GCTrees: Garbage Collecting Snapshots

Chris Dragga

Doug Santry

© 2014 NetApp, Inc. All rights reserved.

GCTrees

Support snapshots for enterprise workloads

Minimize write overhead

Prototype implemented in ext4: gcext4

Reduce overhead up to 68x from state of the art

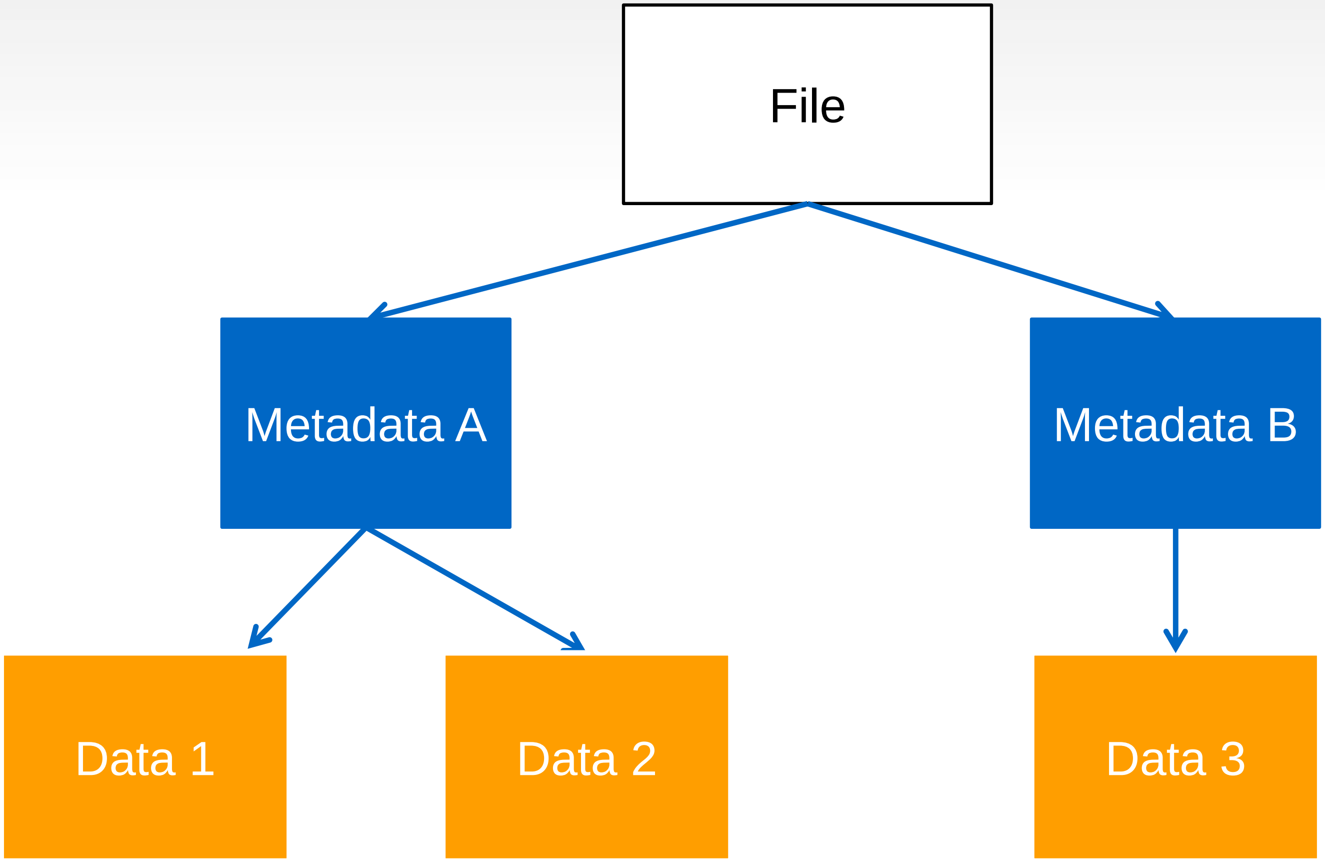
© 2014 NetApp, Inc. All rights reserved.

