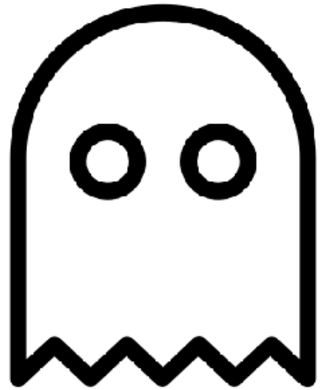
Spooky

Granulating LSM-Tree Compactions Correctly

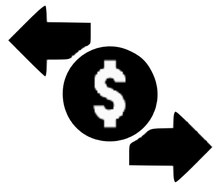


Niv Dayan*
University of Toronto

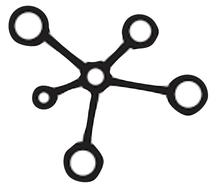
Tamar Weiss, Shmuel Dashevsky, Michael Pan,
Edward Bortnikov, Moshe Twitto *Pliops*

VLDB 2022

OLTP



Graphs



Time series

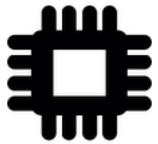
Analytics

data

data

random writes

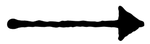
1 2 3 4 5 6



buffer

merge

data



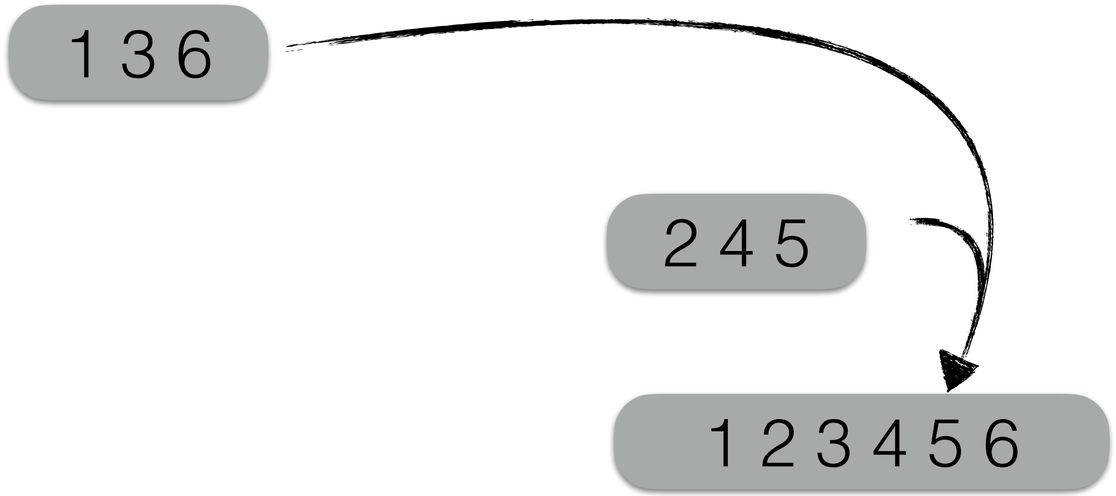
1 3 6

2 4 5

1 2 3 4 5 6

only sequential writes

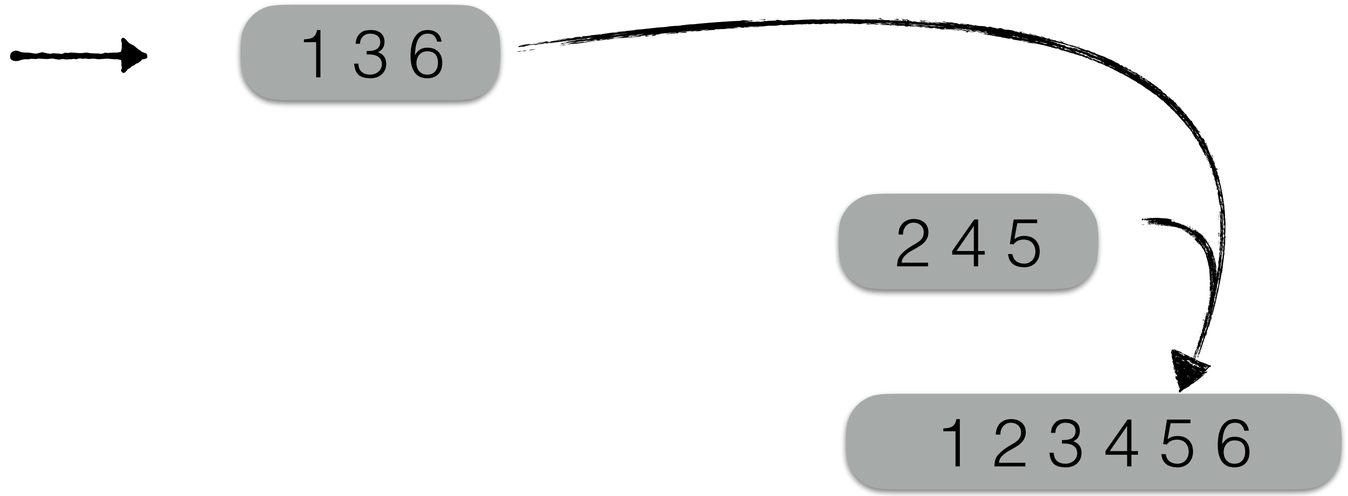
data →

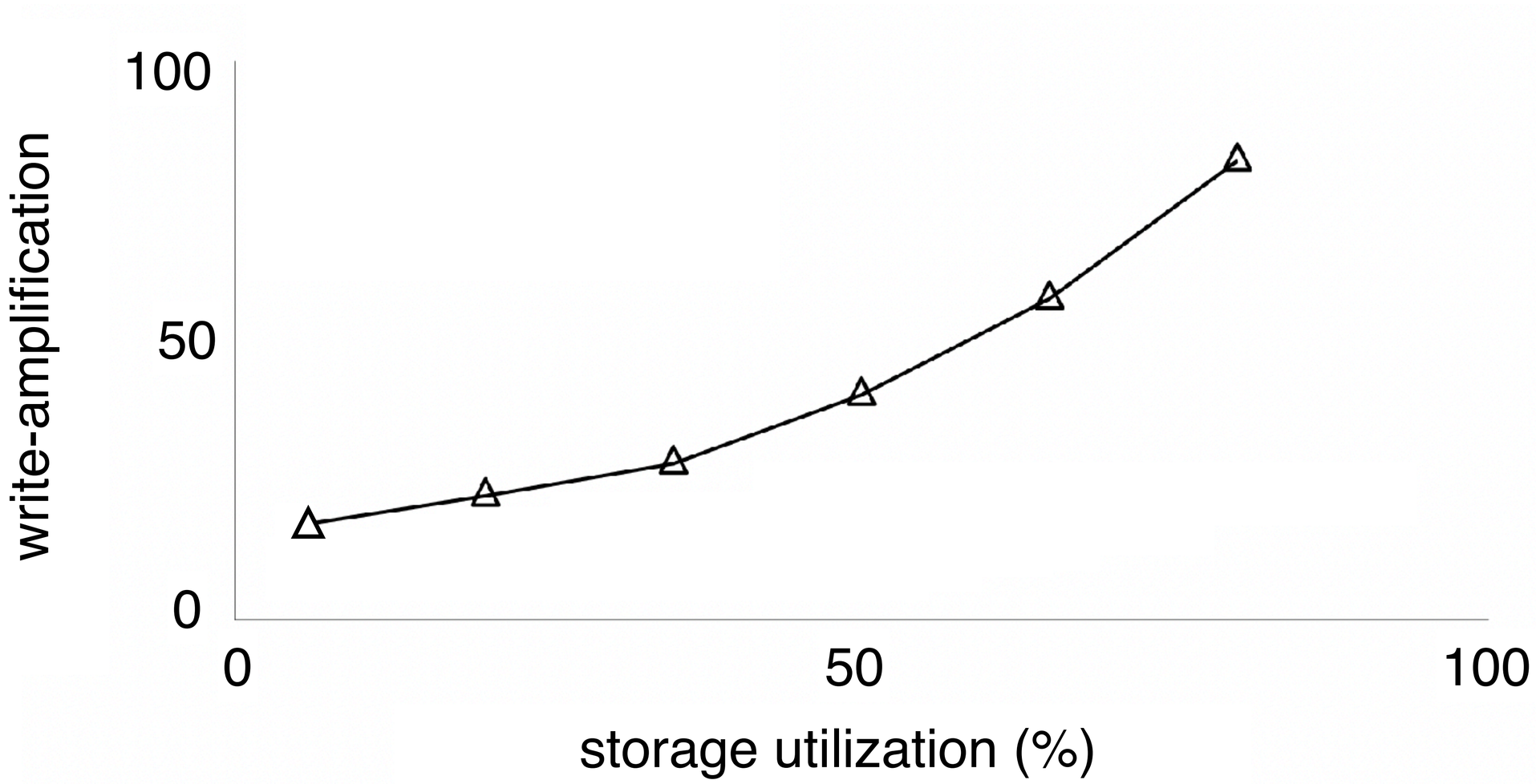


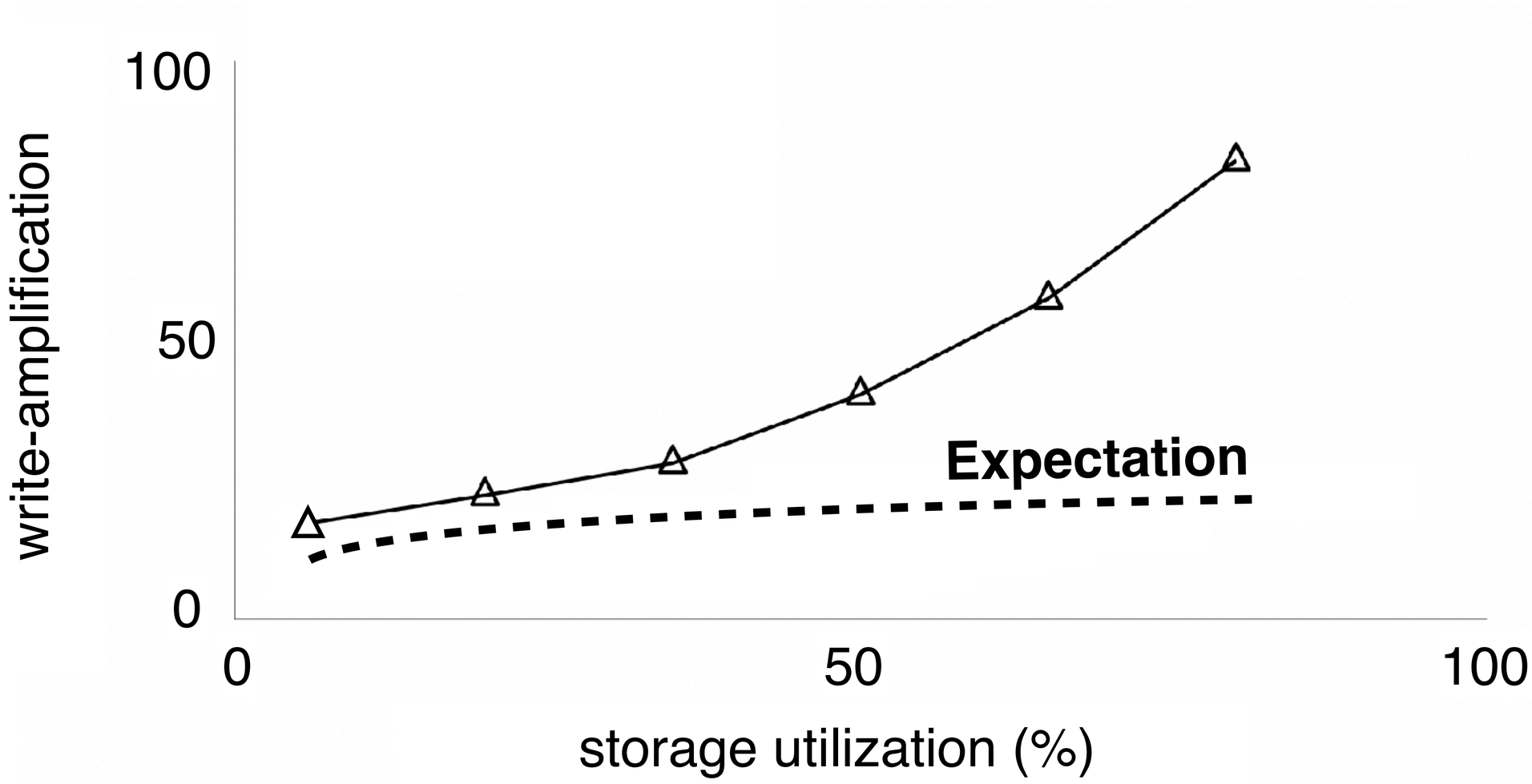
only sequential writes

write-amplification

data




100
50
0
write-amplification
0
50
100
storage utilization (%)


100
50
0
write-amplification
Expectation
0
50
100
storage utilization (%)

