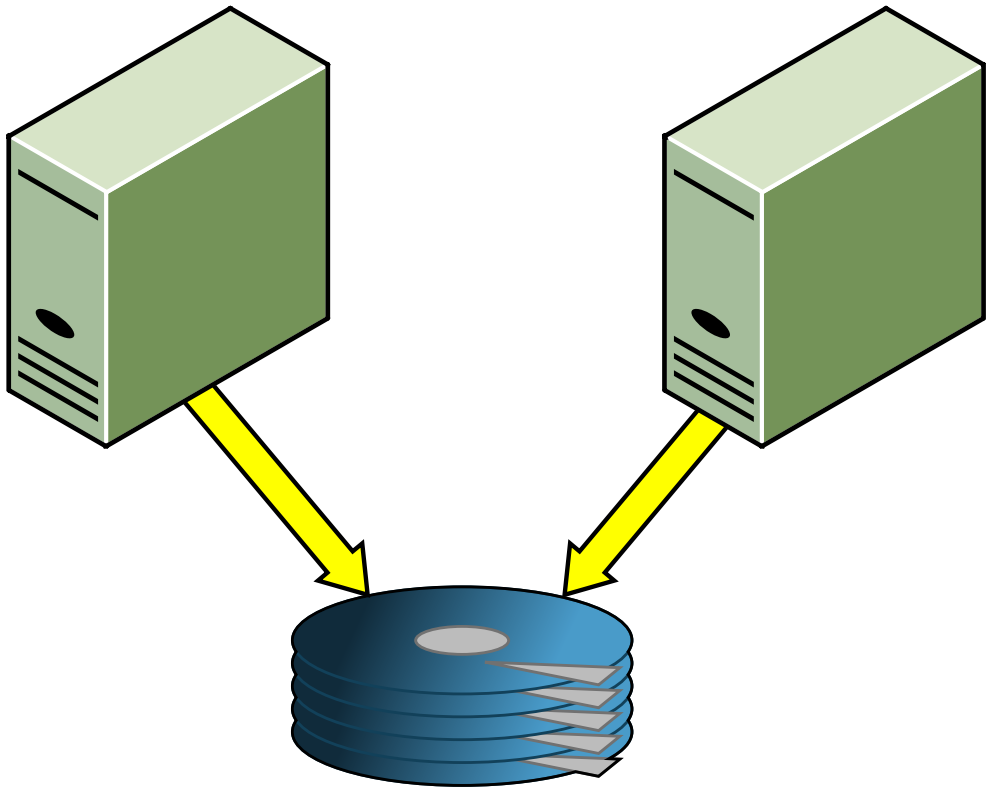
Decentralized Deduplication in SAN Cluster File Systems

Austin T. Clements*

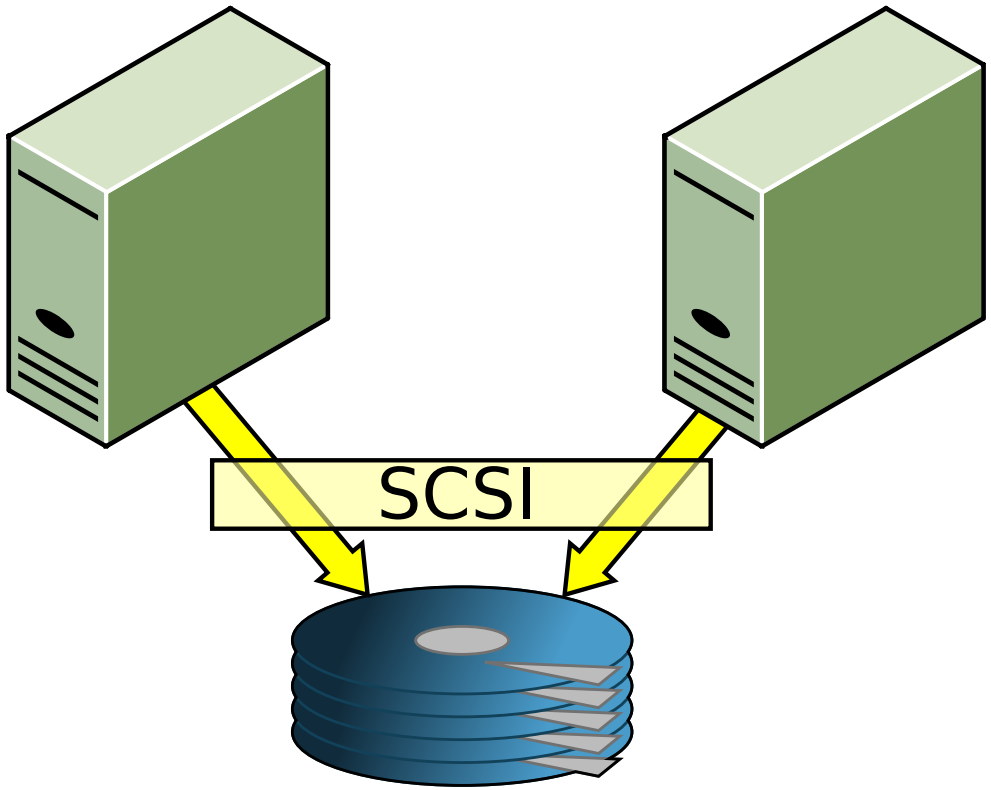
Irfan Ahmad Murali Vilayannur Jinyuan Li