Outline

1) Overview
2) Theory
3) Implementation
4) Evaluation

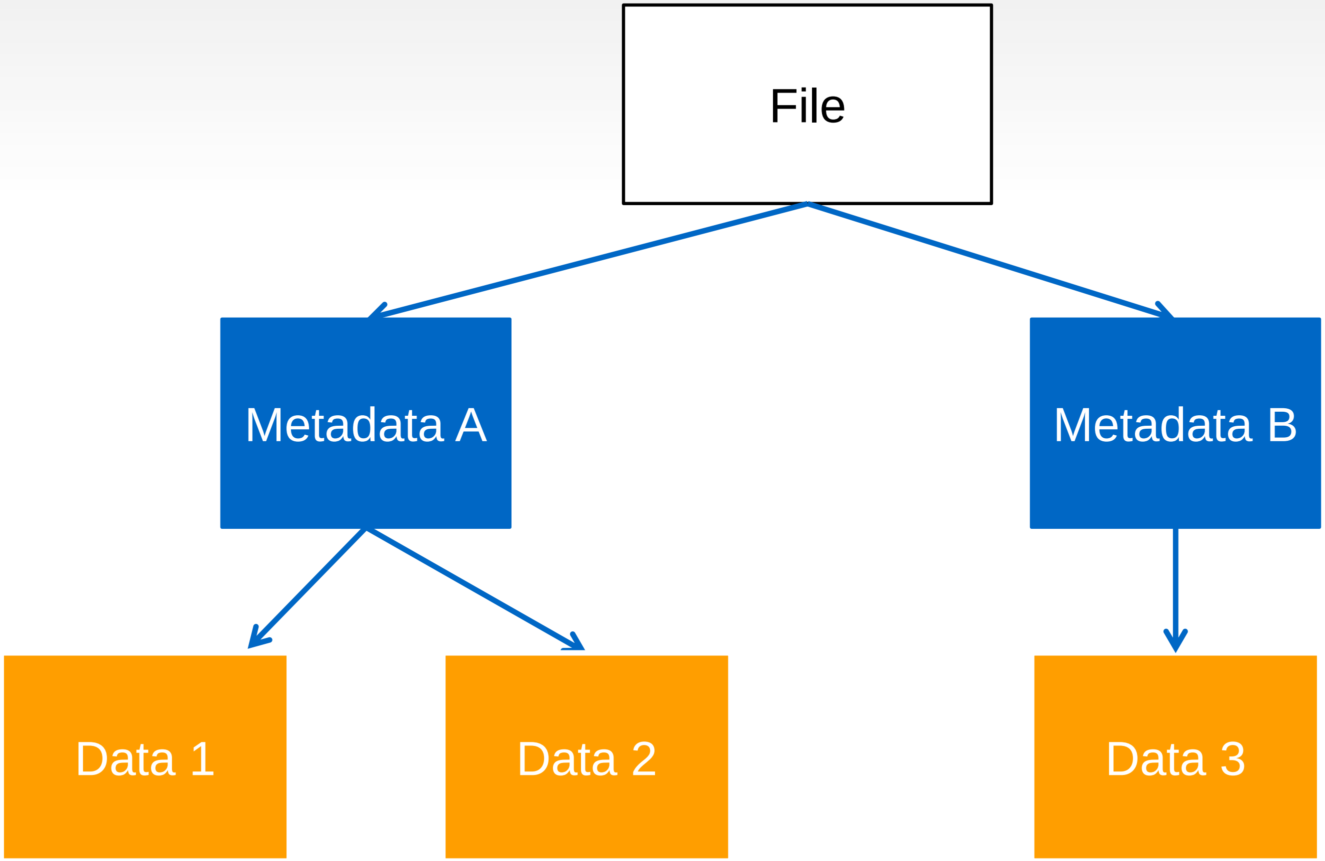
© 2014 NetApp, Inc. All rights reserved.

Snapshot Overview



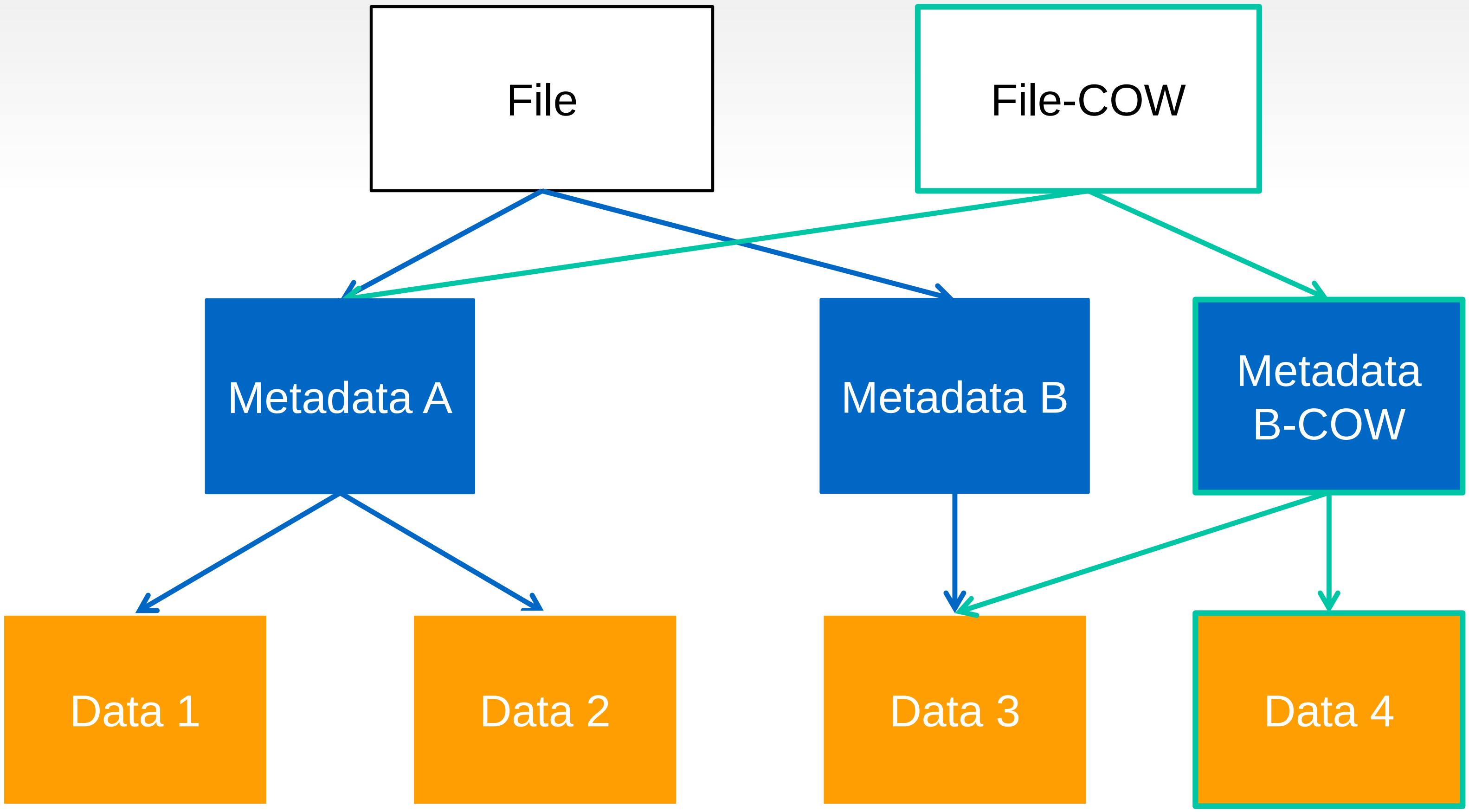
© 2014 NetApp, Inc. All rights reserved.