LSM-Tree

key-value pairs

↓

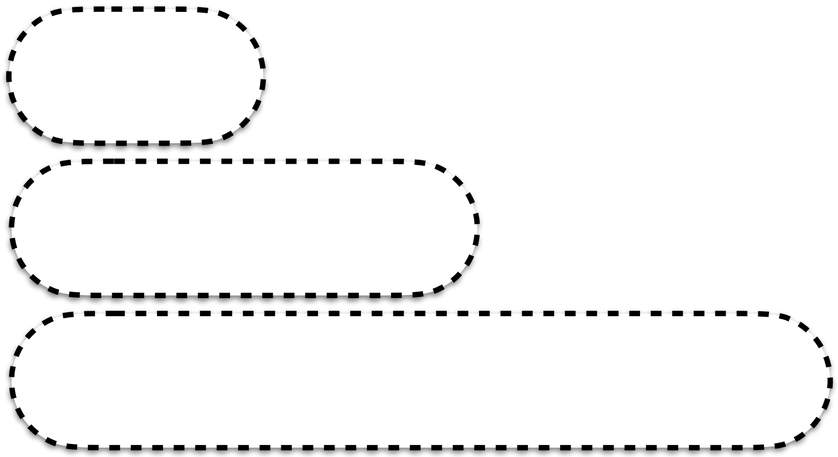
buffer

LSM-Tree

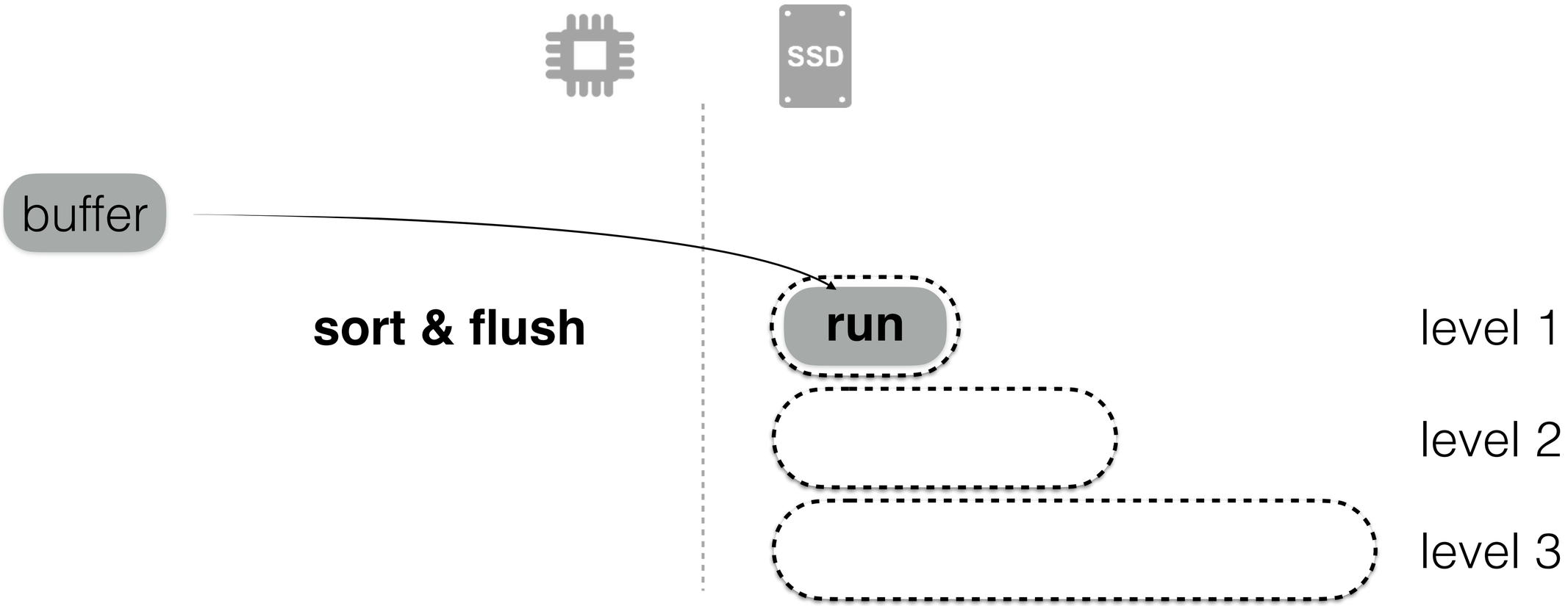
key-value pairs

↓

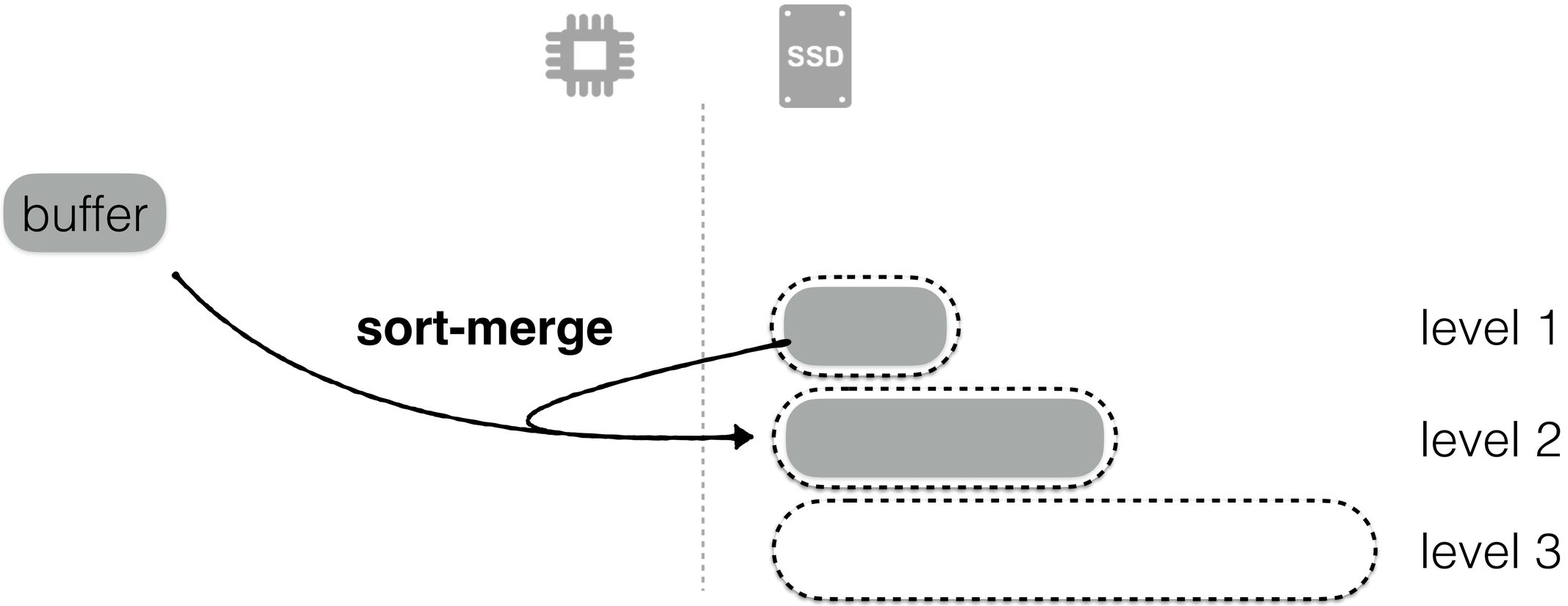
buffer

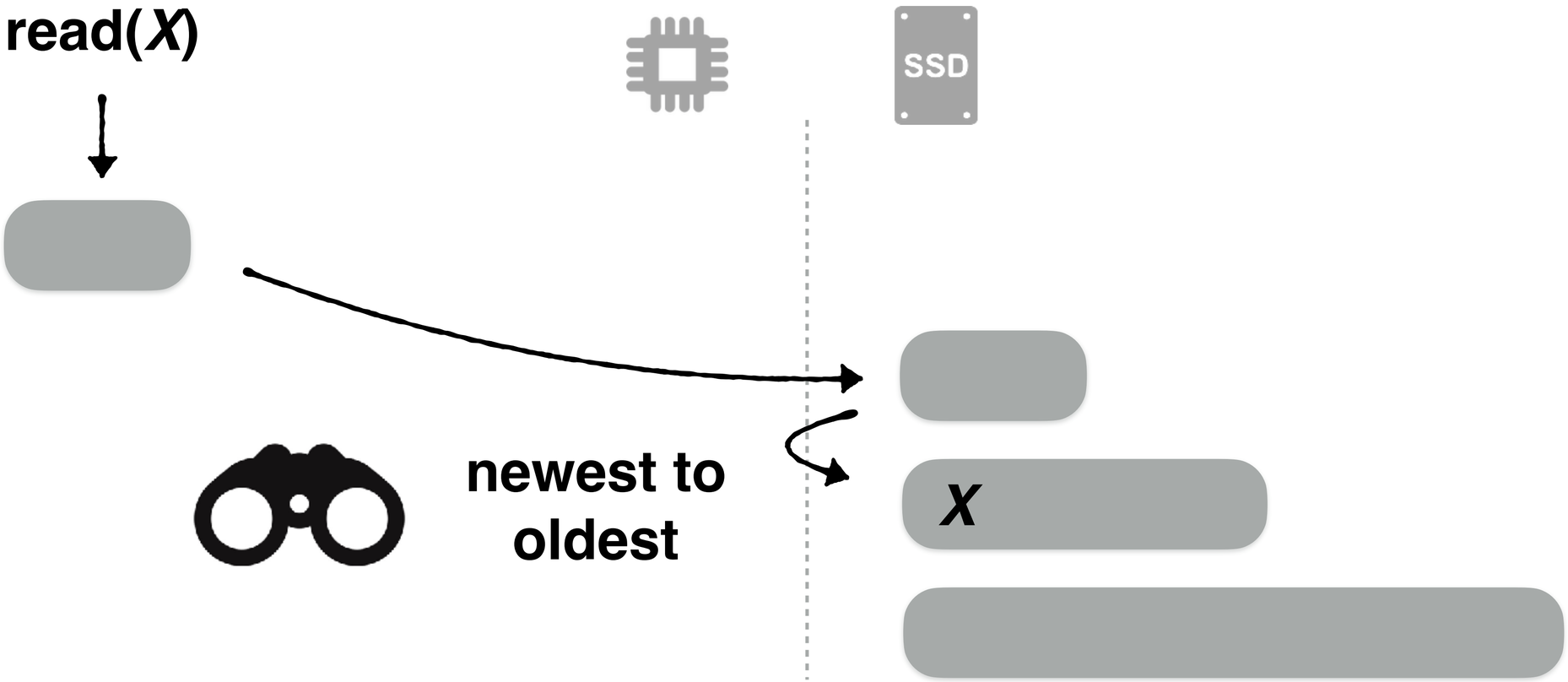


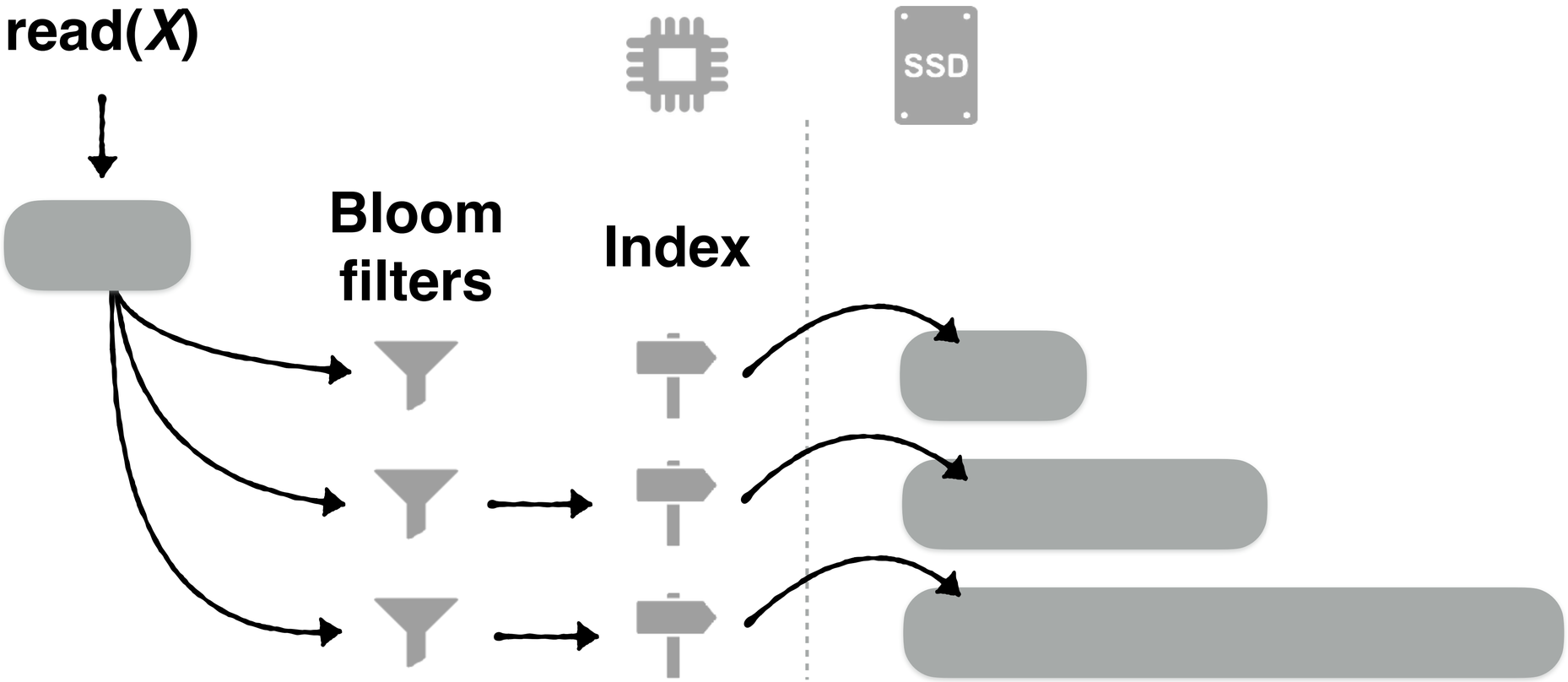
LSM-Tree



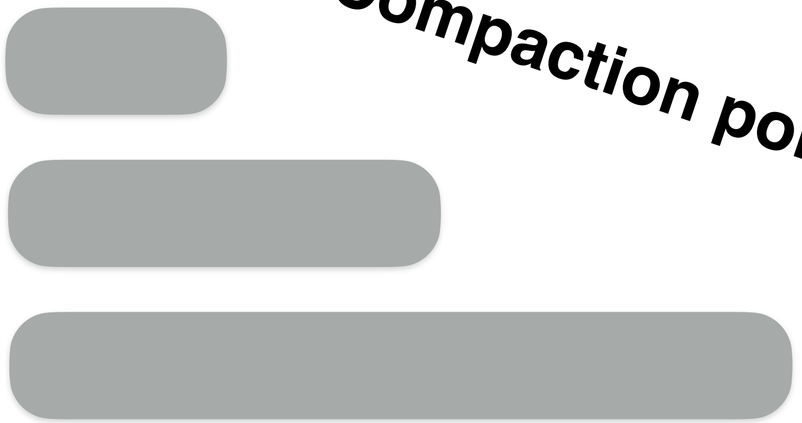
LSM-Tree




read(X)
SSD
newest to
oldest
X


read(X)
SSD
Bloom filters
Index


SSD


Compaction policy

Compaction policy

Write-optimized

Eagerness

Read-optimized

Compaction policy

Write-optimized

Eagerness

Read-optimized

3 1 **Tiering**

2, 7 1, 5

1, 3, 6, 7 2, 3, 5, 9

2, 5 **Leveling**

1, 3, 6, 8

1, 2, 3, 5, 6, 8, 9

Compaction policy

Write-optimized

Eagerness

Read-optimized

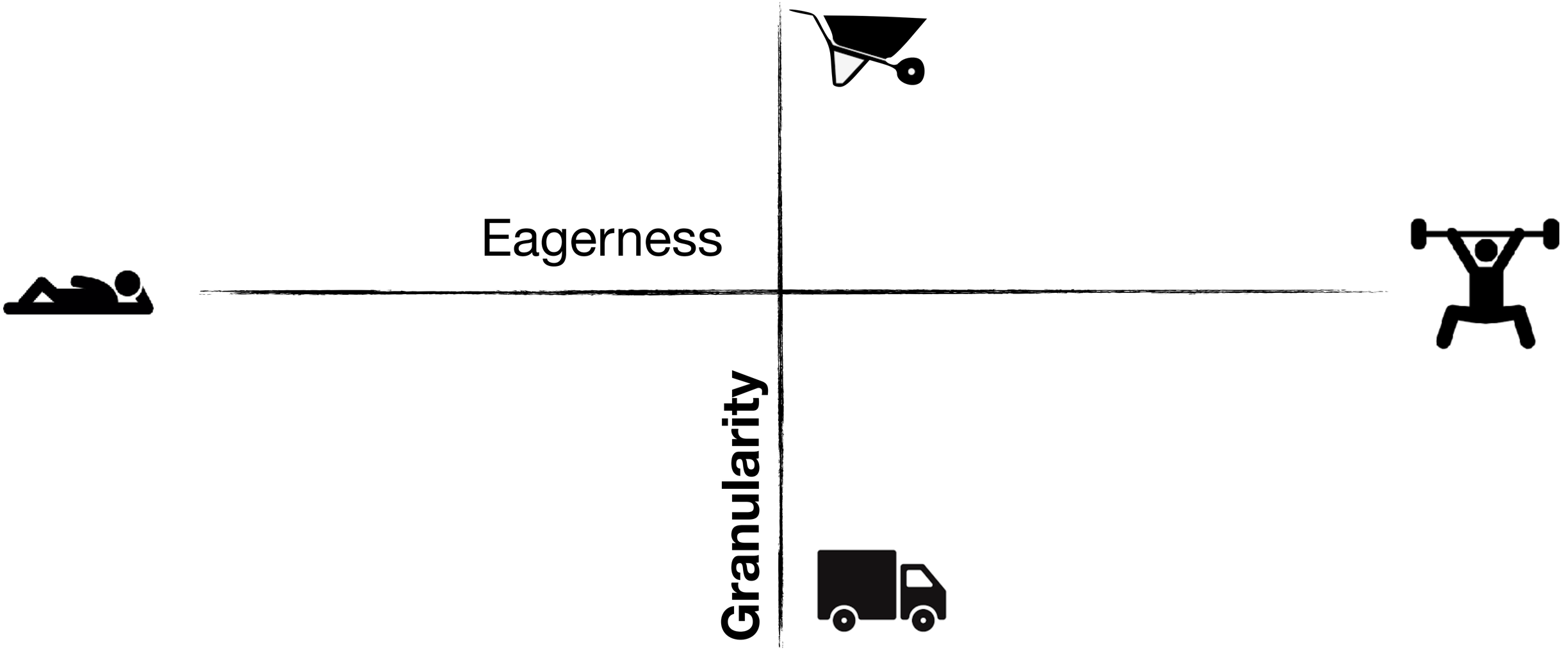
2, 5

Leveling

1, 3, 6, 8

1, 2, 3, 5, 6, 8, 9

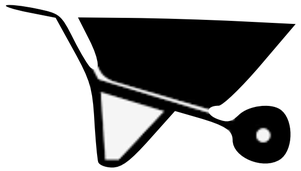
Compaction policy



Compaction Granularity

Full Merge

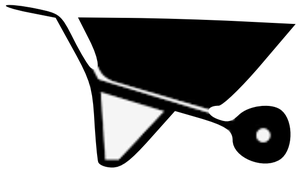
Partial Merge



Compaction Granularity