VMware, Inc. *MIT CSAIL

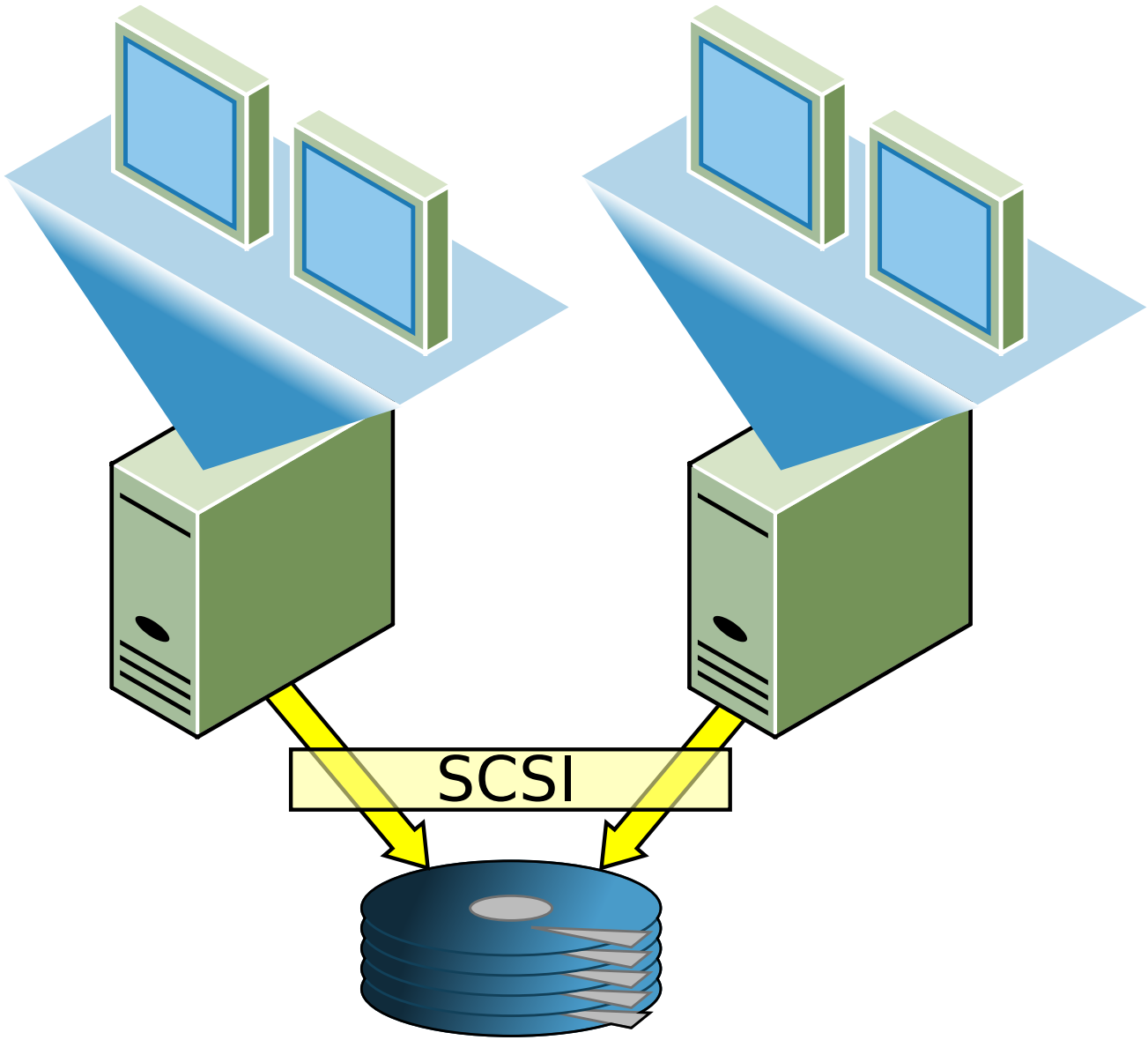
Storage Area Networks



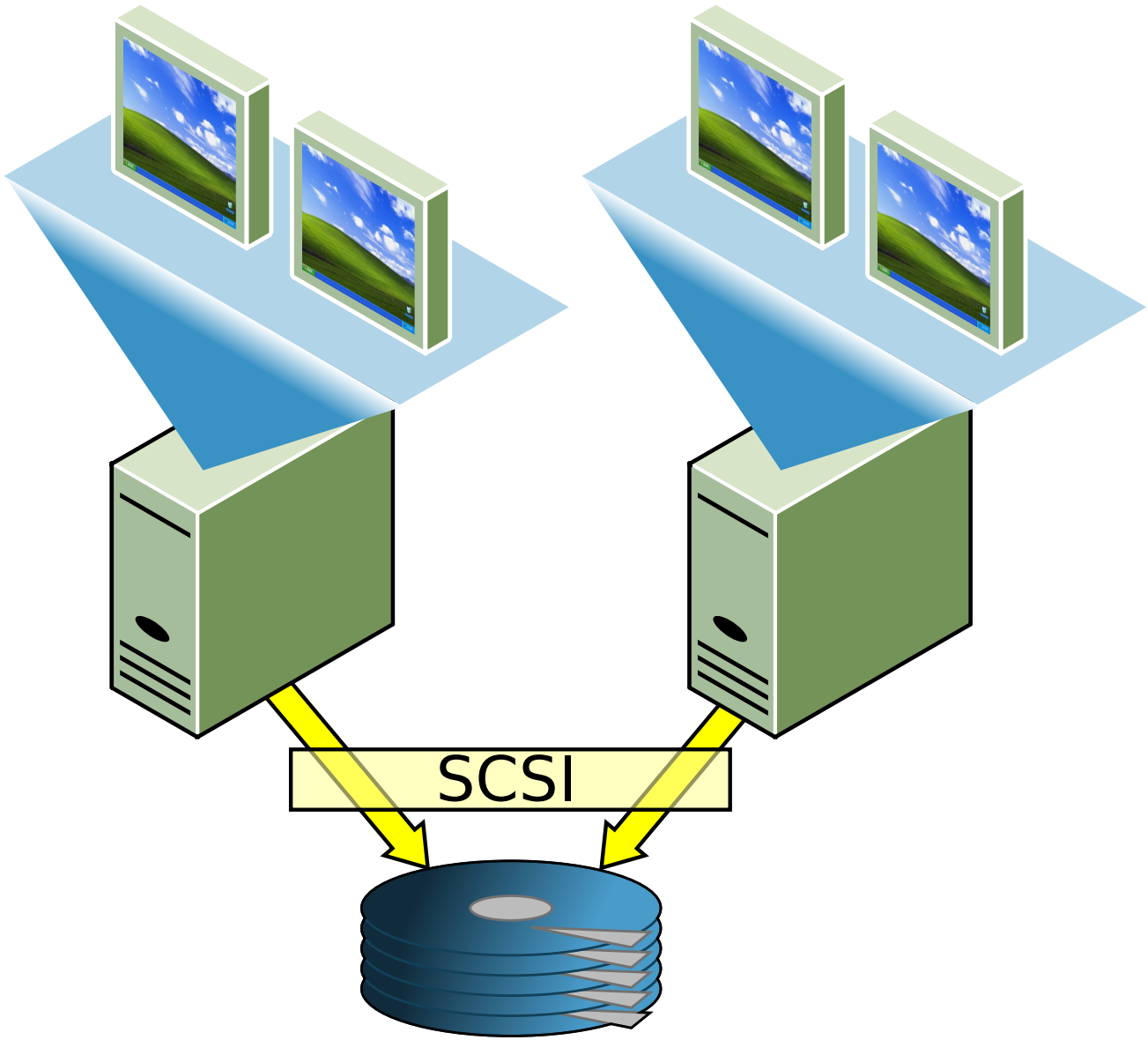
Storage Area Networks



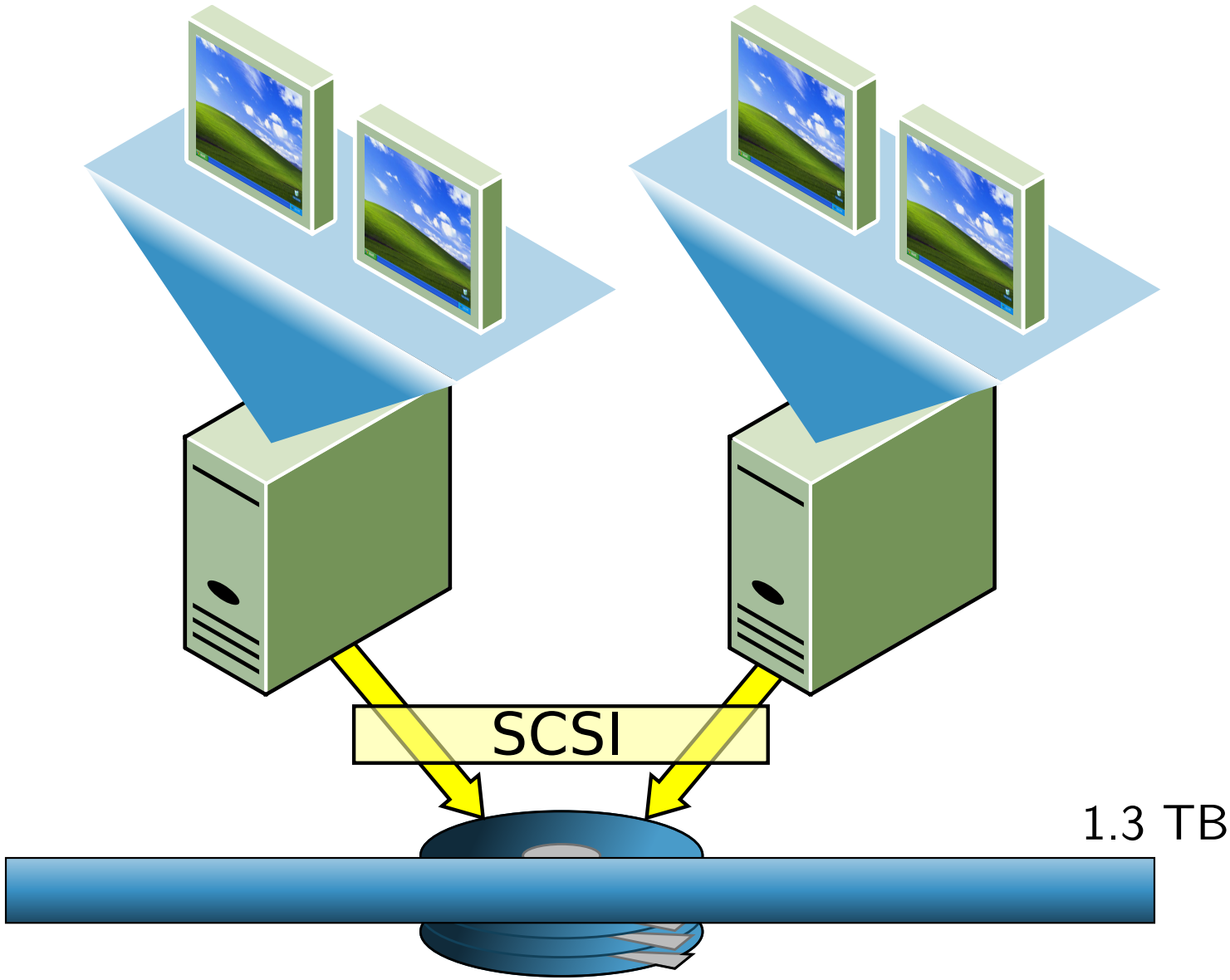
Storage Area Networks



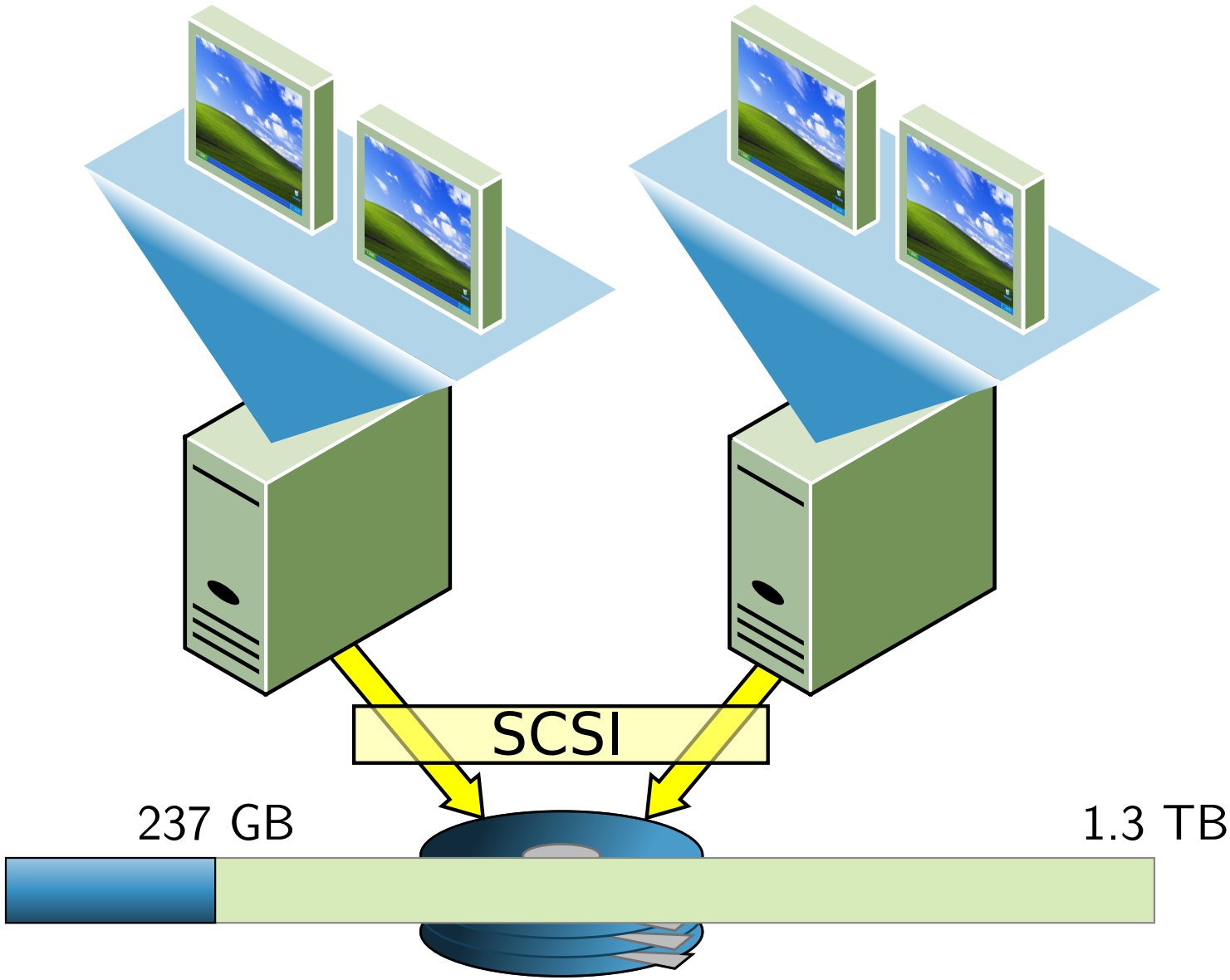
Storage Area Networks



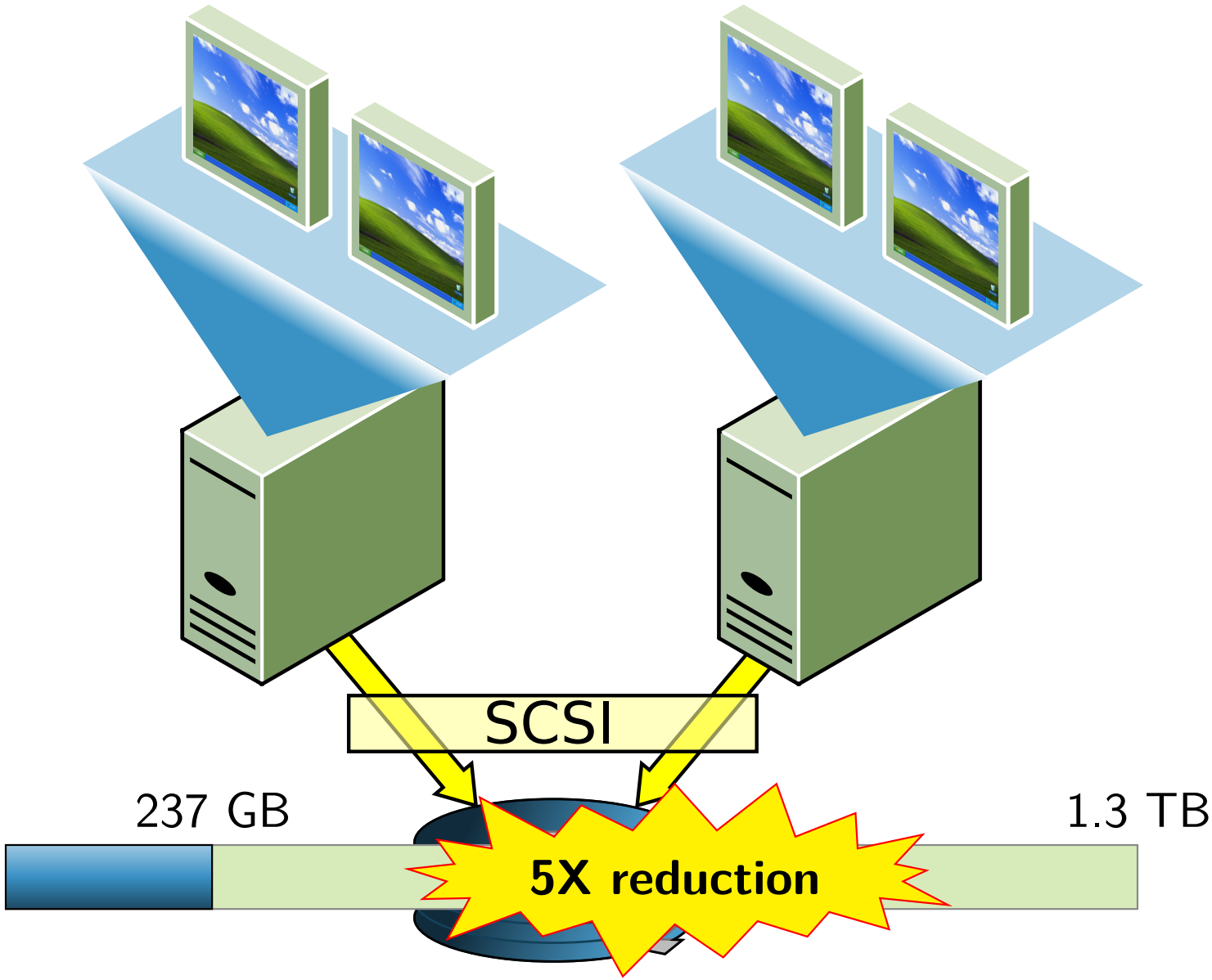
Storage Area Networks



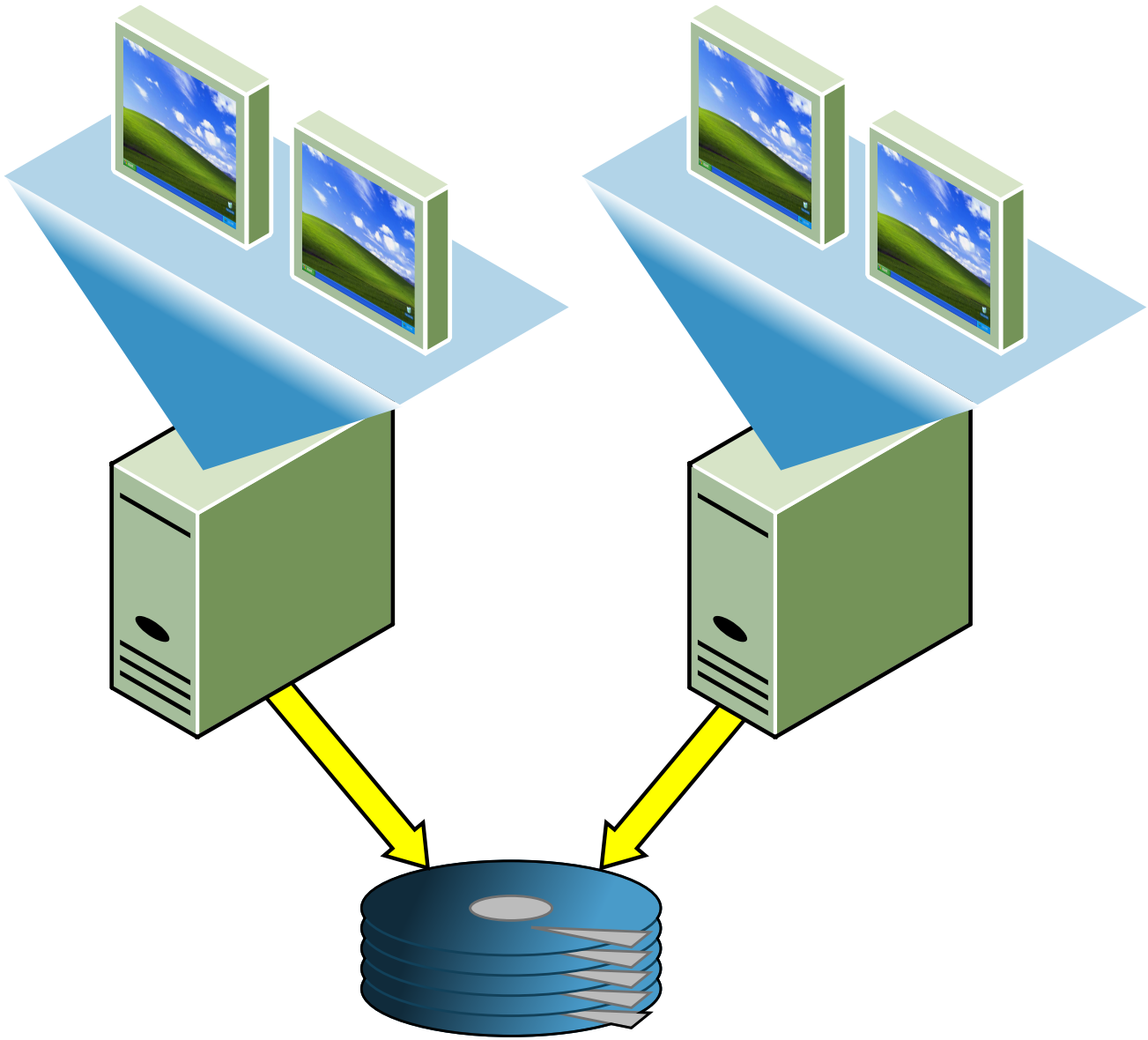
Storage Area Networks



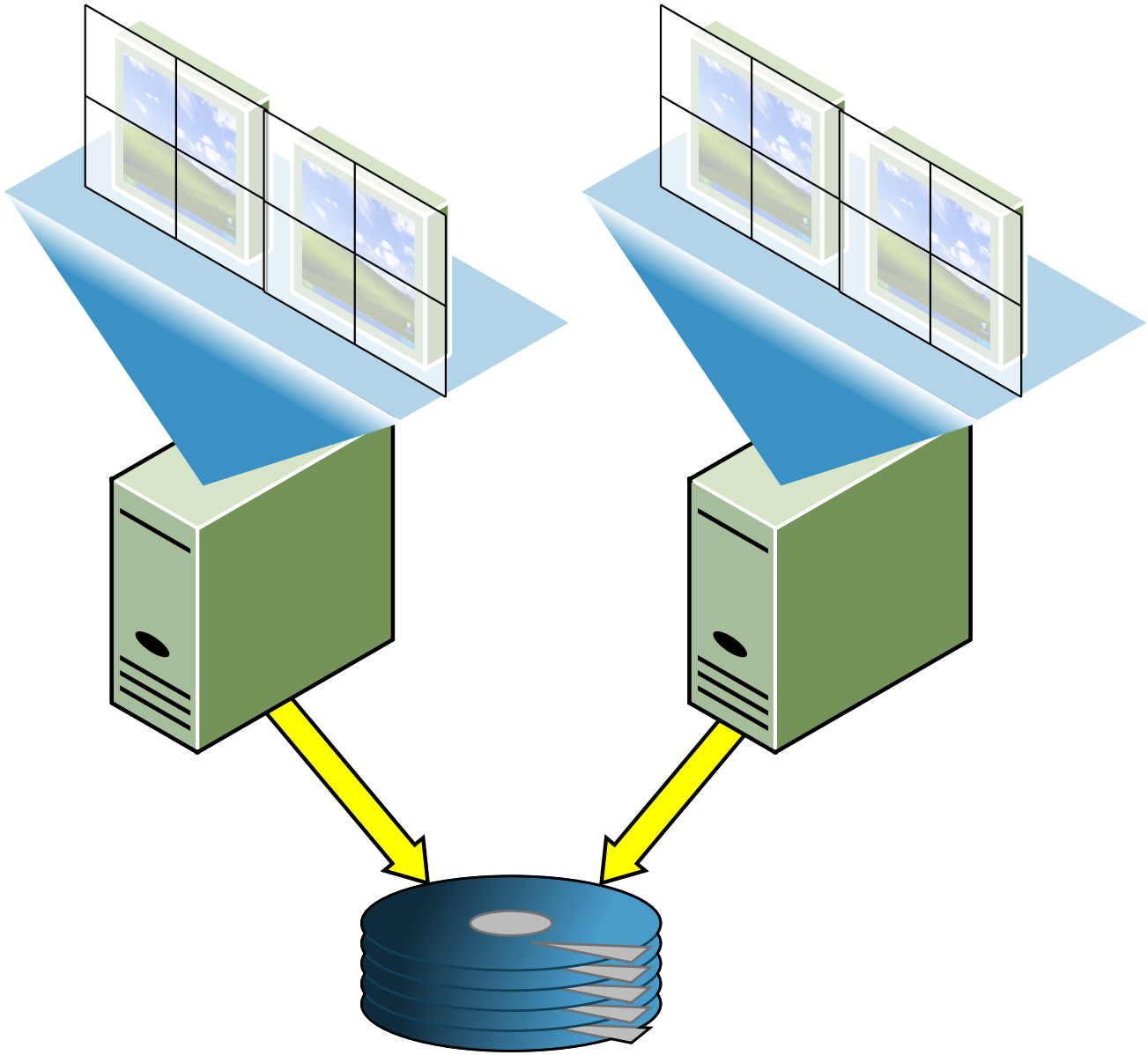
Storage Area Networks



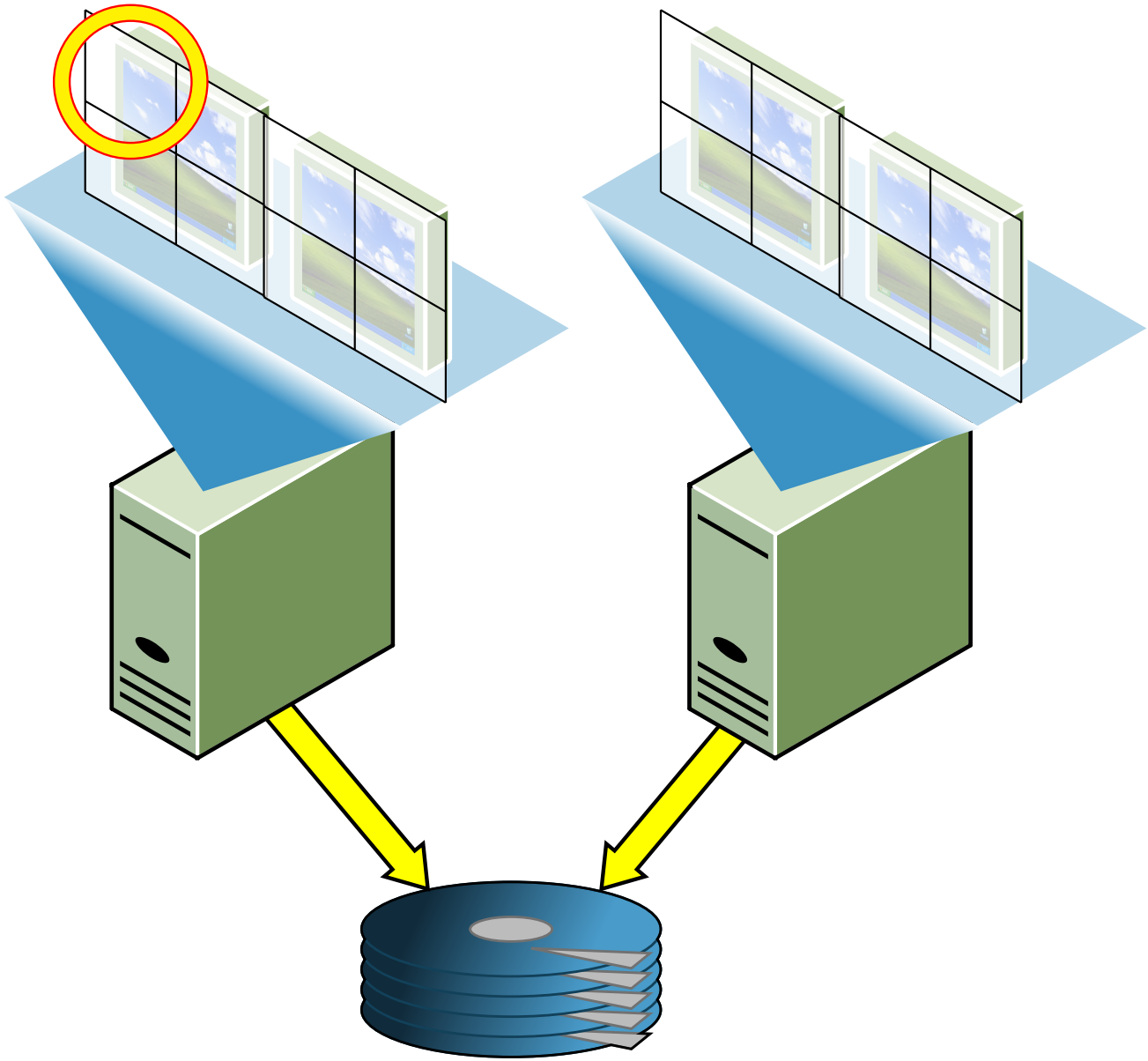
Deduplication



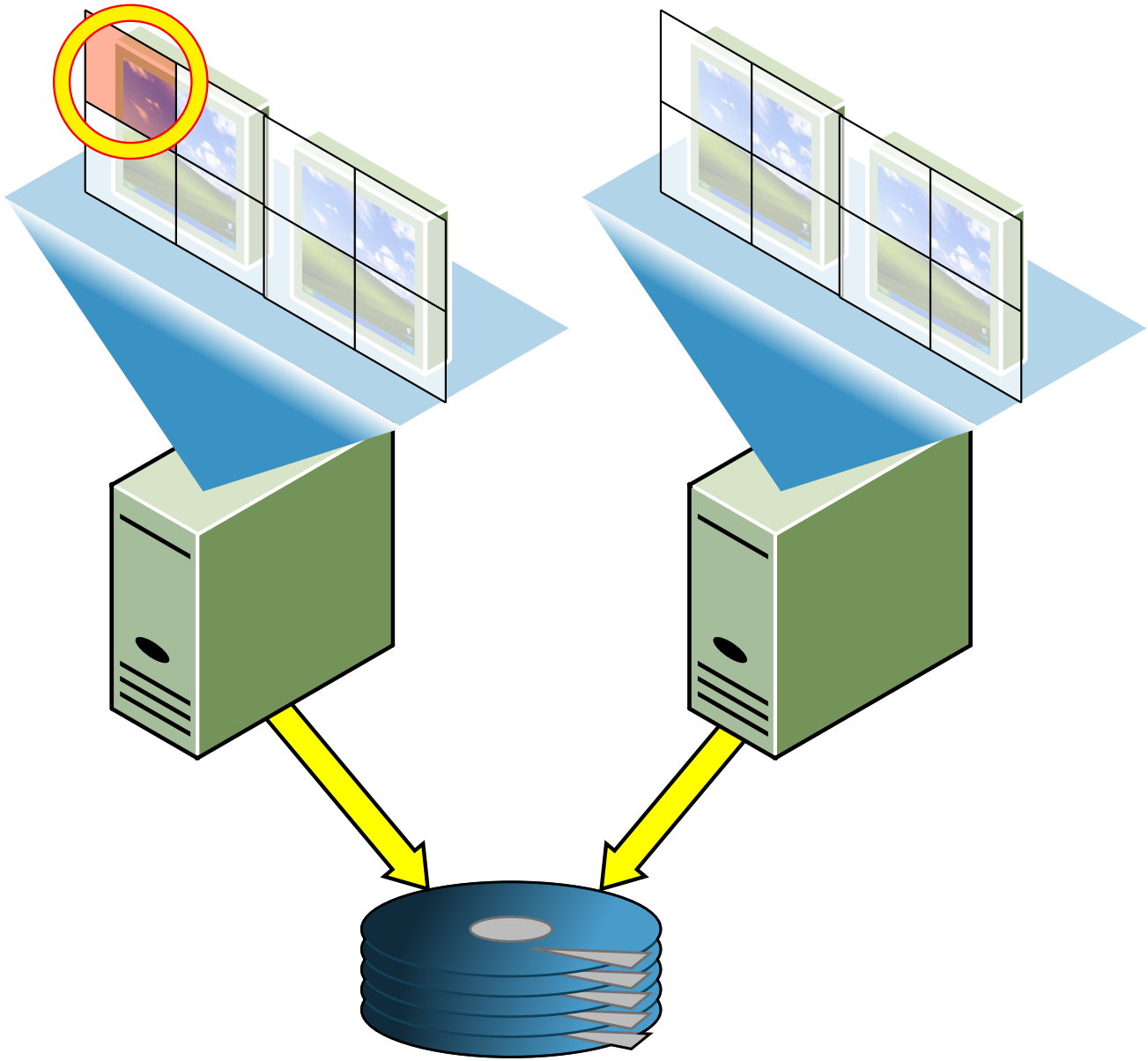
Deduplication



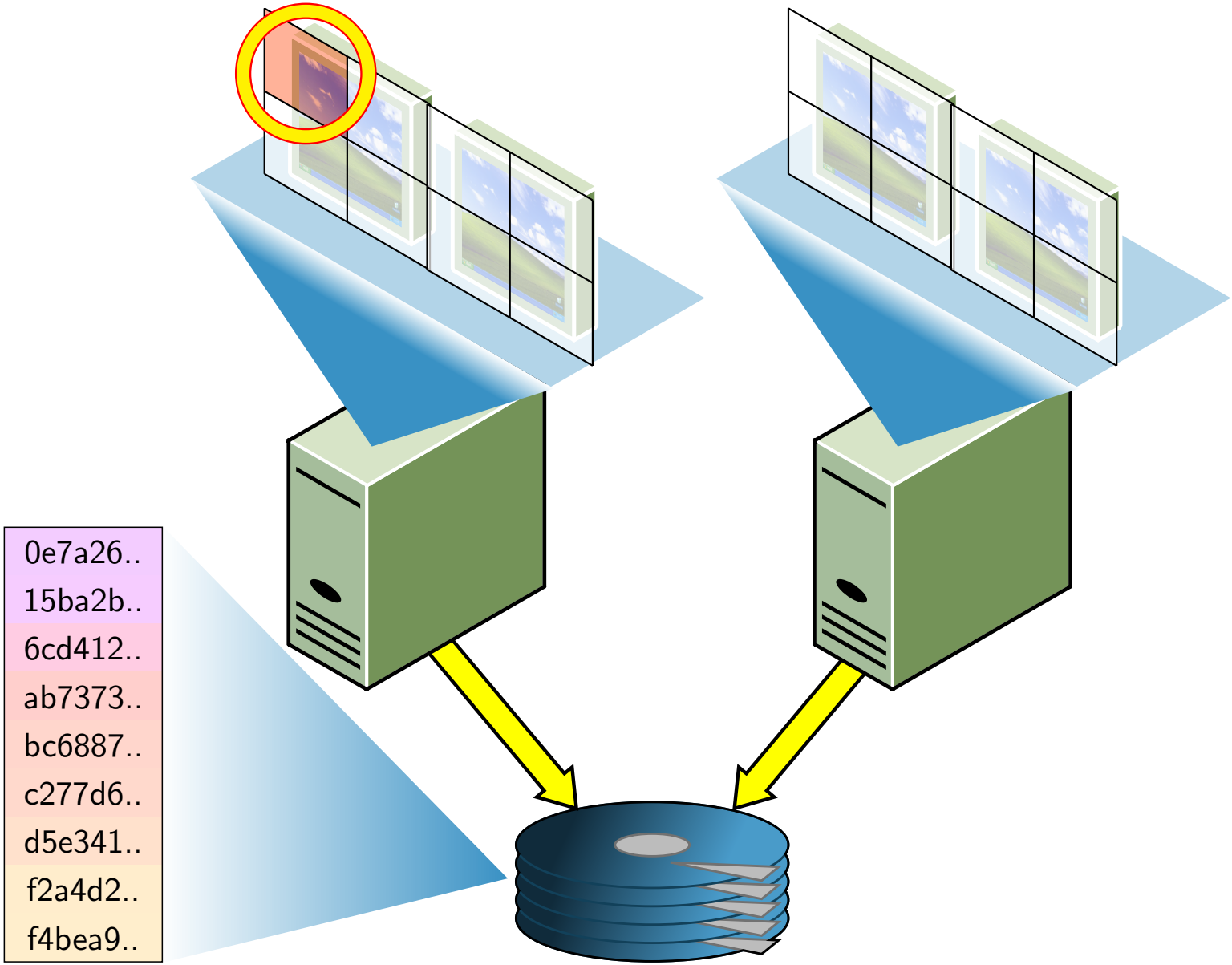
Deduplication



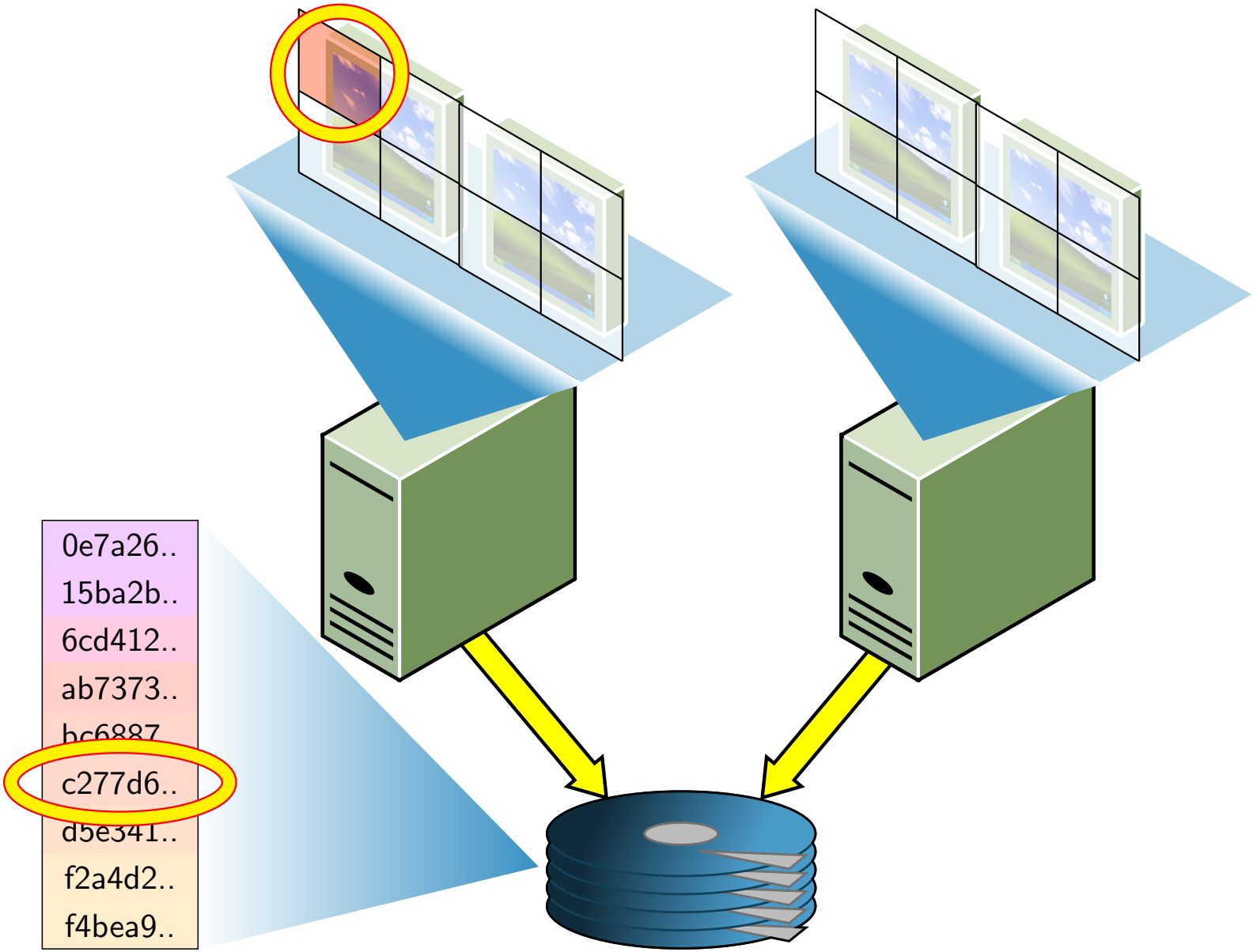
Deduplication



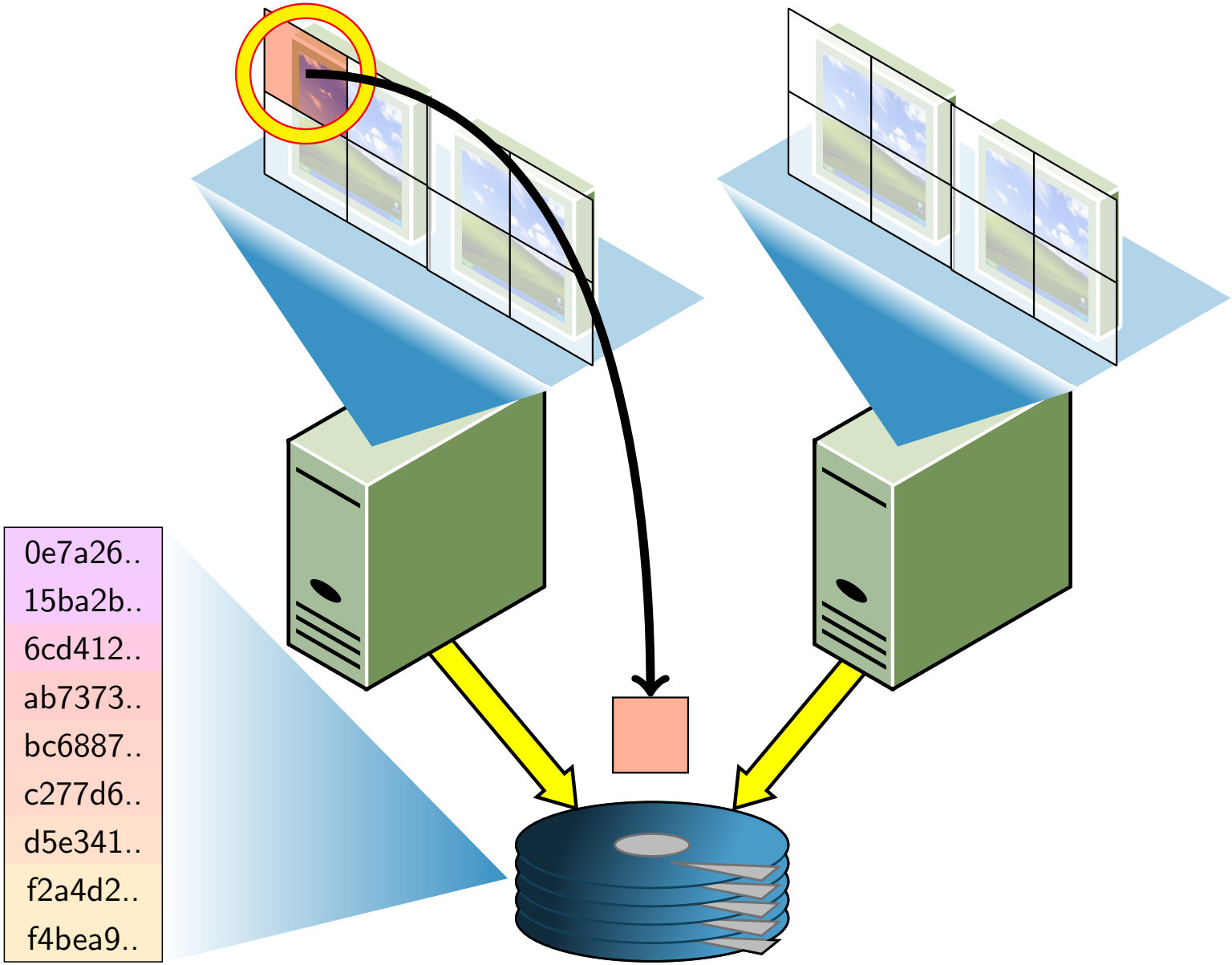
Deduplication



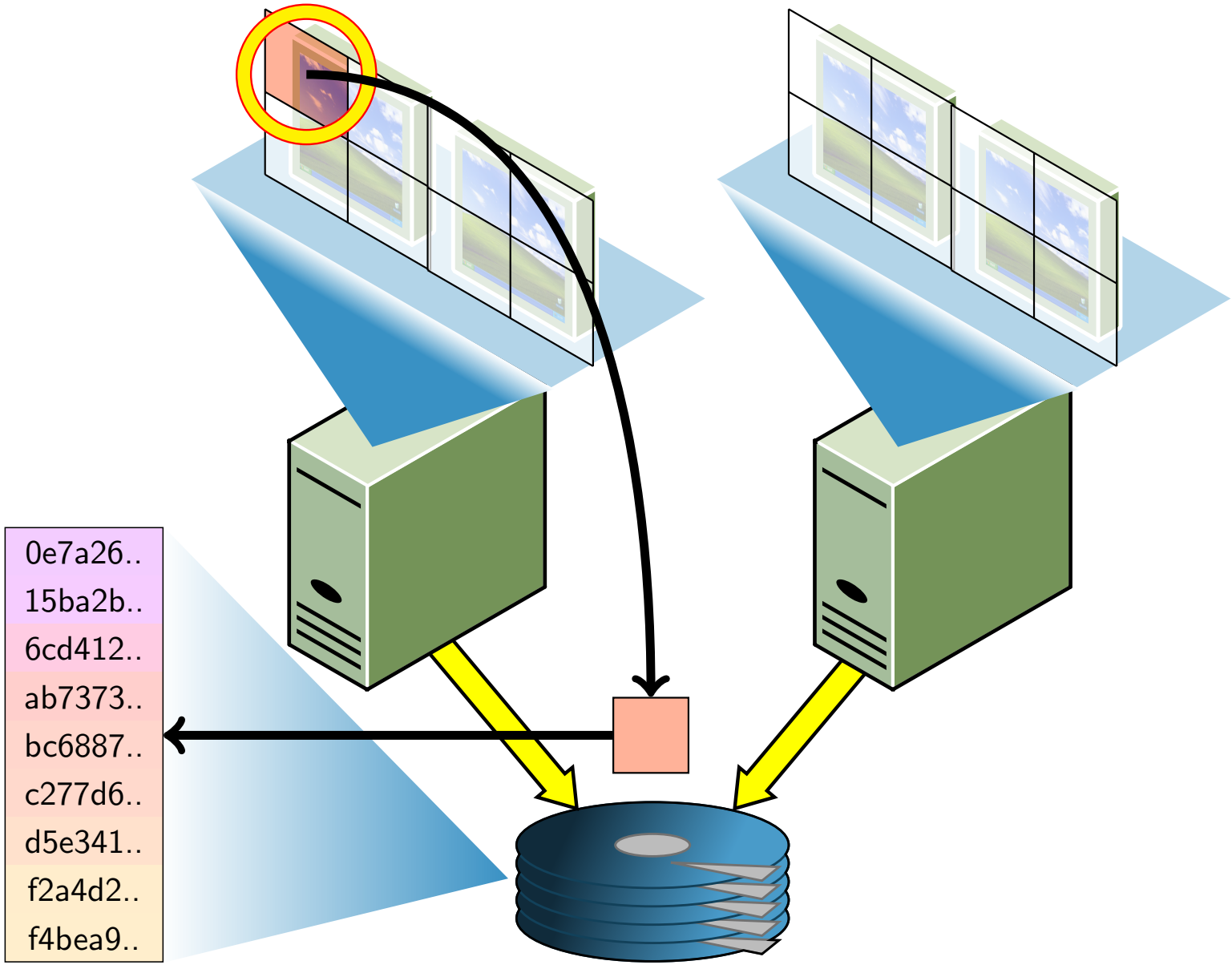
Deduplication



Deduplication



Deduplication



The Problem

Slow — Lots of IO in the write path. Can't cache the index.

The Problem

Slow — Lots of IO in the write path. Can't cache the index.

Very Slow — Writes require allocation and thus coordination.
No hope of disk locality.

The Problem

Slow — Lots of IO in the write path. Can't cache the index.

Very Slow — Writes require allocation and thus coordination.
No hope of disk locality.

Hopelessly Slow — Multi-host lock contention on shared index.

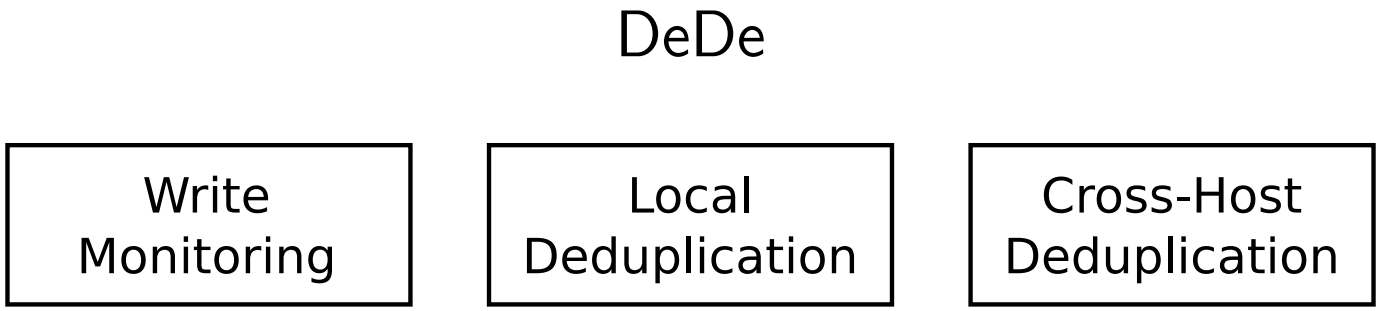
Three-Stage Deduplication

Decentralized Deduplication

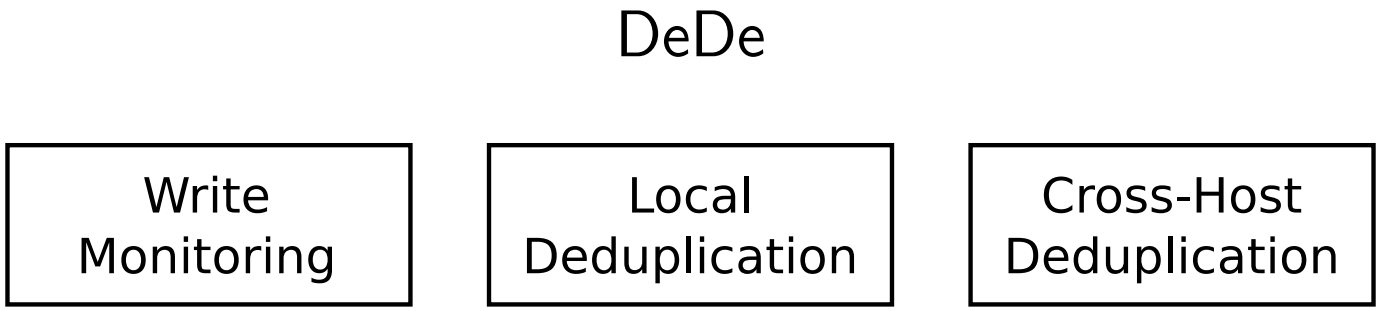
Three-Stage Deduplication

DeDe

Three-Stage Deduplication



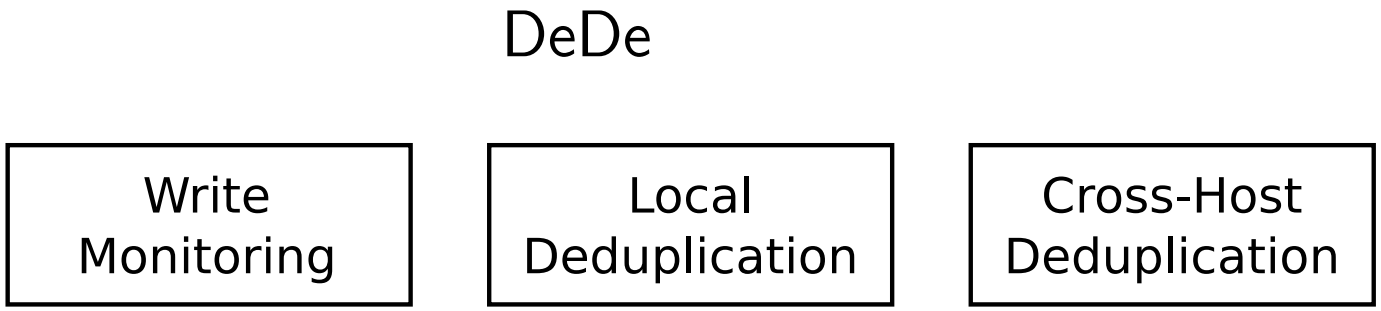
Three-Stage Deduplication



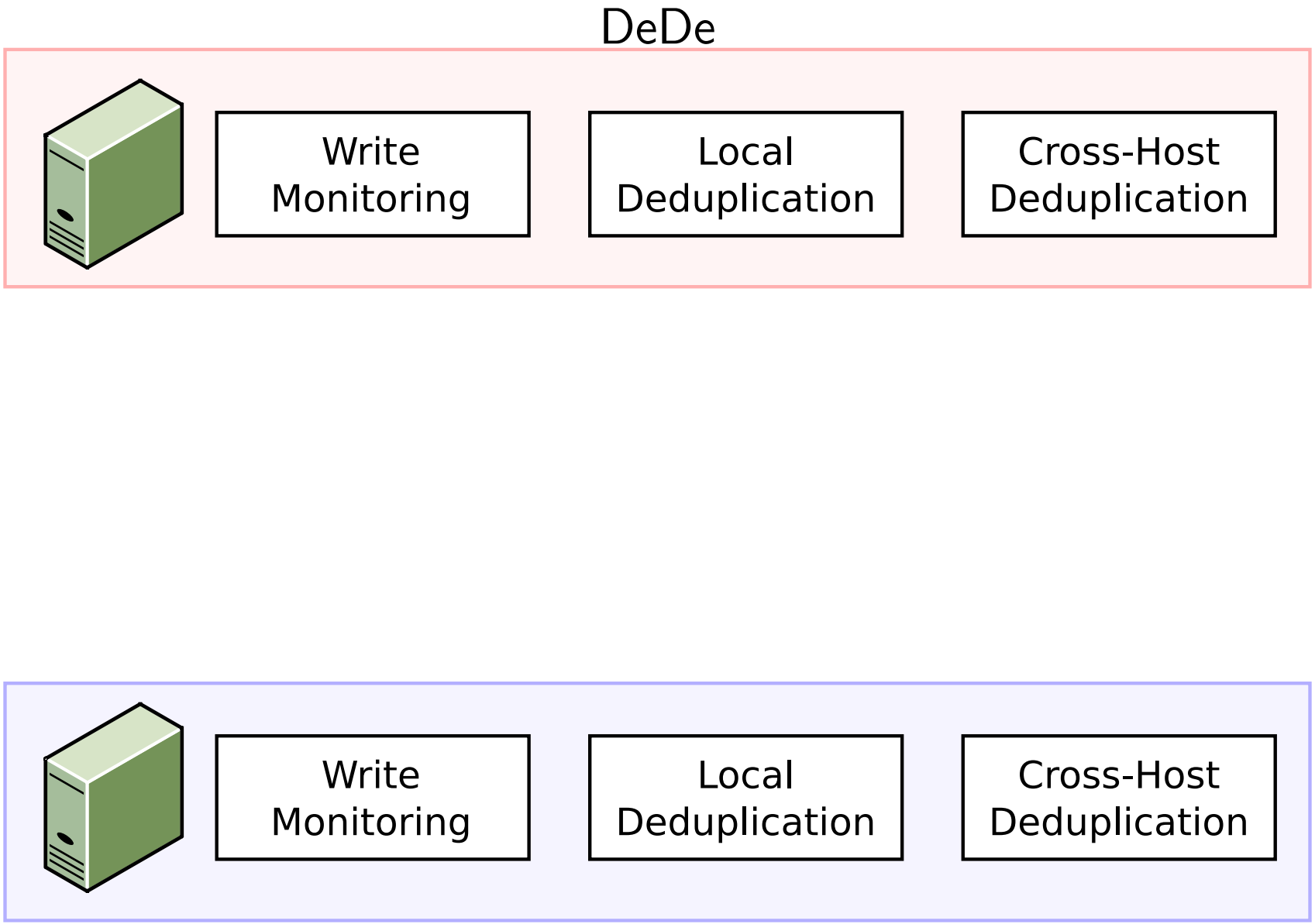
- Out-of-band deduplication of live, primary storage
- Process duplicates efficiently, in large batches
- Minimize contention on the index
- Resilient to stale index information
- Unique blocks remain mutable and sequential

