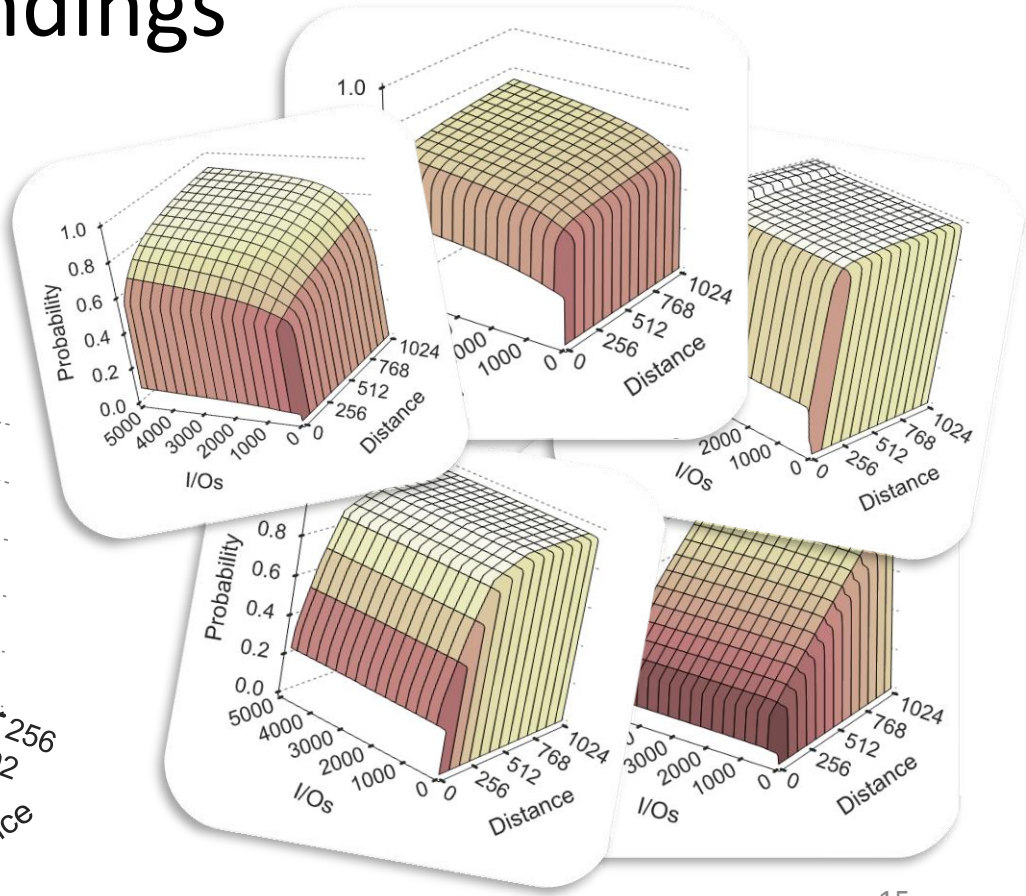



'Casa' workload
(FIU repository)

# Findings

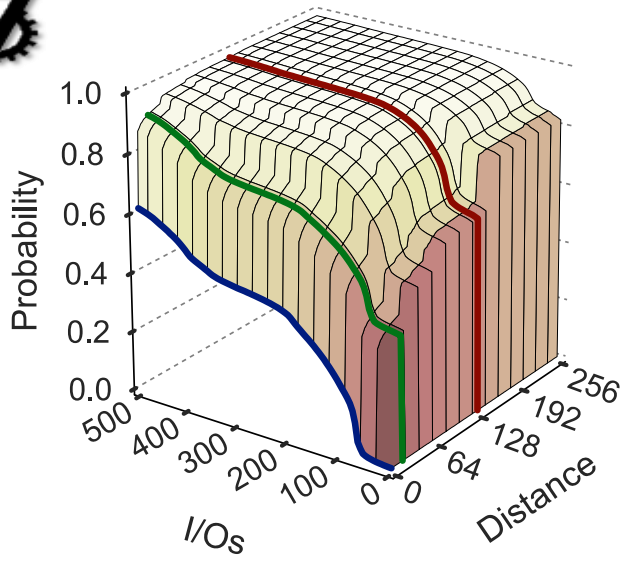


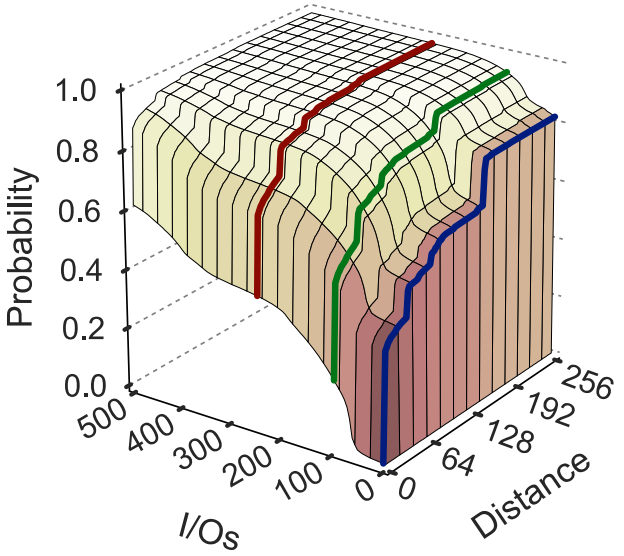




'Casa' workload
(FIU repository)

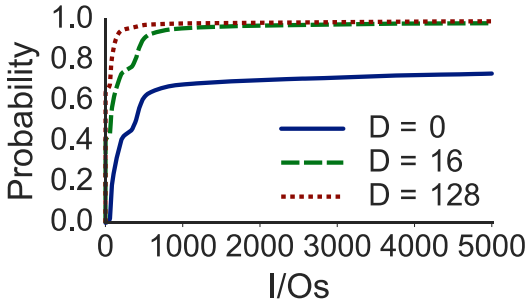
# Findings

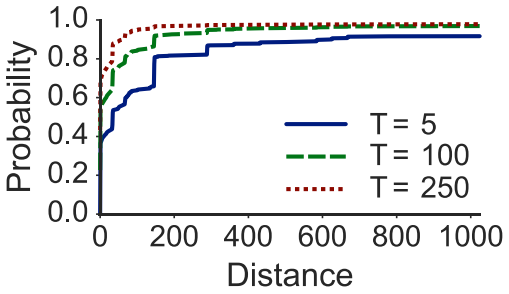



→ Don't wait longer, look further

? How to aggregate the results?



# Conclusions

✓ Temperature ranges, logical locality, (page sizes)

→ Lifetime, write ratio, reads, combinations, correlations…

? What about the way workloads are collected?

? Is there always a streetlight?

? Is that bad?