Copy-on-write



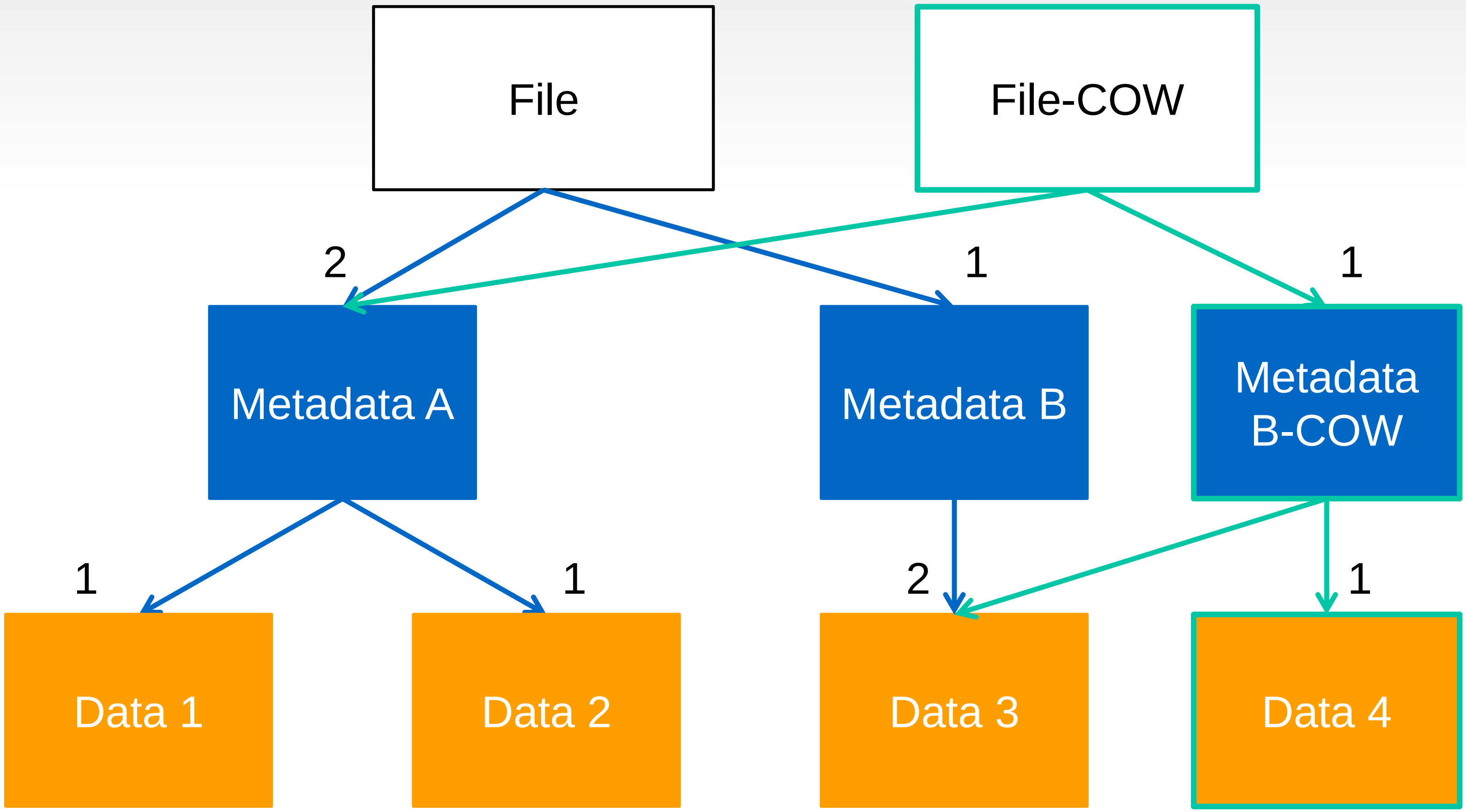
© 2014 NetApp, Inc. All rights reserved.

Copy-on-Write: Appending



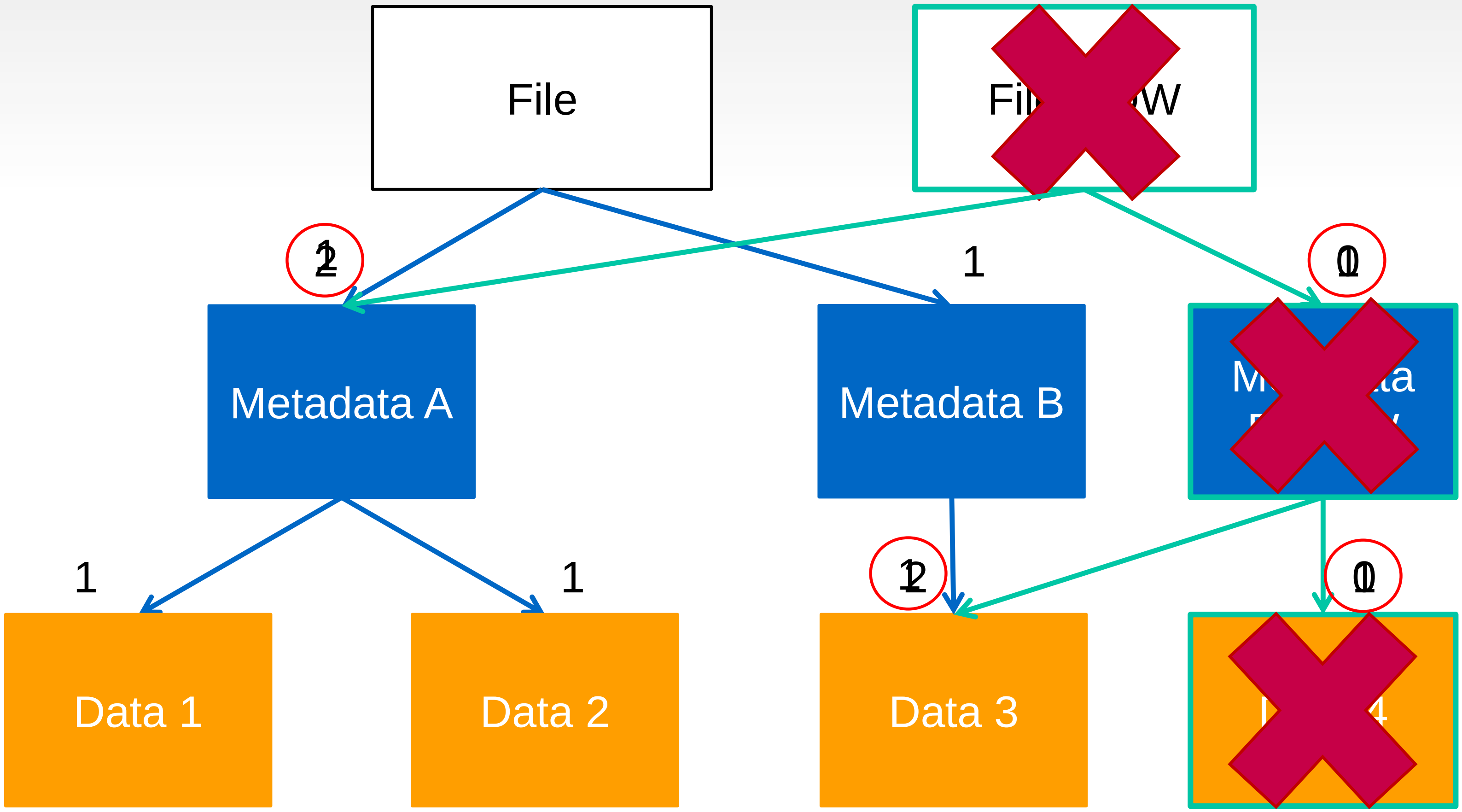
© 2014 NetApp, Inc. All rights reserved.