$\Longrightarrow$ No overhead for blocks that don't benefit from deduplication

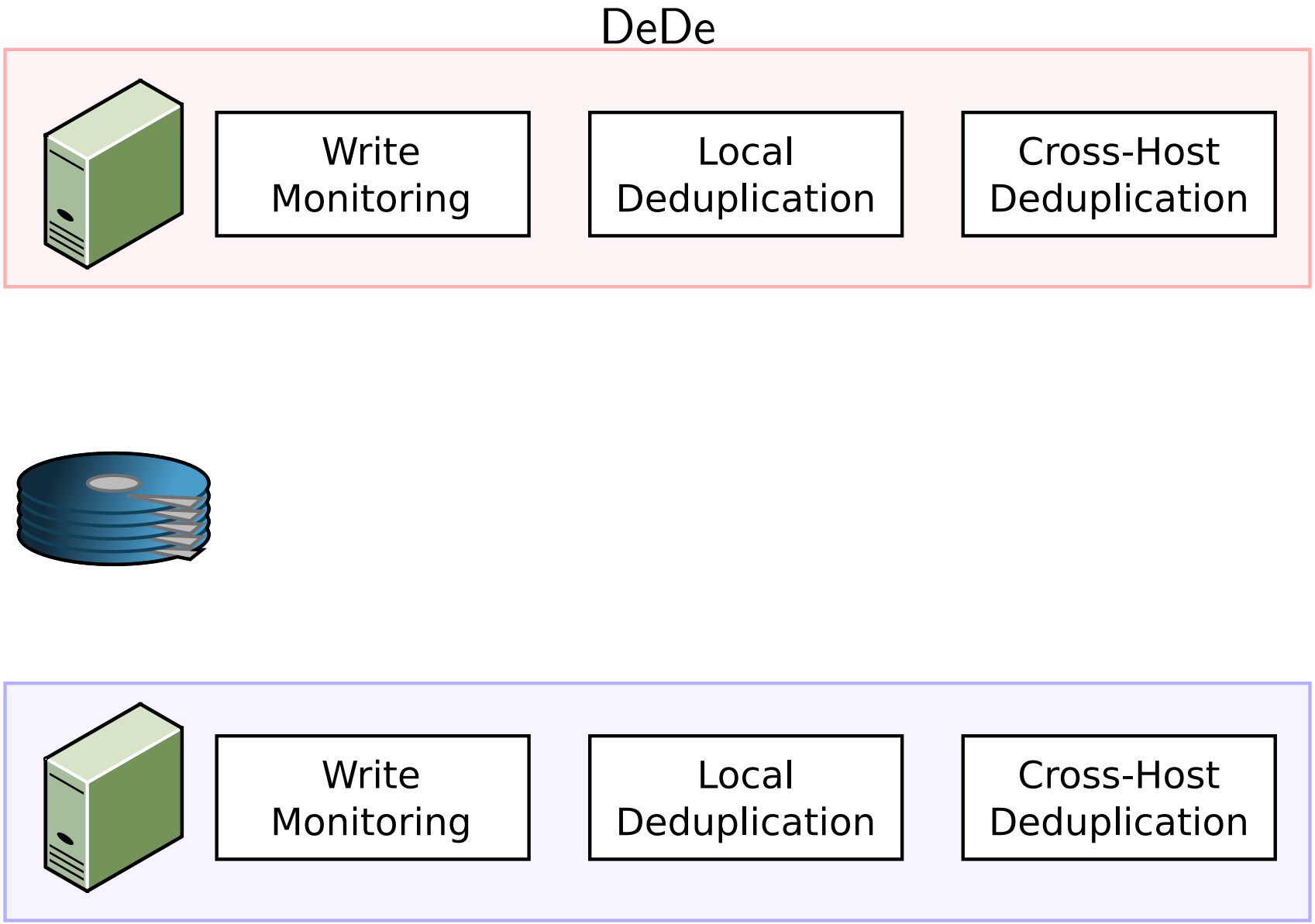
Three-Stage Deduplication



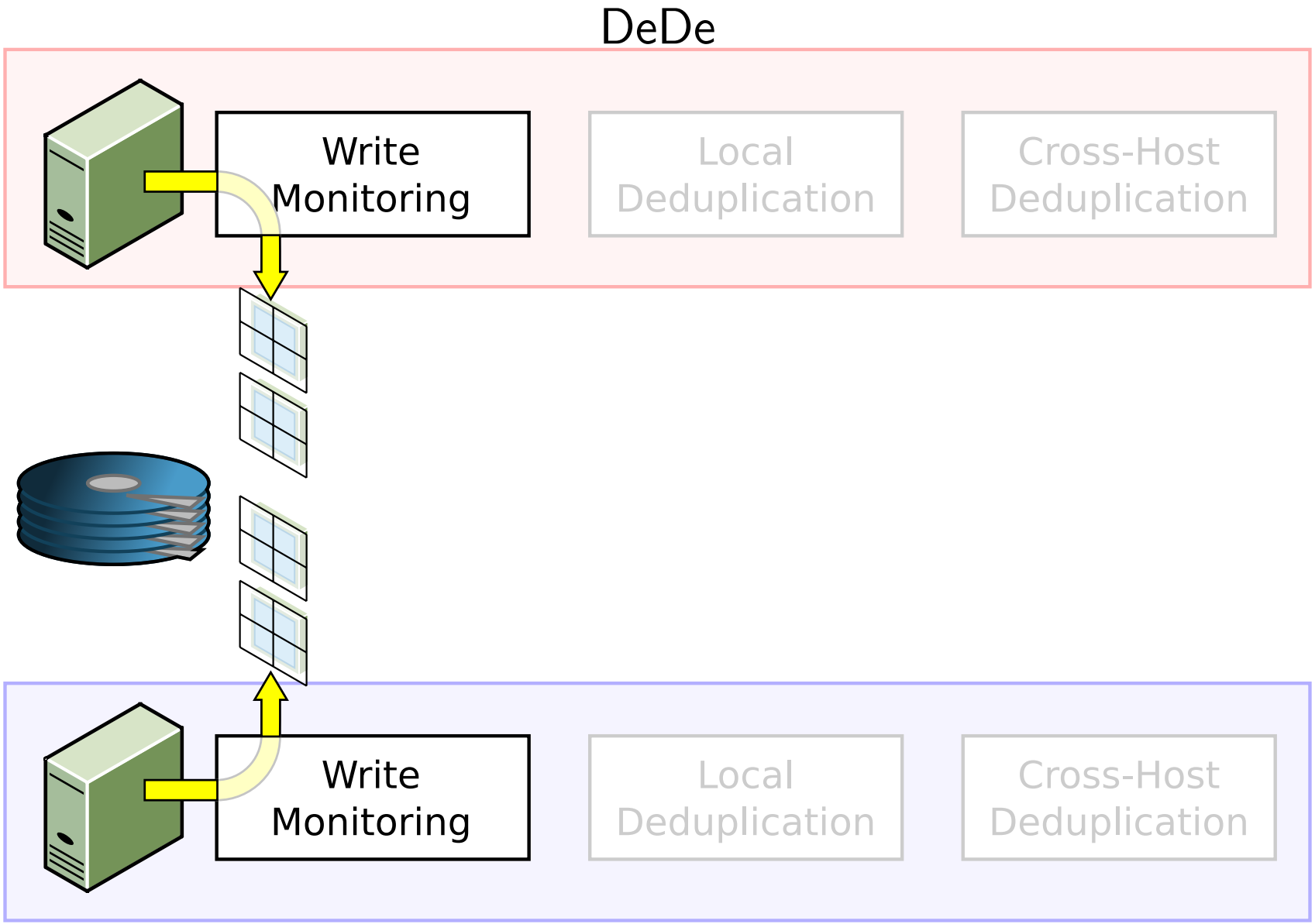
Three-Stage Deduplication



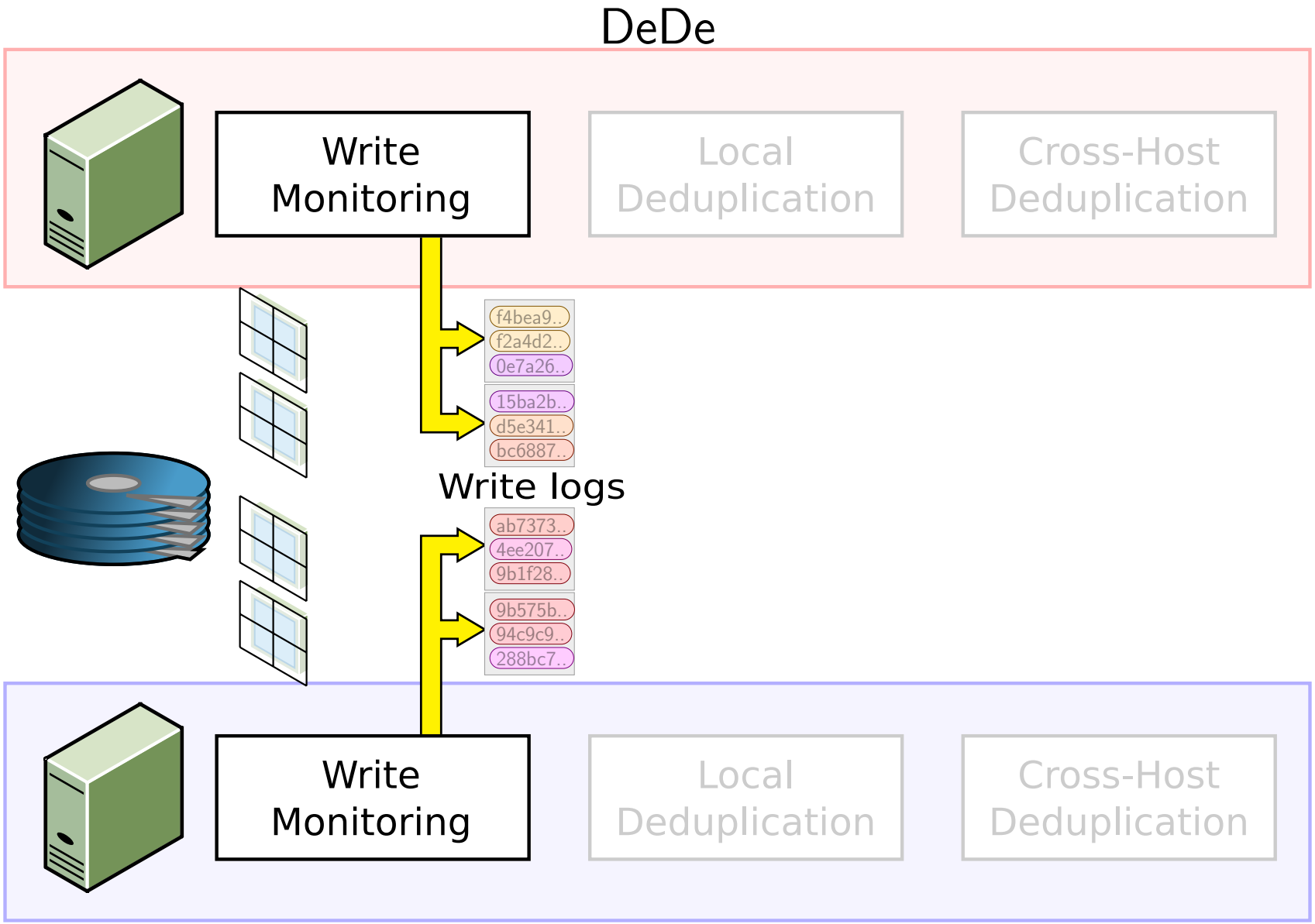
Three-Stage Deduplication



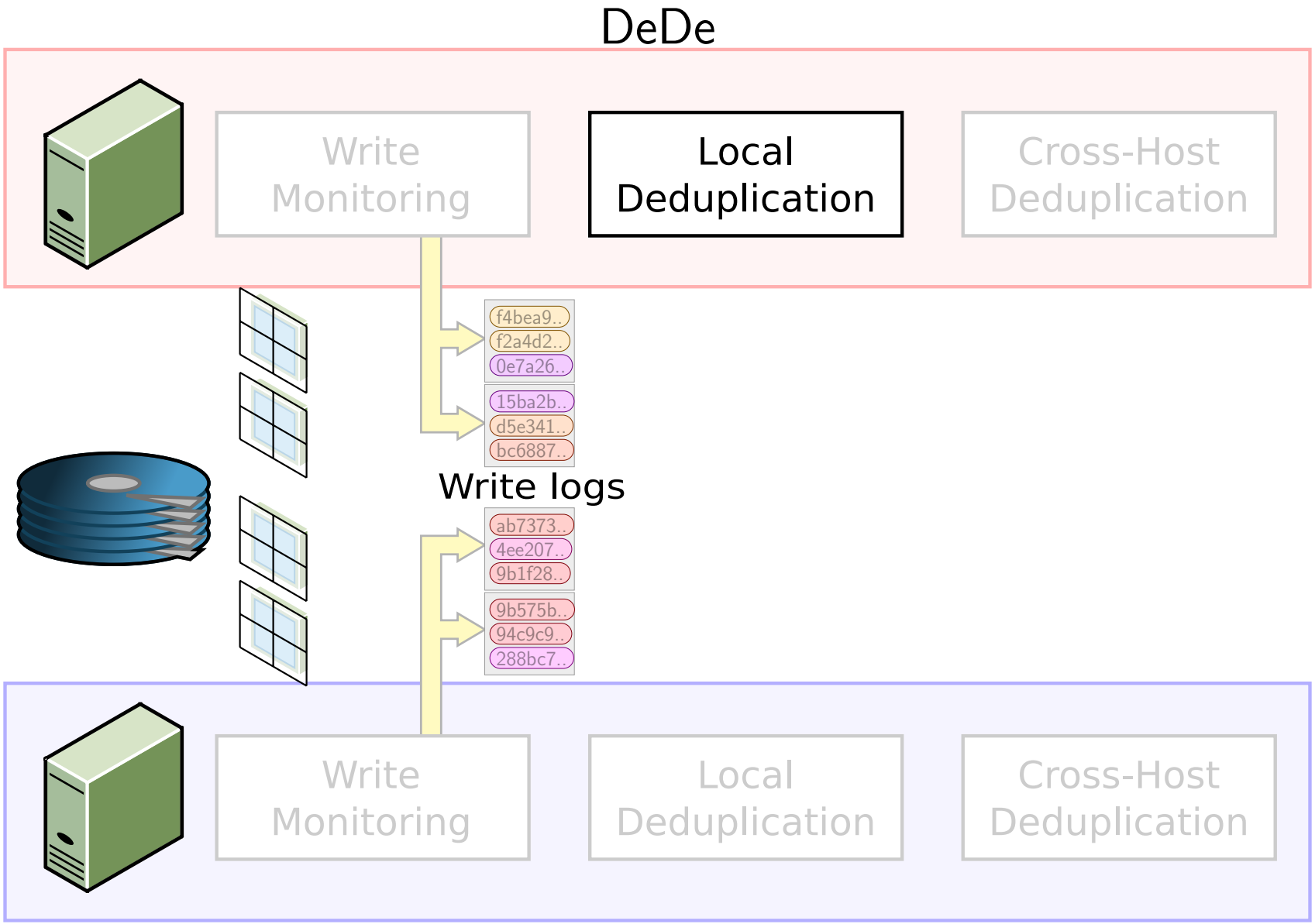
Three-Stage Deduplication



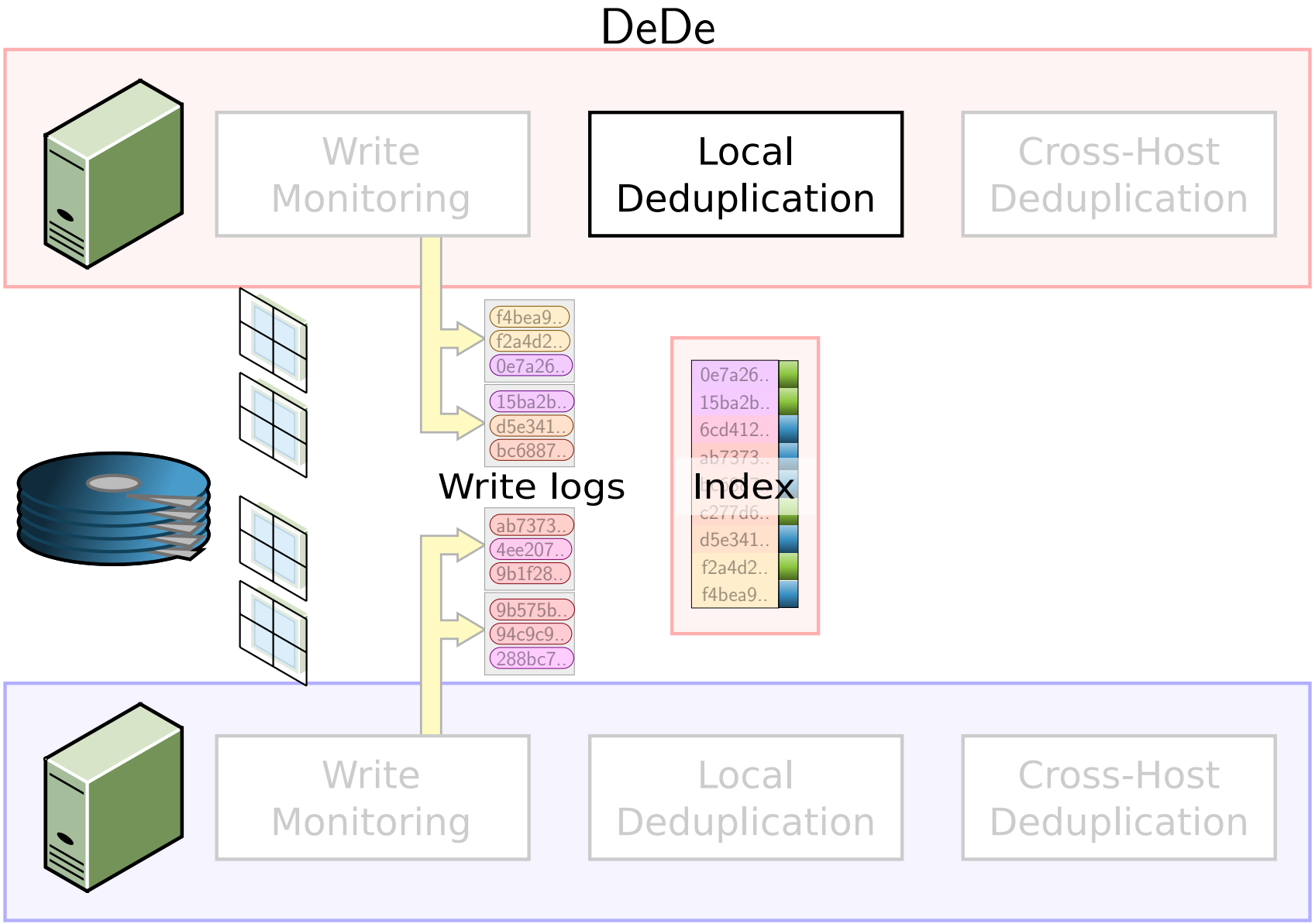
Three-Stage Deduplication



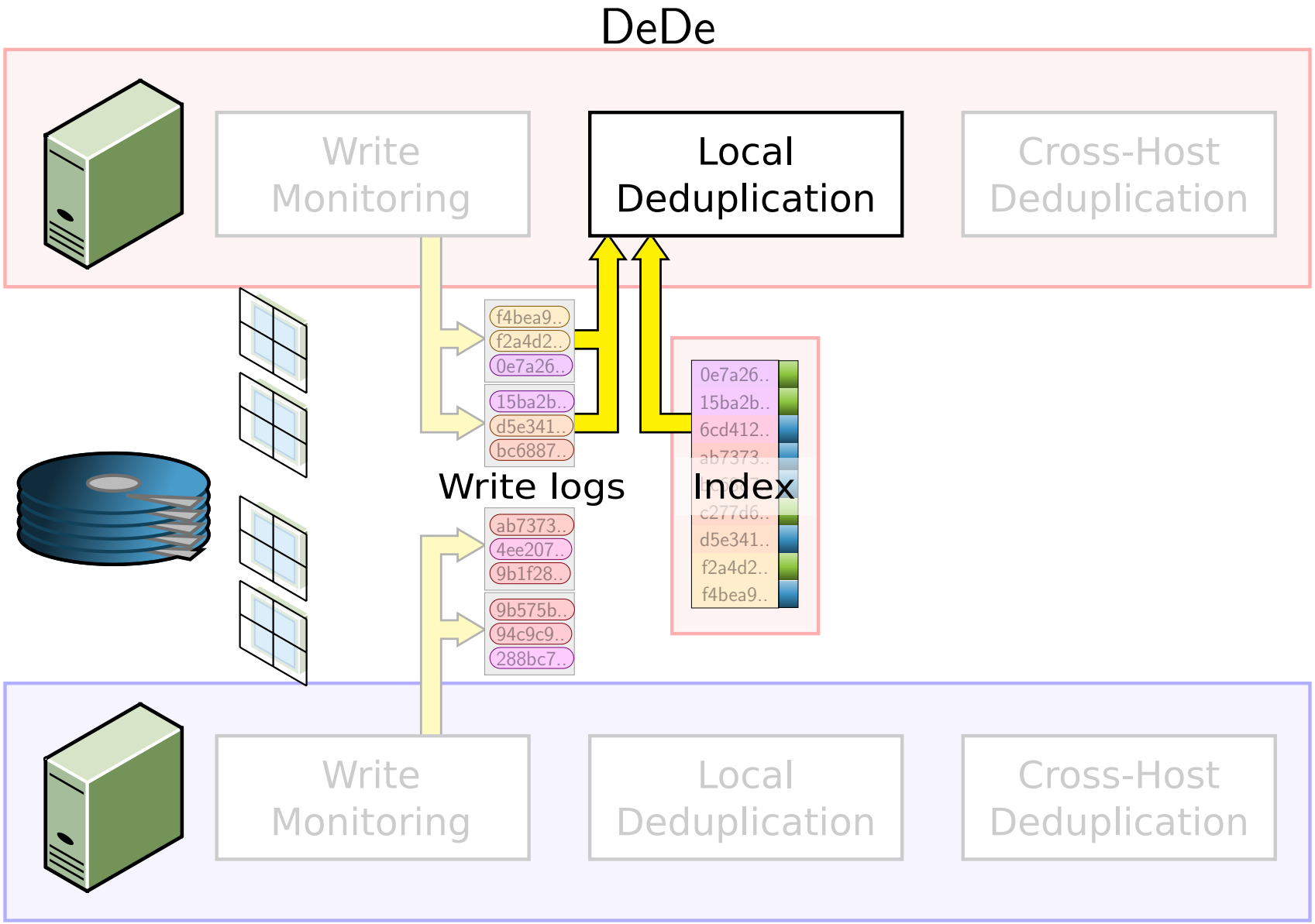
Three-Stage Deduplication



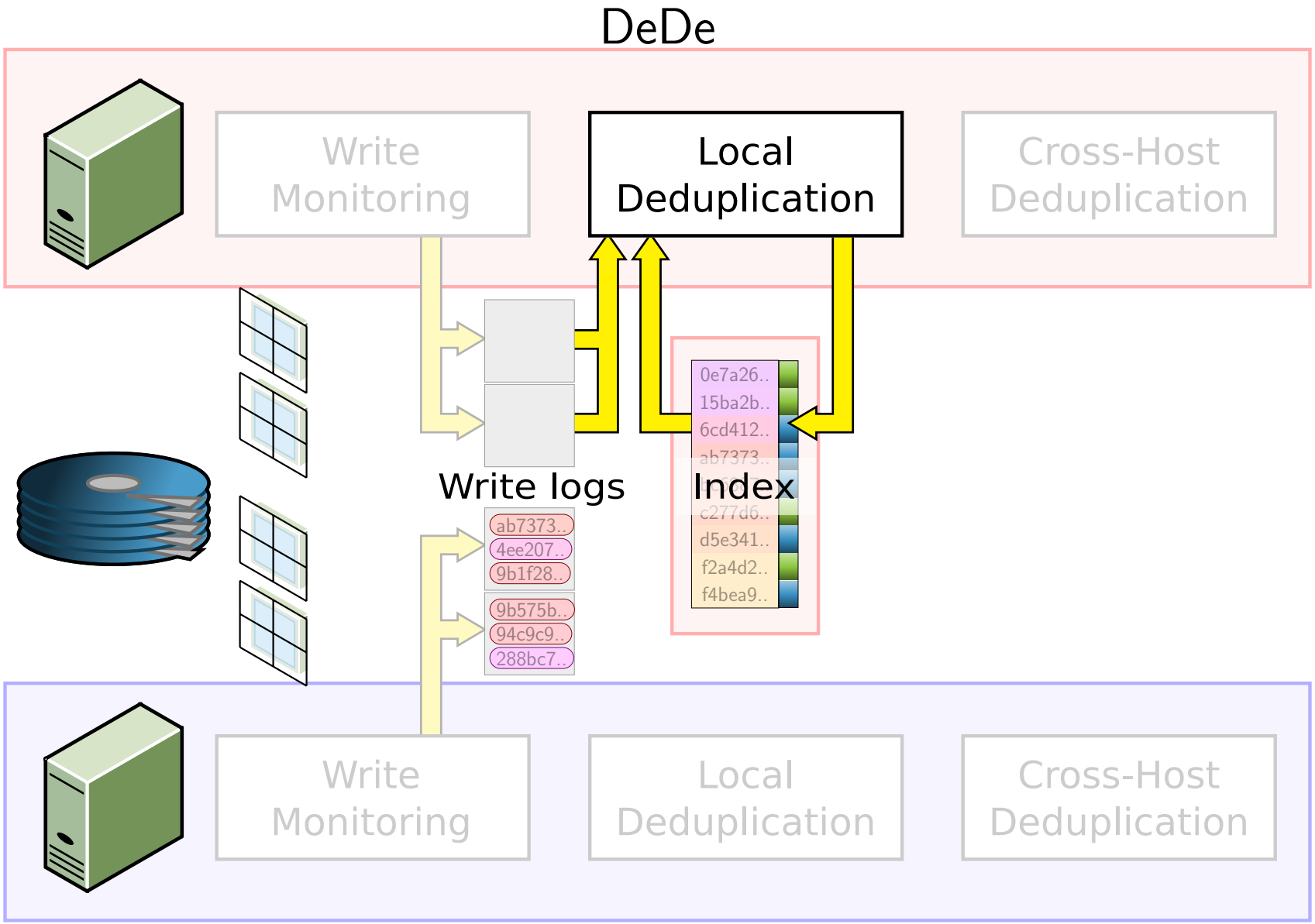
Three-Stage Deduplication



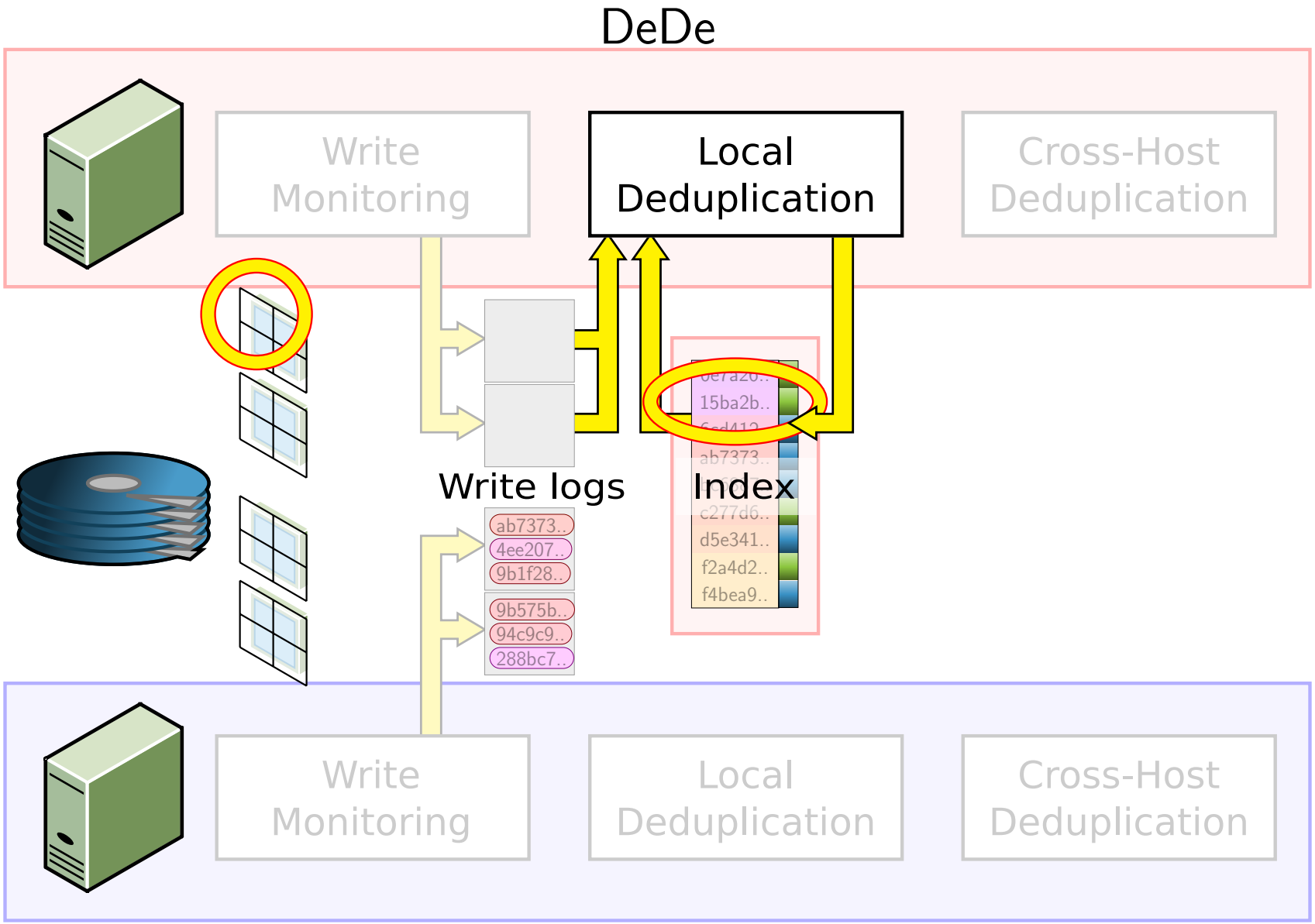
Three-Stage Deduplication



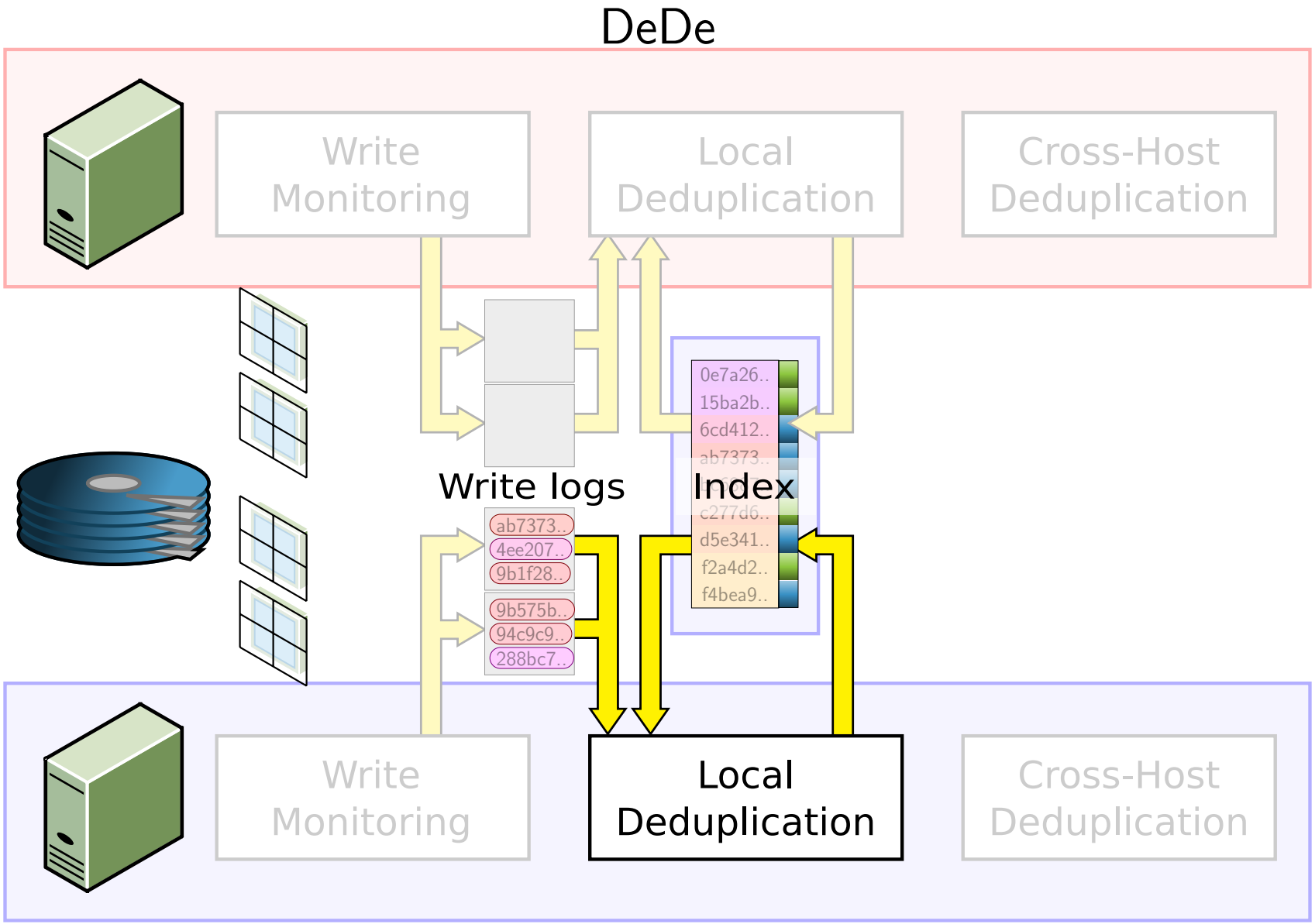
Three-Stage Deduplication



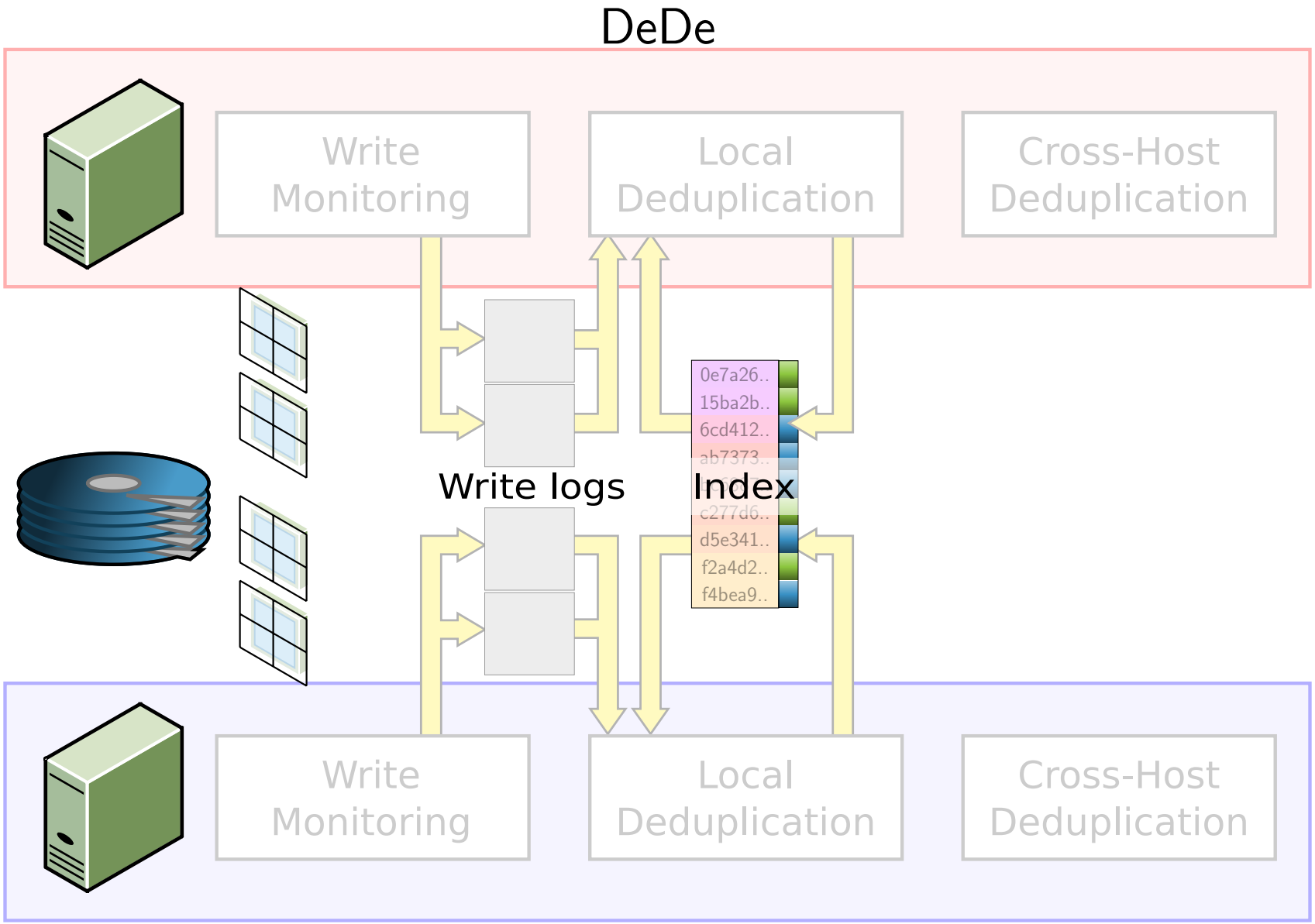
Three-Stage Deduplication



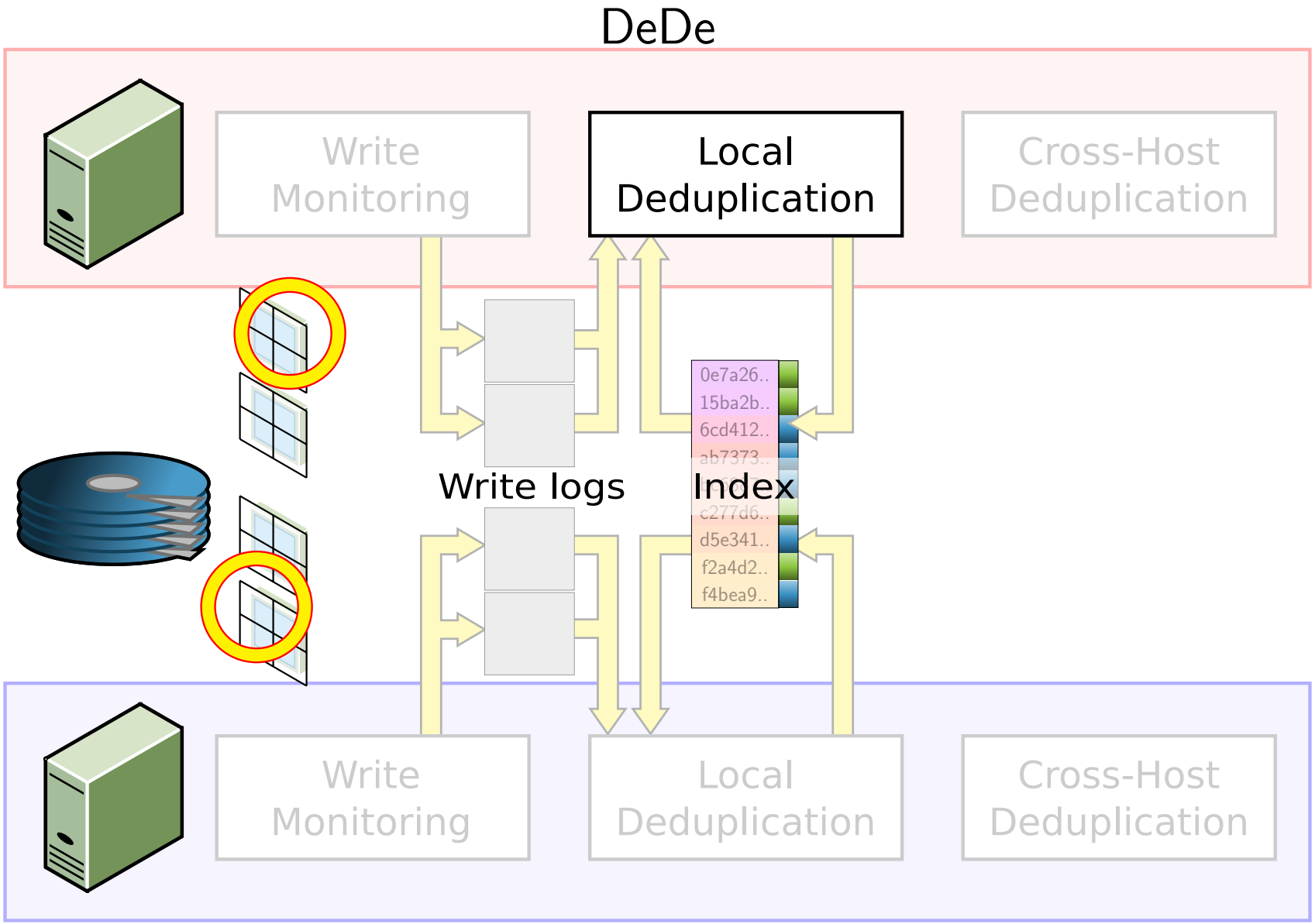
Three-Stage Deduplication



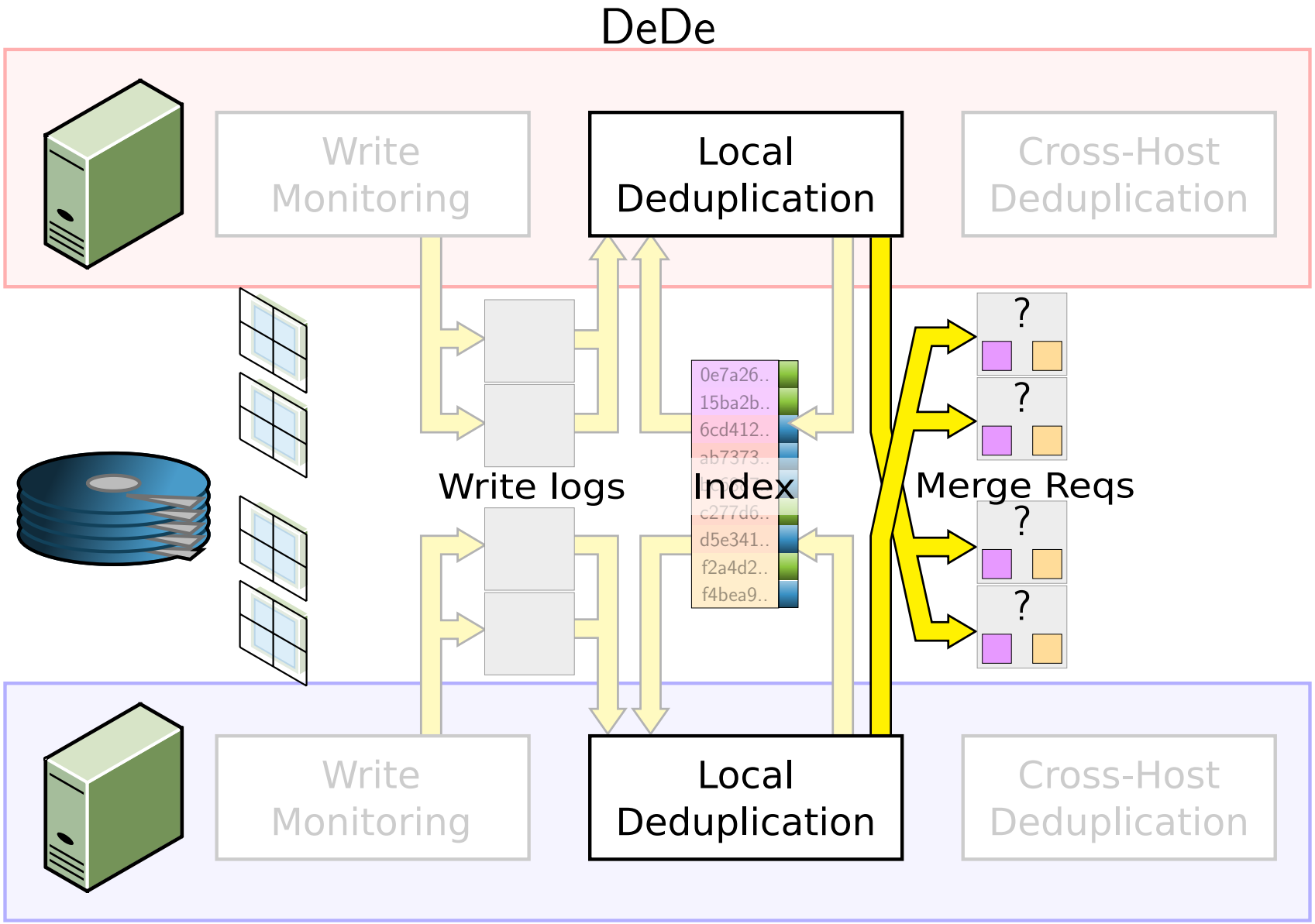