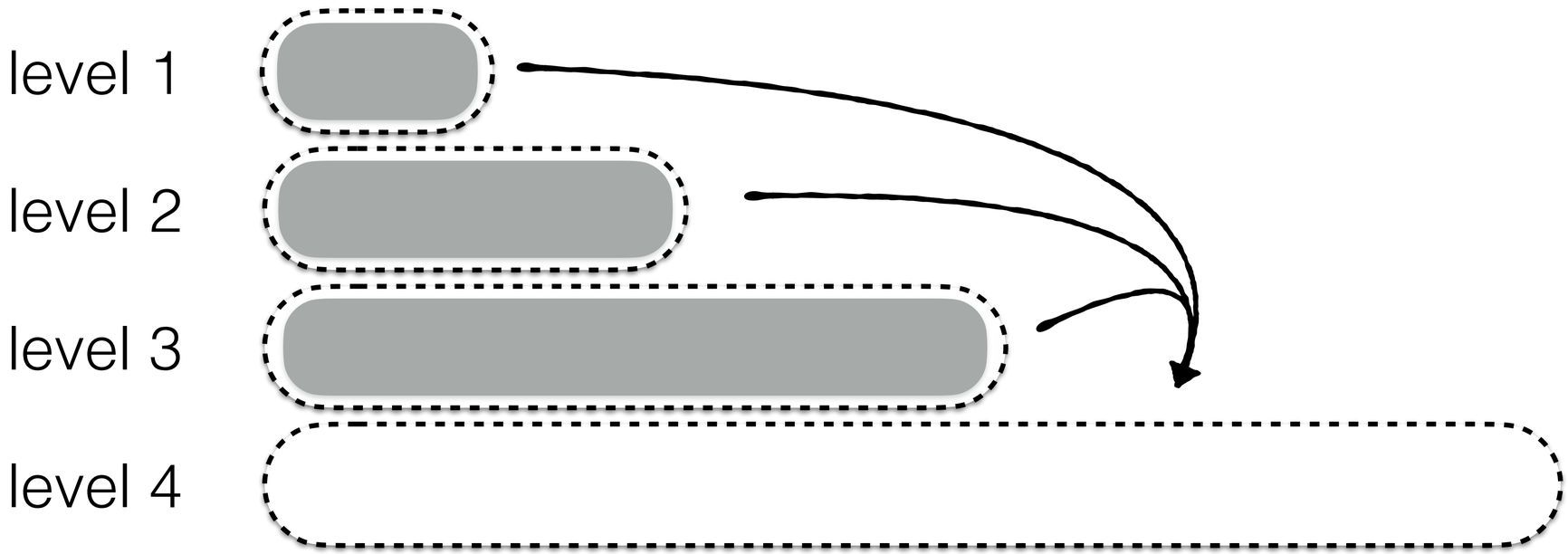
Full Merge

Partial Merge



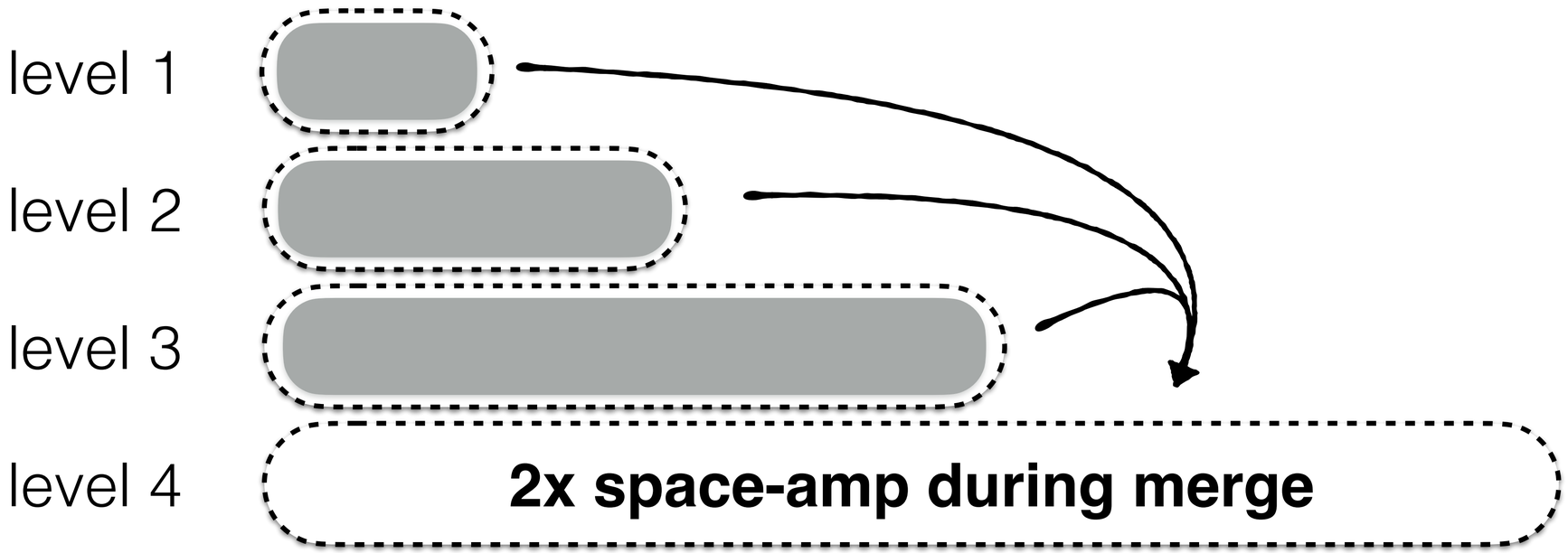
Full Merge

Merge consecutive full levels into first non-full level



Full Merge

Merge consecutive full levels into first non-full level

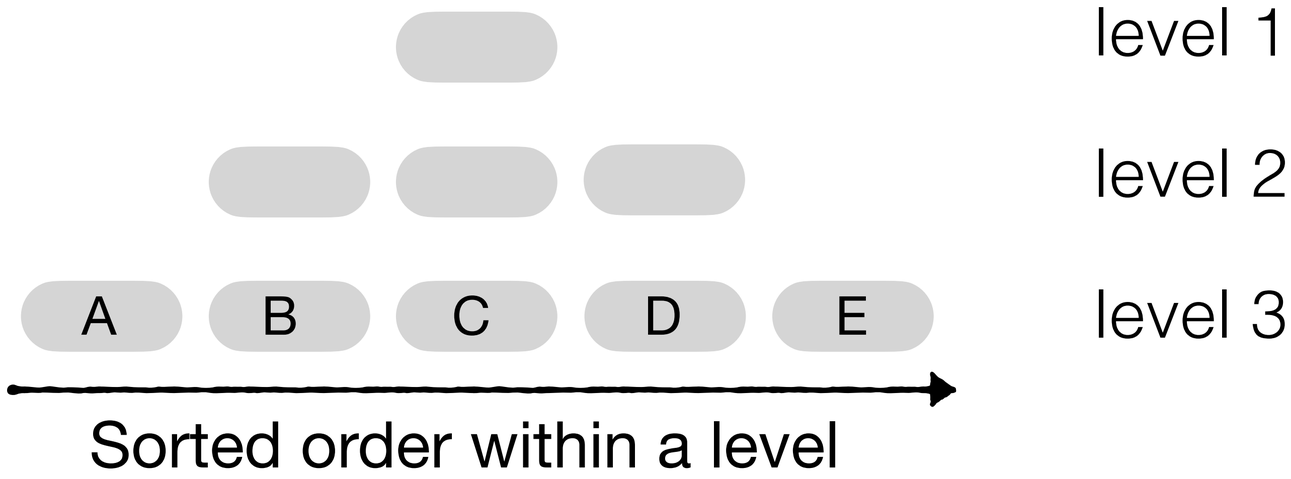


Full Merge

Partial Merge

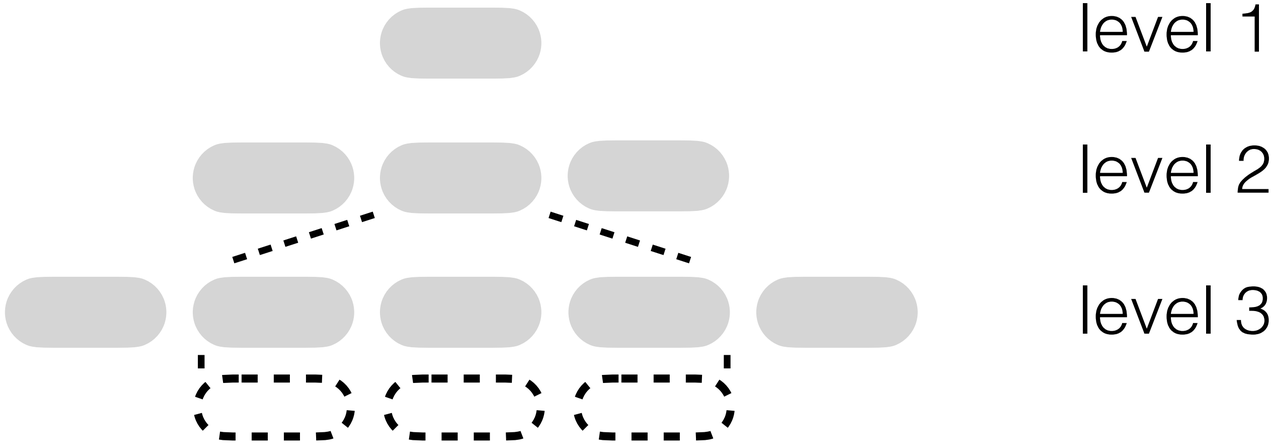
Partial Merge

1. split runs into many files (SSTs) in each level



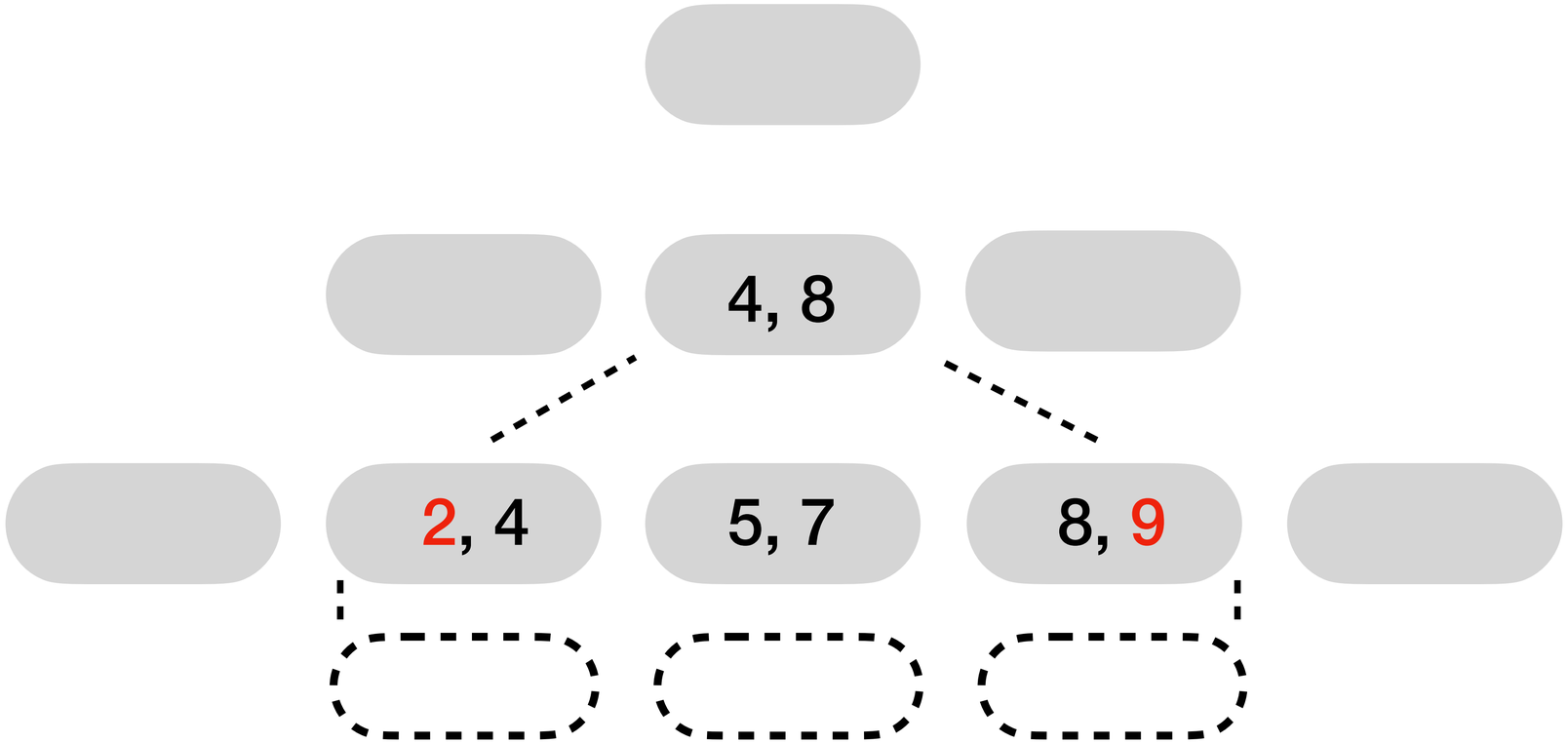
Partial Merge

1. split runs into many files (SSTs) in each level
2. When a level is full, pick SST with smallest intersection into next level



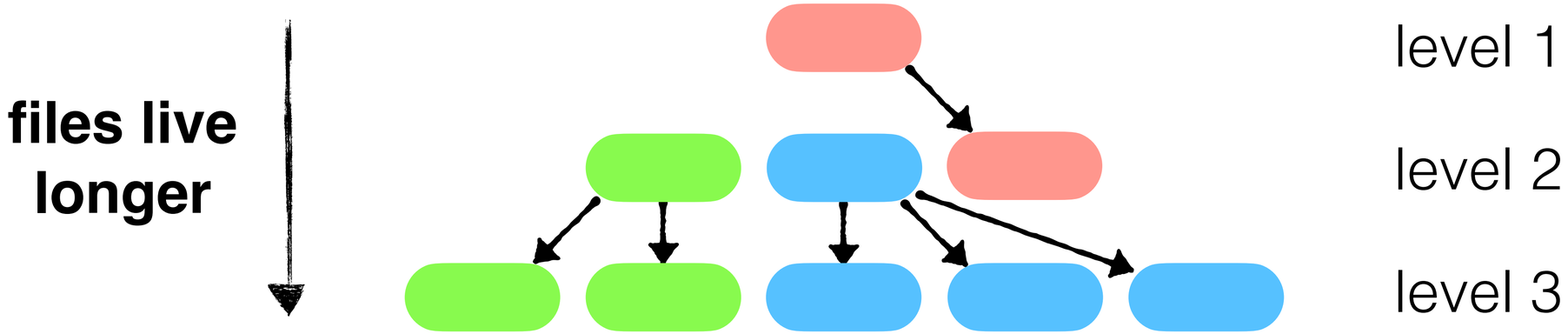
Partial Merge

Problem 1: non-intersecting entries increase write-amplification



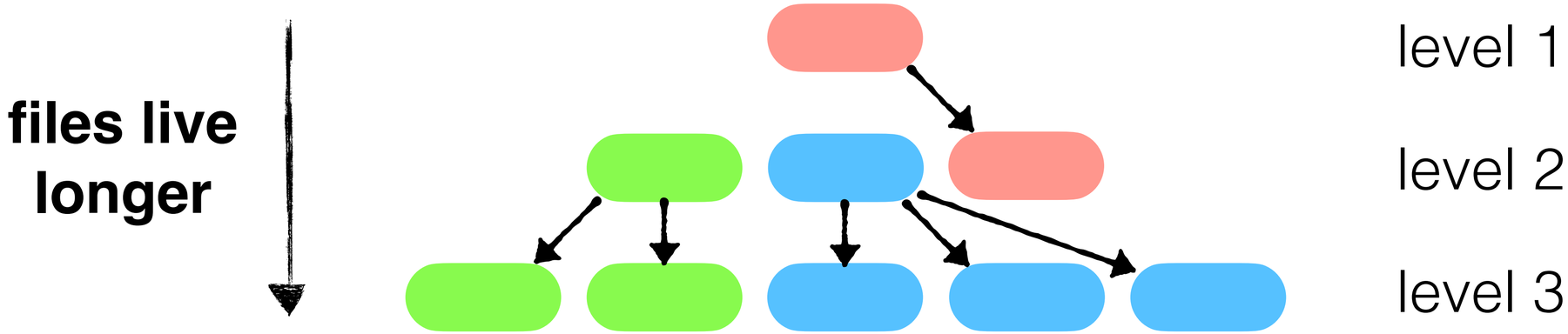
Partial Merge

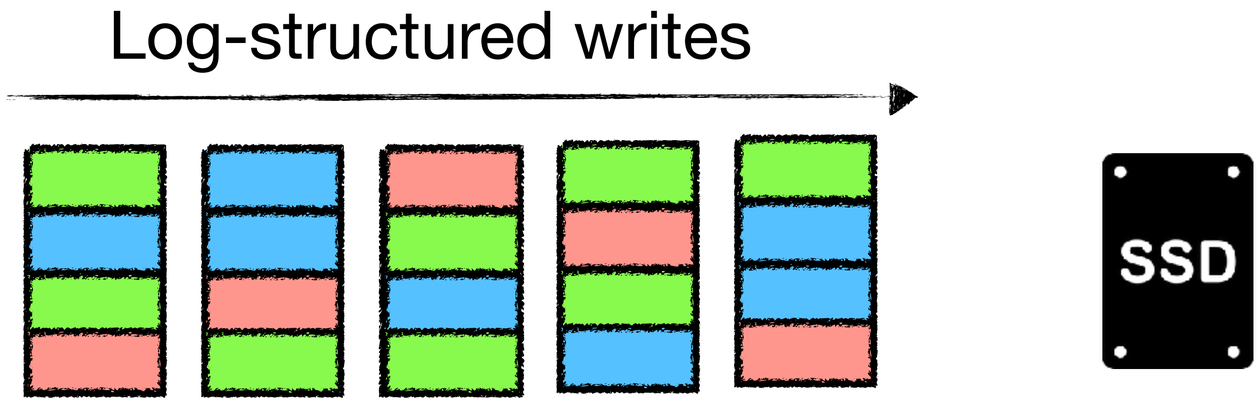
Problem 2: many small simultaneous compactions



Partial Merge

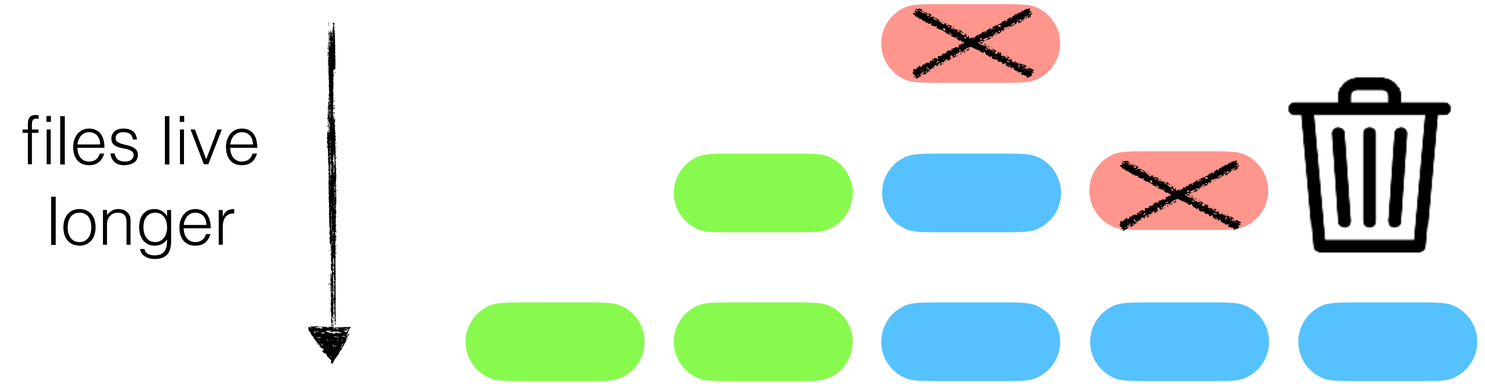
Problem 2: many small simultaneous compactions