Hierarchical Refcounts



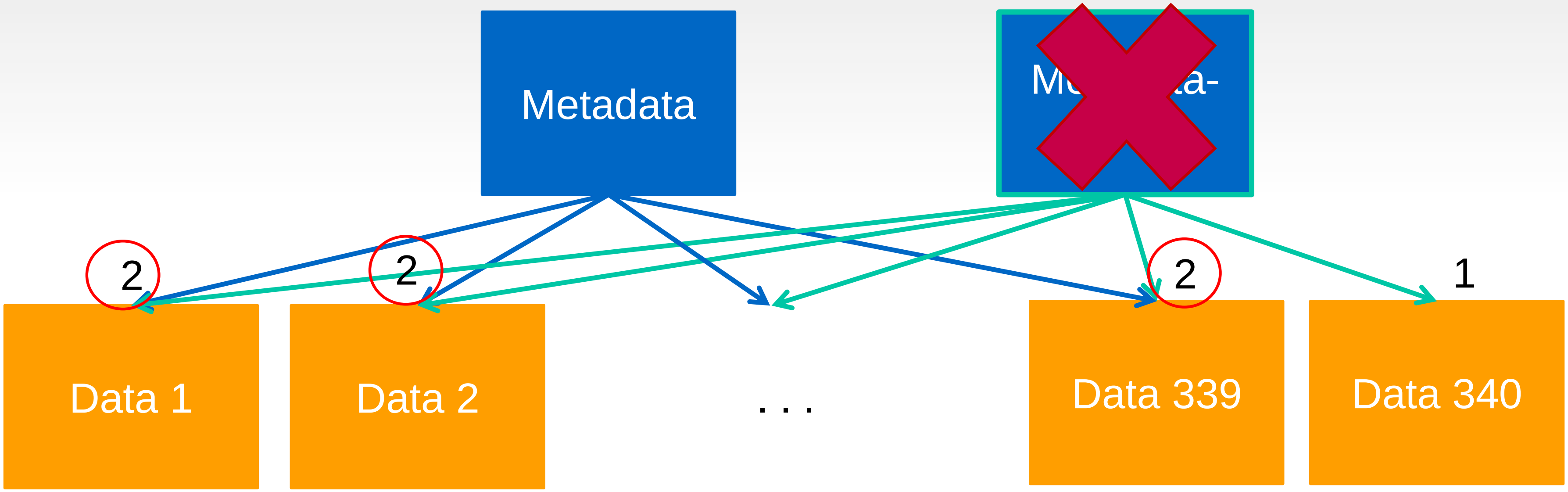
© 2014 NetApp, Inc. All rights reserved.

Hierarchical Refcounts: Deletion



© 2014 NetApp, Inc. All rights reserved.

Problem: Fan-out



Even worse for writes

© 2014 NetApp, Inc. All rights reserved.

Potential Solutions

© 2014 NetApp, Inc. All rights reserved.

Delay

Write refcount changes to log, then checkpoint

Hope that they’ll cancel

No guarantee this will actually happen

Update storm

© 2014 NetApp, Inc. All rights reserved.