Three-Stage Deduplication



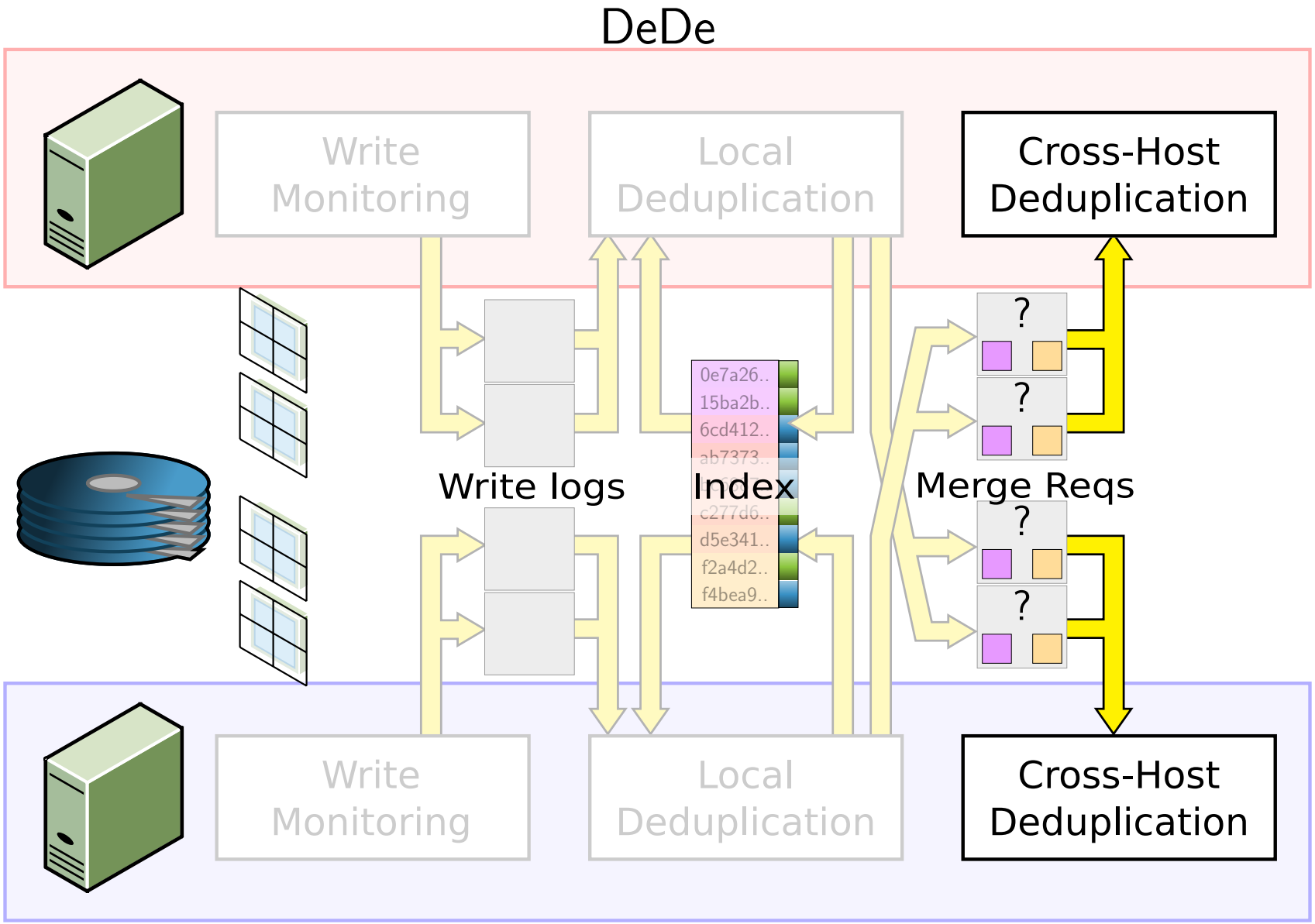
Three-Stage Deduplication



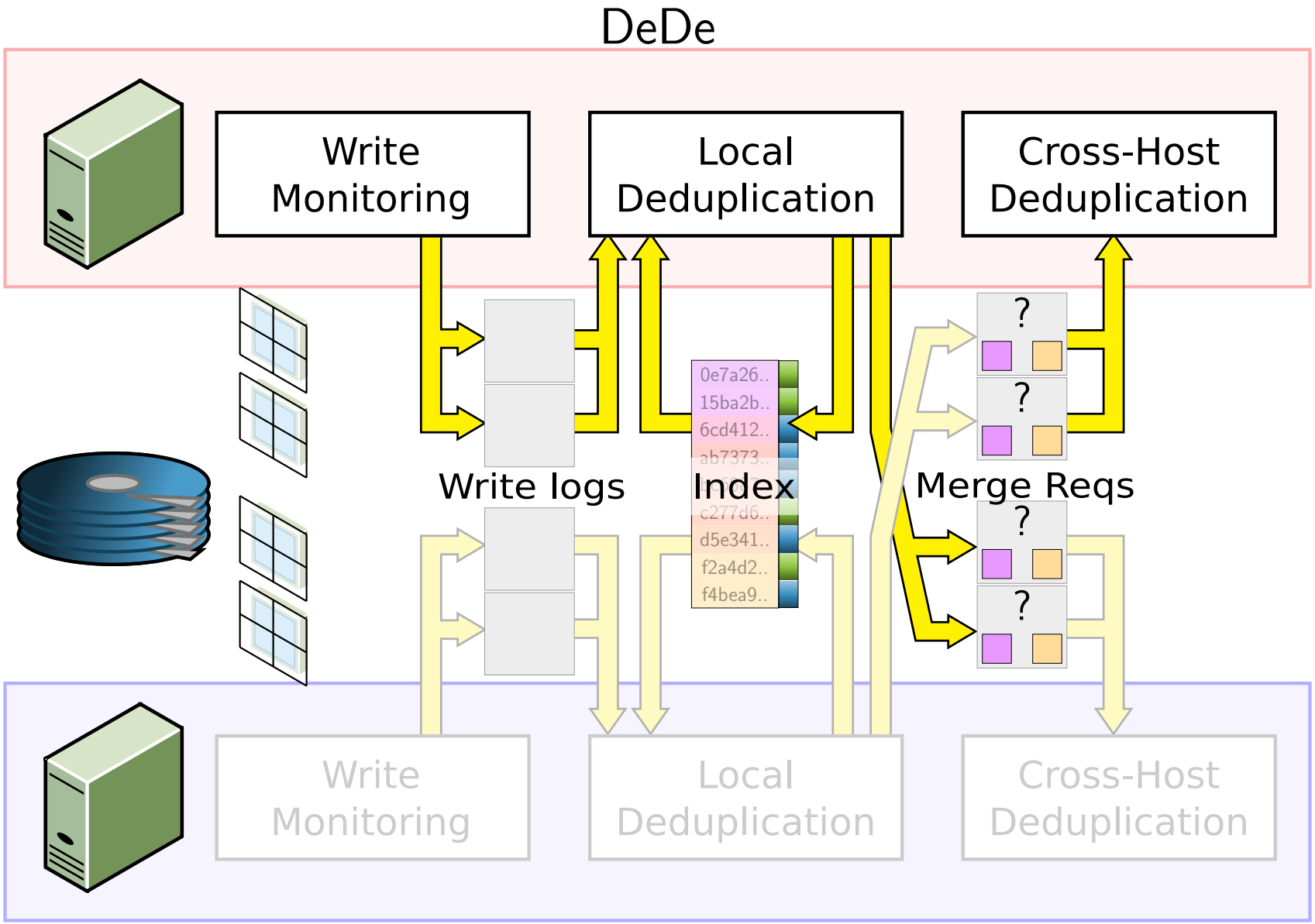
Three-Stage Deduplication



Three-Stage Deduplication

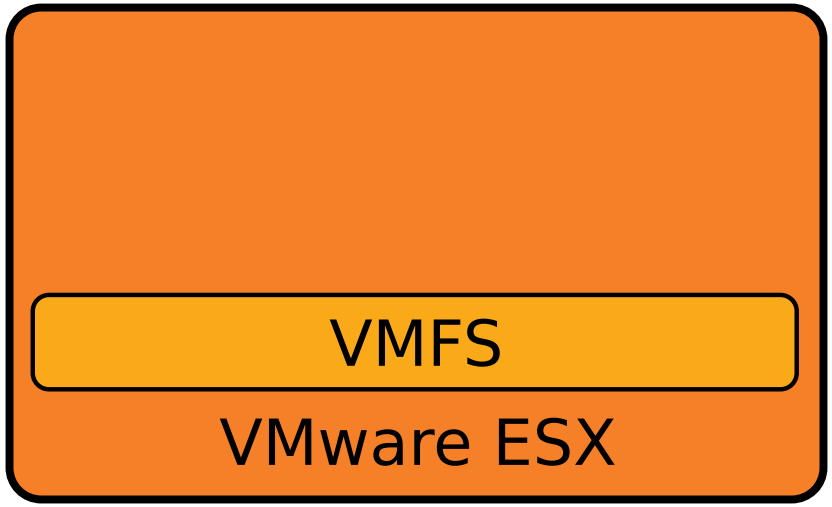


Three-Stage Deduplication

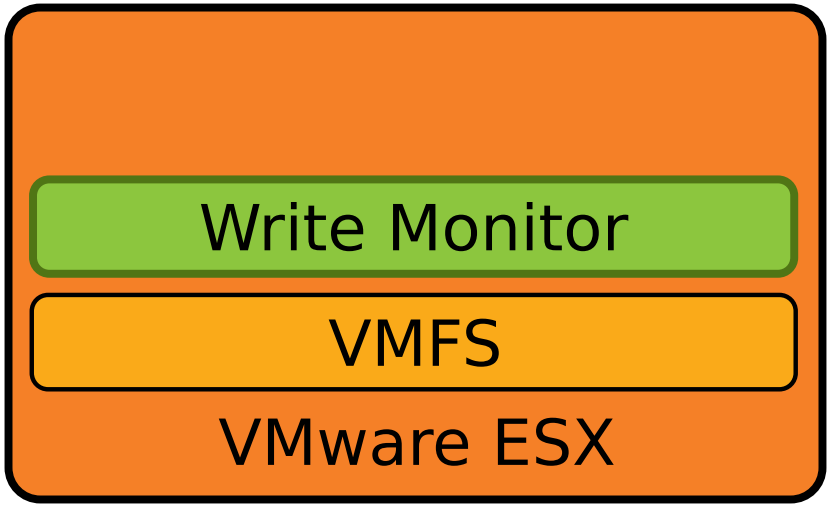


Implementation

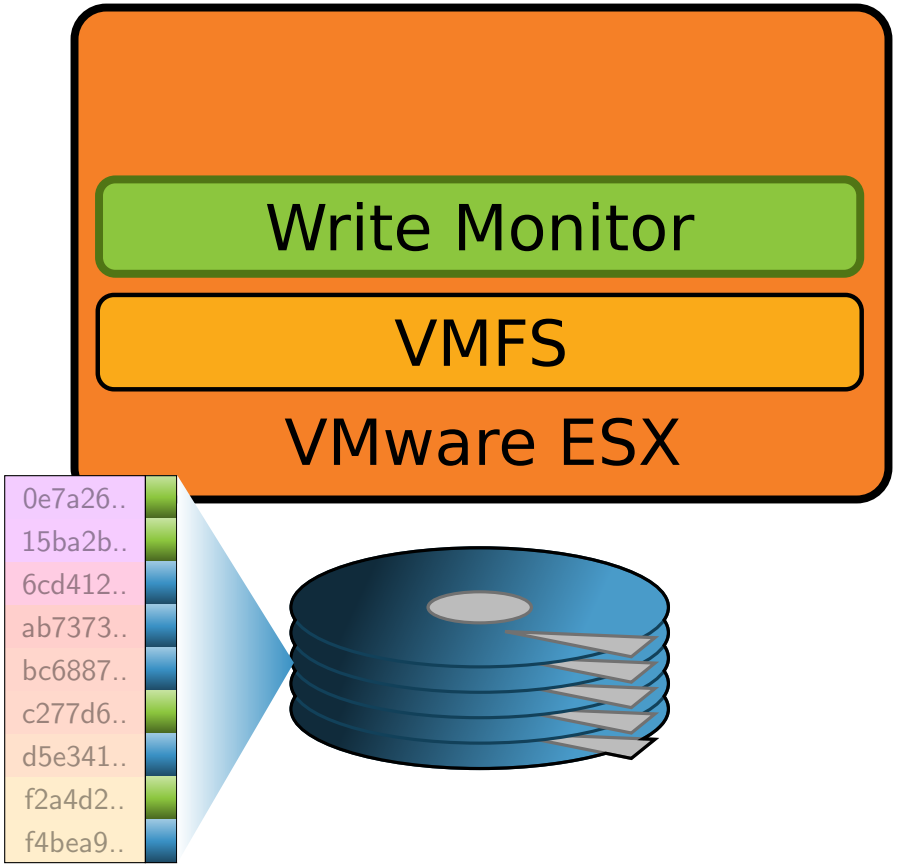
Implementation



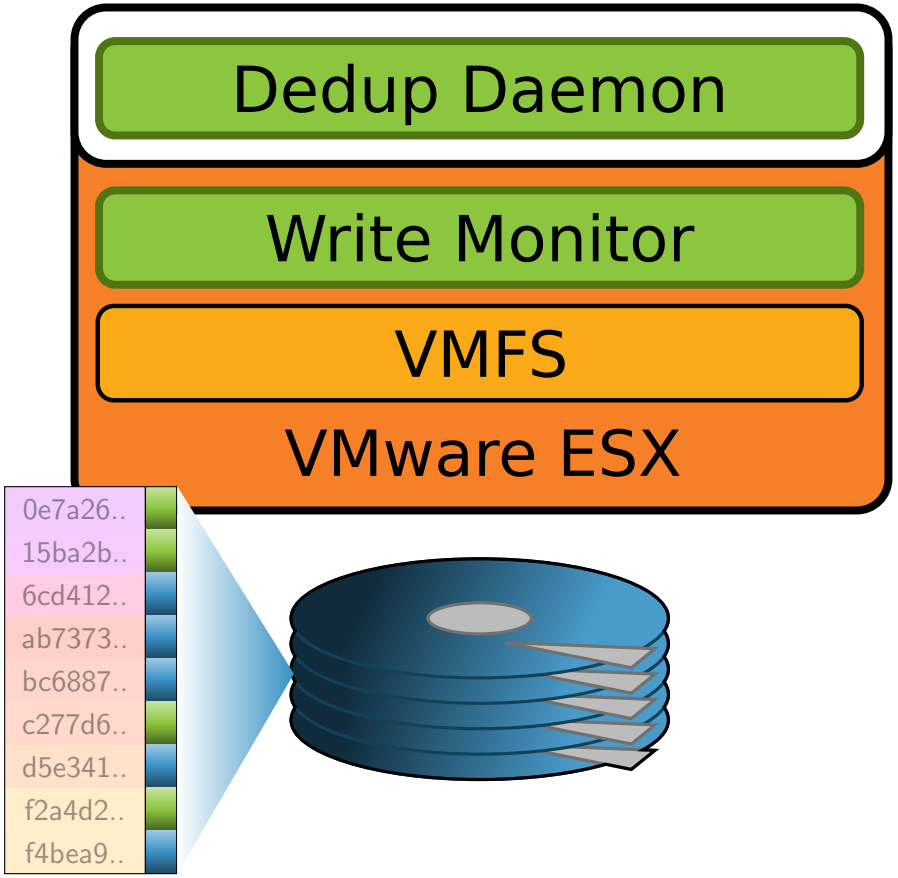
Implementation



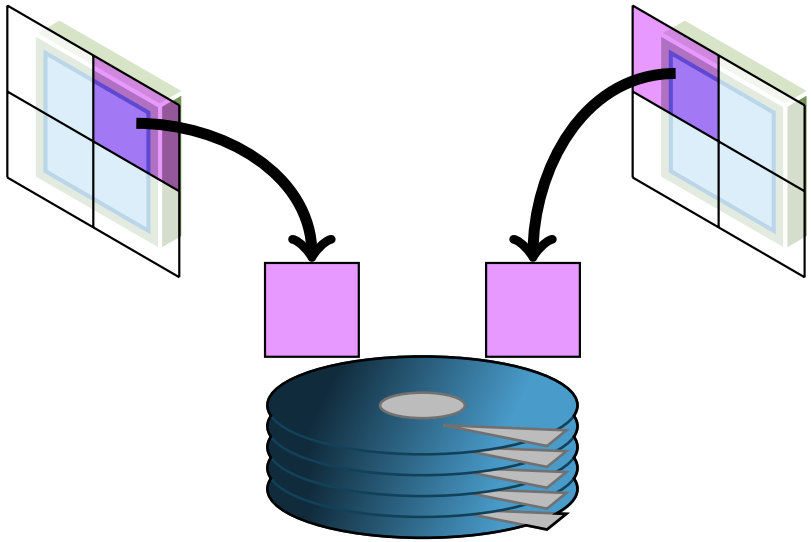
Implementation



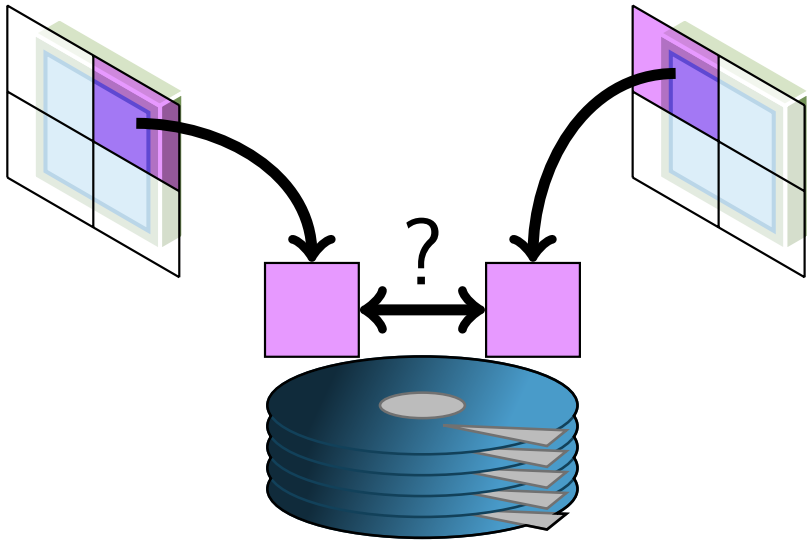
Implementation



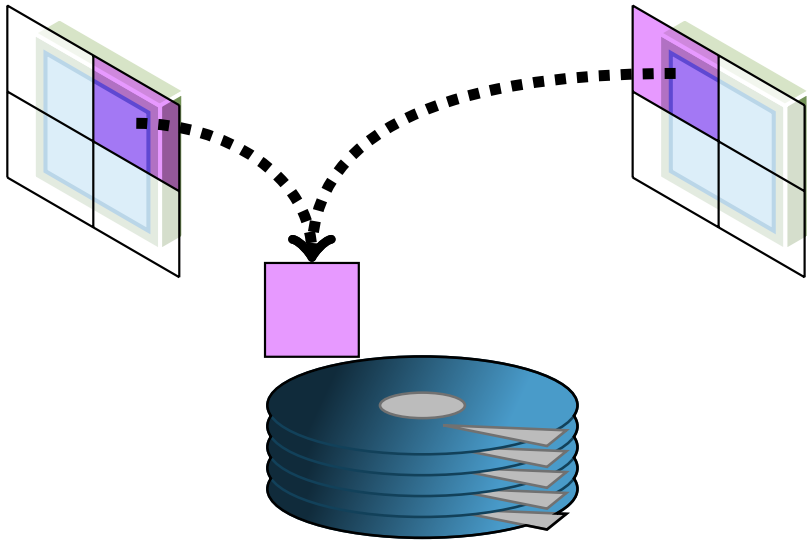
Compare-and-share



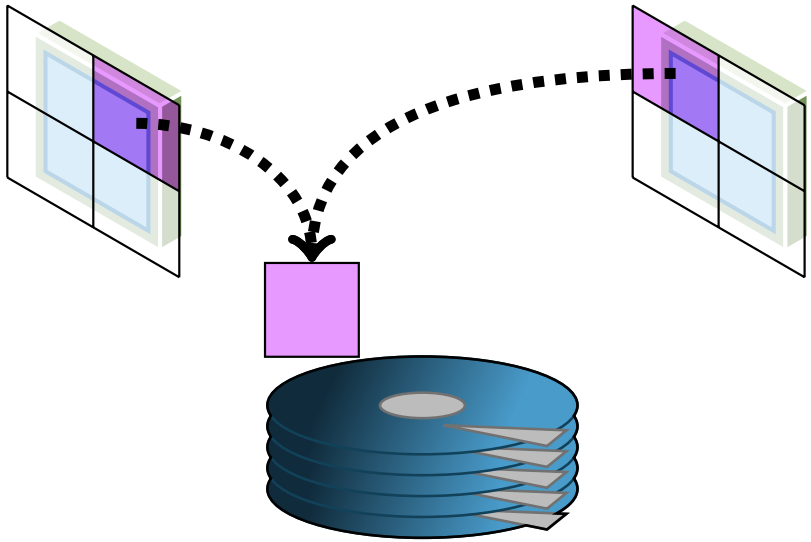
Compare-and-share



Compare-and-share

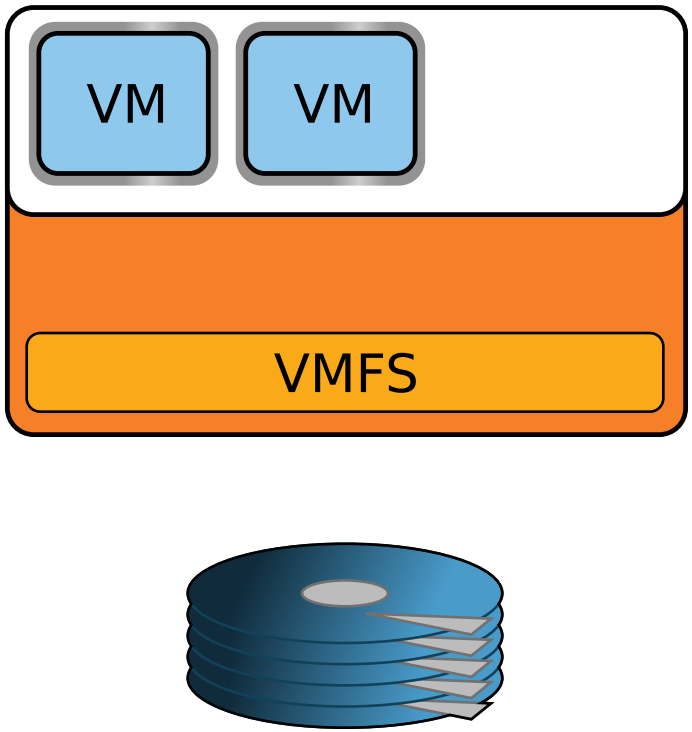


Compare-and-share

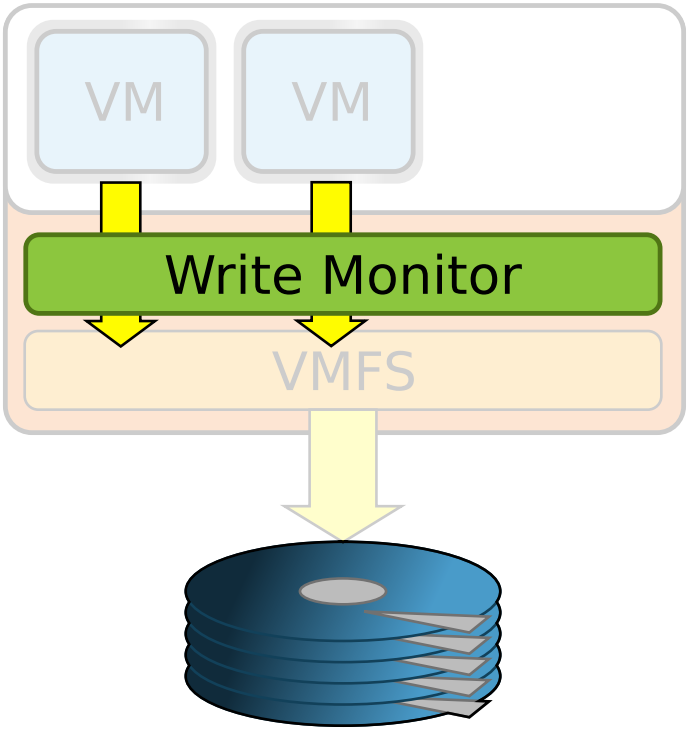


DeDe finds duplicates. VMFS eliminates them.