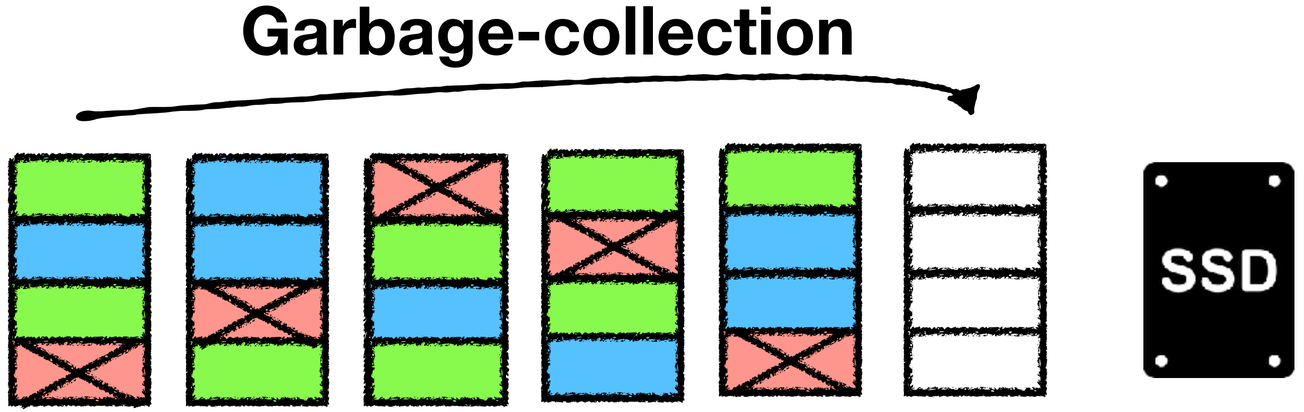




Partial Merge

Problem 2: many small simultaneous compactions

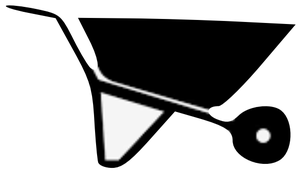




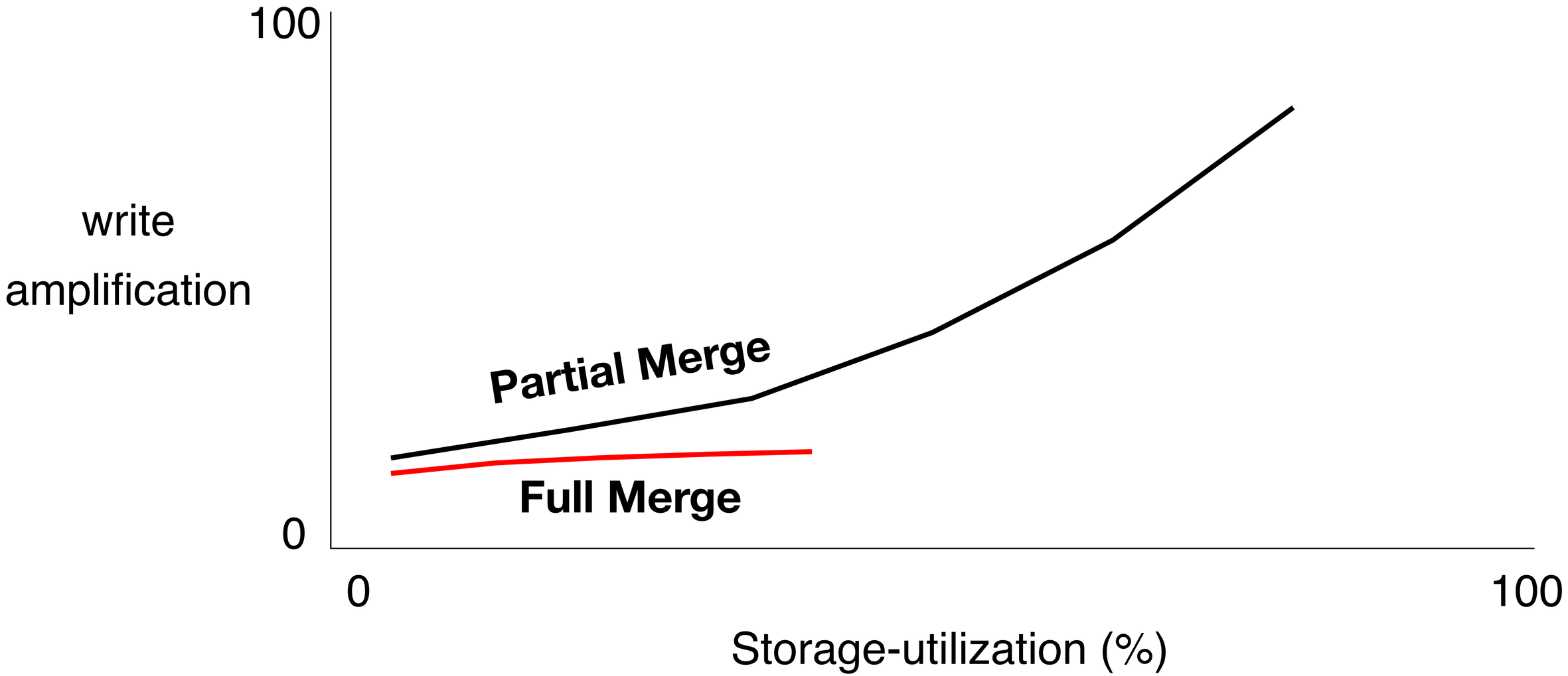
Full Merge

space amplification

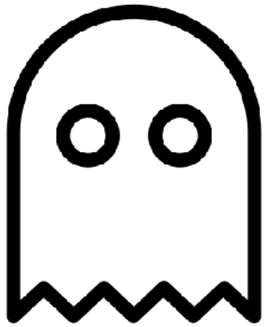
Partial Merge



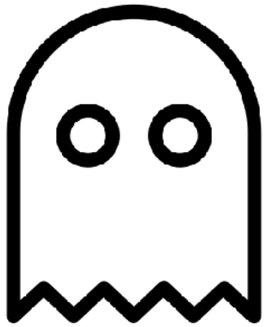
write amplification


100
write
amplification
Partial Merge
Full Merge
0
0
100
Storage-utilization (%)

Spooky



Spooky: **p**artitioned c**o**mpaction f**o**r **k**e**y**-value stores

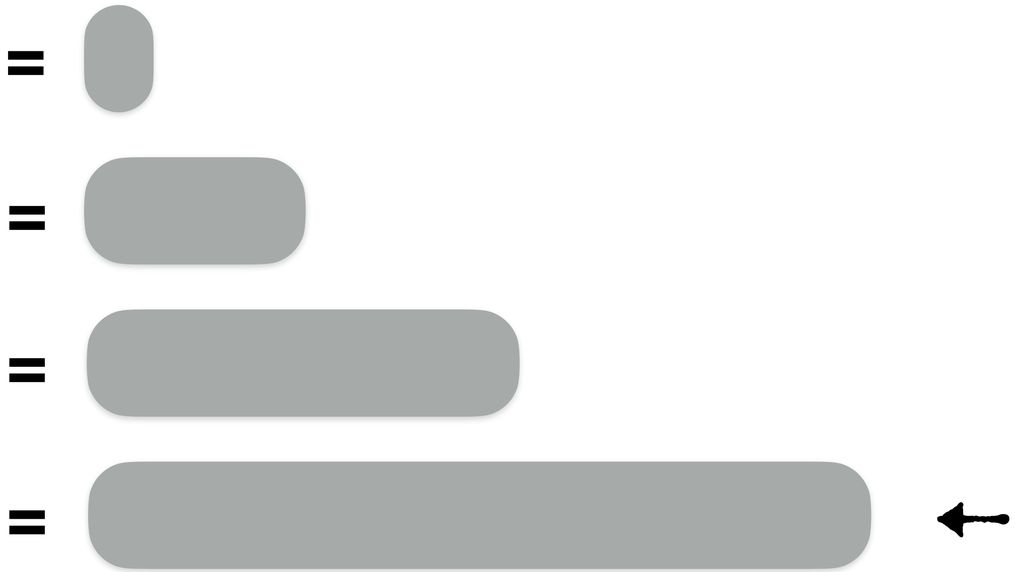


Spooky's intuition

transient
space amplification

Spooky's intuition

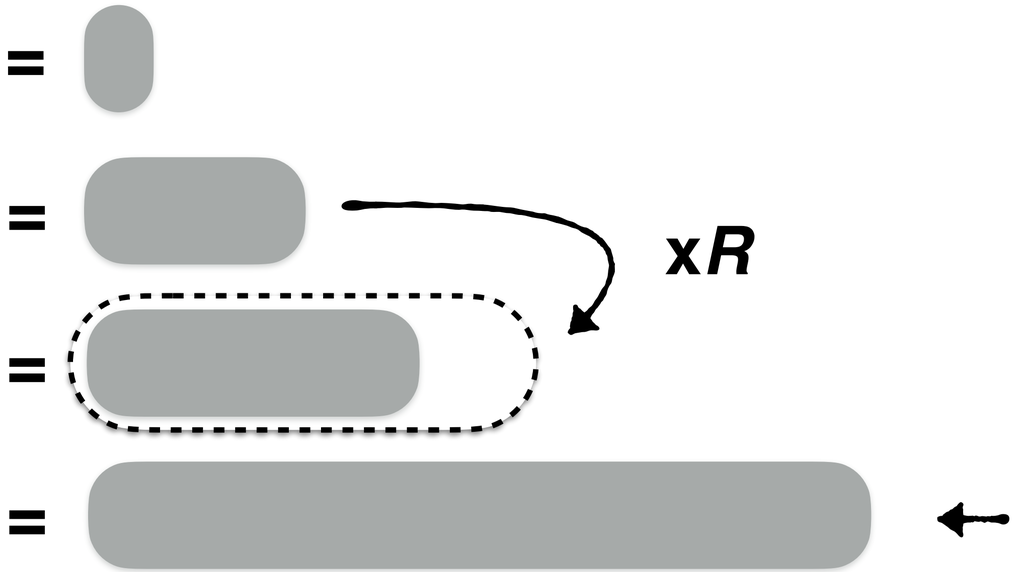
write amplification



transient
space amplification

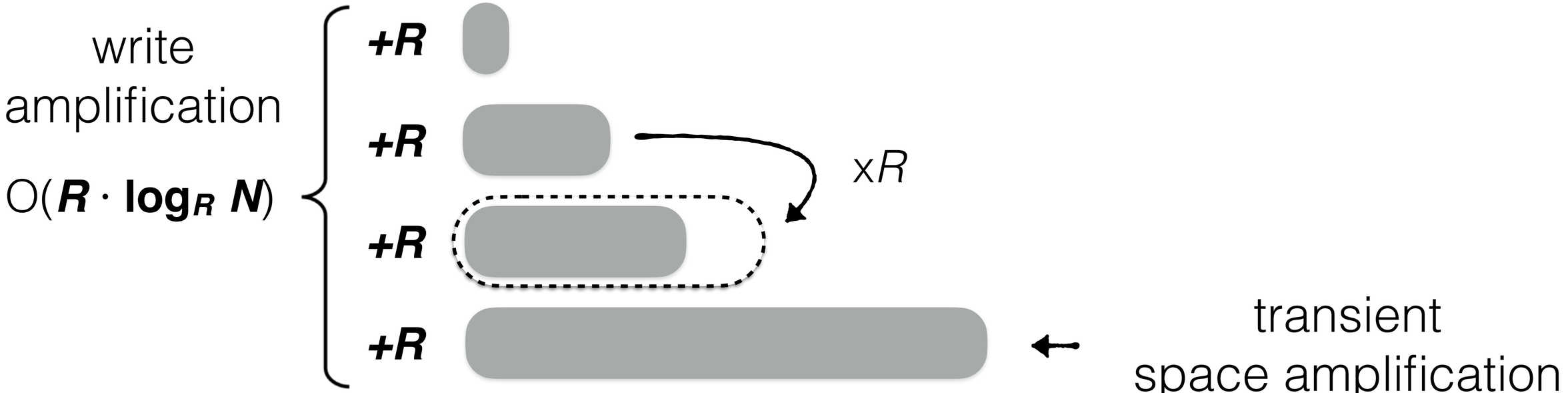
Spooky's intuition

write amplification



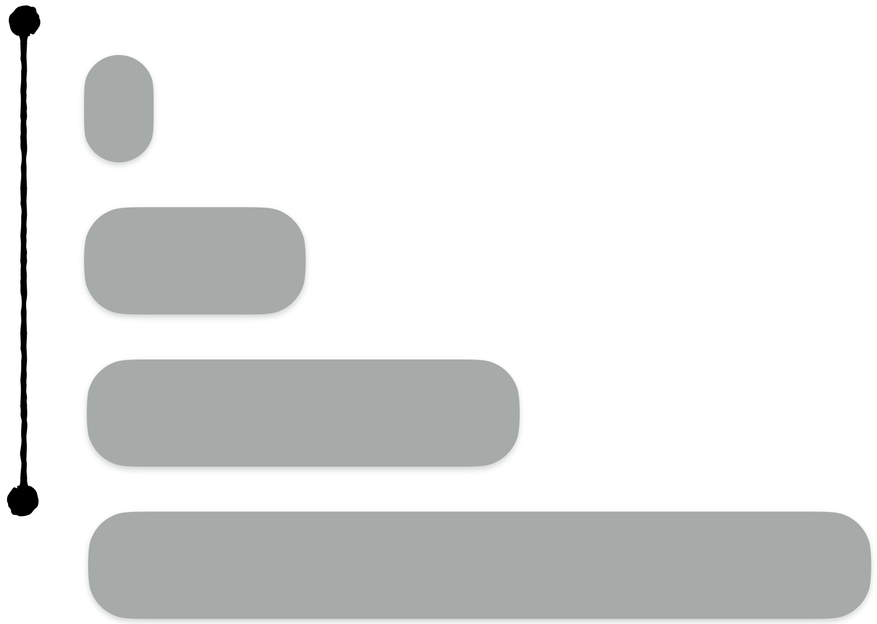
transient
space amplification

Spooky's intuition



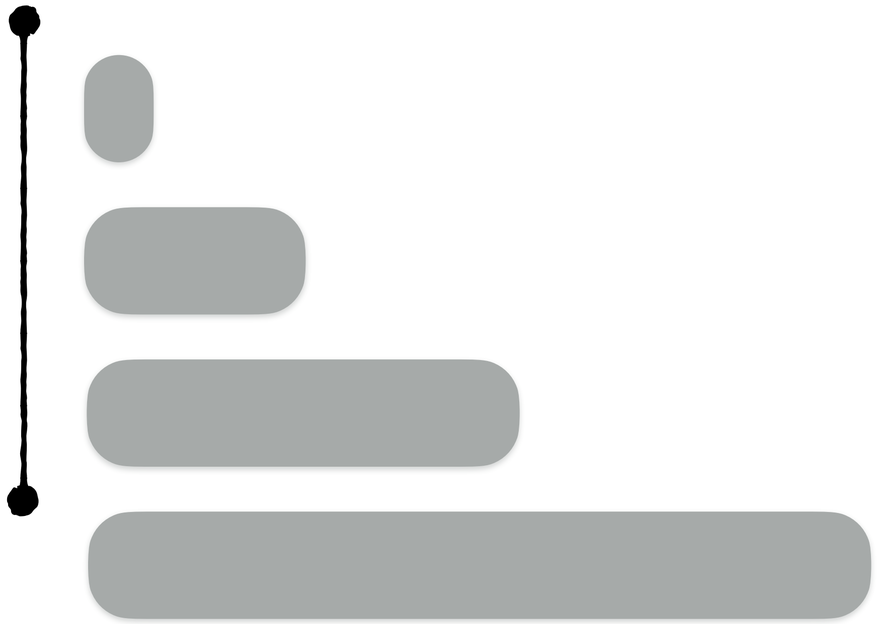
Spooky’s intuition

Full merge
(write optimized)



Spooky's intuition

Full merge
(write optimized)