GCTrees

Don’t count references at all

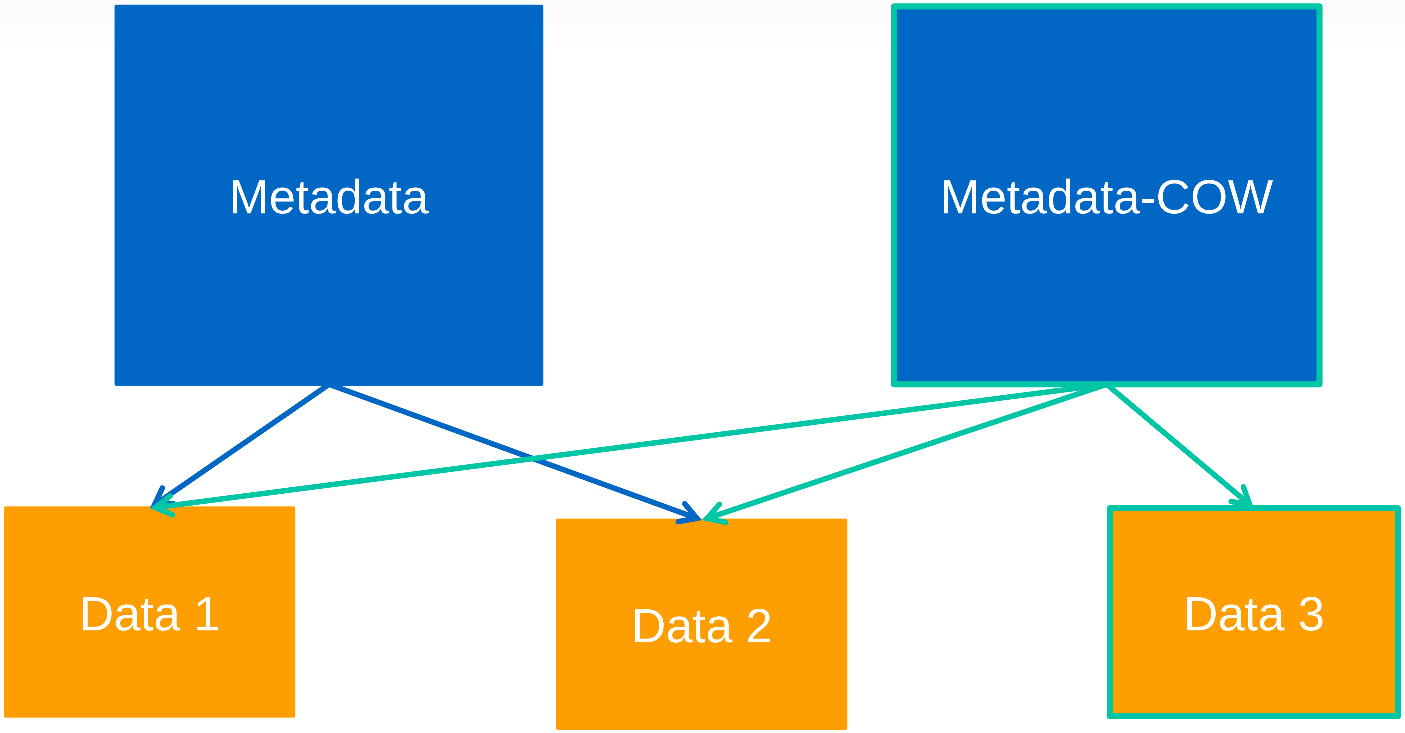
Track lineage

Use garbage collection (GC) to determine what can be freed

Keep deletion, write overhead minimal

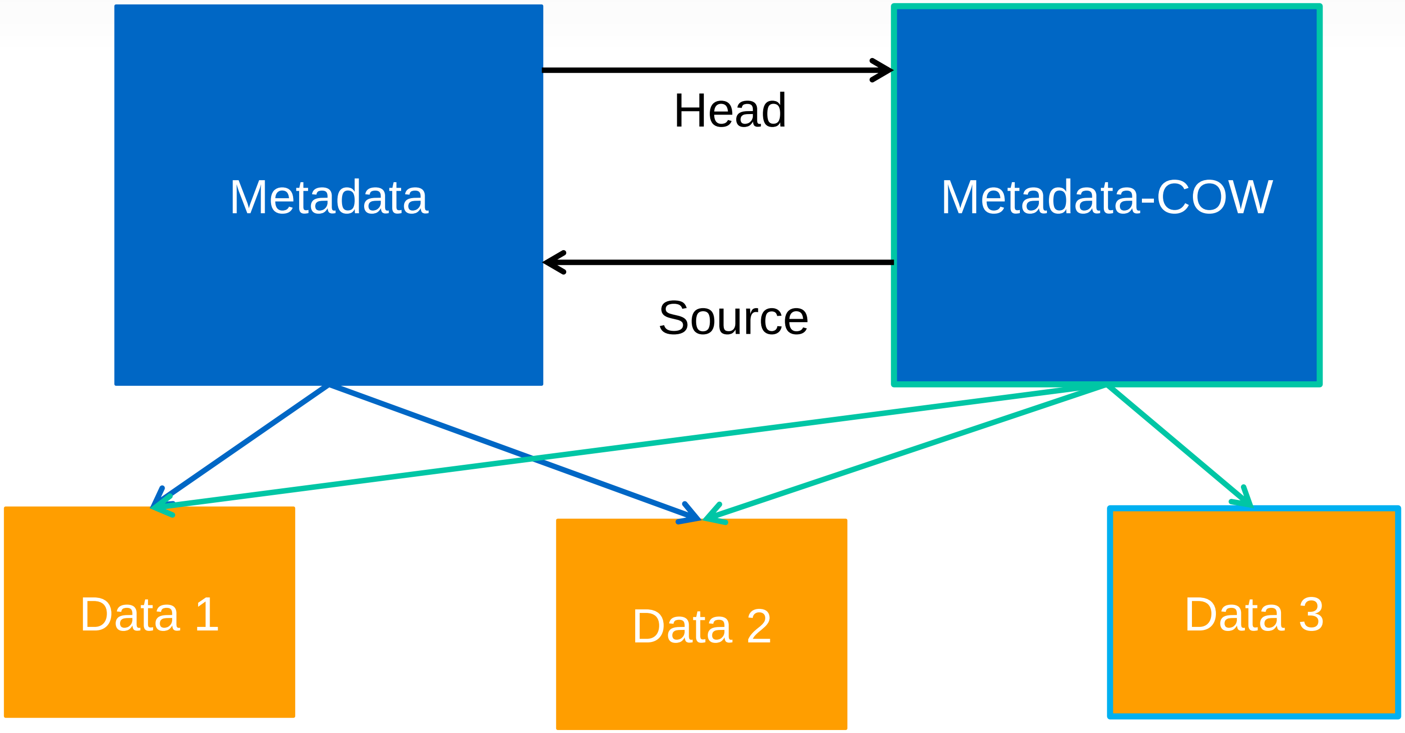
© 2014 NetApp, Inc. All rights reserved.