Write Monitor

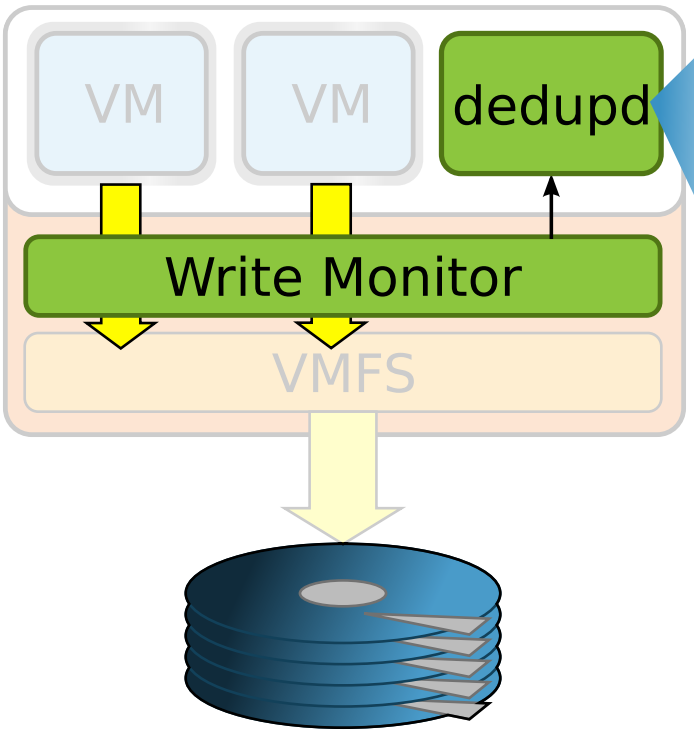


Write Monitor



- A lightweight kernel module monitors writes, computes hashes

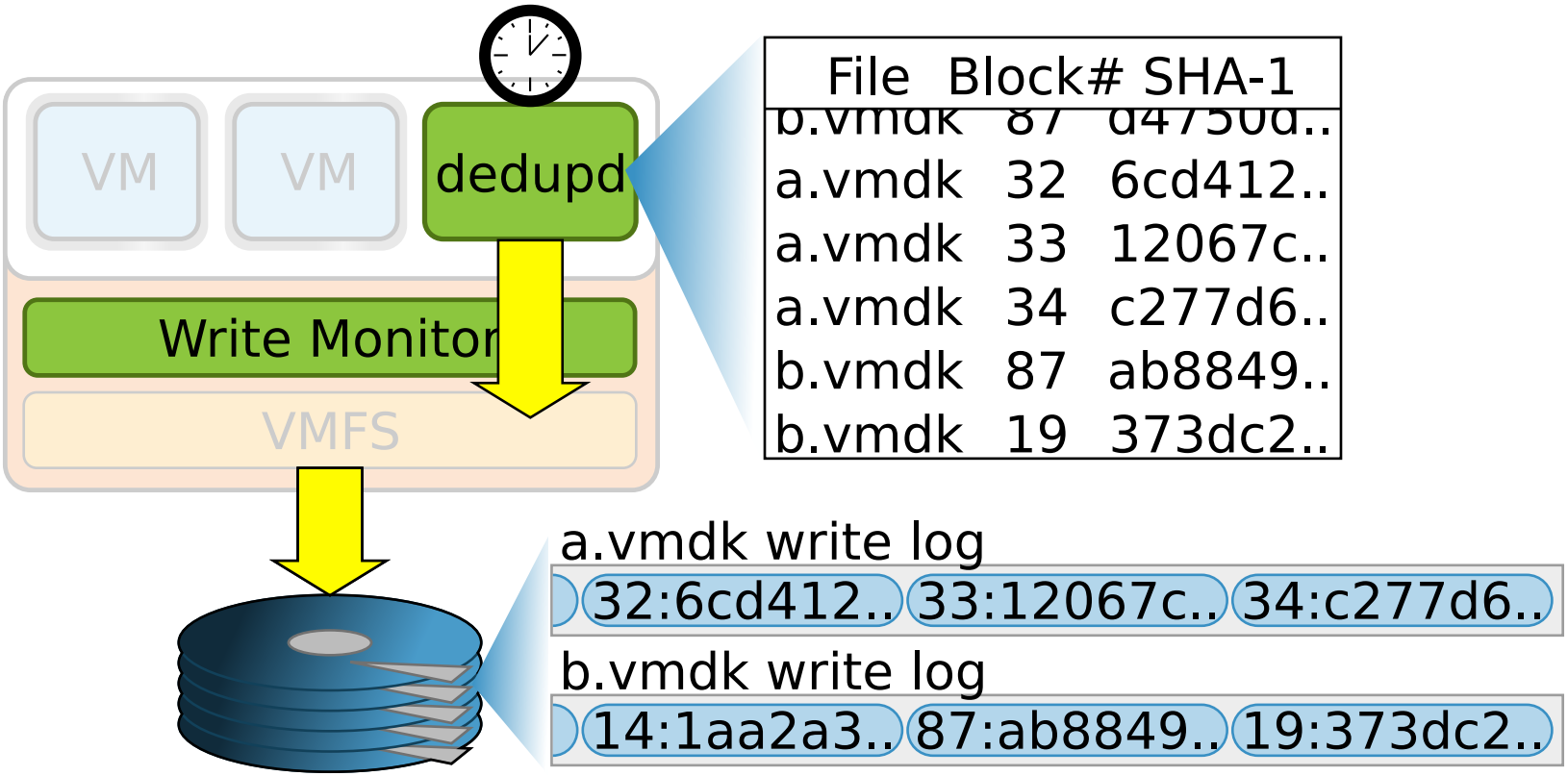
Write Monitor



File	Block#	SHA-1
b.vmdk	87	d4750d..
a.vmdk	32	6cd412..
a.vmdk	33	12067c..
a.vmdk	34	c277d6..
b.vmdk	87	ab8849..
b.vmdk	19	373dc2..

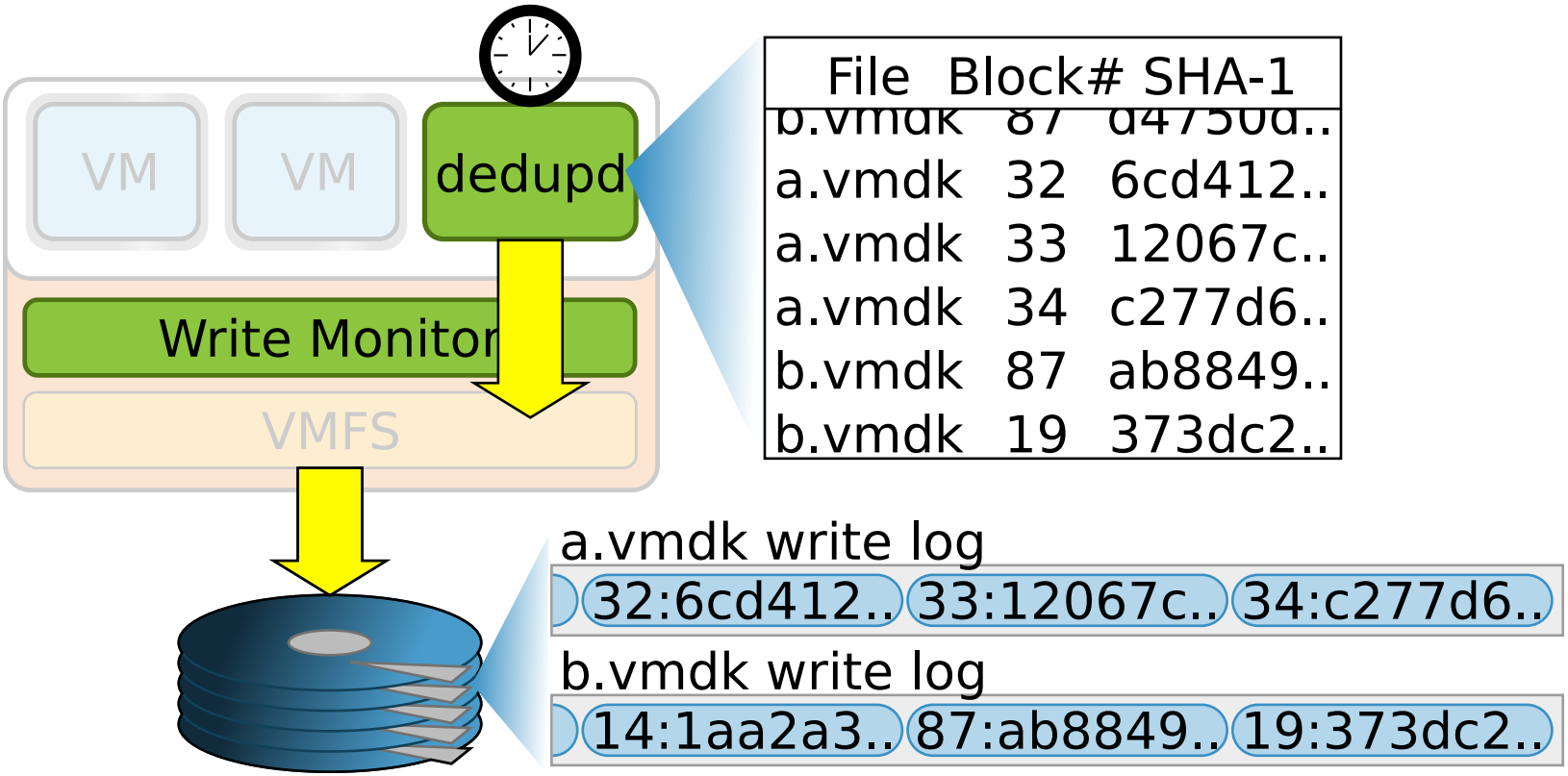
- A lightweight kernel module monitors writes, computes hashes
- It buffers the write log in userspace before writing it to disk

Write Monitor



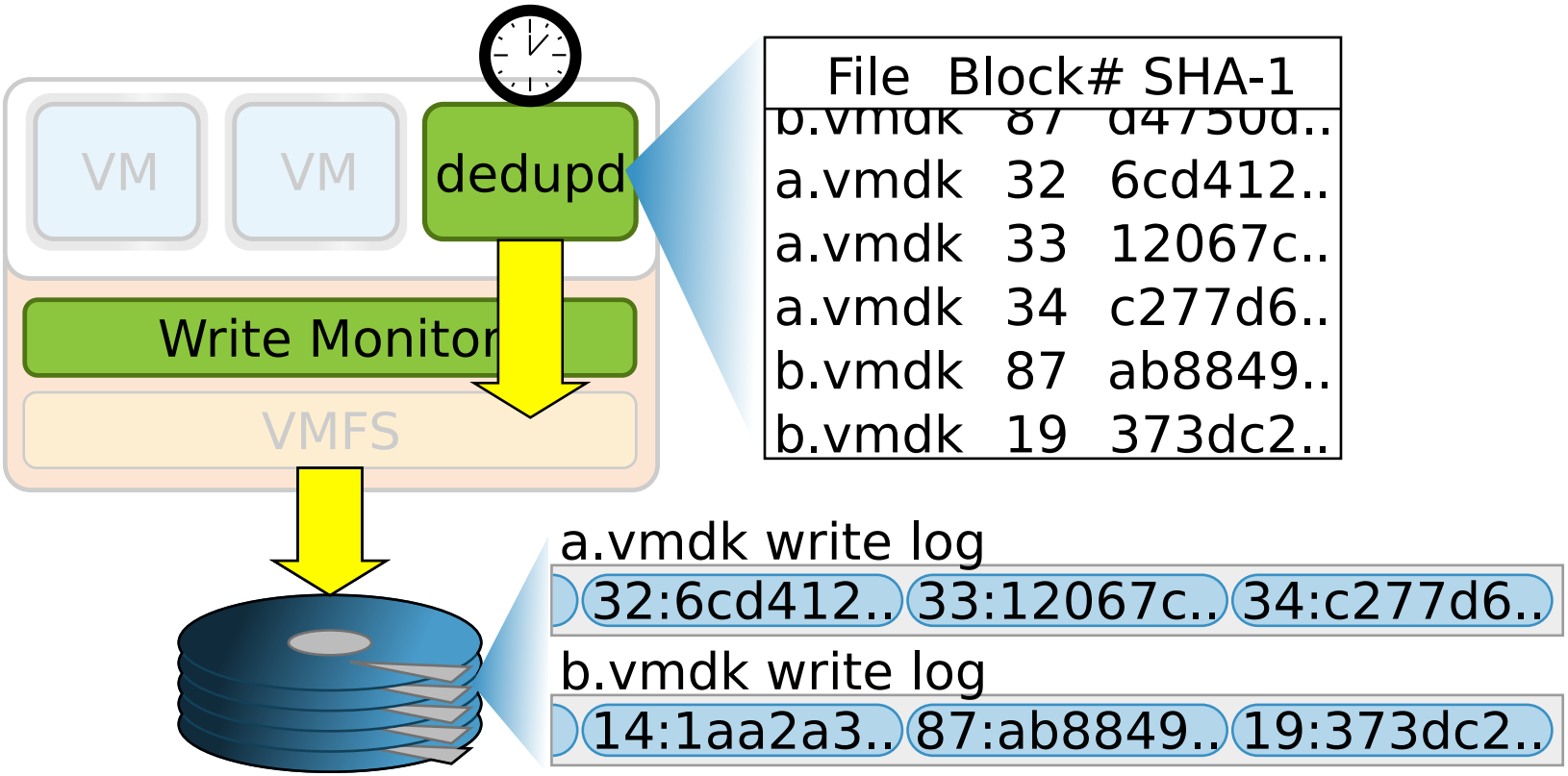
- A lightweight kernel module monitors writes, computes hashes
- It buffers the write log in userspace before writing it to disk

Write Monitor



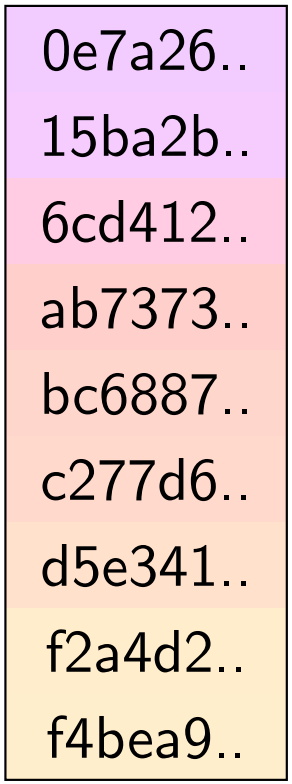
- A lightweight kernel module monitors writes, computes hashes
- It buffers the write log in userspace before writing it to disk
- Safe to buffer the log because index is resilient

Write Monitor



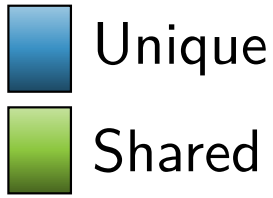
- A lightweight kernel module monitors writes, computes hashes
- It buffers the write log in userspace before writing it to disk
- Safe to buffer the log because index is resilient
- 150 MB of regular writes $\longrightarrow$ 1 MB sequential log write

The Index



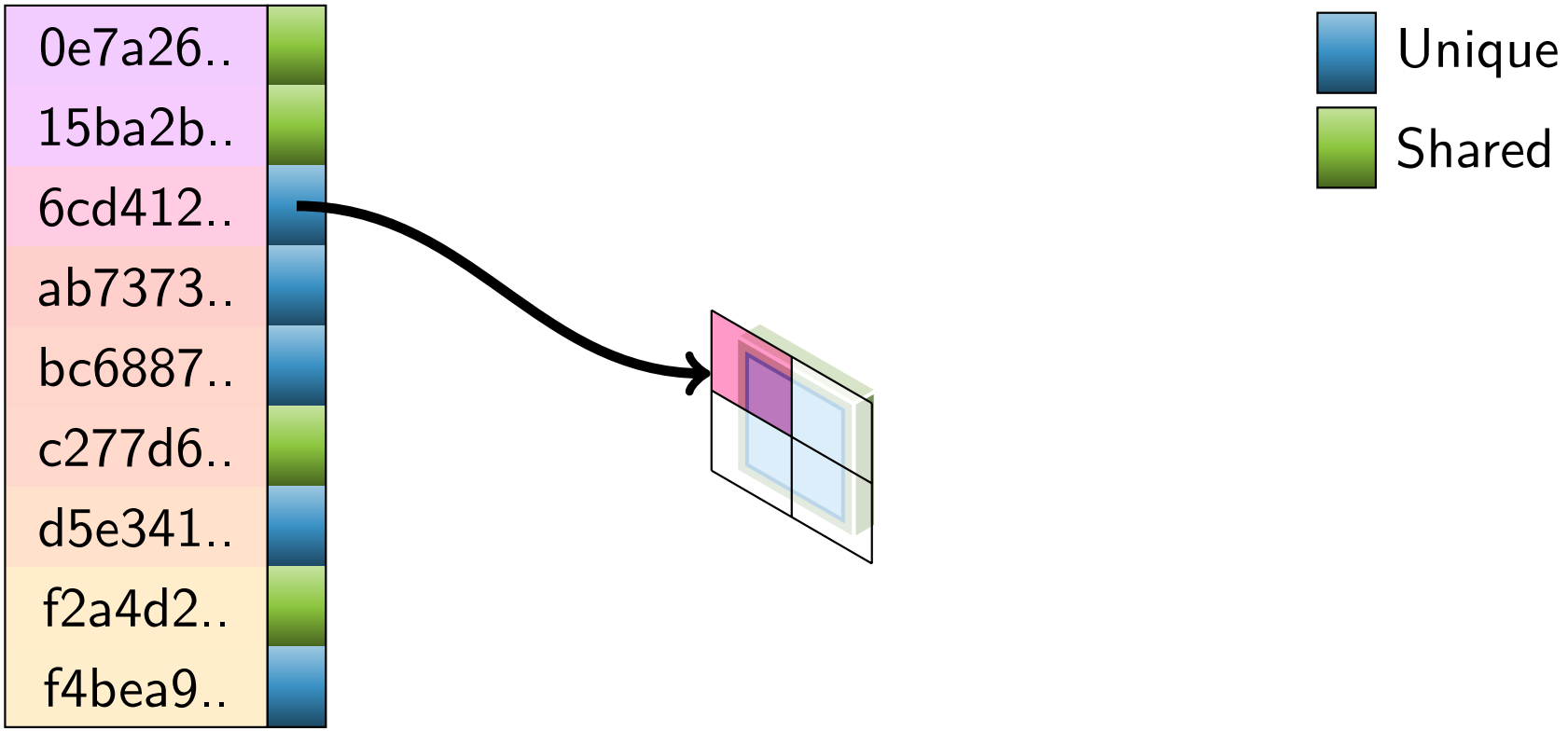
- Map from hashes to block locators, list sorted by hash

The Index



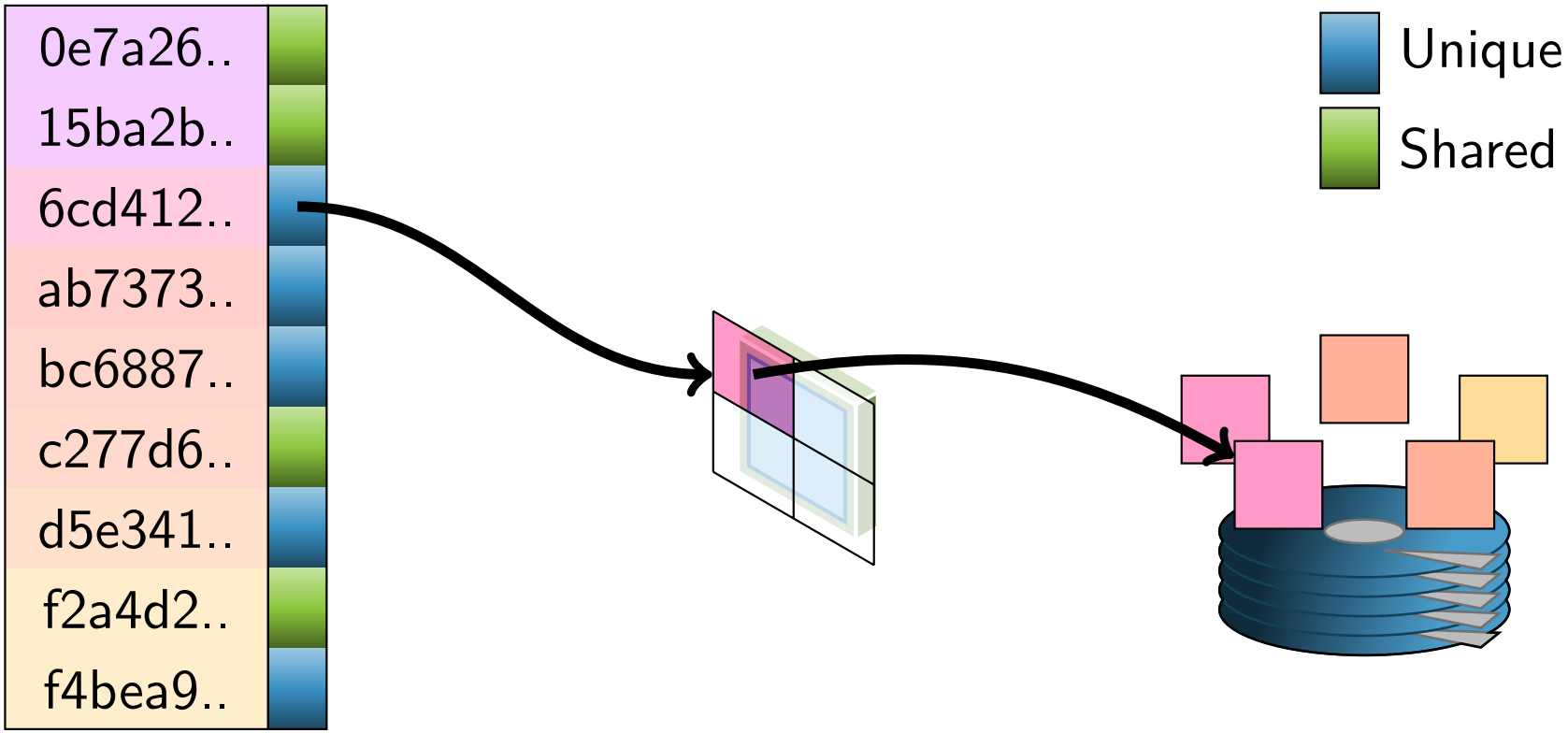
- Map from hashes to block locators, list sorted by hash

The Index



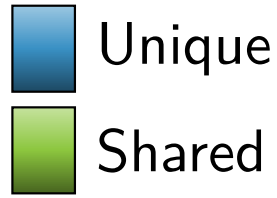
- Map from hashes to block locators, list sorted by hash
- Unique blocks are located in files and remain mutable

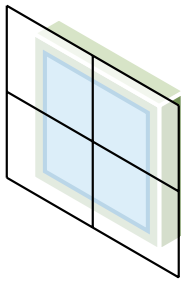
The Index

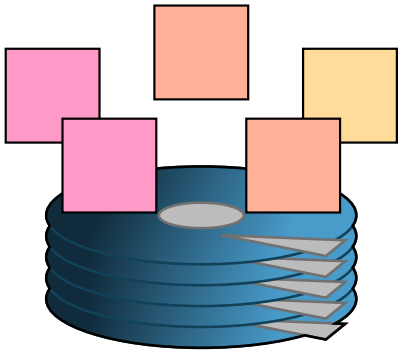


- Map from hashes to block locators, list sorted by hash
- Unique blocks are located in files and remain mutable

The Index

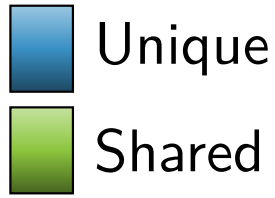


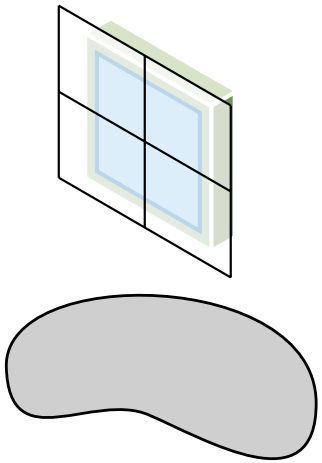


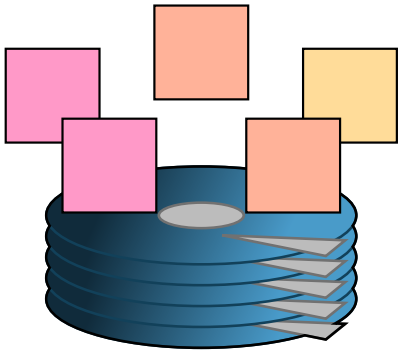


- Map from hashes to block locators, list sorted by hash
- Unique blocks are located in files and remain mutable

The Index

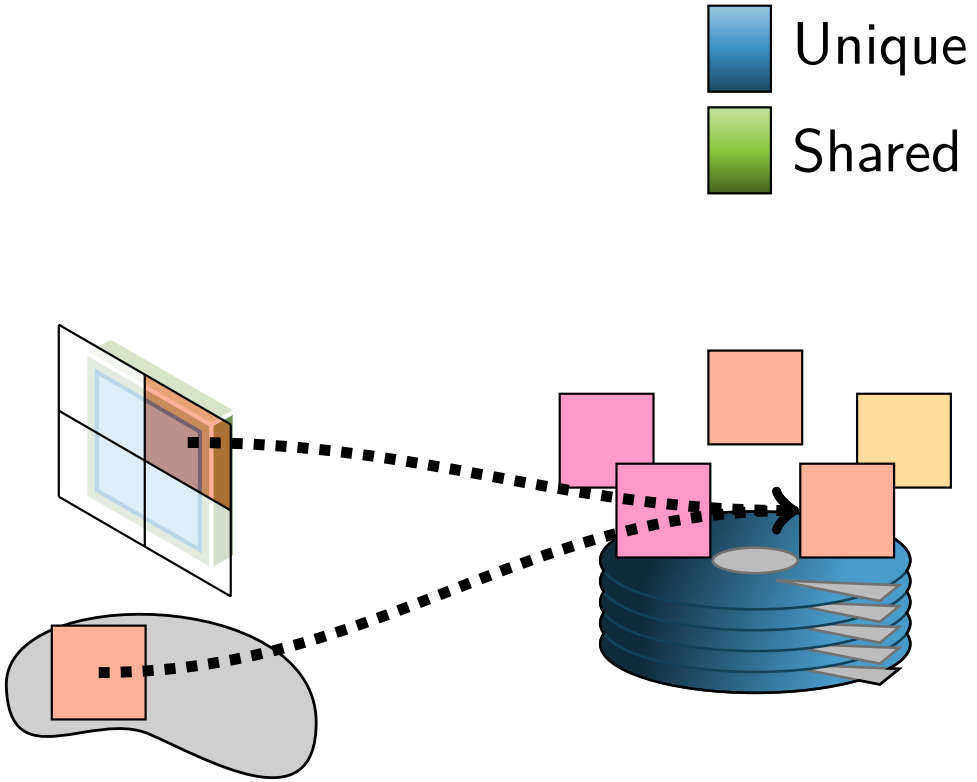






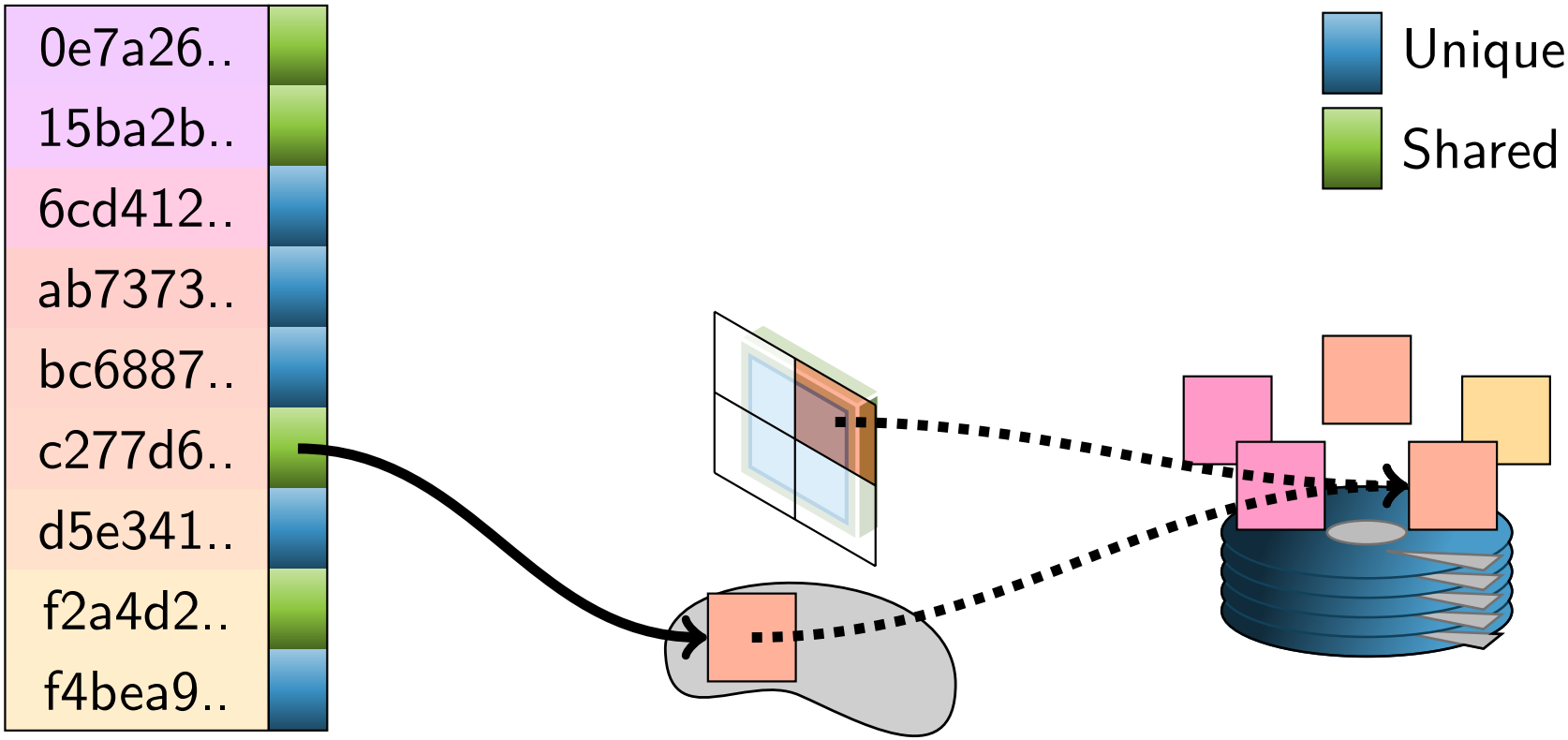
- Map from hashes to block locators, list sorted by hash
- Unique blocks are located in files and remain mutable
- A virtual arena stores COW references to all shared blocks