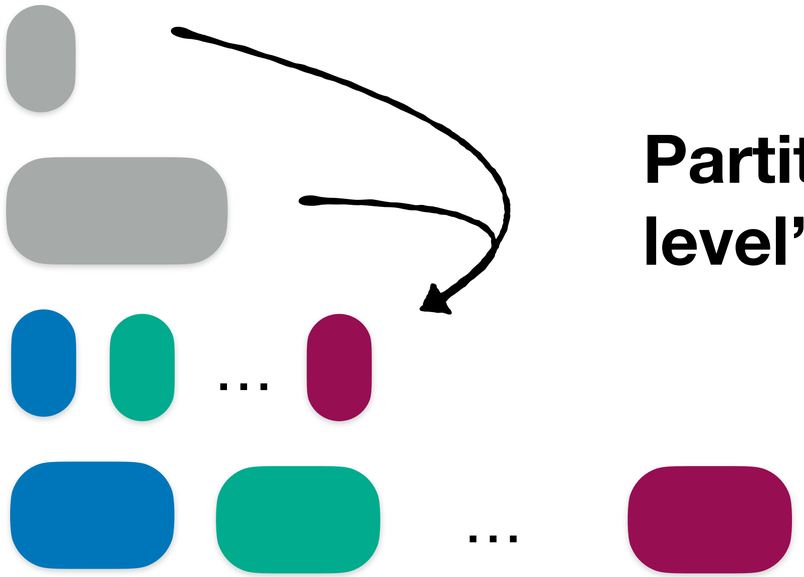
Partial Merge Variant
(space optimized)

Spooky's intuition

Full merge
(write optimized)

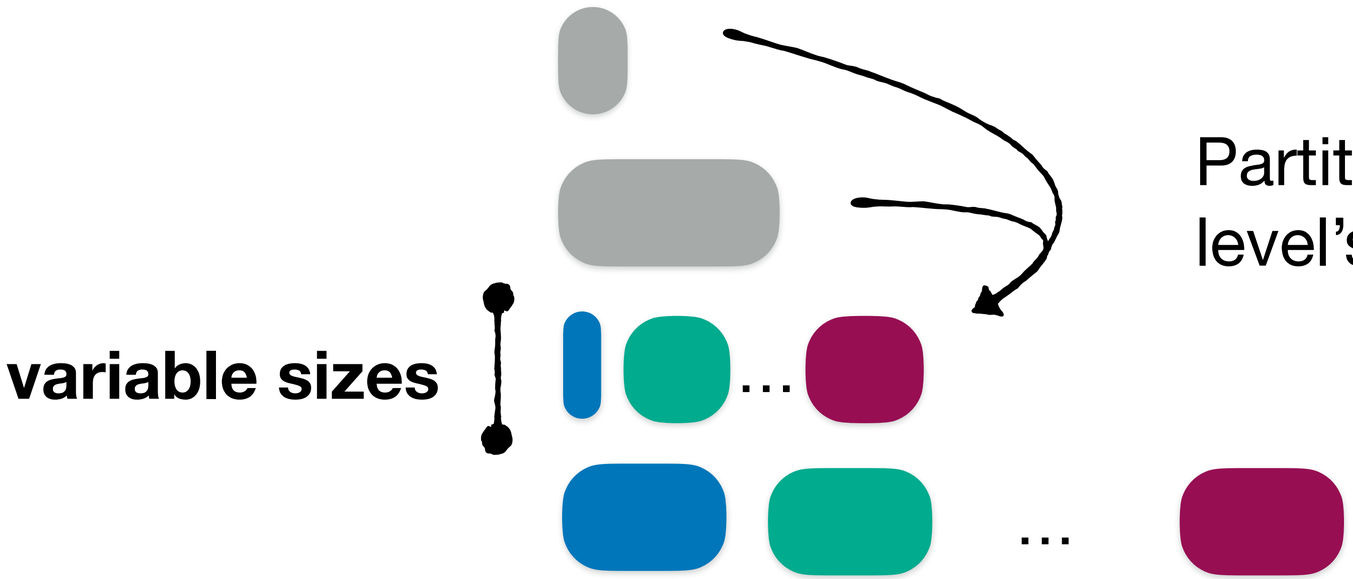
Partial Merge Variant
(space optimized)

Spooky

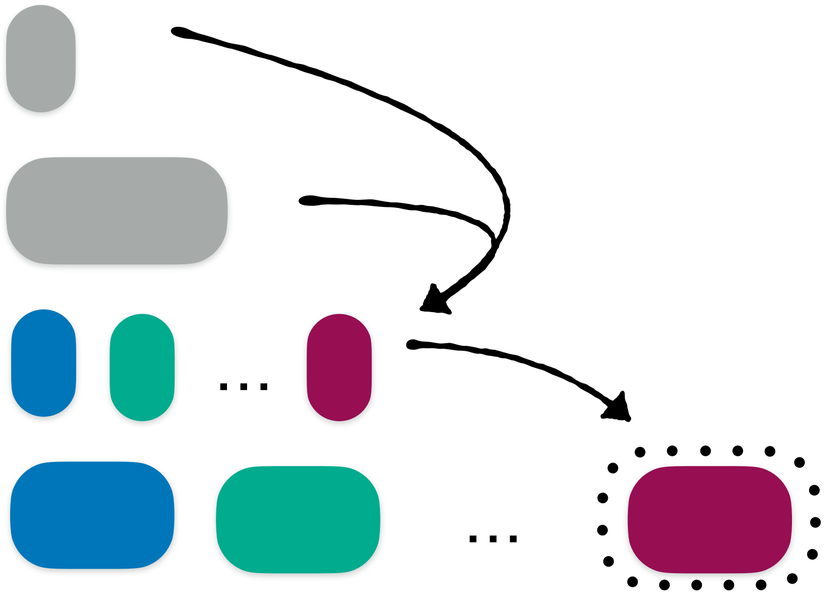


Partition based on largest level's file boundaries.

Spooky


variable sizes
Partition based on largest level's file boundaries.