Basic Metadata Structure



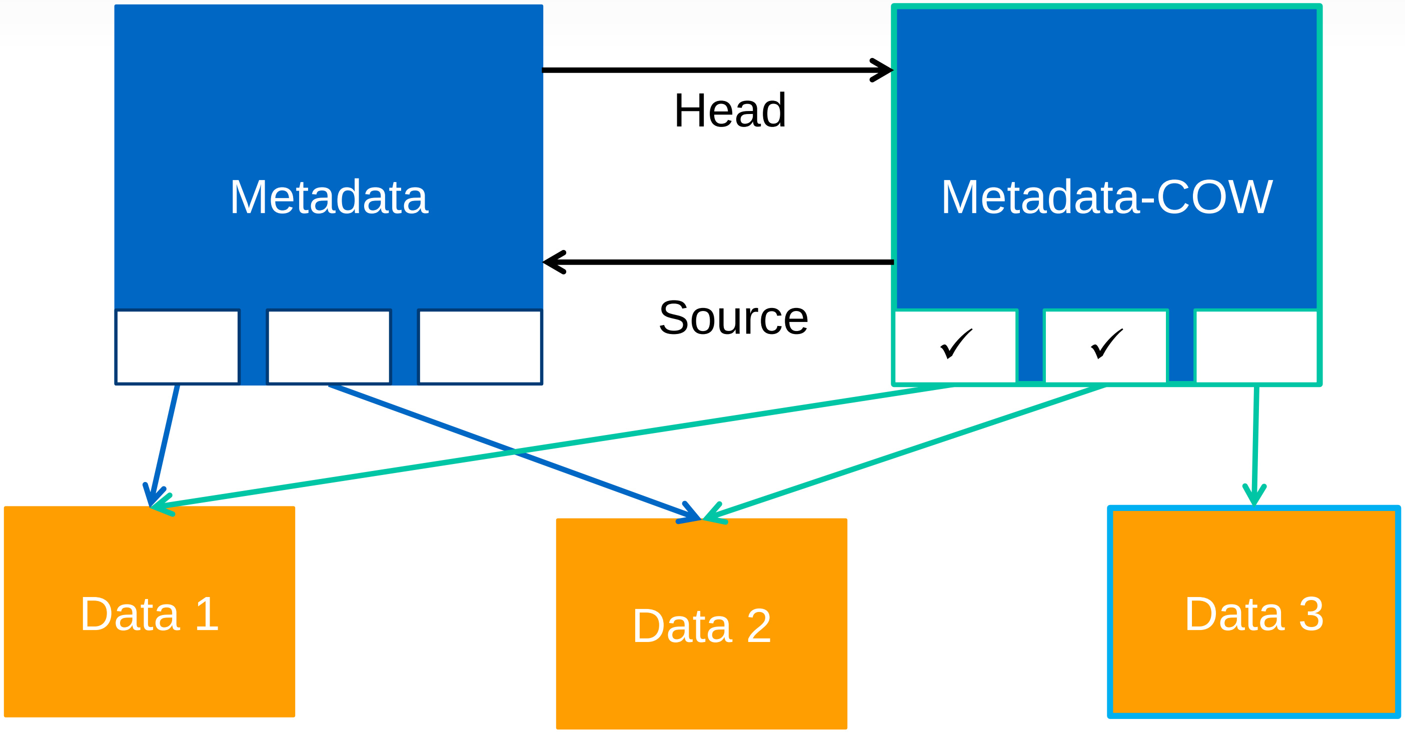
© 2014 NetApp, Inc. All rights reserved.

Basic Metadata Structure



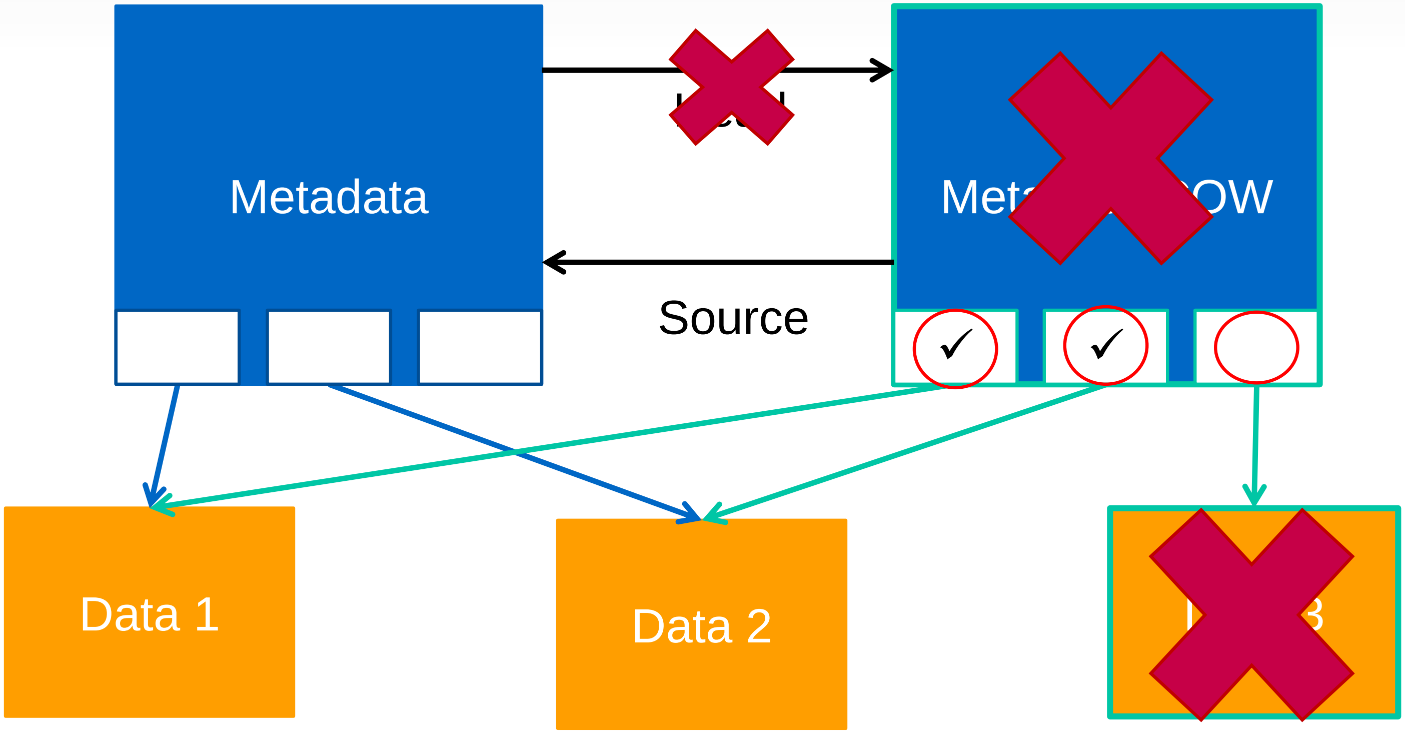
© 2014 NetApp, Inc. All rights reserved.