The Index



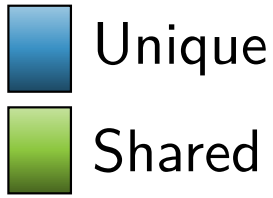
- Map from hashes to block locators, list sorted by hash
- Unique blocks are located in files and remain mutable
- A virtual arena stores COW references to all shared blocks

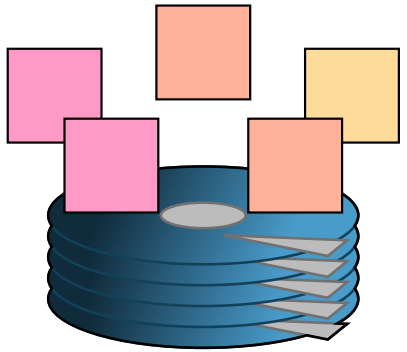
The Index



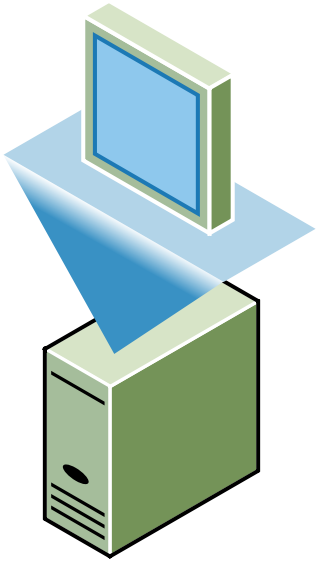
- Map from hashes to block locators, list sorted by hash
- Unique blocks are located in files and remain mutable
- A virtual arena stores COW references to all shared blocks

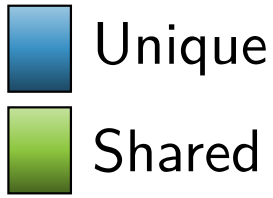
Indexing and Duplicate Elimination

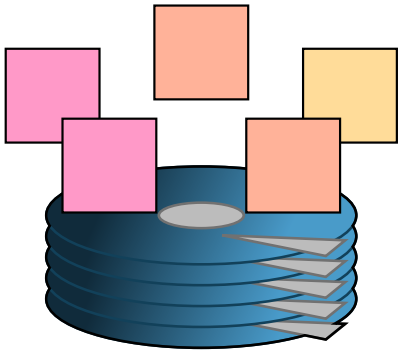




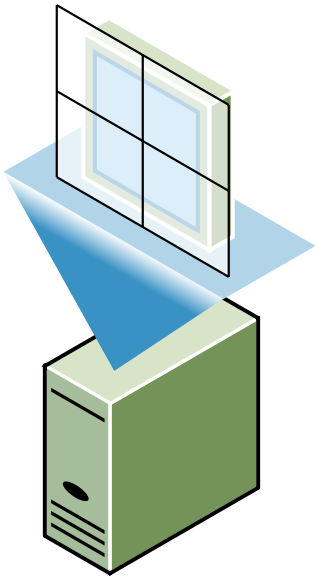
Indexing and Duplicate Elimination

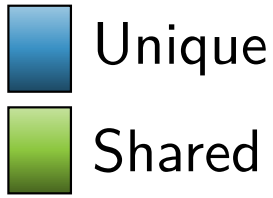


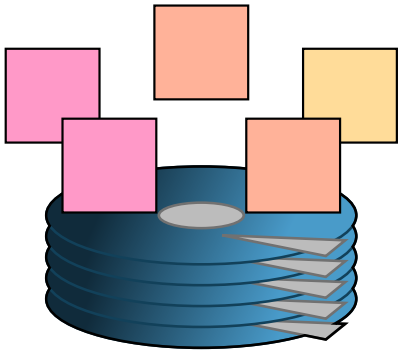




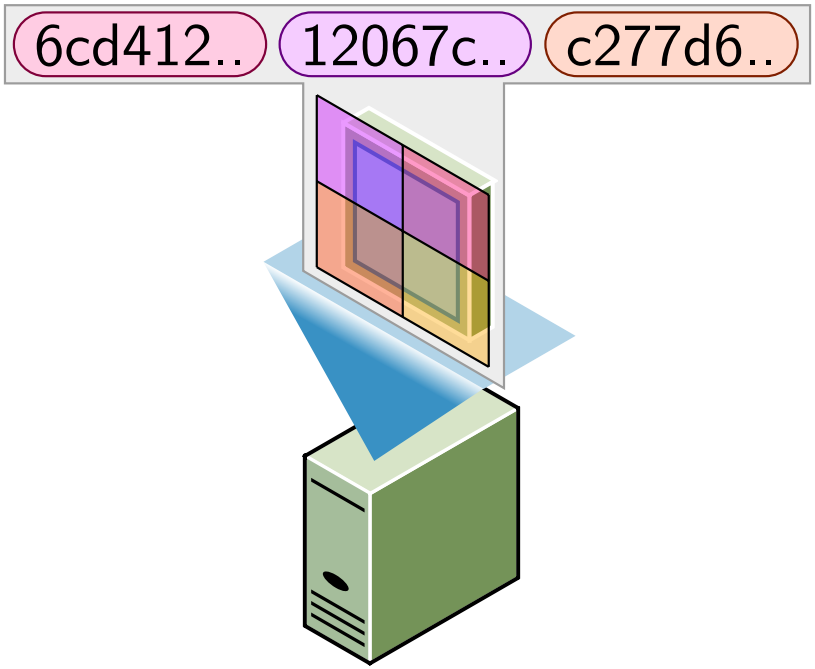
Indexing and Duplicate Elimination

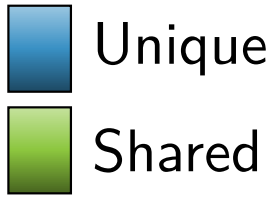


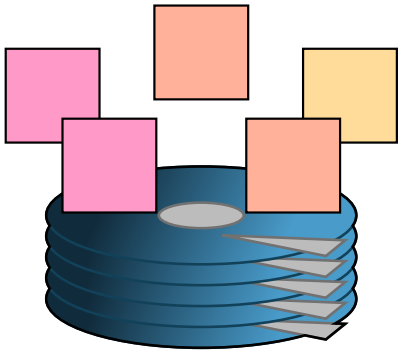




Indexing and Duplicate Elimination

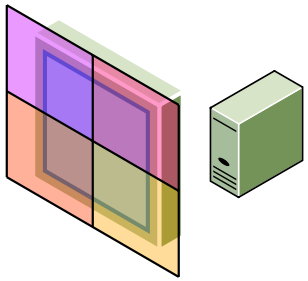


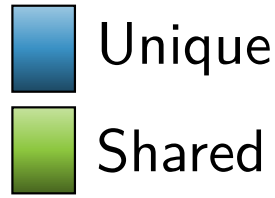


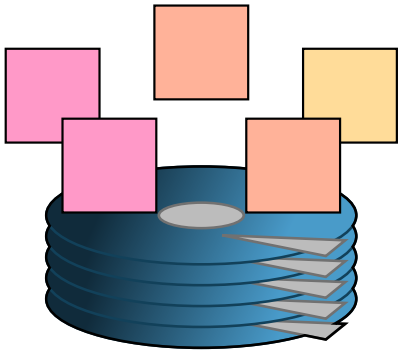


Indexing and Duplicate Elimination

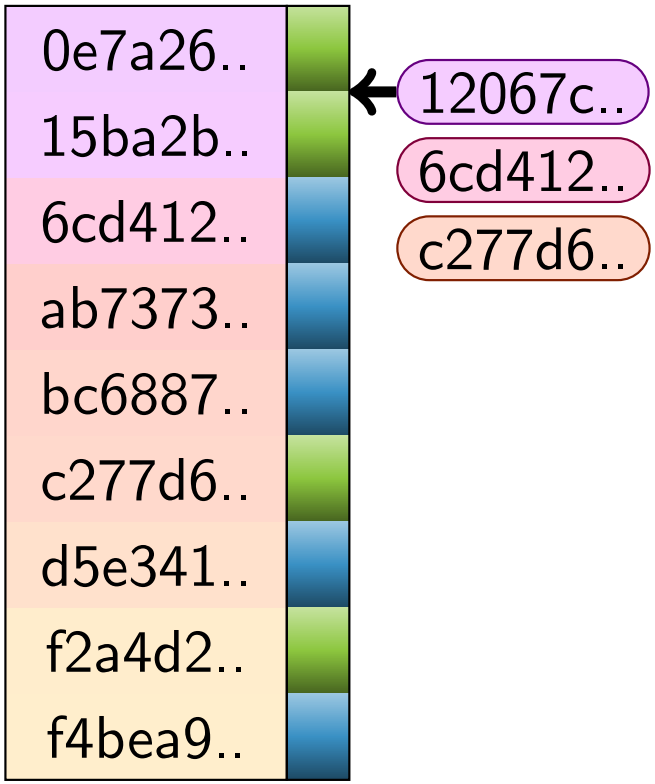
12067c..
6cd412..
c277d6..

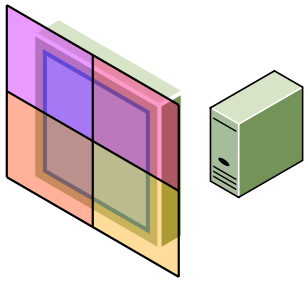


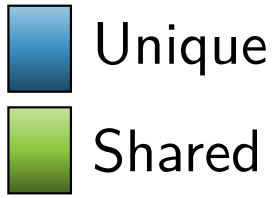


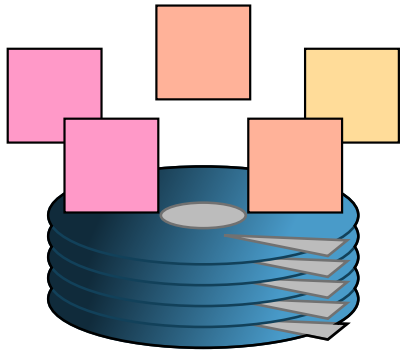


Indexing and Duplicate Elimination

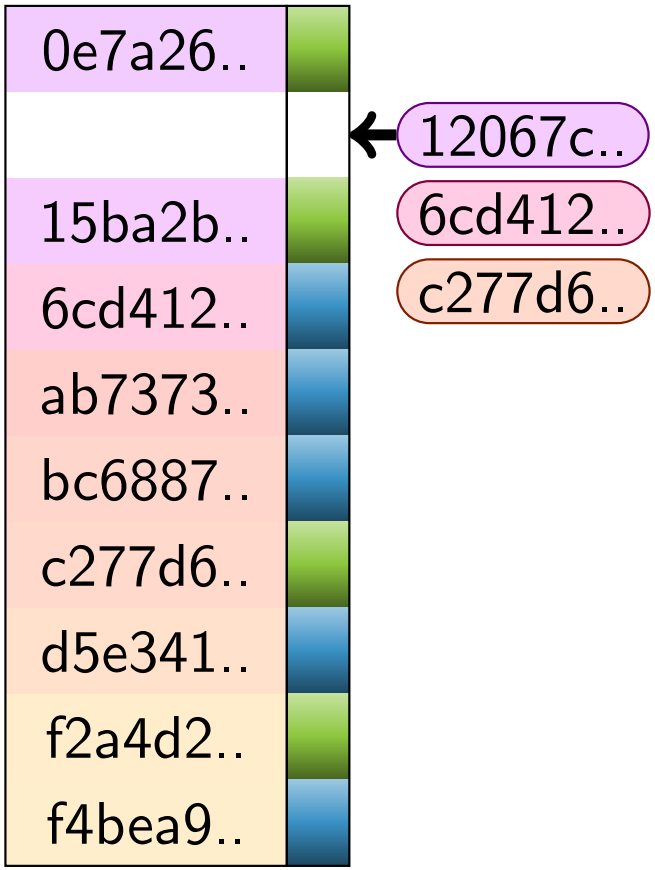


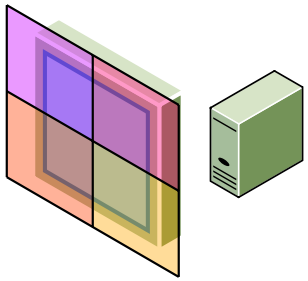


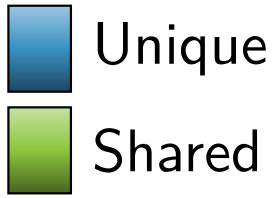


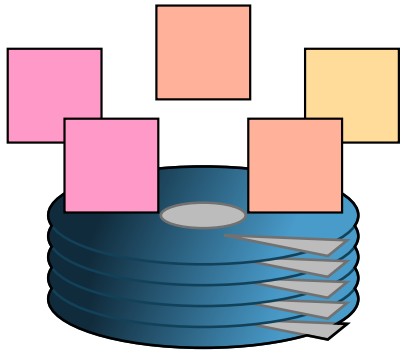


Indexing and Duplicate Elimination



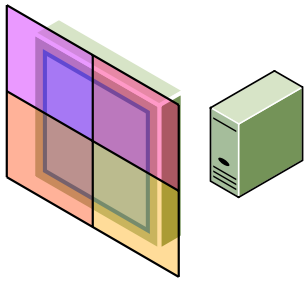


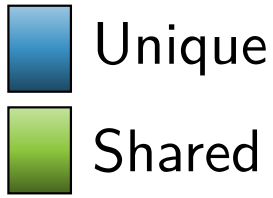


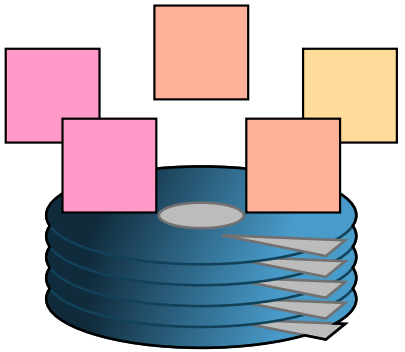


Indexing and Duplicate Elimination

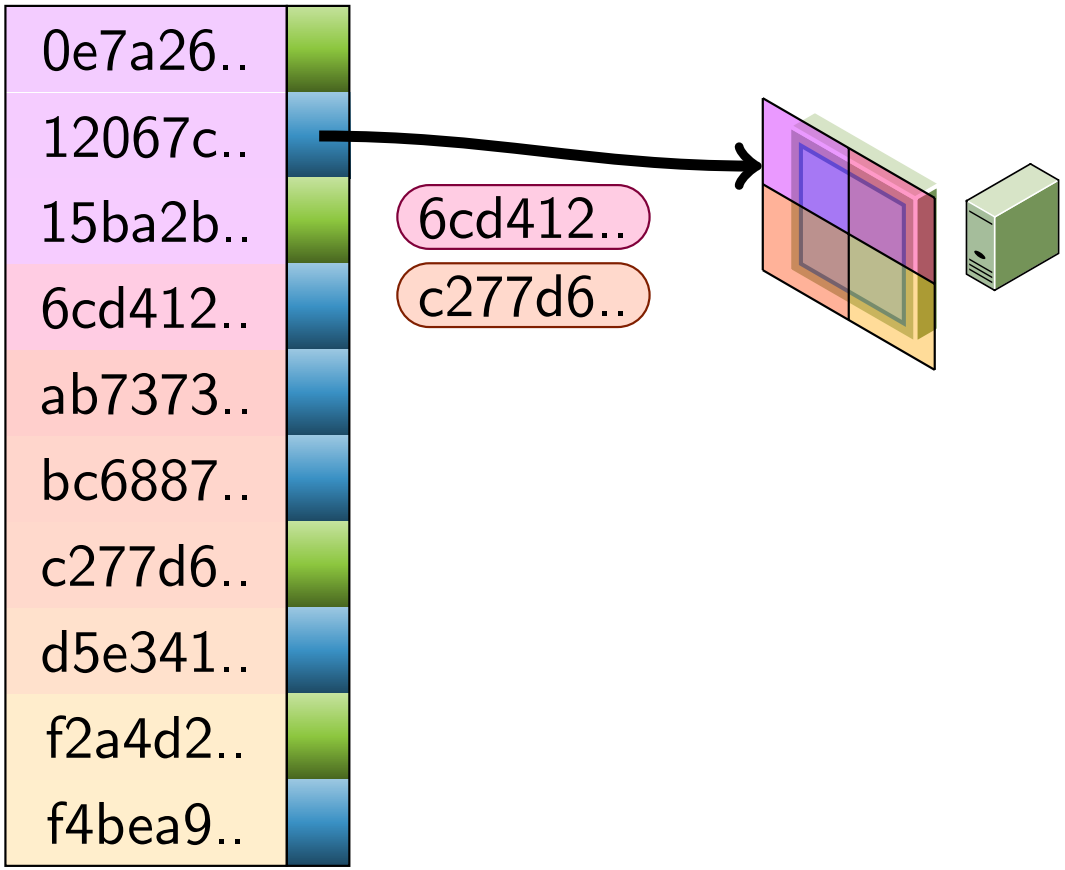
6cd412..
c277d6..