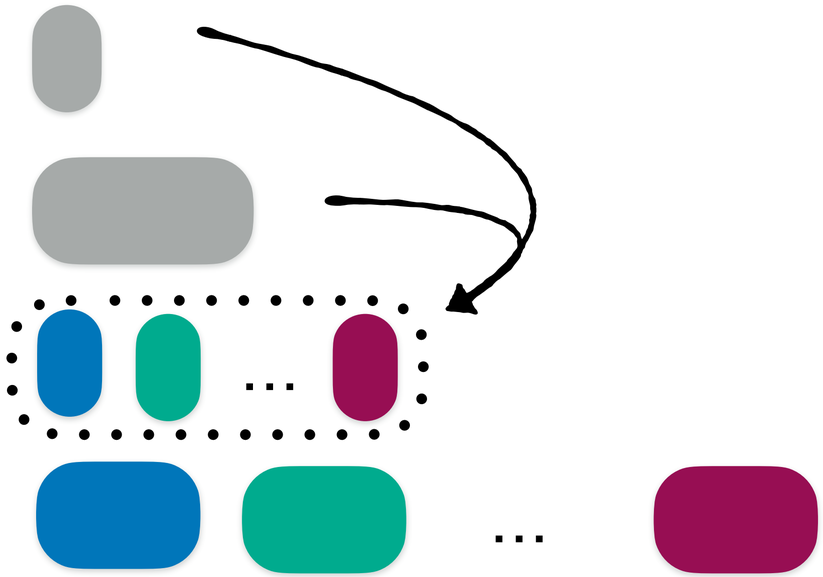
Spooky



Merge one
partition at a time

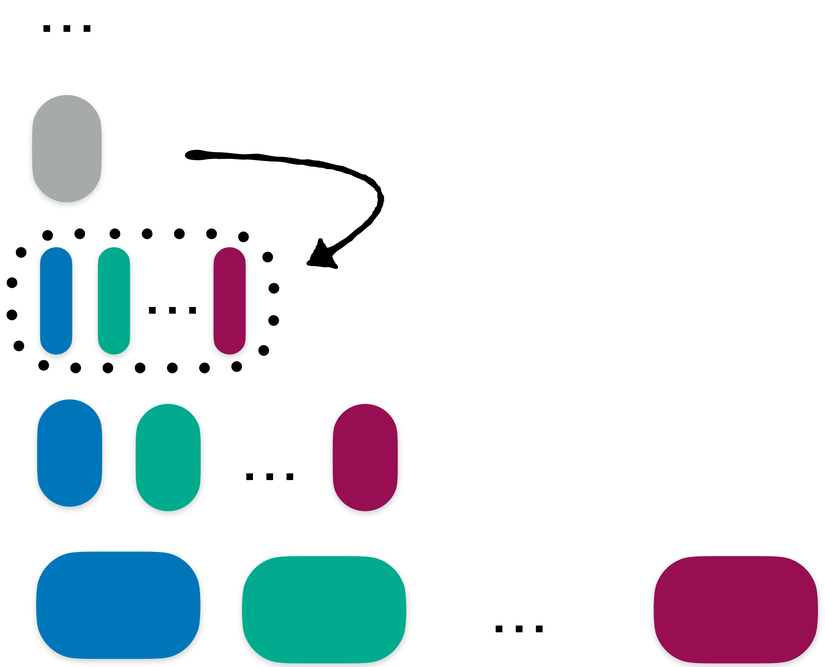
Spooky

transient
space-amp: ***N/R***

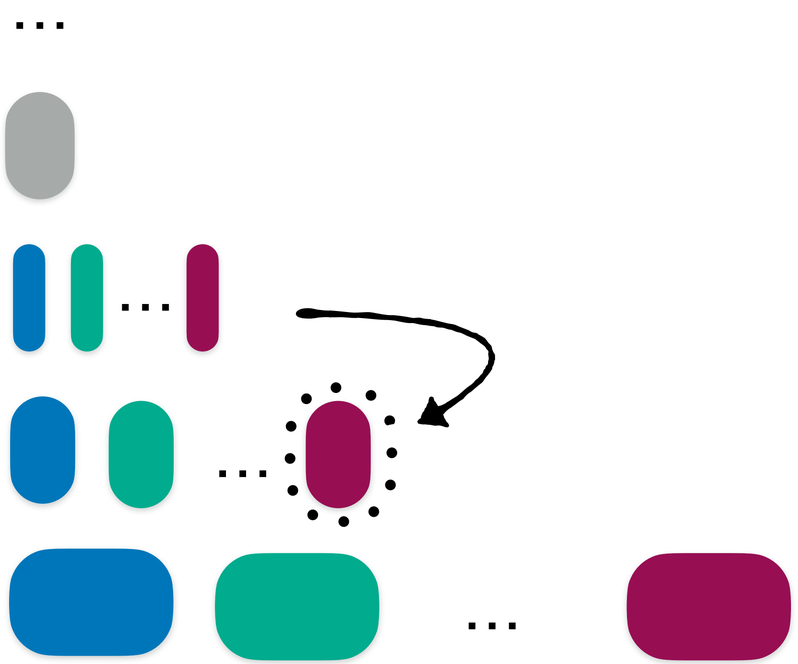


Spooky

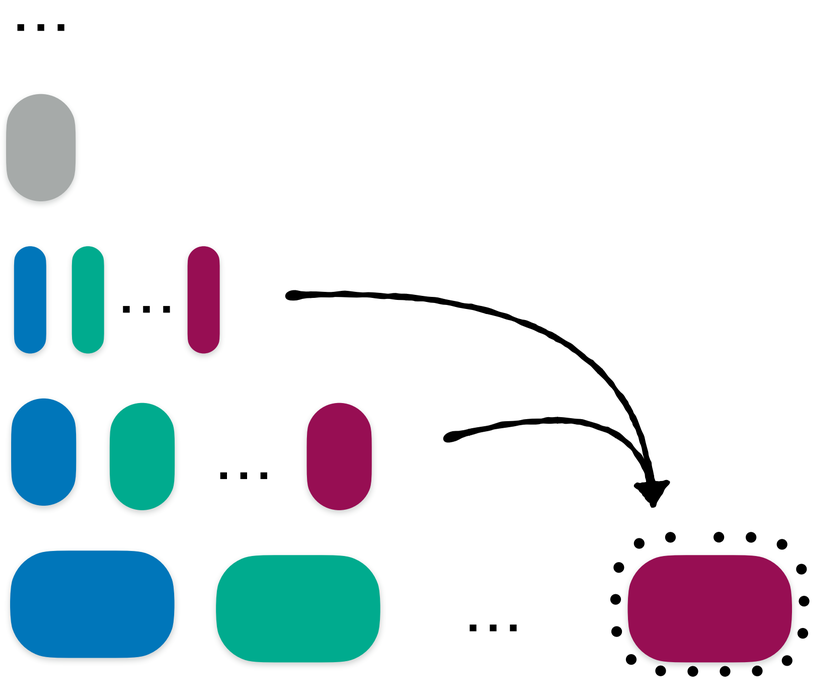
transient
space-amp: N/R^2



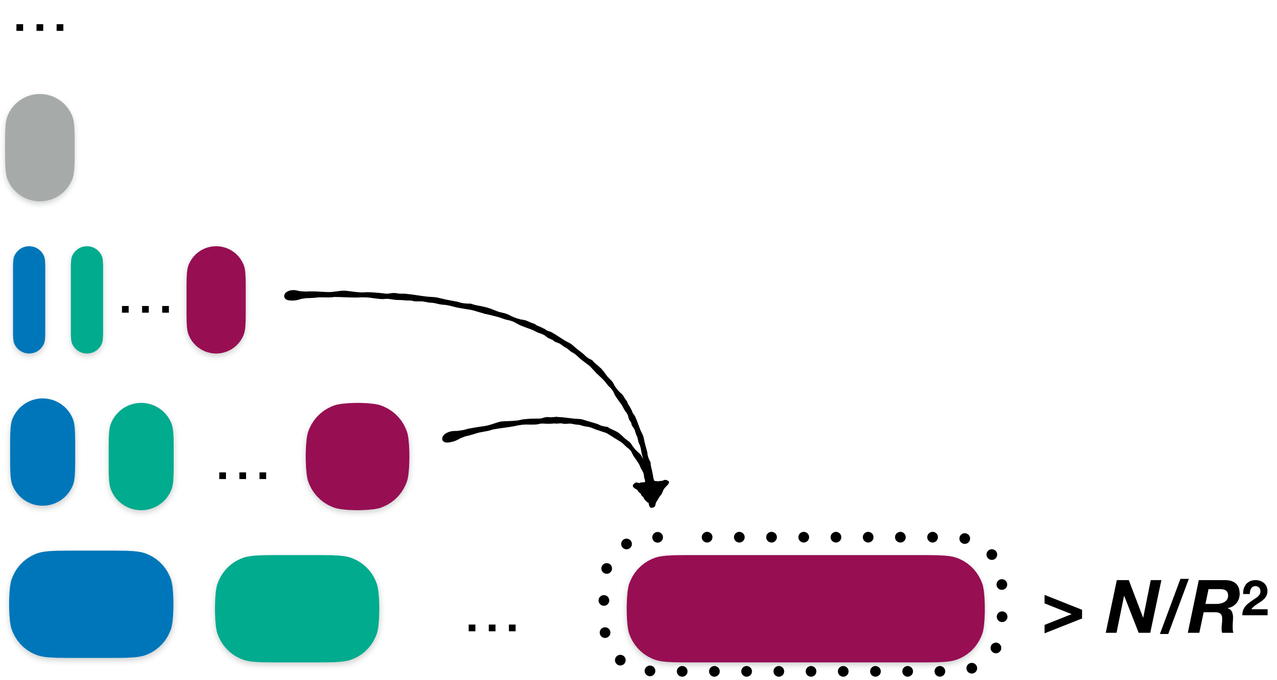
Spooky



Spooky

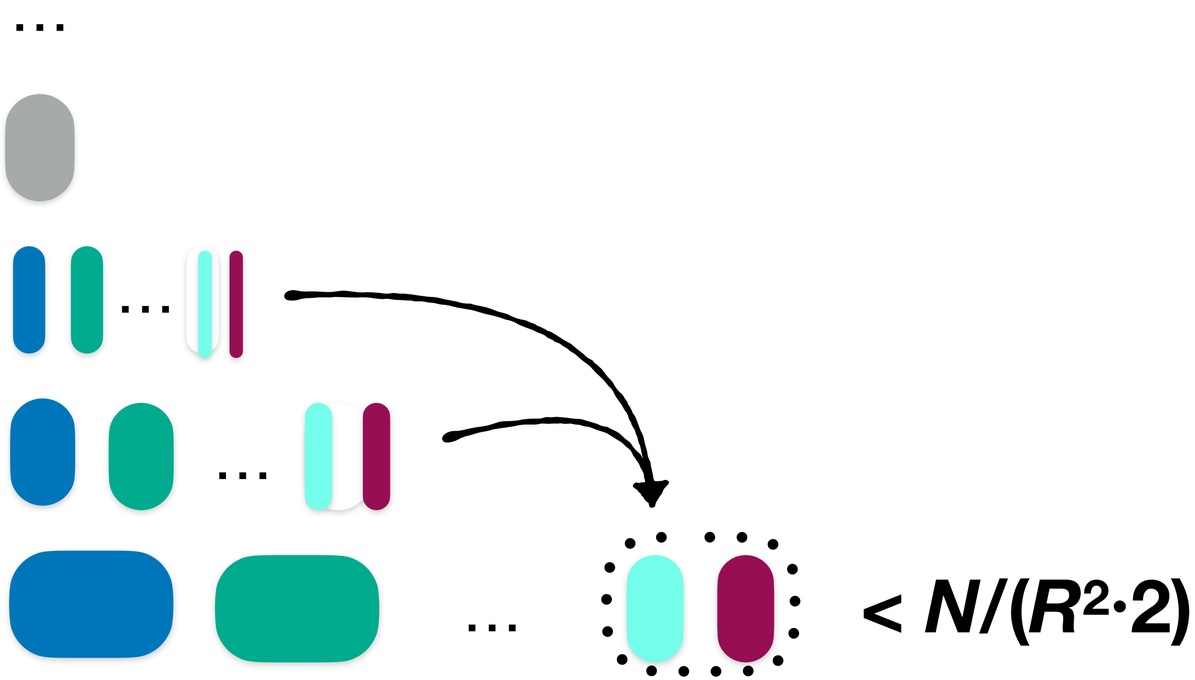


Spooky



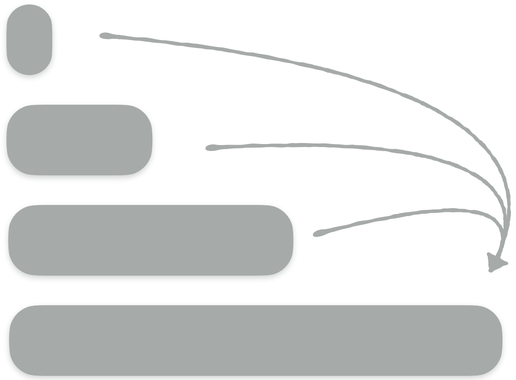
Spooky

Spooky

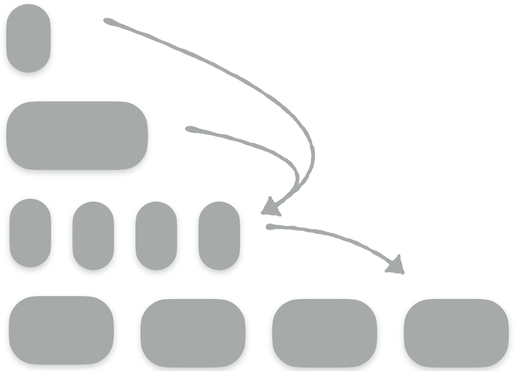


Spooky

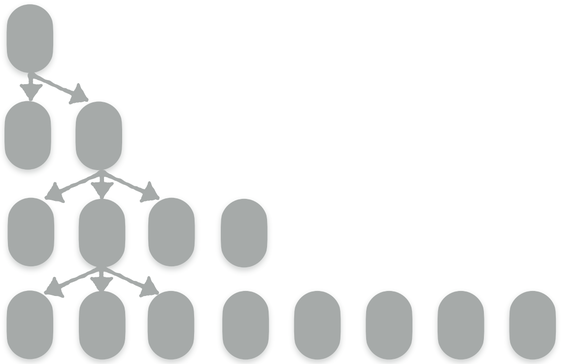
Full



Spooky

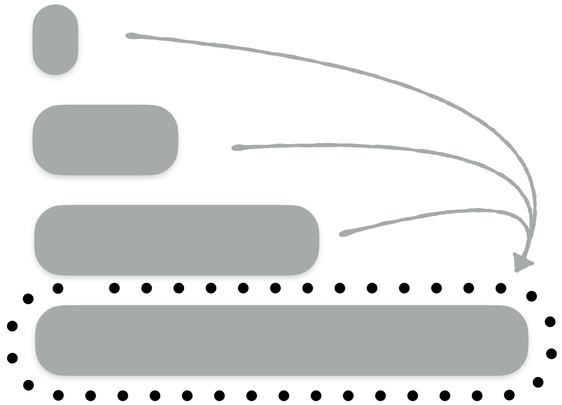


Partial

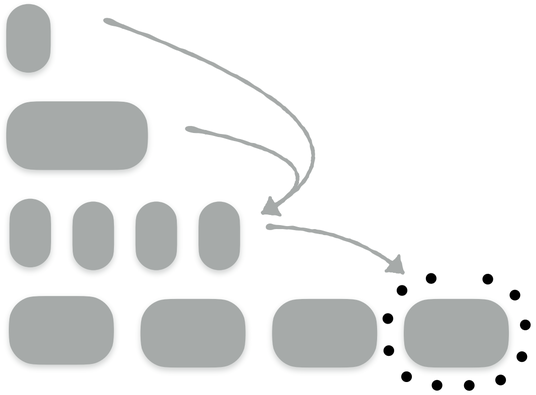


Space-Amplification

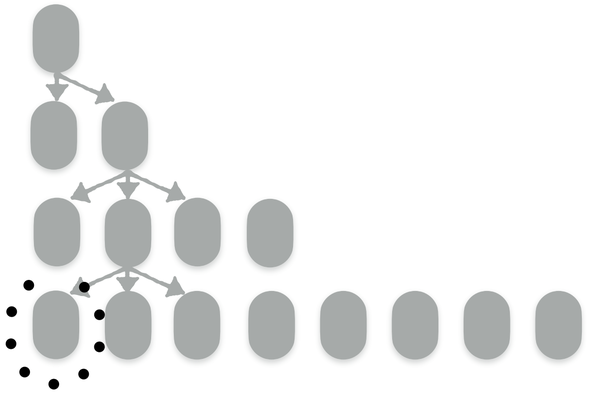
Full

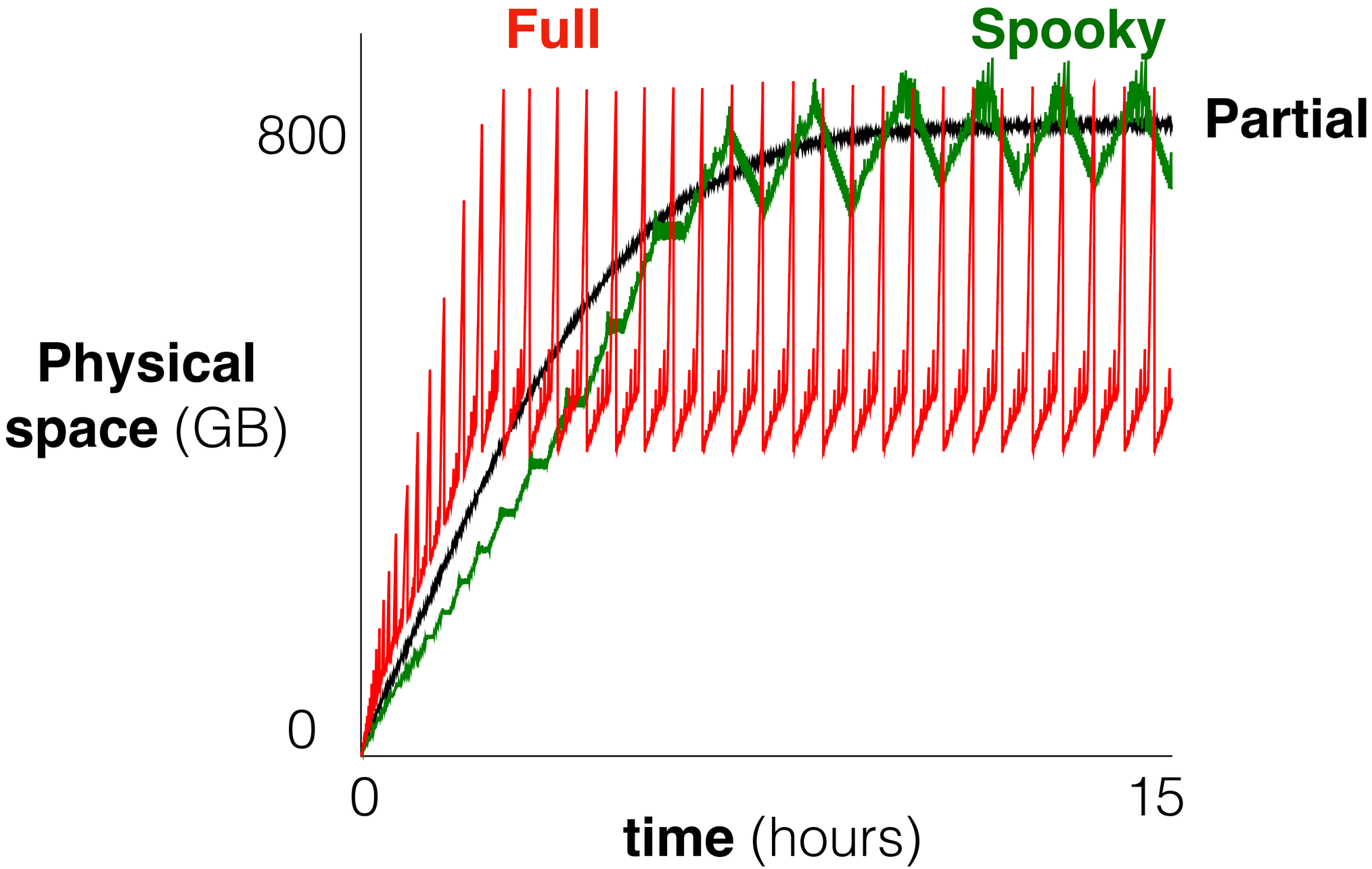


Spooky



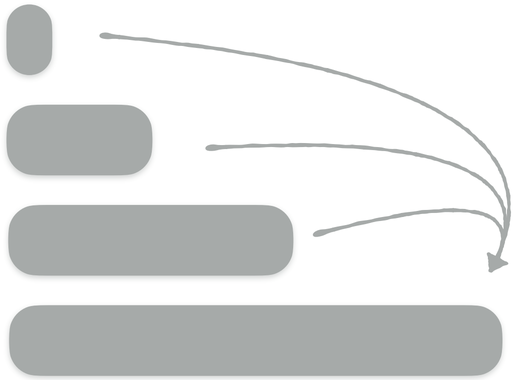
Partial



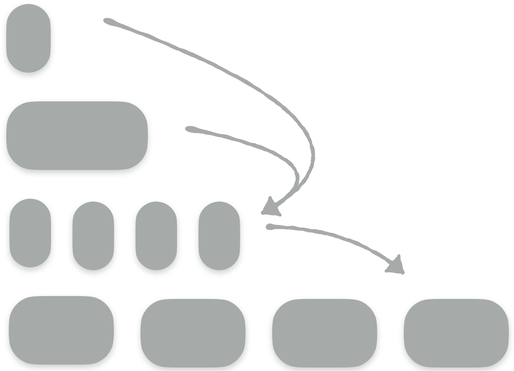

Full
Spooky
Partial
800
0
Physical space (GB)
0
15
time (hours)