Basic Metadata Structure



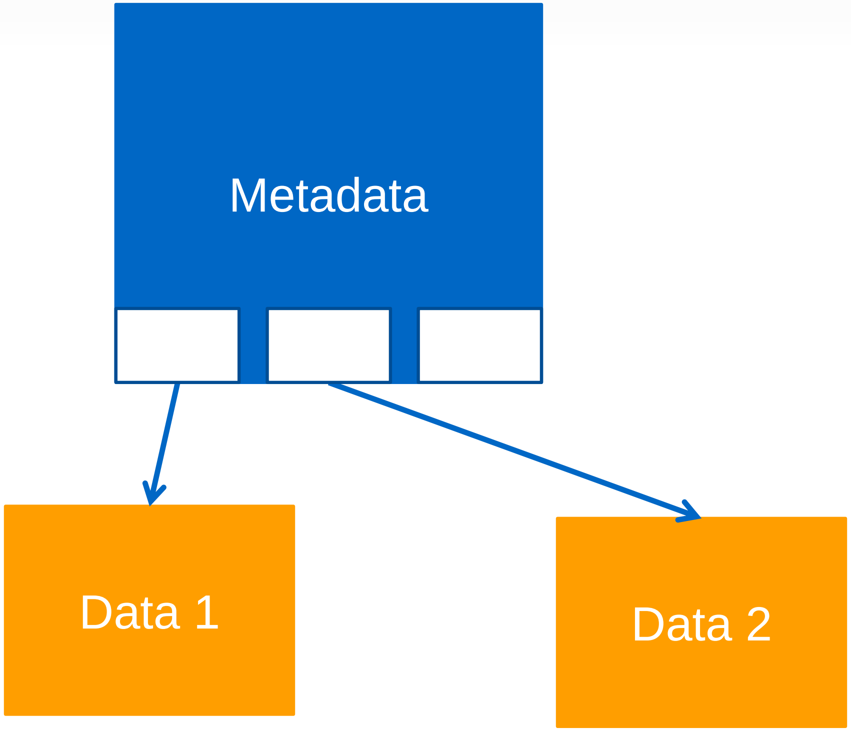
© 2014 NetApp, Inc. All rights reserved.

Deletion: Child



© 2014 NetApp, Inc. All rights reserved.

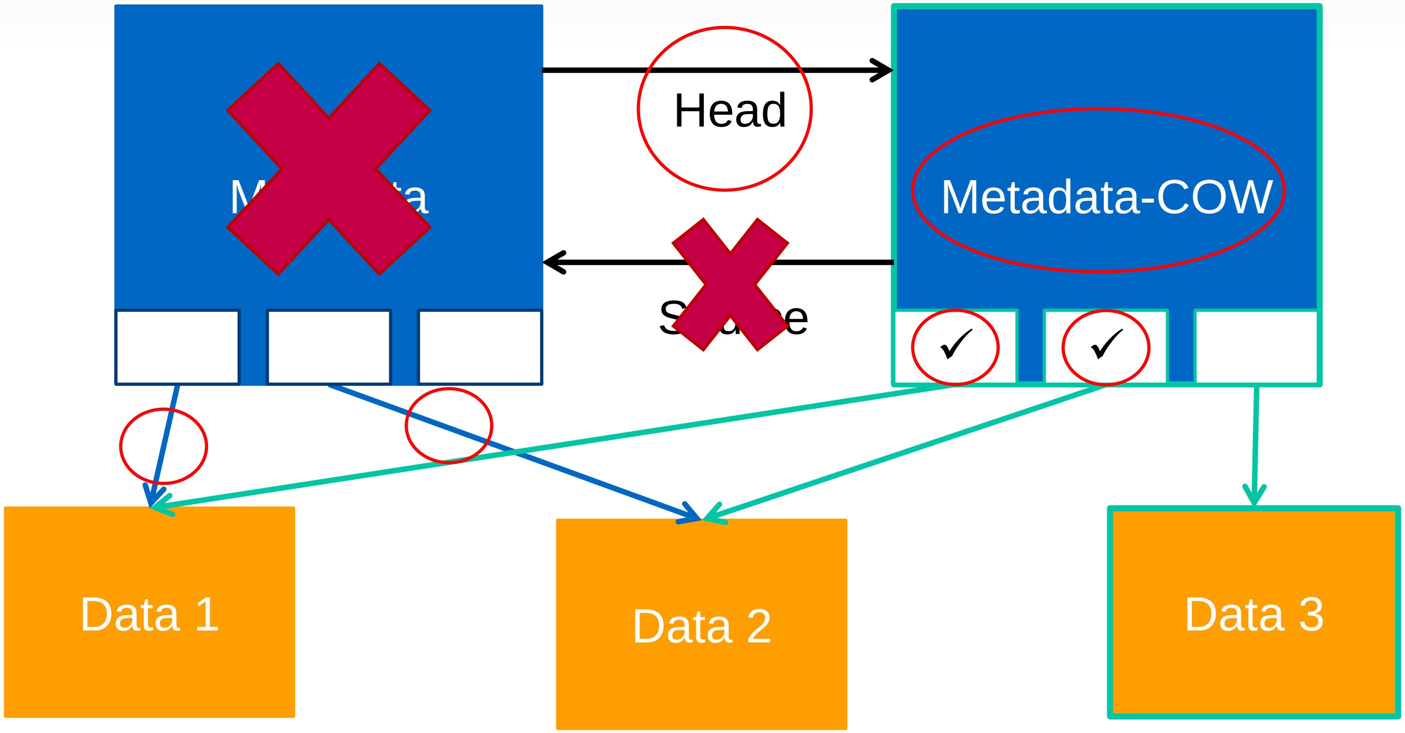
Deletion: Child



1 extra write

© 2014 NetApp, Inc. All rights reserved.