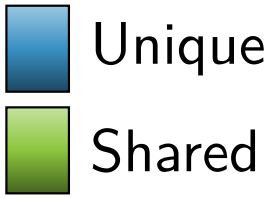


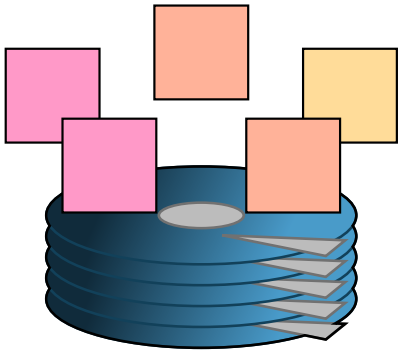




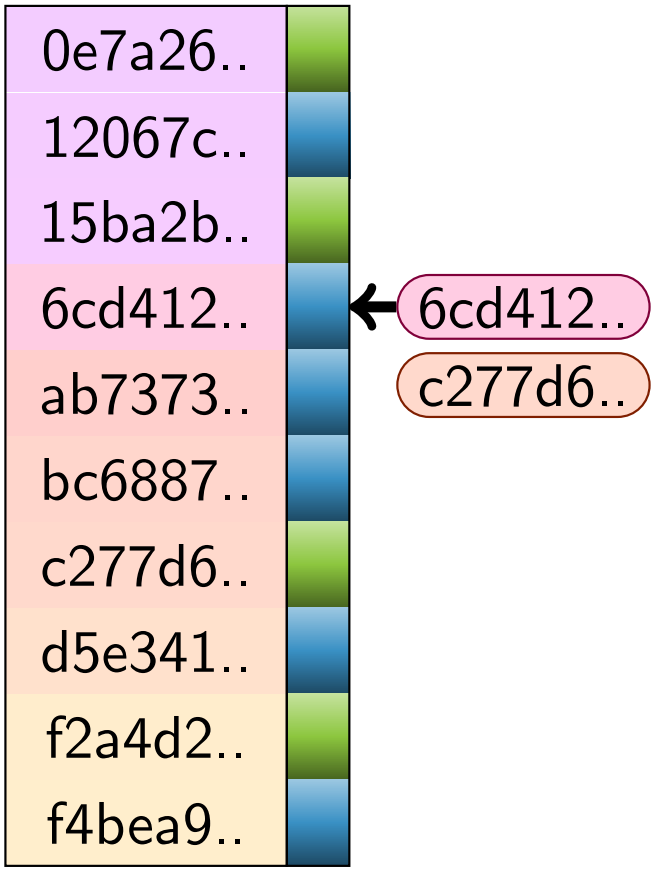
Indexing and Duplicate Elimination

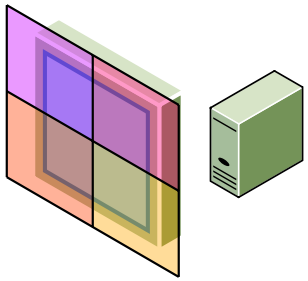


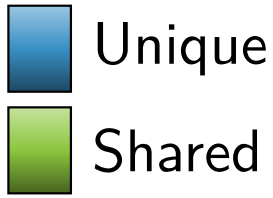


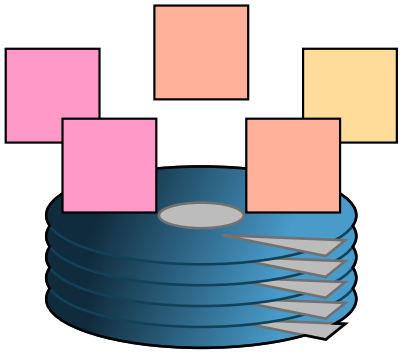


Indexing and Duplicate Elimination

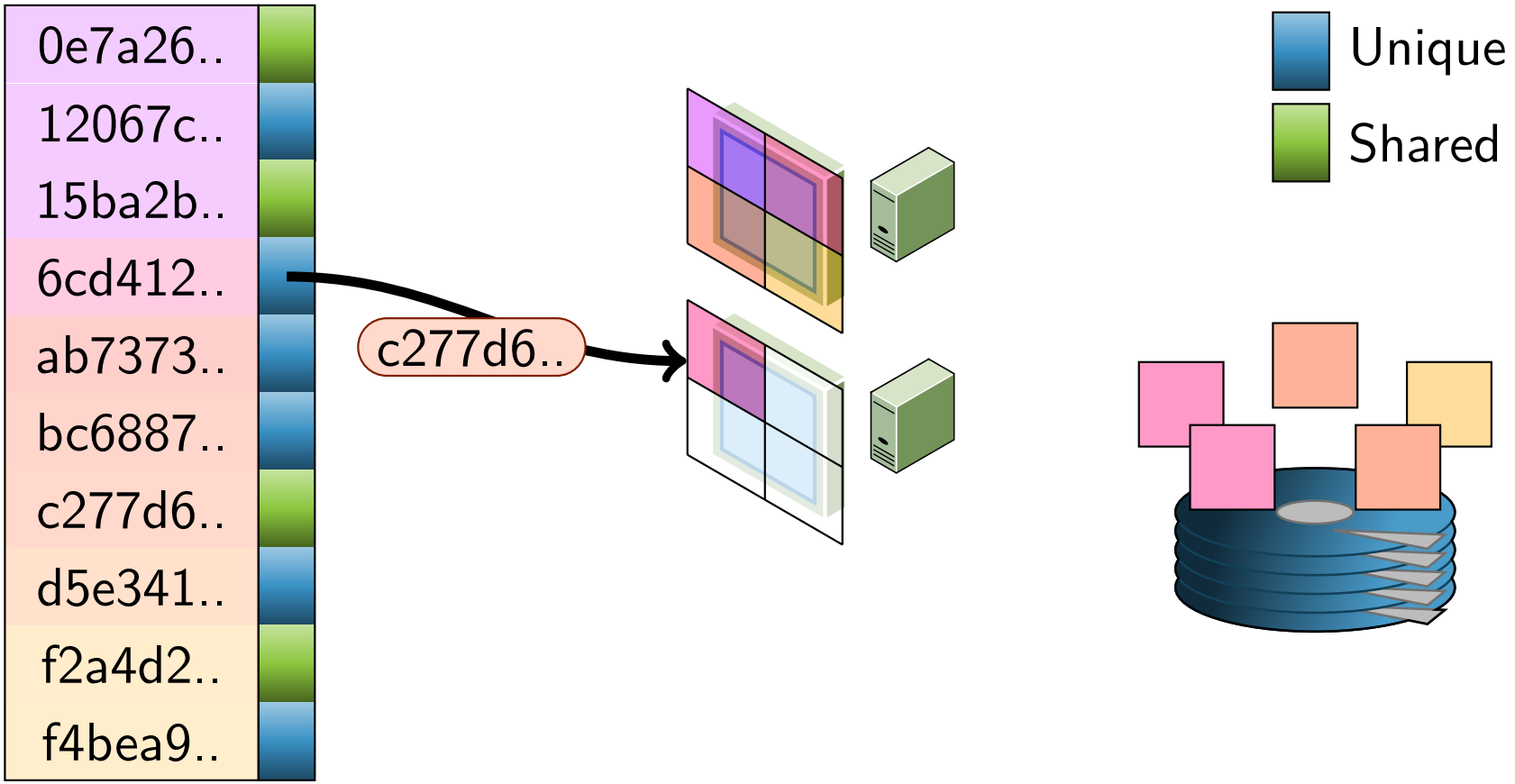




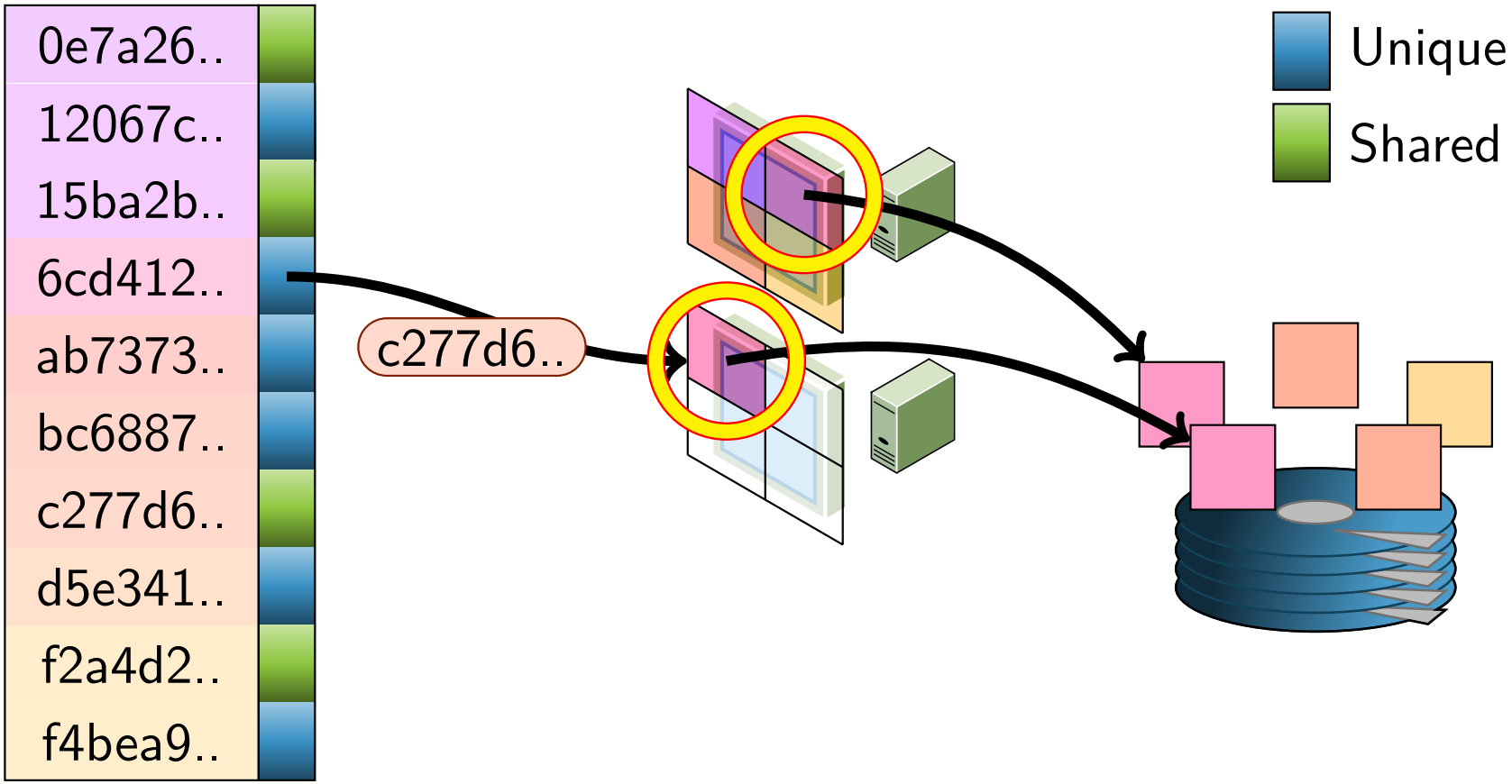




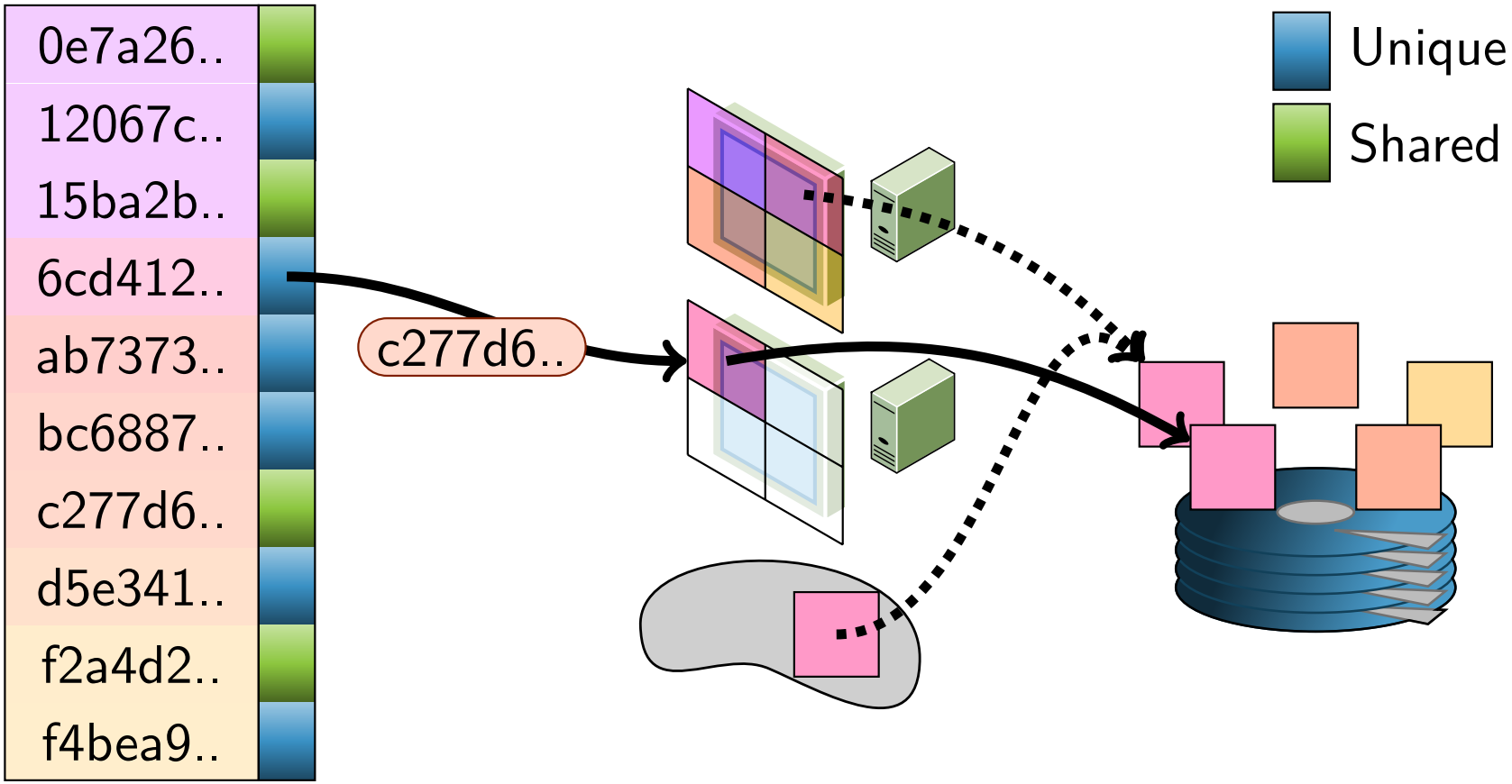
Indexing and Duplicate Elimination



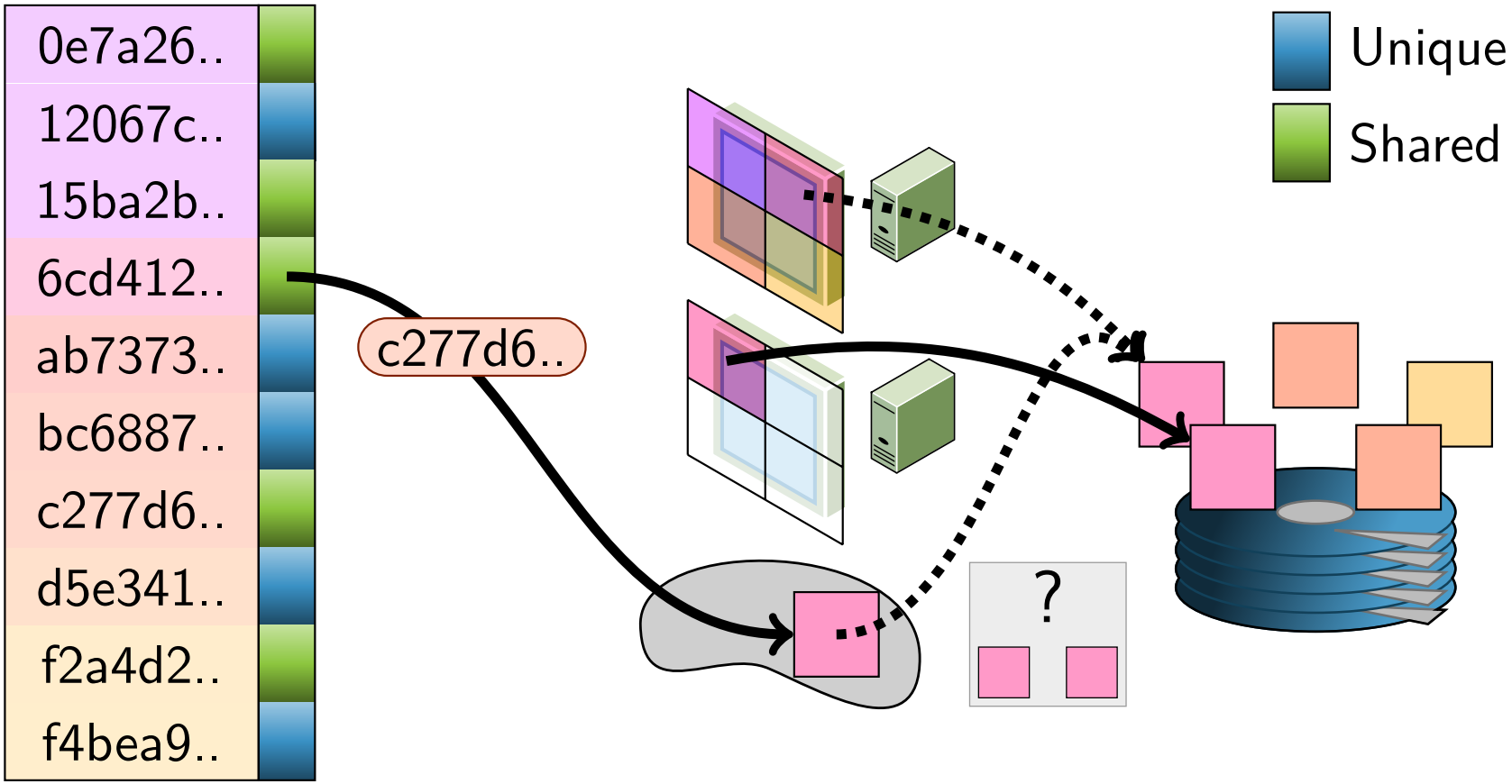
Indexing and Duplicate Elimination



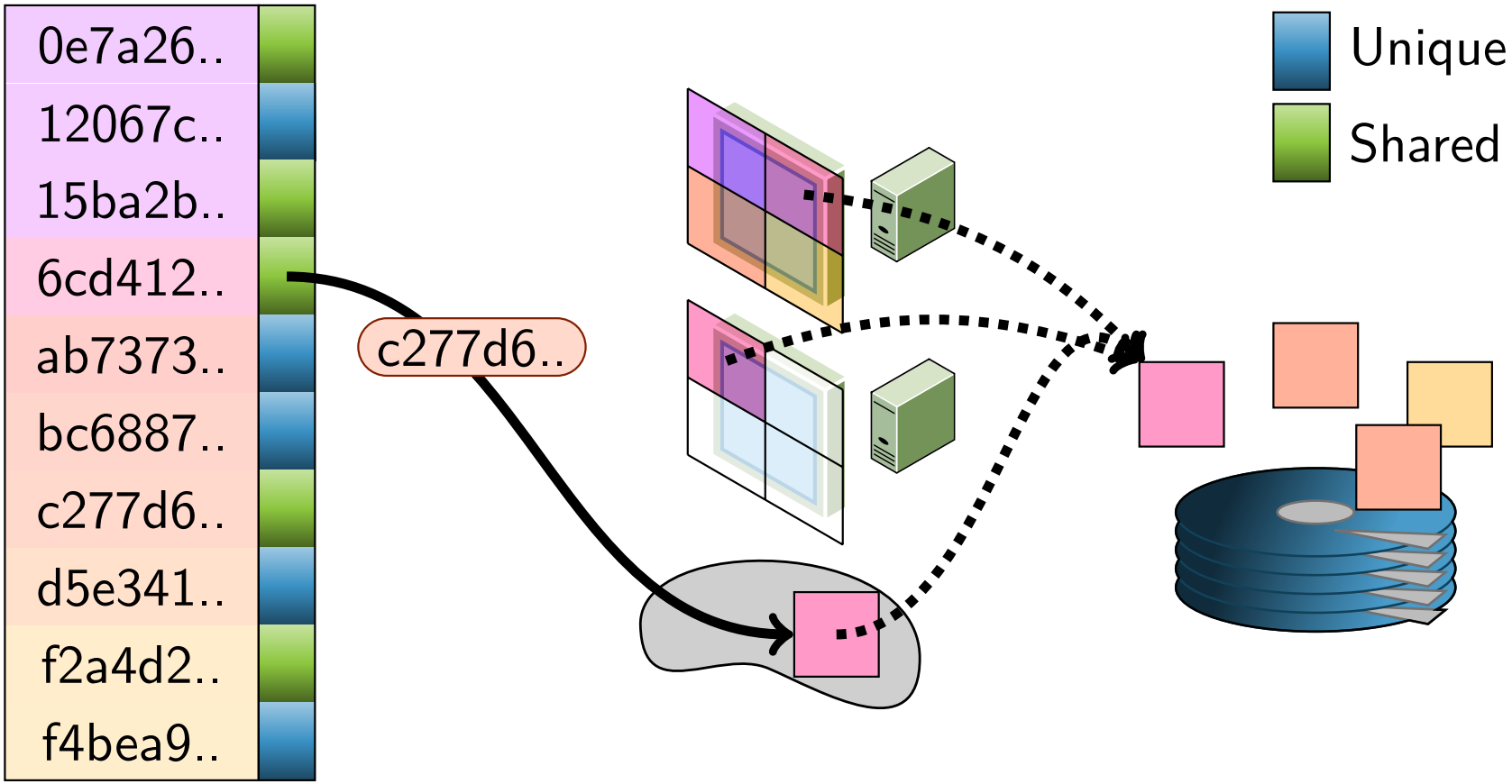
Indexing and Duplicate Elimination



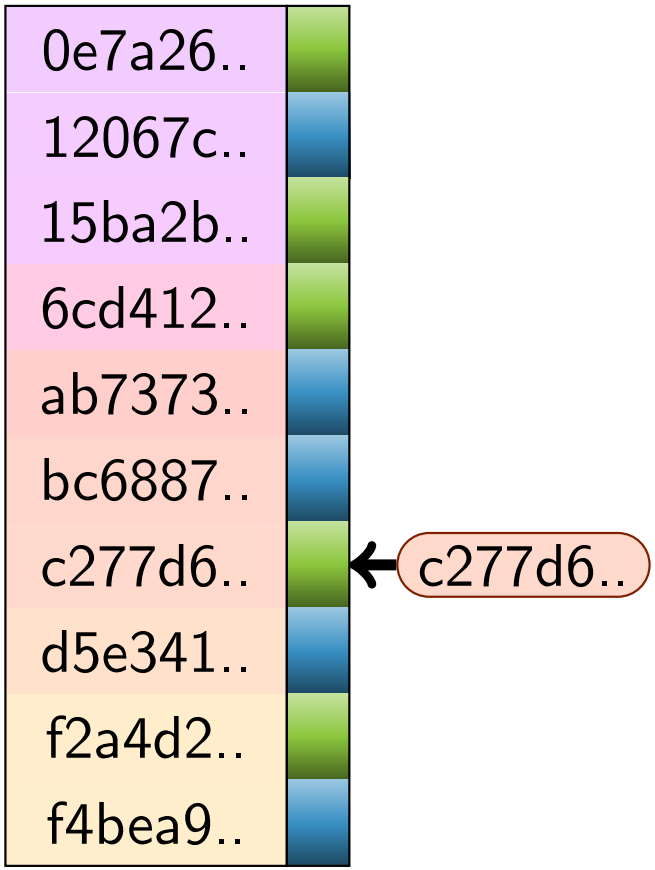
Indexing and Duplicate Elimination

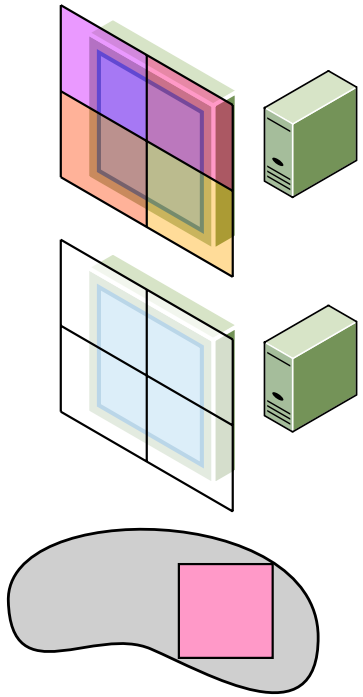


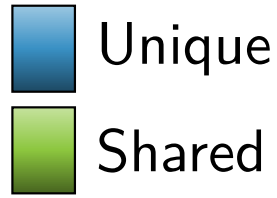
Indexing and Duplicate Elimination

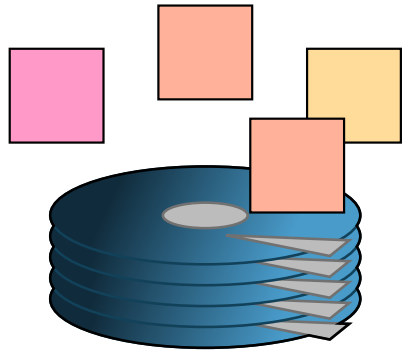


Indexing and Duplicate Elimination

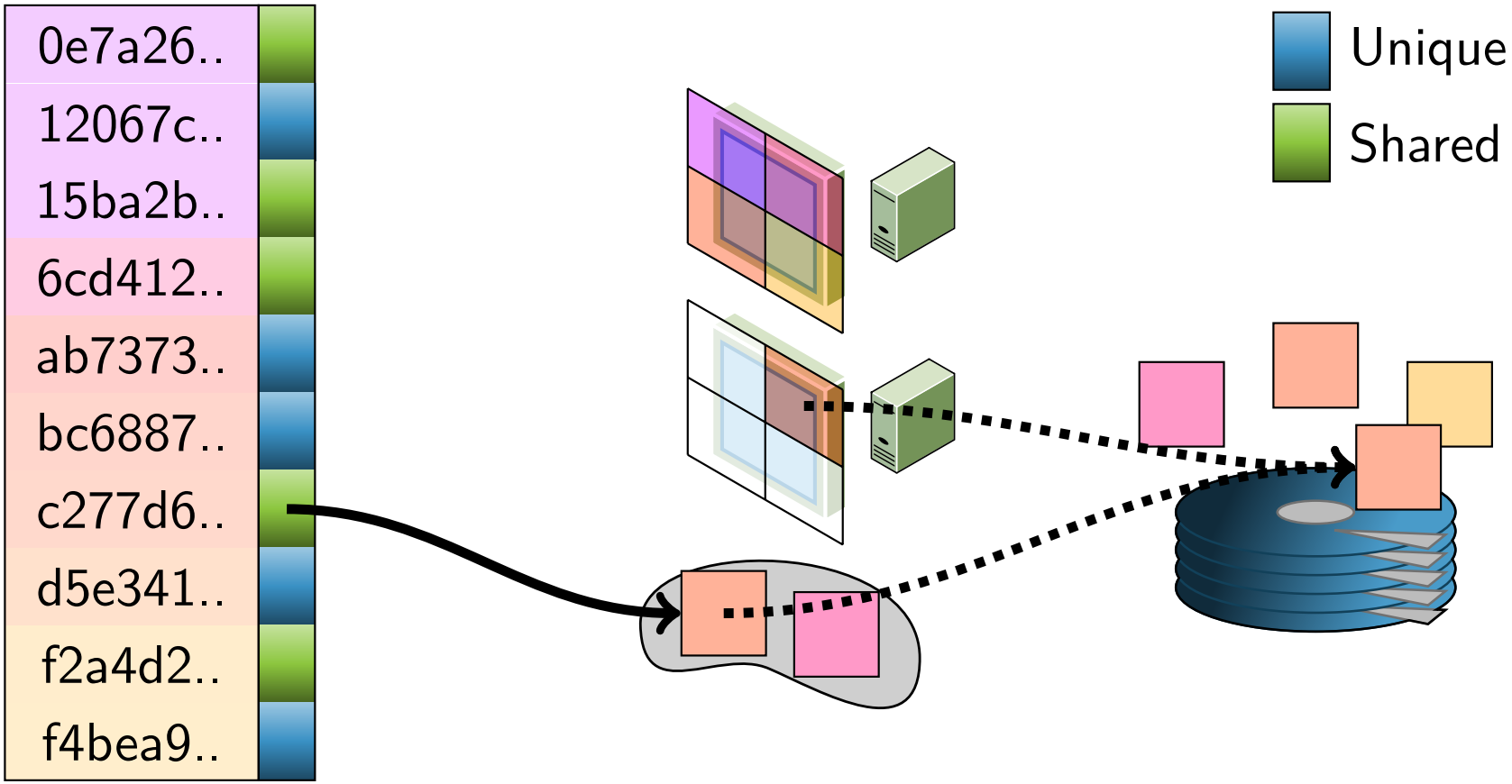




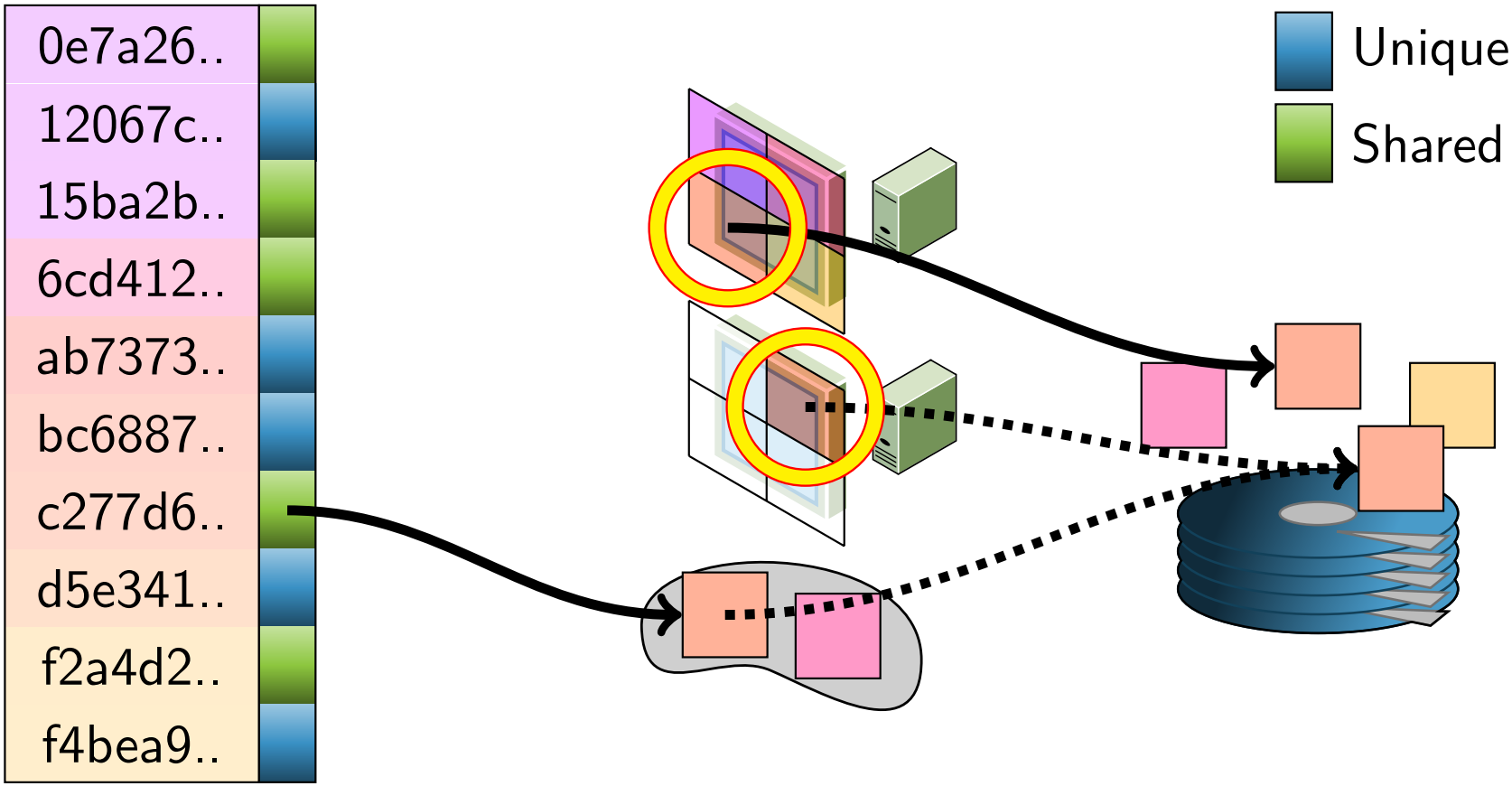




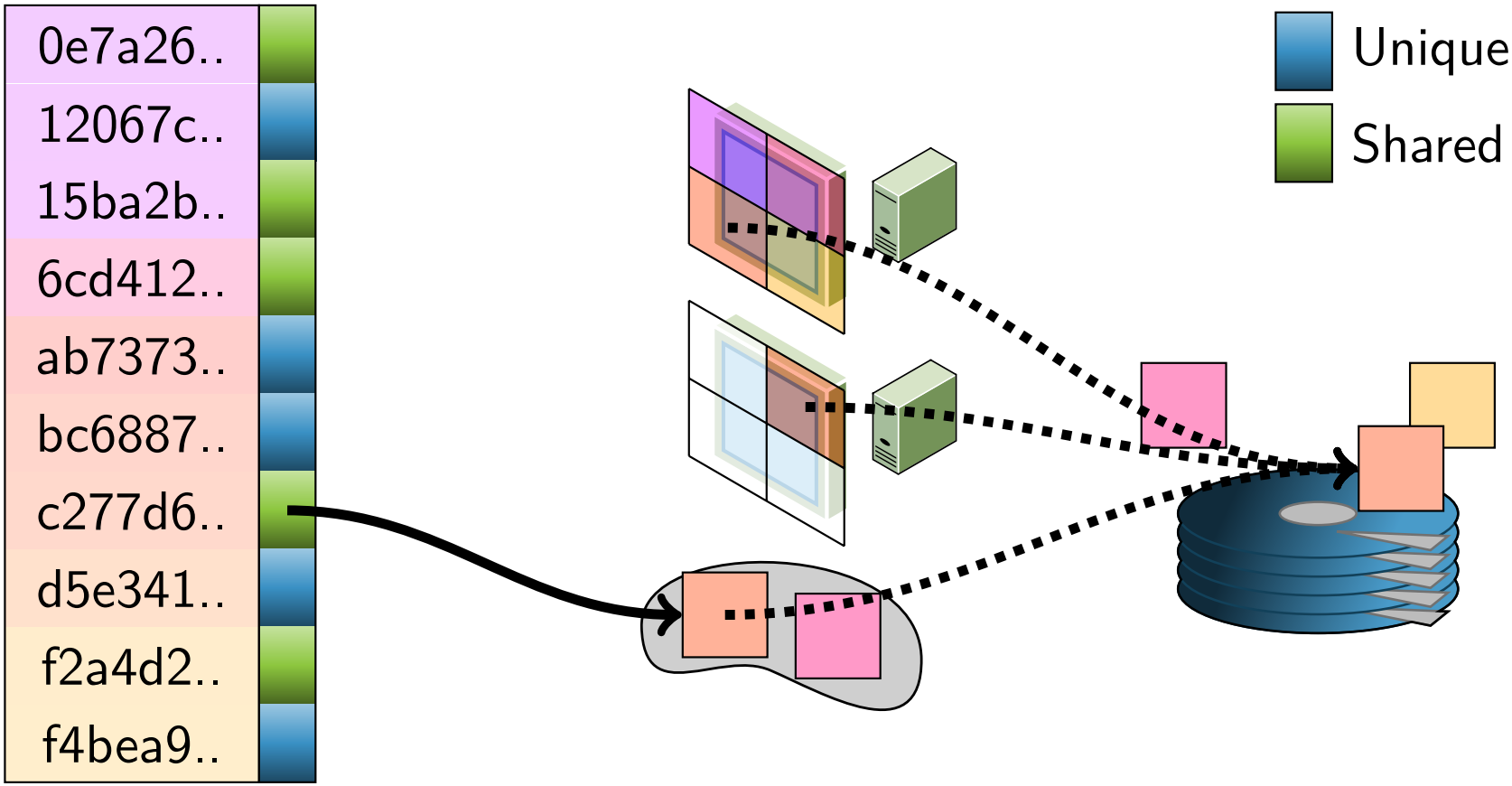
Indexing and Duplicate Elimination



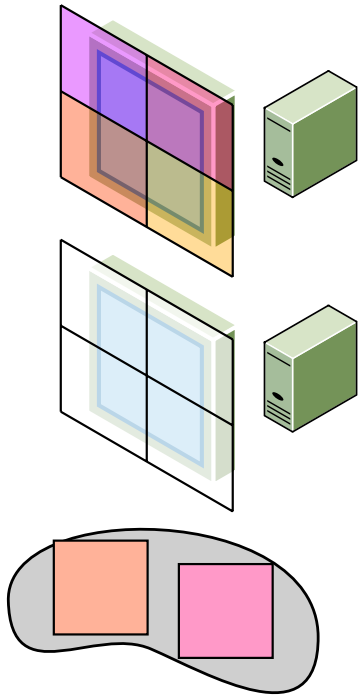
Indexing and Duplicate Elimination

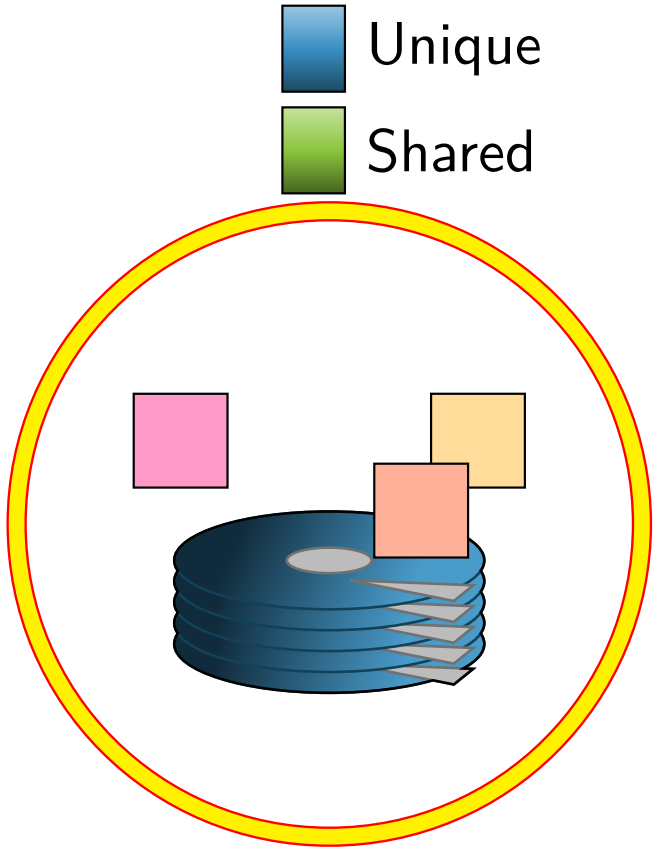


Indexing and Duplicate Elimination



Indexing and Duplicate Elimination





Phew.

Evaluation

Evaluation

How much space does DeDe save?

Evaluation

How much space does DeDe save?

How much overhead does DeDe introduce?

Evaluation

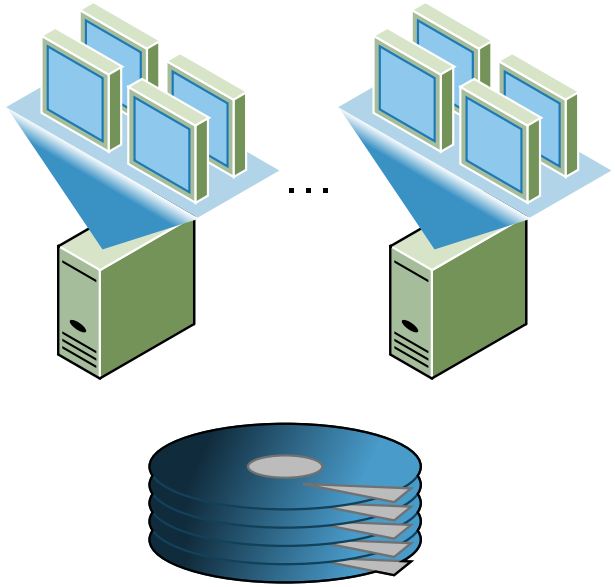
How much space does DeDe save?

How much overhead does DeDe introduce?

How fast can DeDe deduplicate?

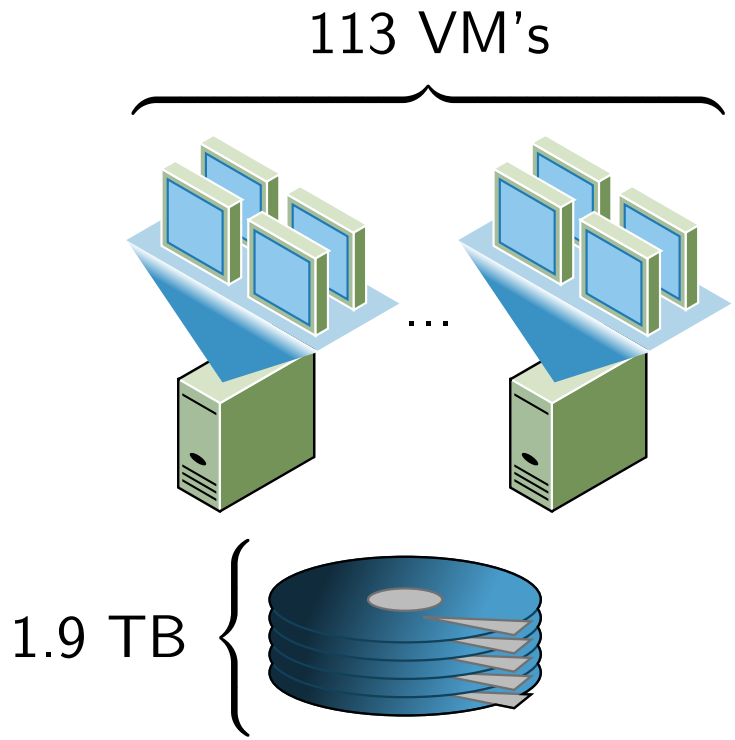
Space Savings: VDI Cluster

- Corporate Virtual Desktop Infrastructure cluster
- Desktop XP VM's
- 6–12 months of active use
- Originally cloned from small number of base images



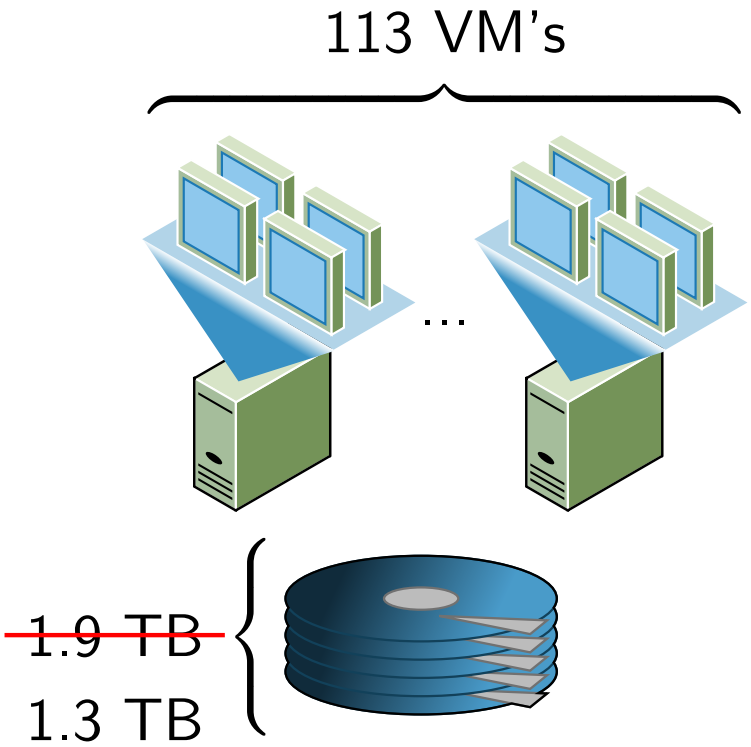
Space Savings: VDI Cluster

- Corporate Virtual Desktop Infrastructure cluster
- Desktop XP VM's
- 6–12 months of active use
- Originally cloned from small number of base images



Space Savings: VDI Cluster

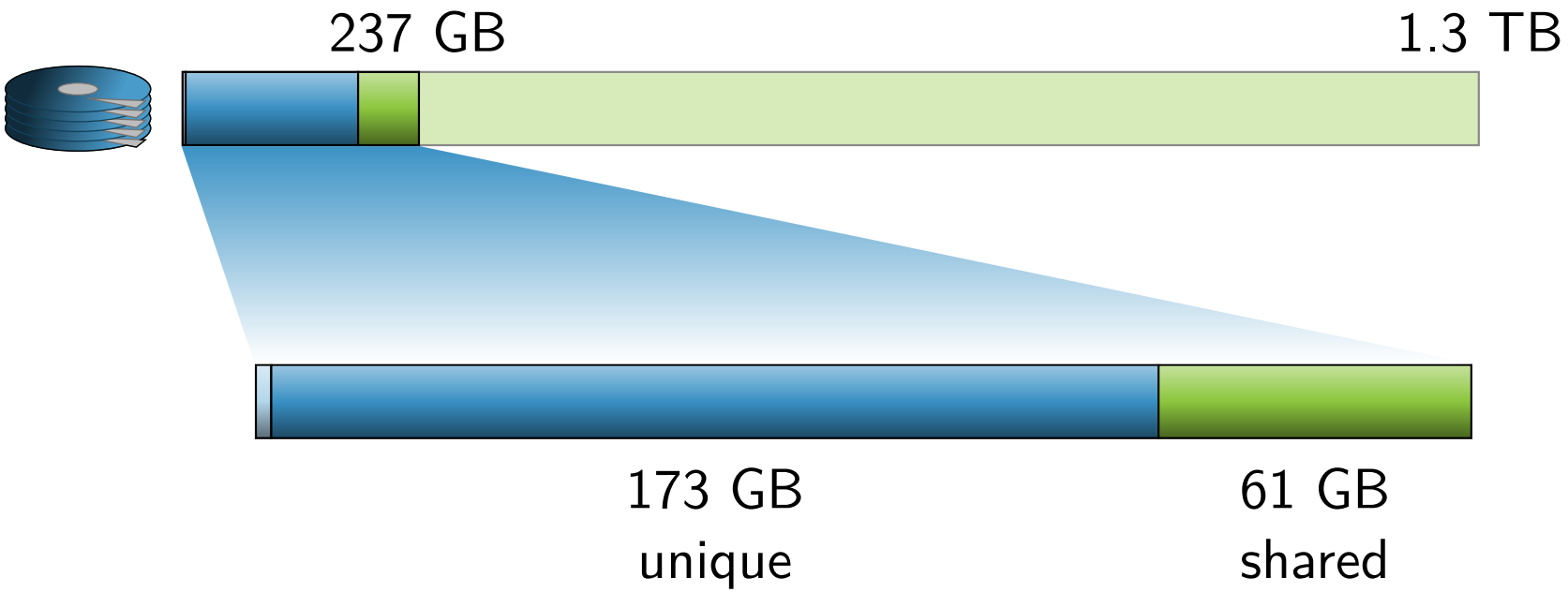
- Corporate Virtual Desktop Infrastructure cluster
- Desktop XP VM's
- 6–12 months of active use
- Originally cloned from small number of base images