Compaction Overheads

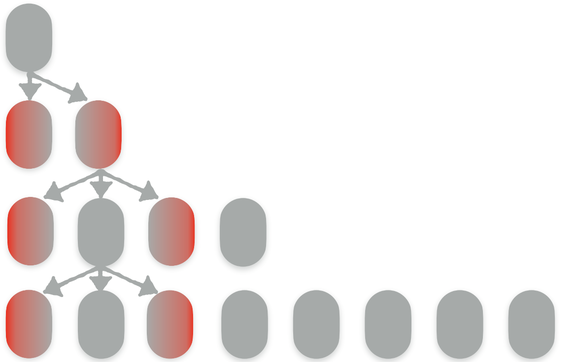
Full

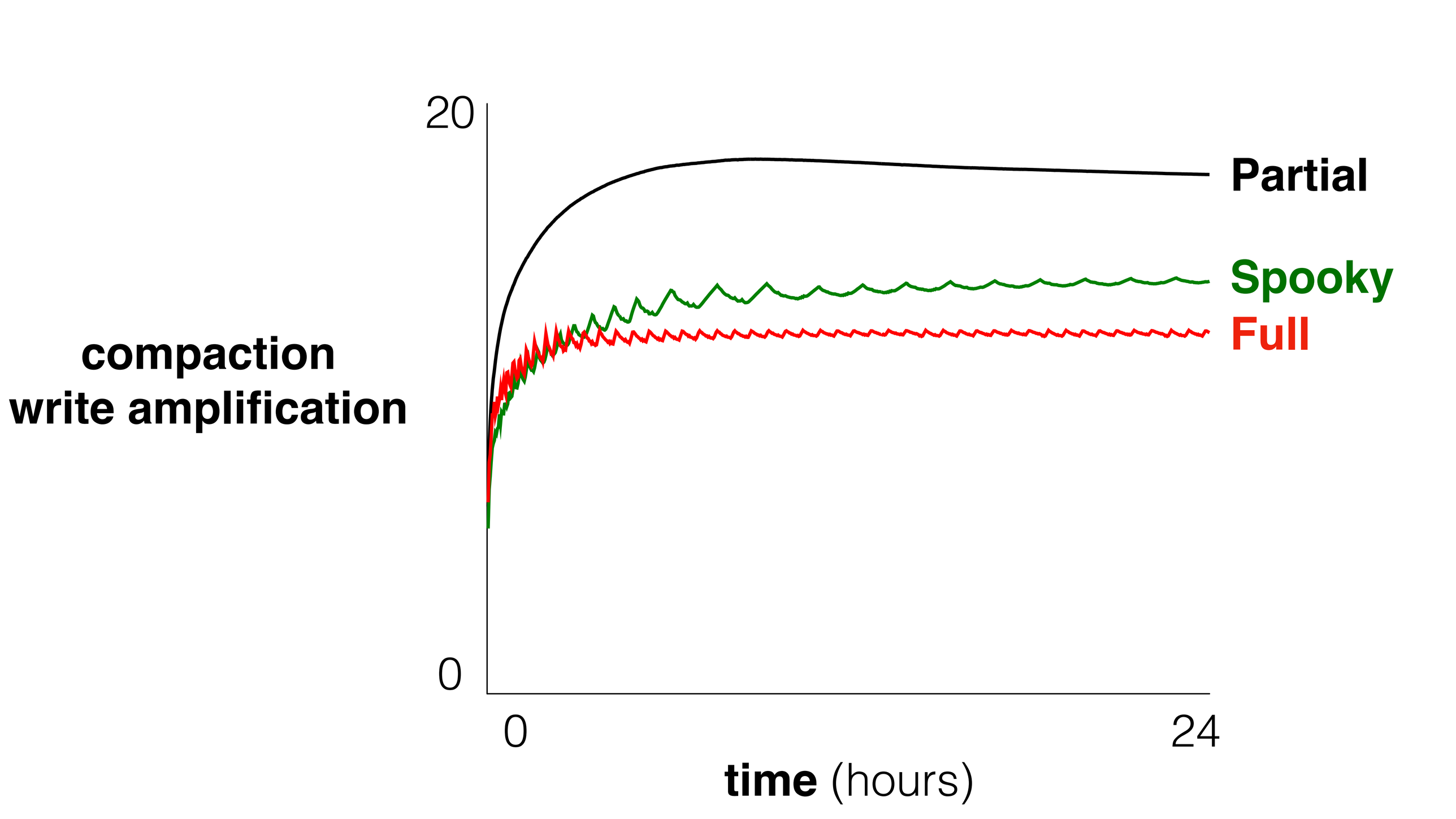


Spooky



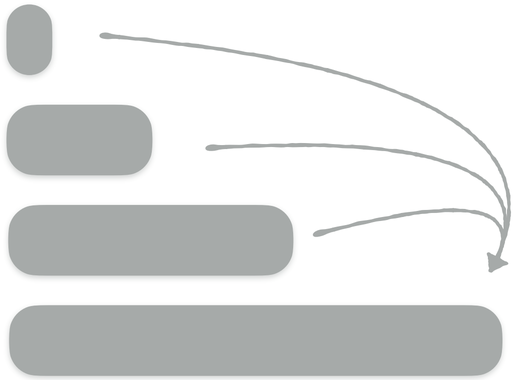
Partial



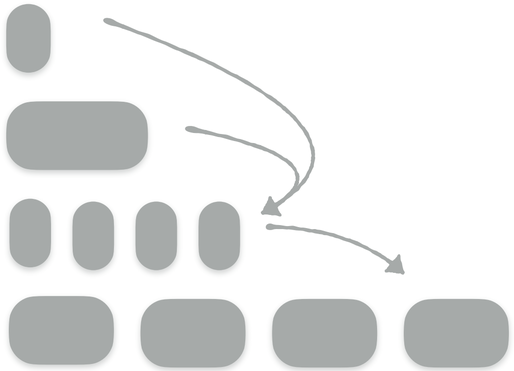

compaction
write amplification
20
0
0
24
time (hours)
Partial
Spooky
Full

SSD Garbage-Collection Overheads

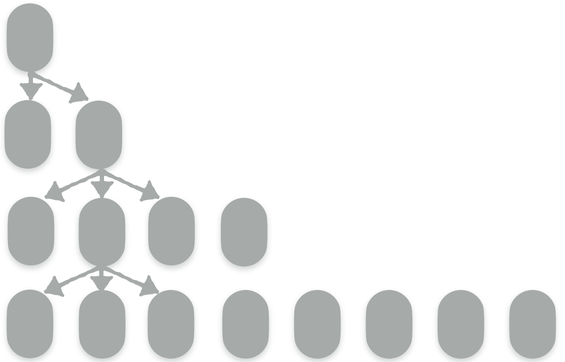
Full



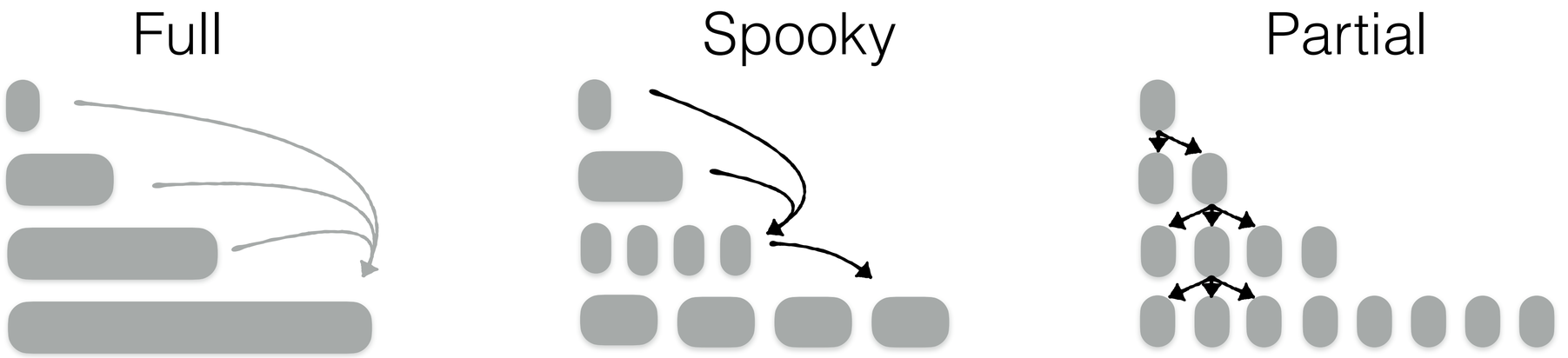
Spooky



Partial

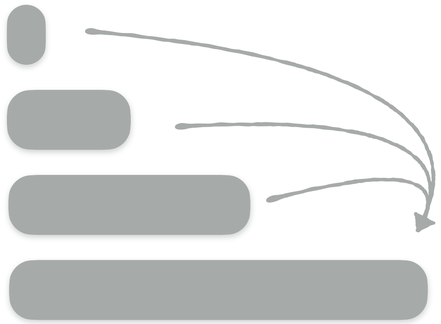


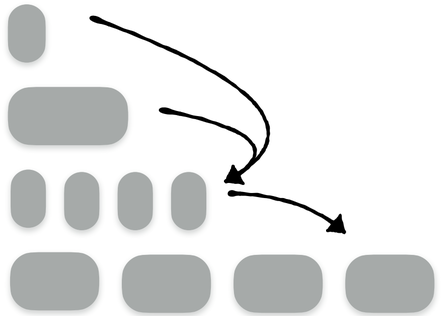
SSD Garbage-Collection Overheads

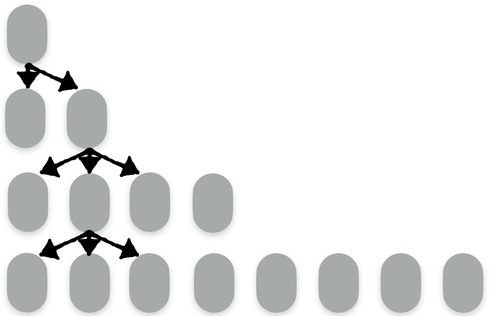


	Full	Spooky	Partial
# simultaneous compactions:	**few**	**few**	**many**

SSD Garbage-Collection Overheads

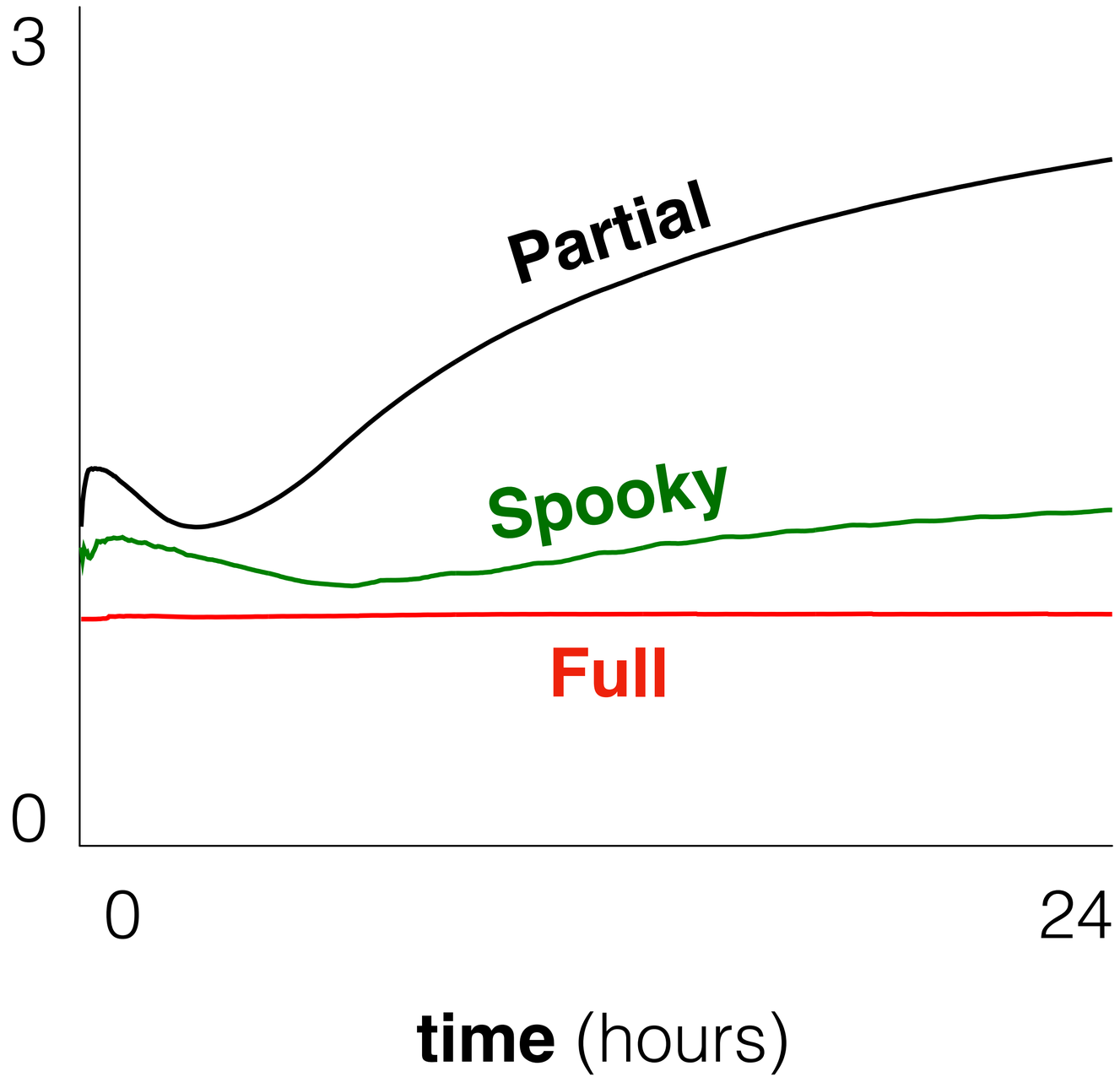






# simultaneous compactions:	few	few	many
File sizes:	**large**	**large**	**small**

Garbage Collection
write amplification



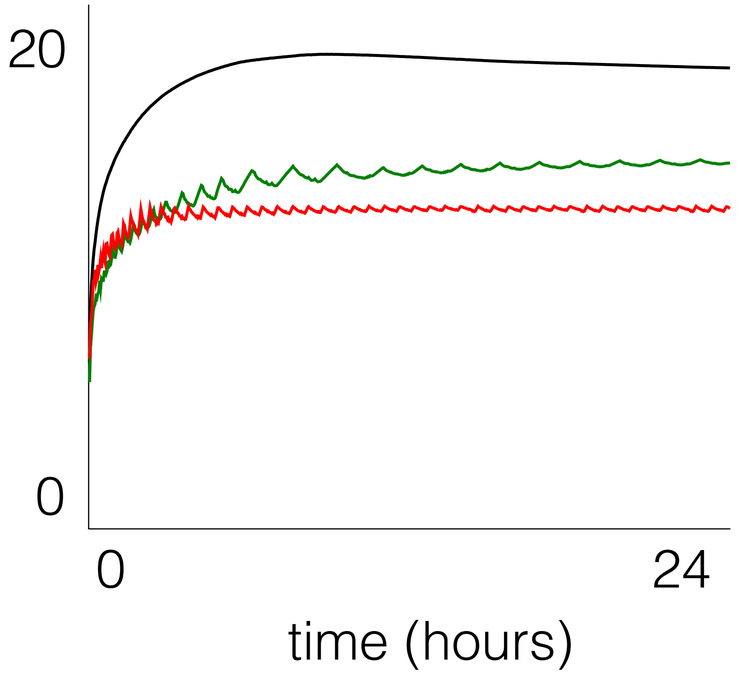
Compaction
write amplification

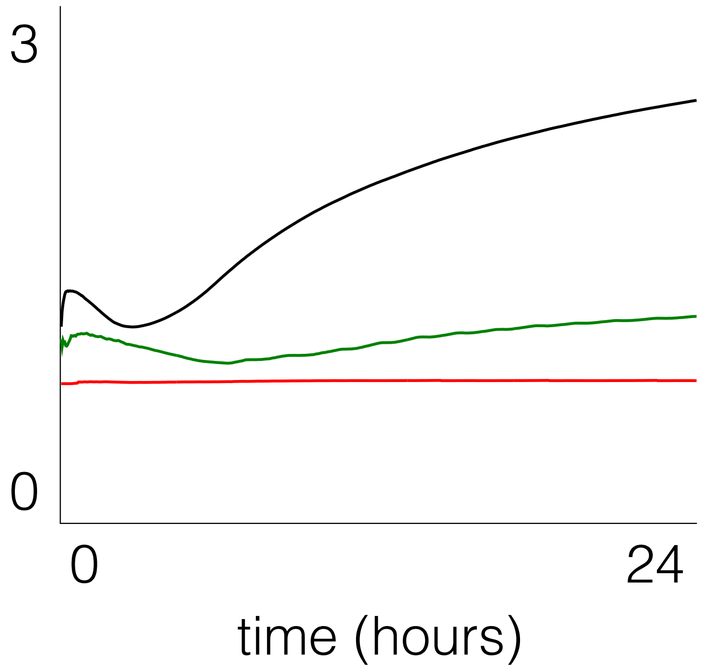
X

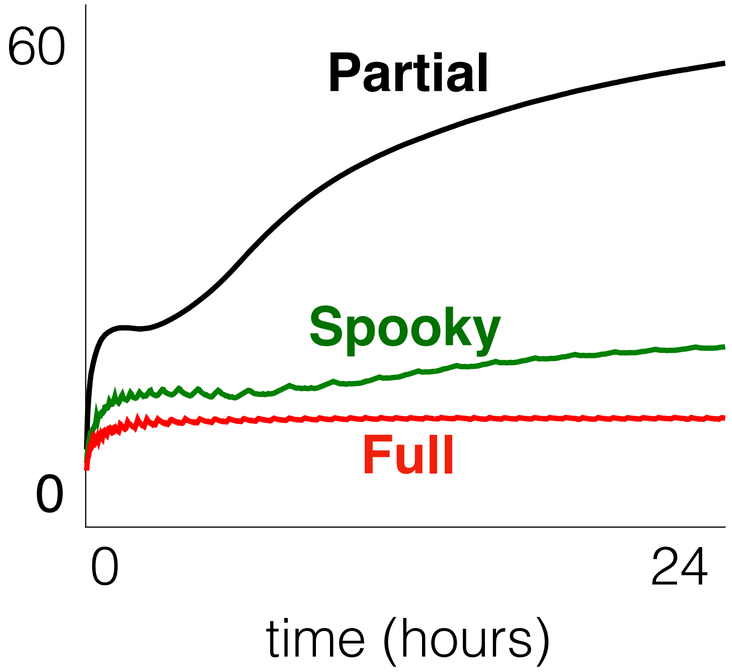
Garbage Collection
write amplification

=

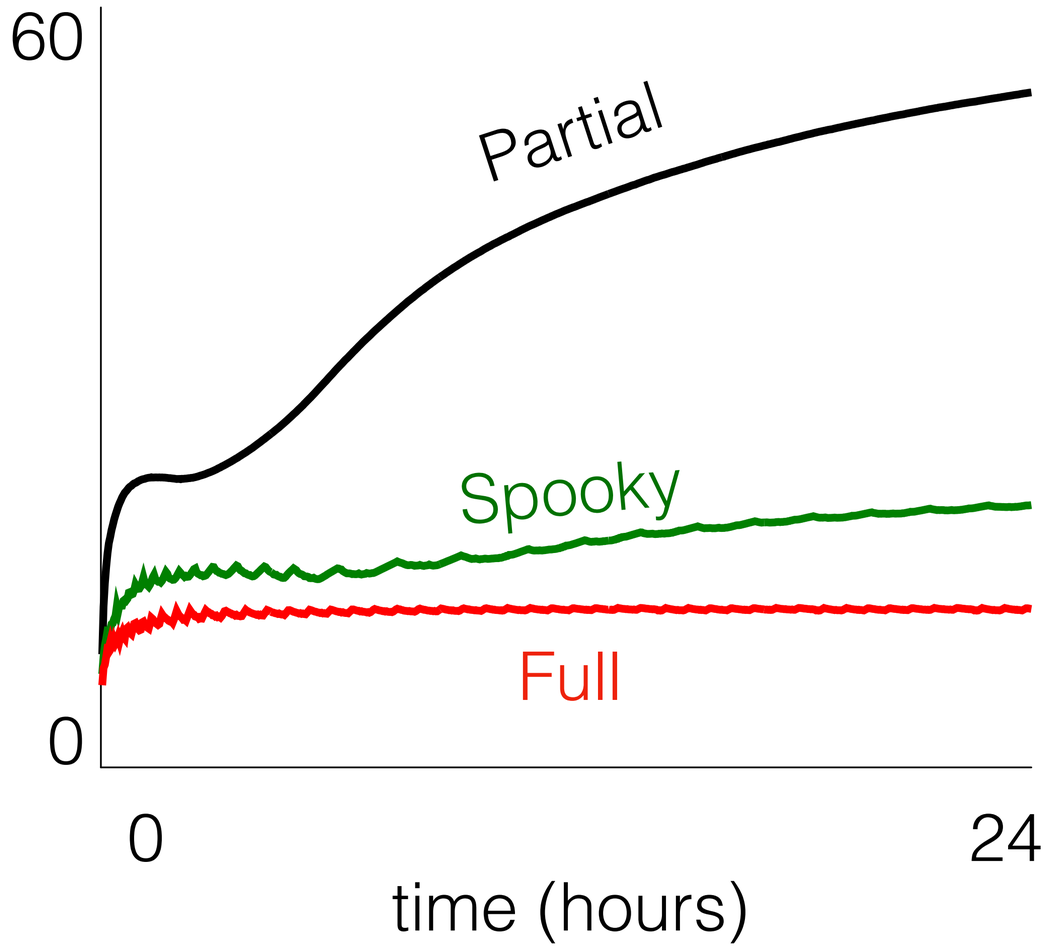
Total
write amplification

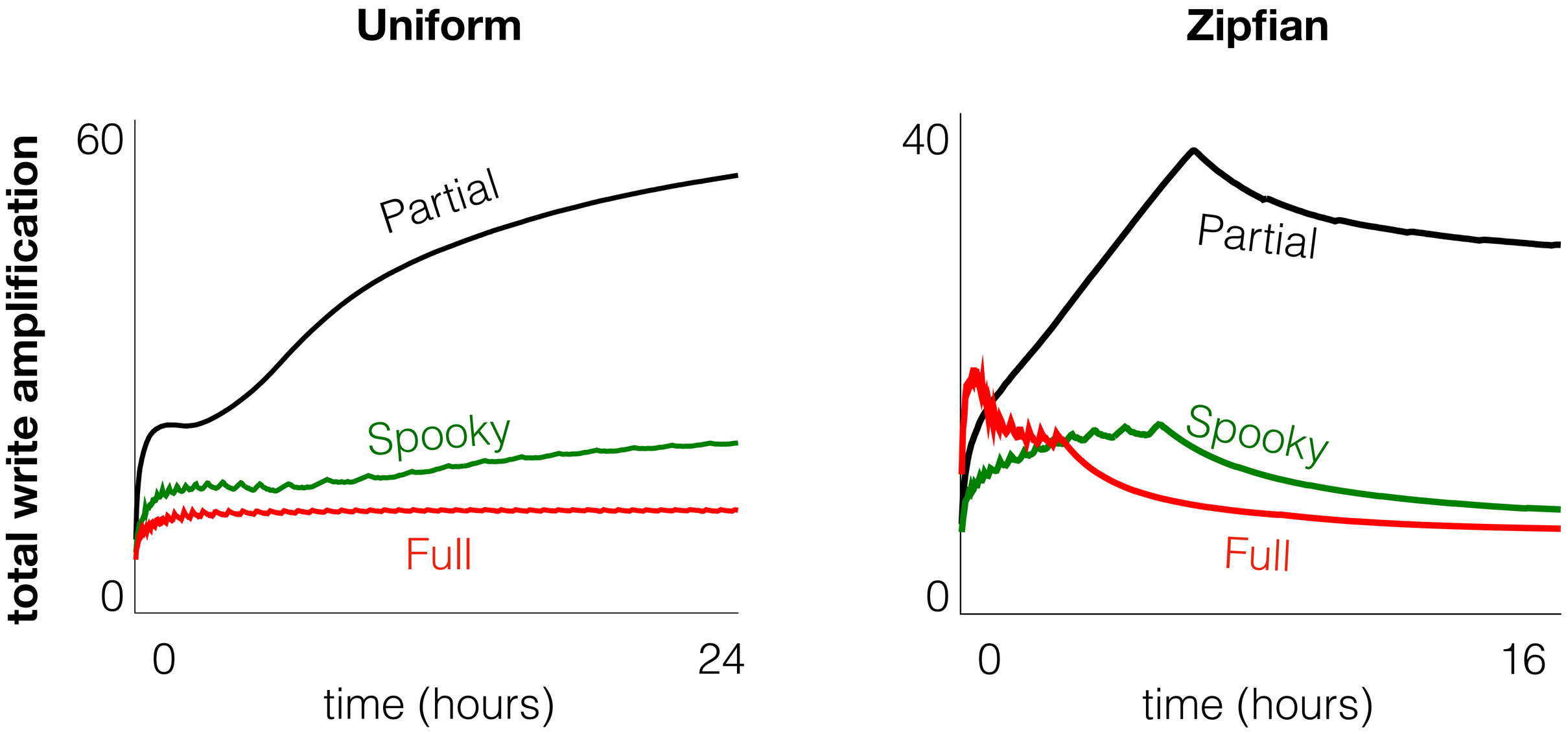


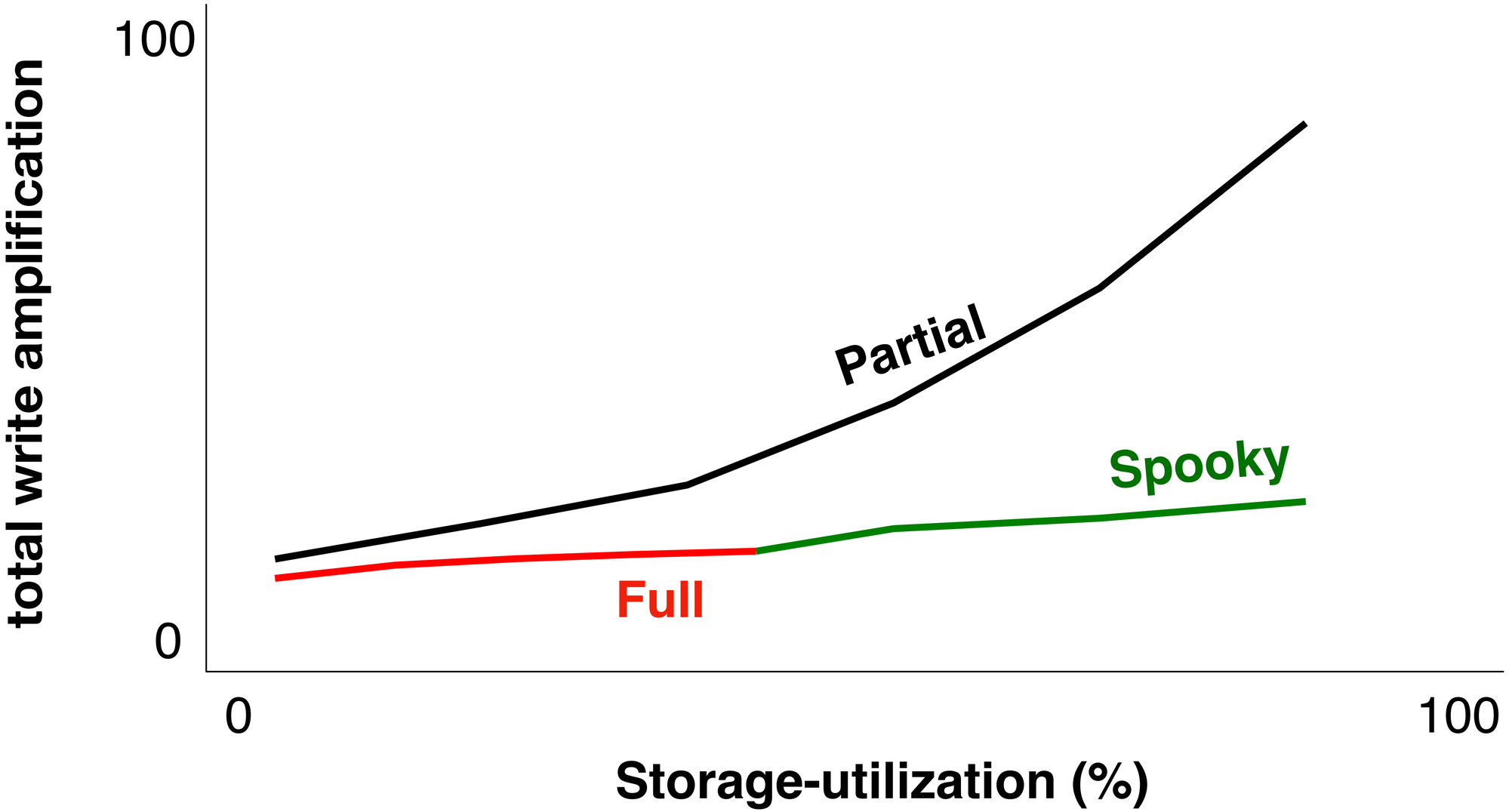


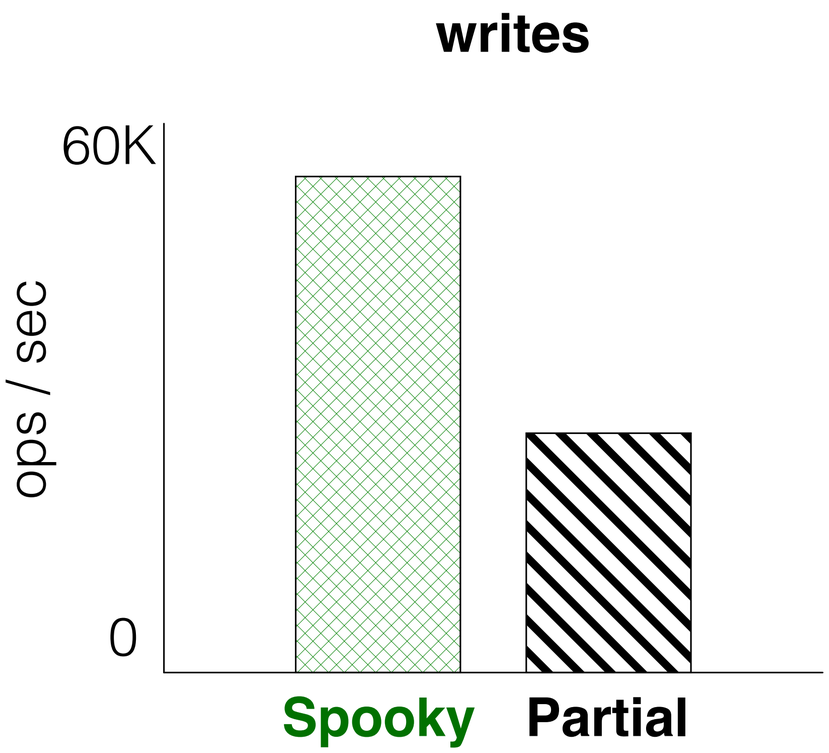


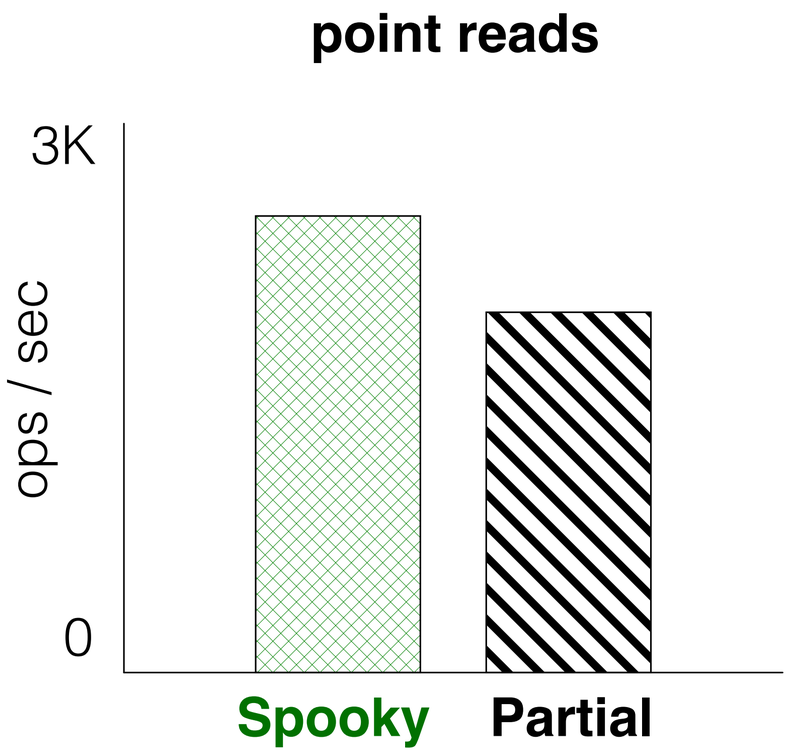