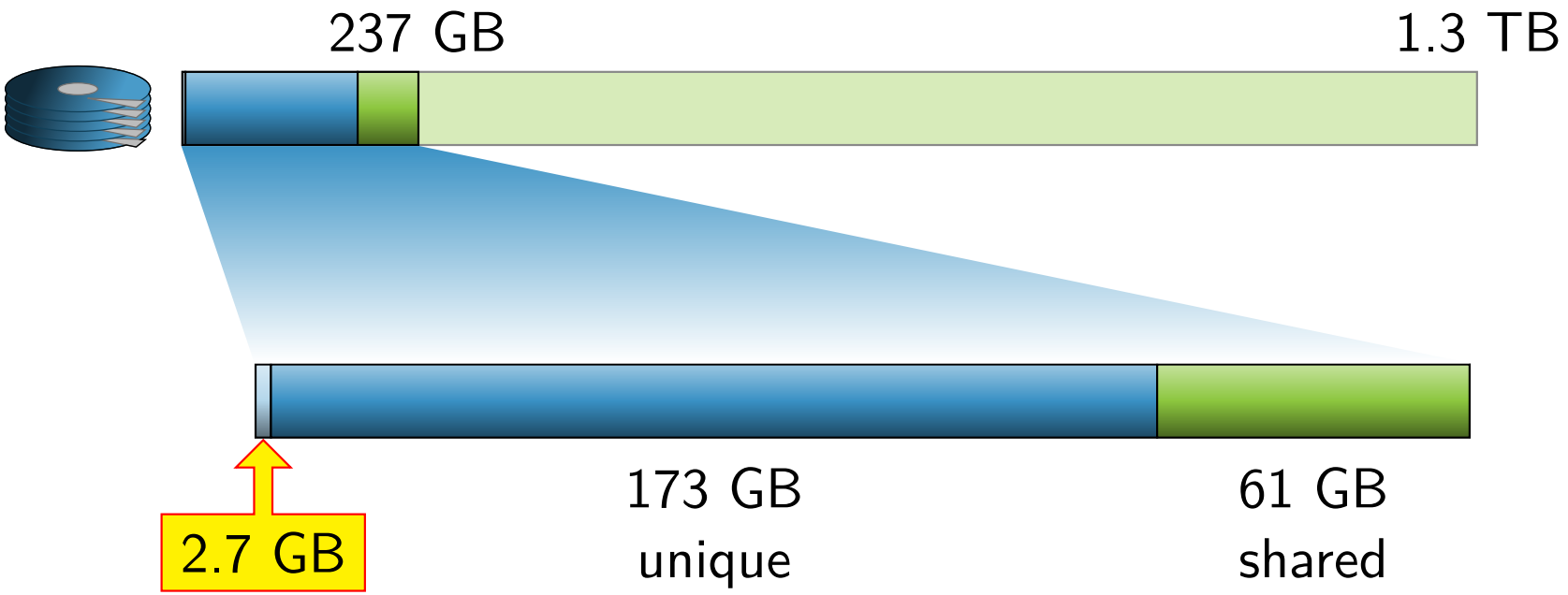
Space Savings: VDI Cluster

Space Savings: VDI Cluster

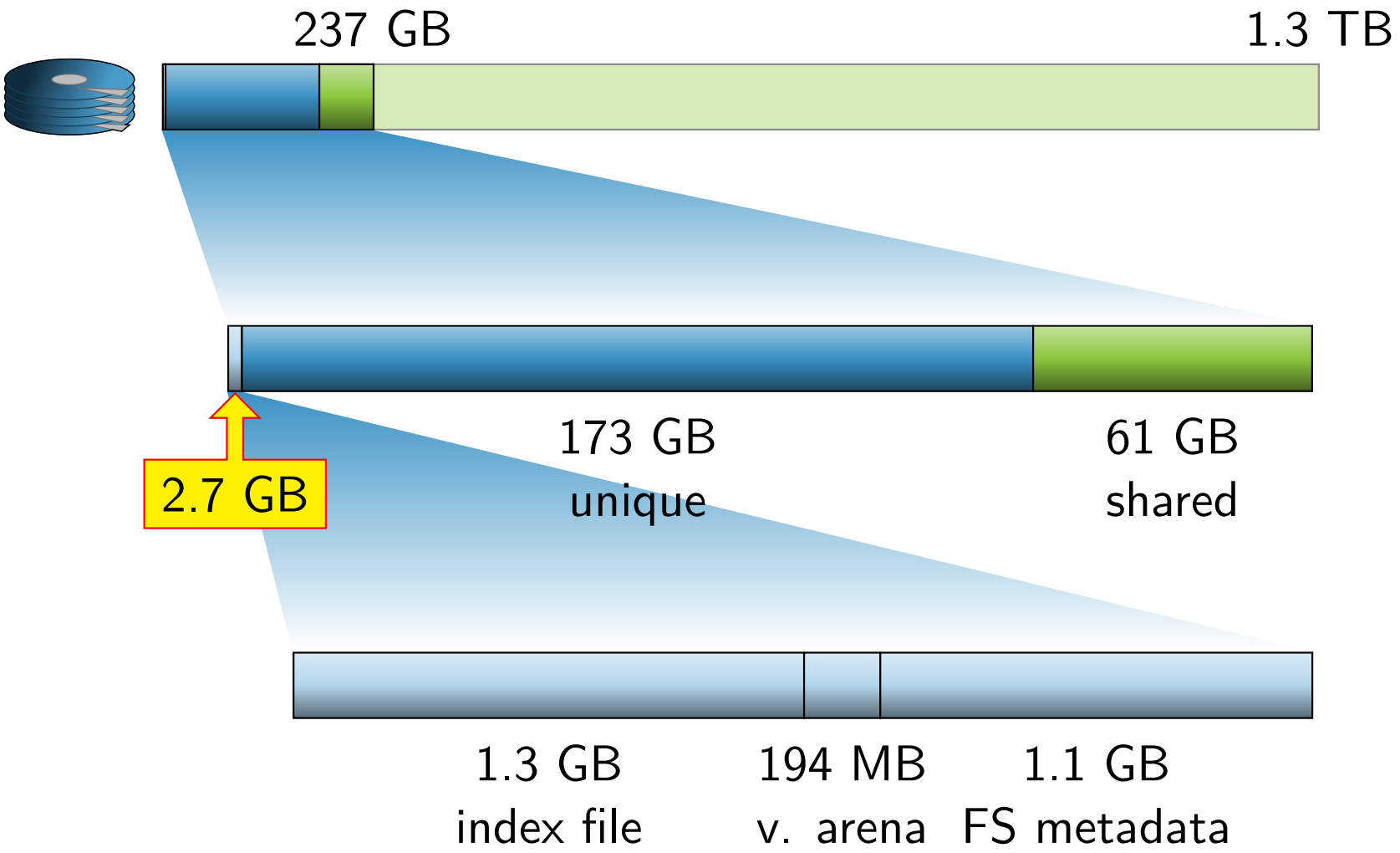
Space Savings: VDI Cluster



Space Savings: VDI Cluster



Space Savings: VDI Cluster

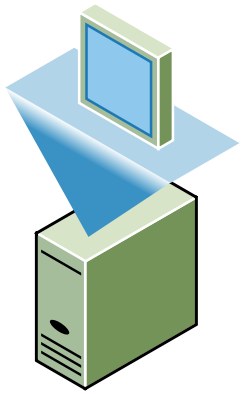


Runtime Effects

- Write monitoring
- Disk array caching

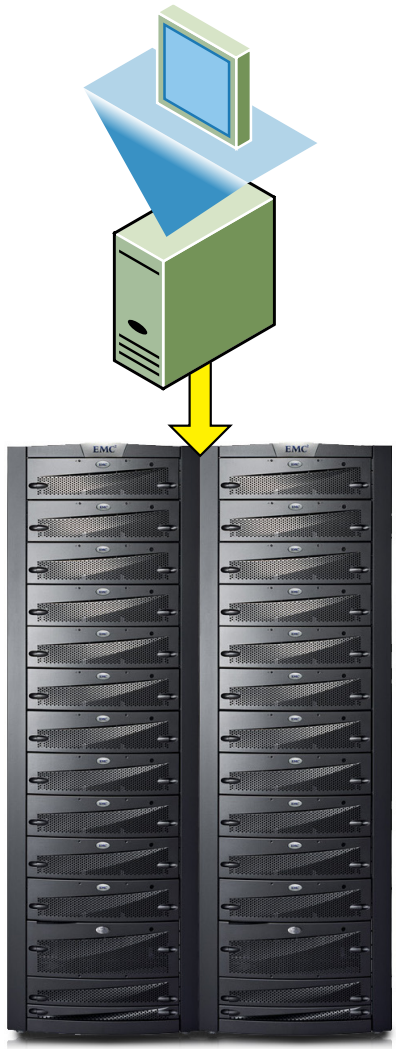
Runtime Effects

- Write monitoring
- Disk array caching



Runtime Effects

- Write monitoring
- Disk array caching



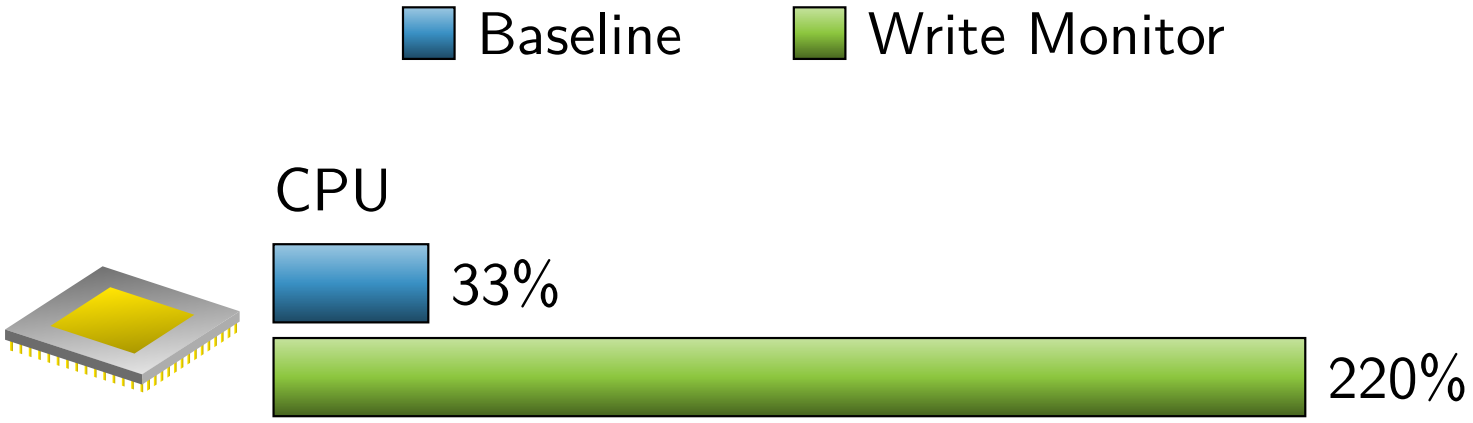
EMC CLARiiON CX3-40

Runtime Overhead: Write Monitoring

Worst-case benchmark: 100% sequential write IO, No computation

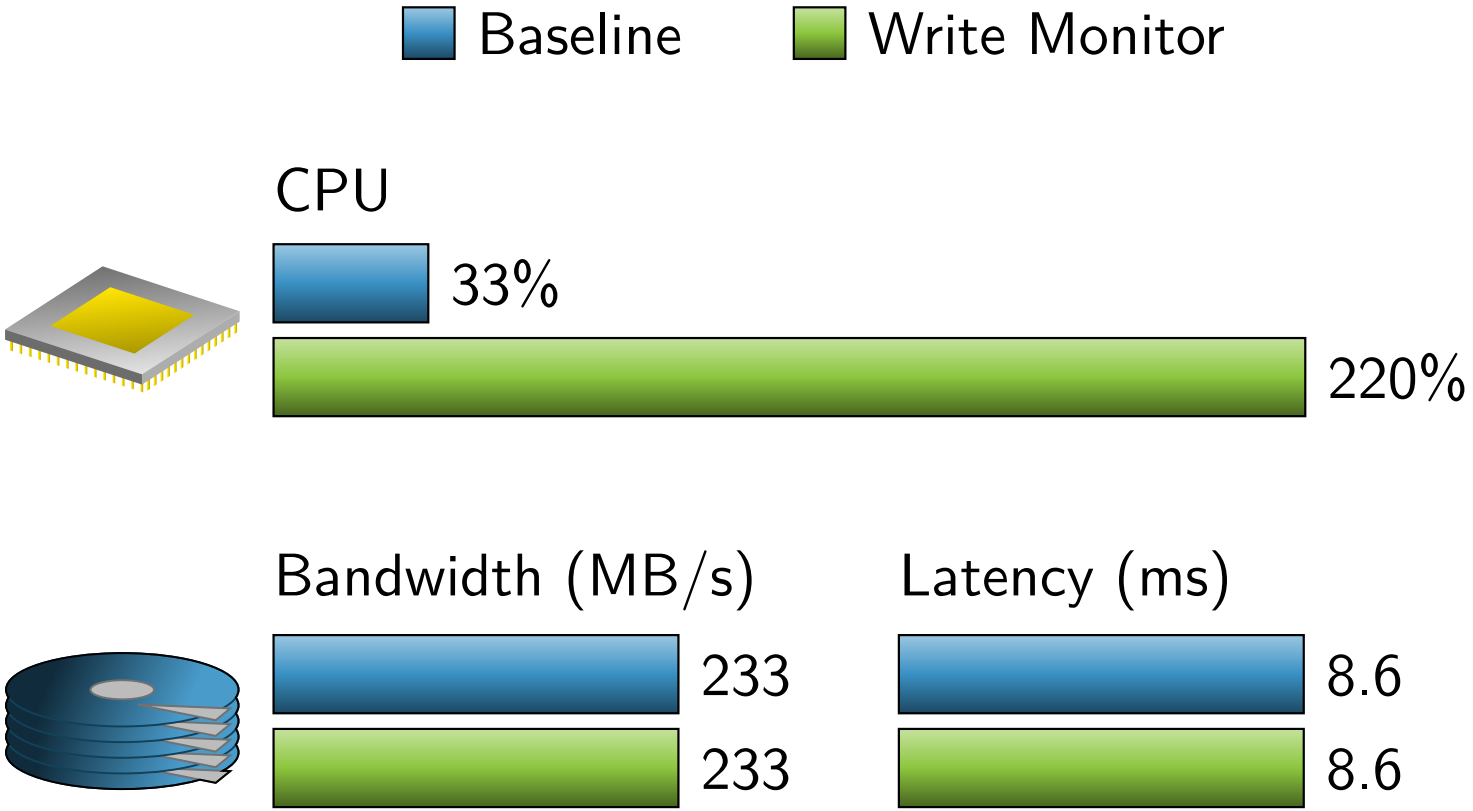
Runtime Overhead: Write Monitoring

Worst-case benchmark: 100% sequential write IO, No computation



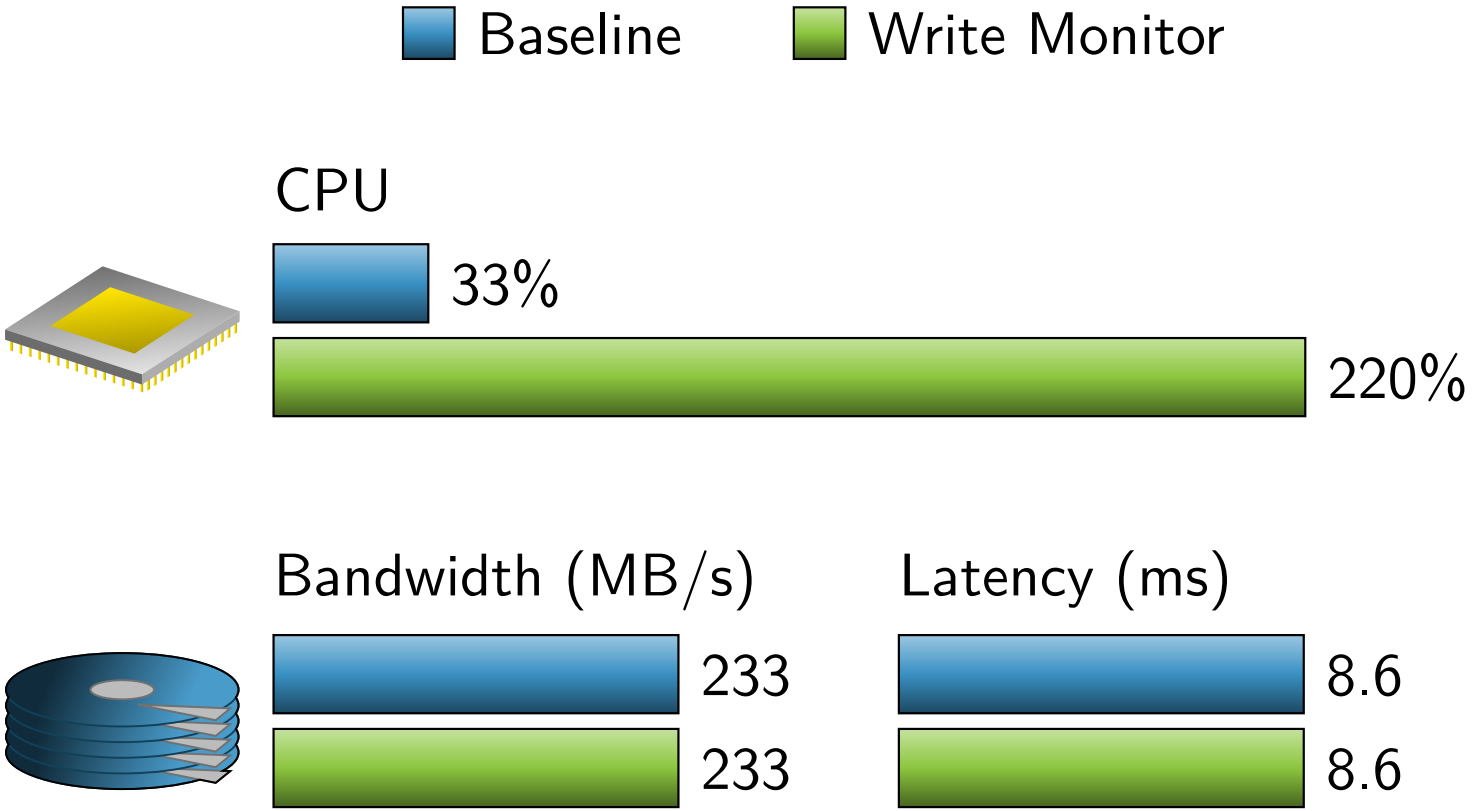
Runtime Overhead: Write Monitoring

Worst-case benchmark: 100% sequential write IO, No computation



Runtime Overhead: Write Monitoring

Worst-case benchmark: 100% sequential write IO, No computation



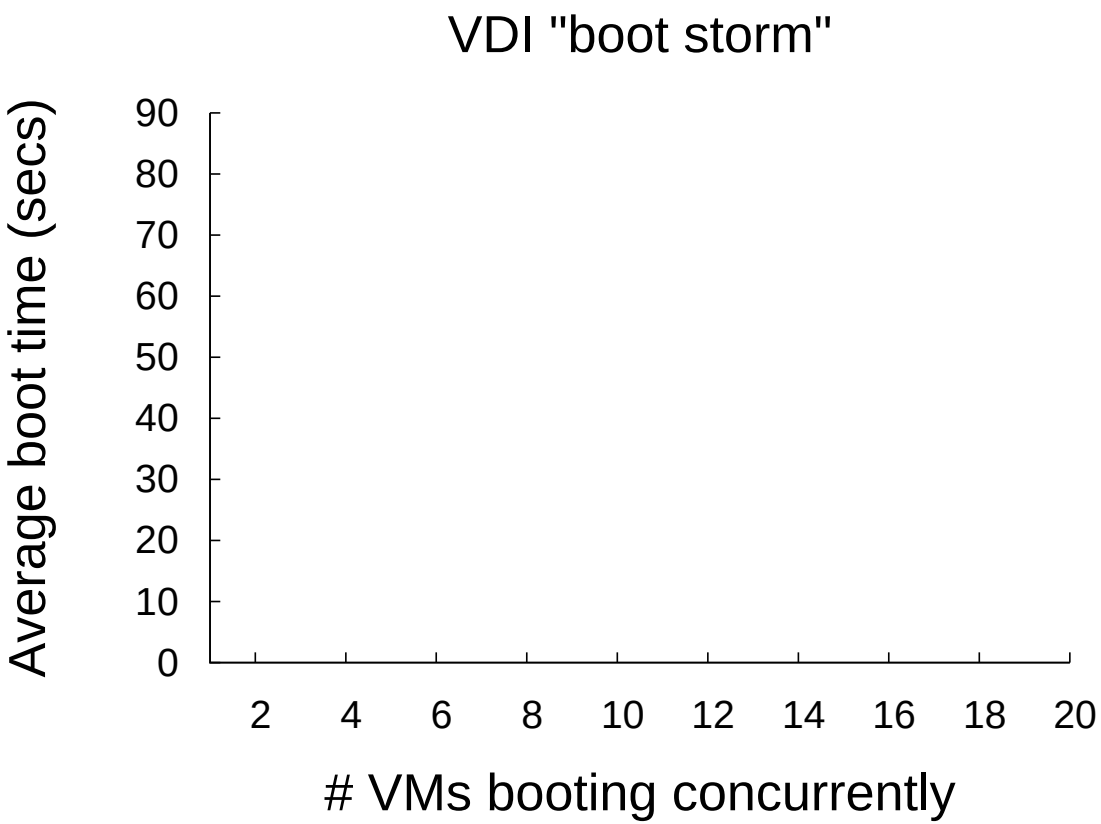
(See paper for database application benchmark)

Runtime Gains: Disk Array Caching

Reduced storage footprint ⇨ Better caching ⇨ Less IO

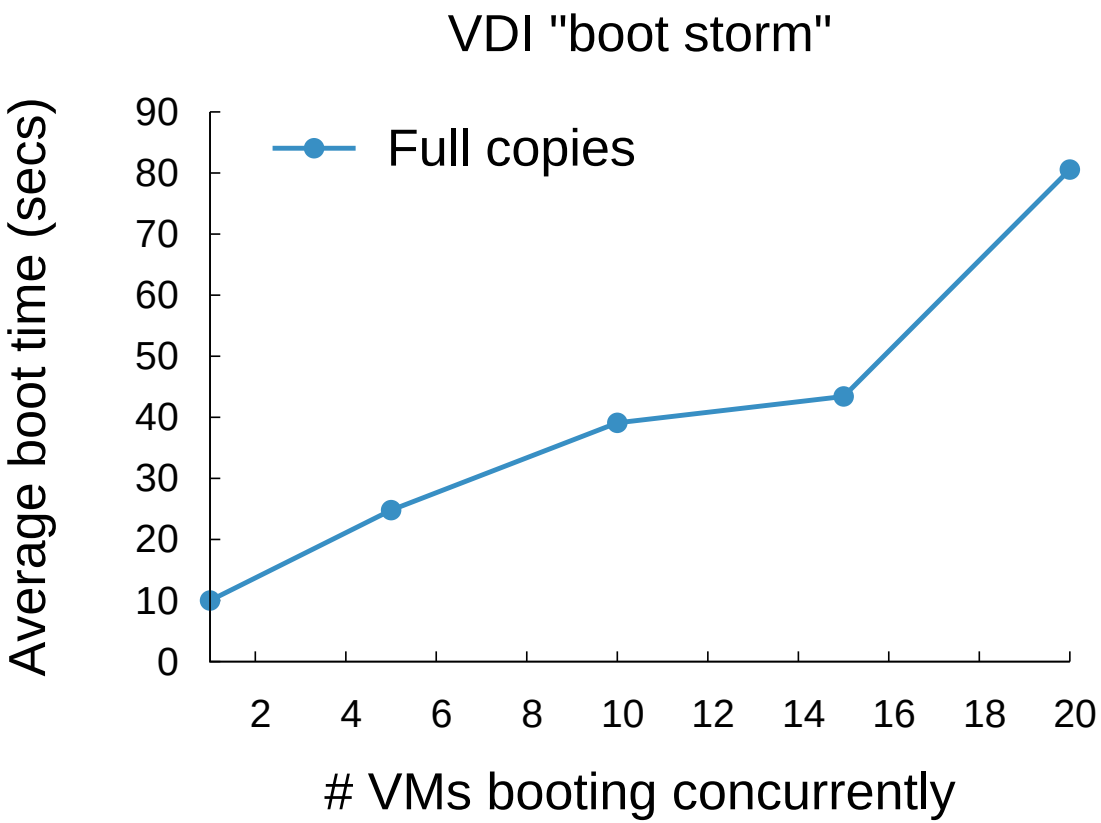
Runtime Gains: Disk Array Caching

Reduced storage footprint ⇨ Better caching ⇨ Less IO



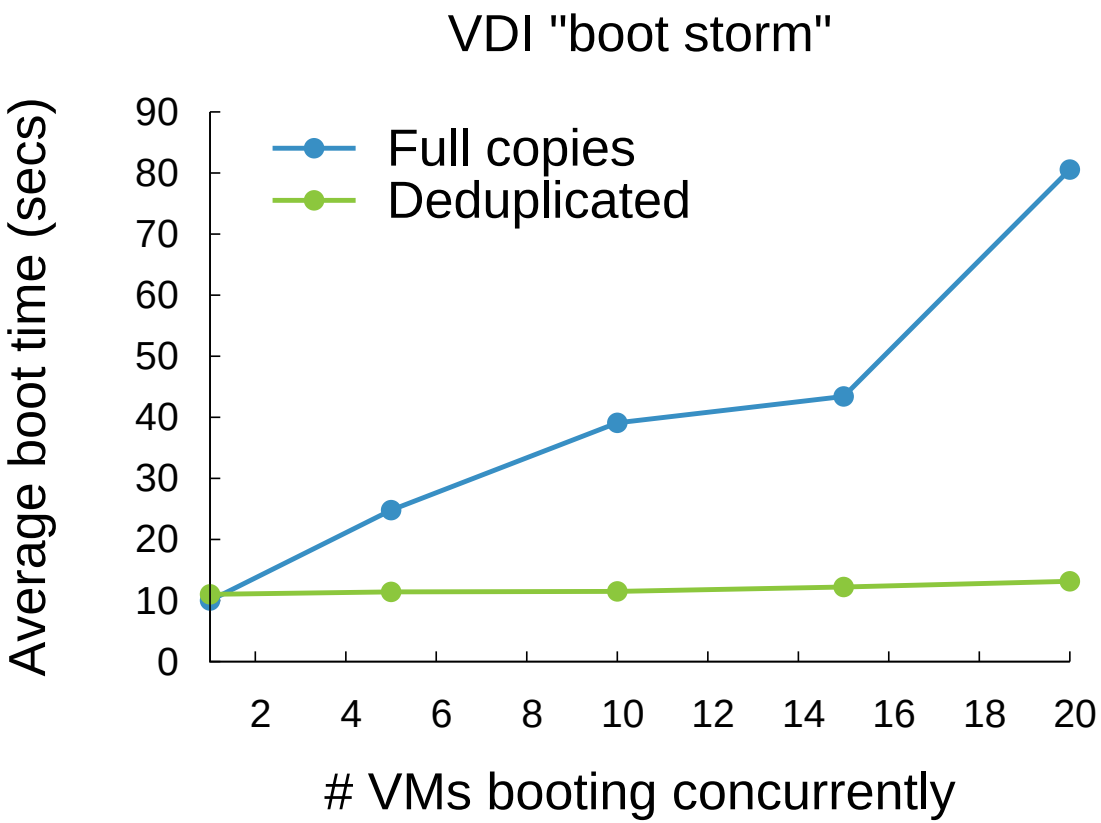
Runtime Gains: Disk Array Caching

Reduced storage footprint ⇨ Better caching ⇨ Less IO



Runtime Gains: Disk Array Caching

Reduced storage footprint ⇨ Better caching ⇨ Less IO



Out-of-band Deduplication Rate

Index scan
COW sharing

Out-of-band Deduplication Rate

Index scan	6.6 GB/sec
COW sharing	

⇨ Virtually no cost for unique blocks

Out-of-band Deduplication Rate

Index scan	6.6 GB/sec
COW sharing	2.6 MB/sec

⇨ Virtually no cost for unique blocks

Out-of-band Deduplication Rate

Index scan	6.6 GB/sec
COW sharing	2.6 MB/sec

(It's a prototype!)

⇨ Virtually no cost for unique blocks

Out-of-band Deduplication Rate

Index scan	6.6 GB/sec
COW sharing	2.6 MB/sec

(It's a prototype!)

⇨ Virtually no cost for unique blocks

⇨ 9 GB of *new shared blocks* per hour

(And provisioning can be special-cased)

Related Work

- Centralized archival
  - Venti
  - Data Domain
  - Foundation
- Centralized primary storage
  - NetApp ASIS
  - Microsoft Single Instance Store
- Distributed
  - Farsite
- SAN with Coordinator
  - DDE

Conclusion

Decentralized, out-of-band, live file system deduplication.

Conclusion

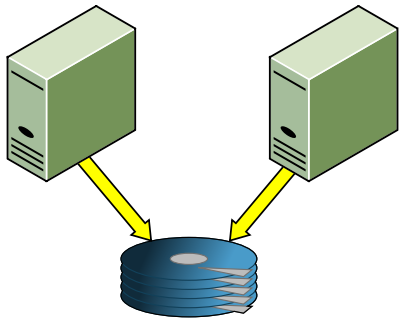
Decentralized, out-of-band, live file system deduplication.

Deduplication is effective.

Conclusion

Decentralized, out-of-band, live file system deduplication.

Deduplication is effective.

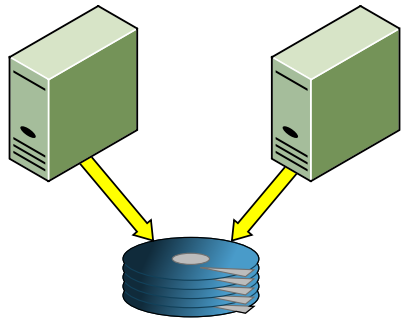


Deduplication is hard.

Conclusion

Decentralized, out-of-band, live file system deduplication.

Deduplication is effective.



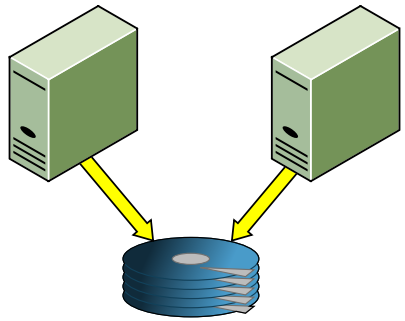
Deduplication is hard.

Three-stage deduplication has only modest performance overhead.

Conclusion

Decentralized, out-of-band, live file system deduplication.

Deduplication is effective.



Deduplication is hard.

Write Monitoring	Local Deduplication	Cross-Host Deduplication

Three-stage deduplication has only modest performance overhead.

Thank you.