total write amplification

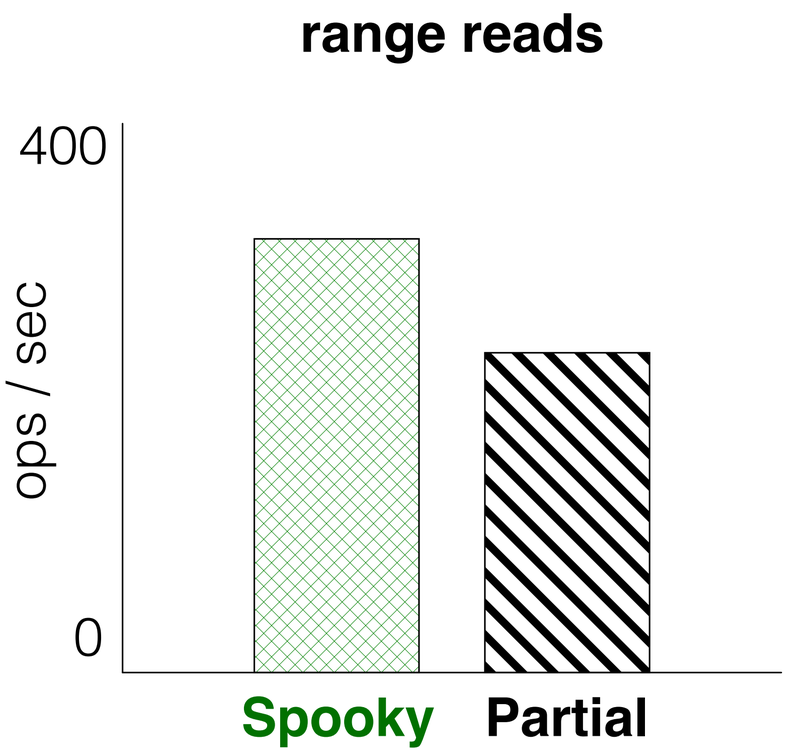



Uniform
Zipfian
total write amplification
60
0
40
0
Partial
Spooky
Full
Partial
Spooky
Full
0
24
0
16
time (hours)
time (hours)


100
0
total write amplification
Partial
Spooky
Full
0
100
Storage-utilization (%)


writes
60K
ops / sec
0
Spooky
Partial


point reads
3K
ops / sec
0
Spooky
Partial


range reads
400
ops / sec
0
Spooky
Partial

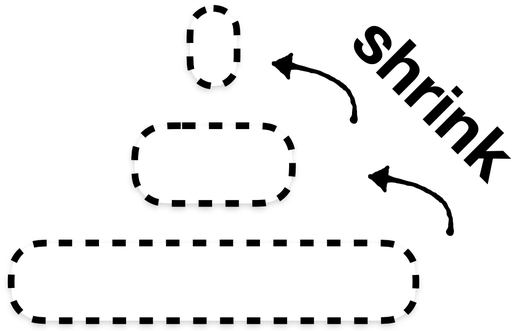
more in paper

more in paper

Any compaction eagerness

more in paper

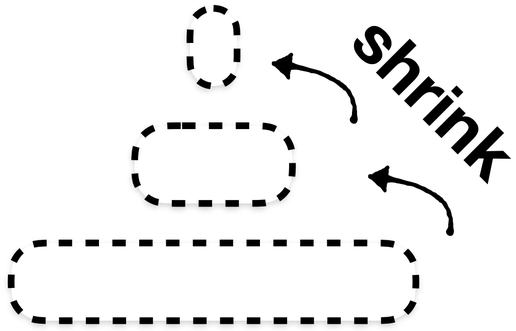
Any compaction eagerness



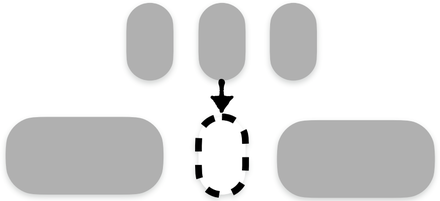
dynamic capacity adaptation

more in paper

Any compaction eagerness



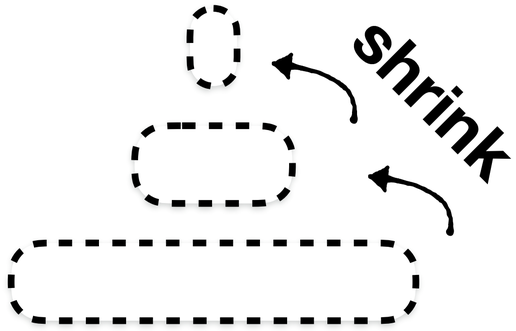
dynamic capacity adaptation



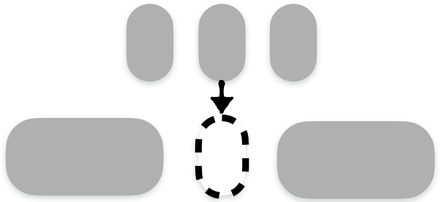
Trivial Moves

more in paper

Any compaction eagerness



dynamic capacity adaptation

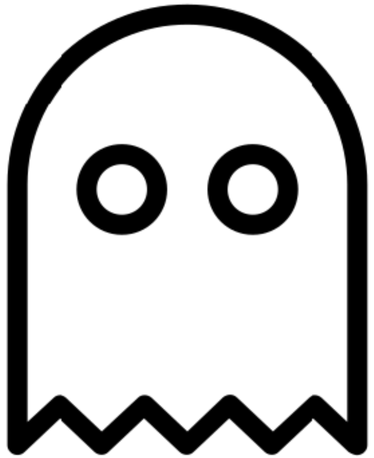


Trivial Moves

X+Y=Z

Cost models

Spooky



write-amplification

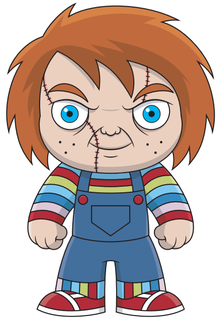
space-amplification

Monkey
SIGMOD 2017

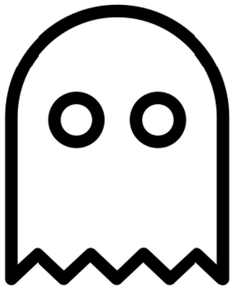
Dostoevsky
SIGMOD 2018

LSM-bush
SIGMOD 2019

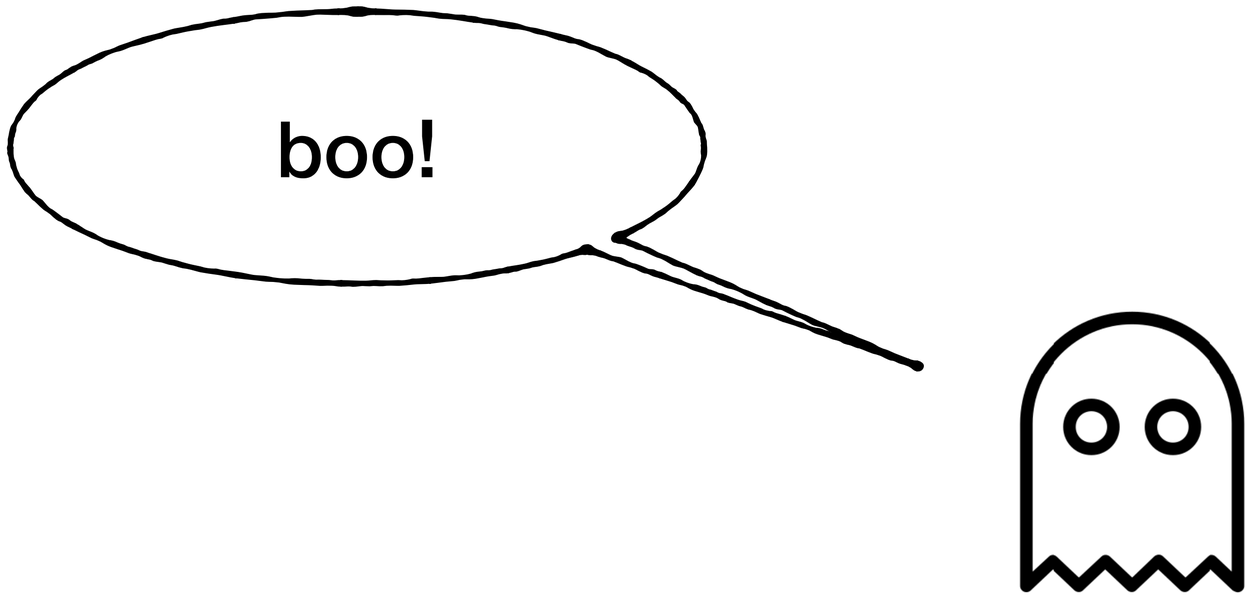
Rosetta
SIGMOD 2020



Chucky
SIGMOD 2021



Spooky
VLDB 2022


boo!