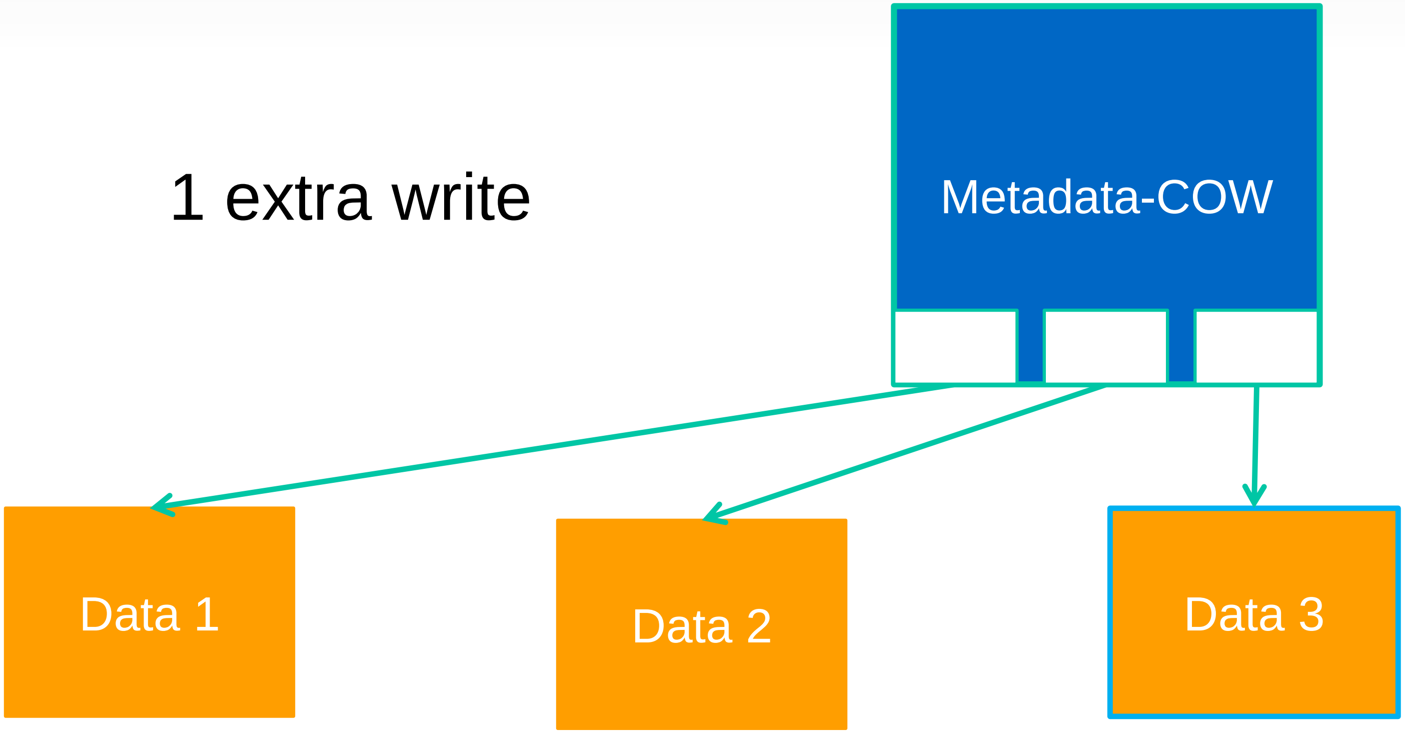
Deletion: Parent



© 2014 NetApp, Inc. All rights reserved.

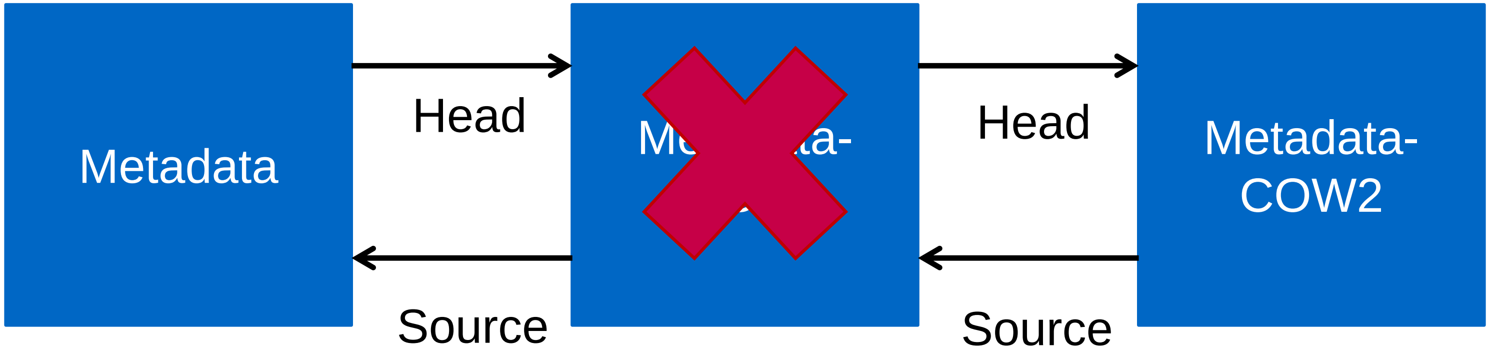
Deletion: Parent

1 extra write



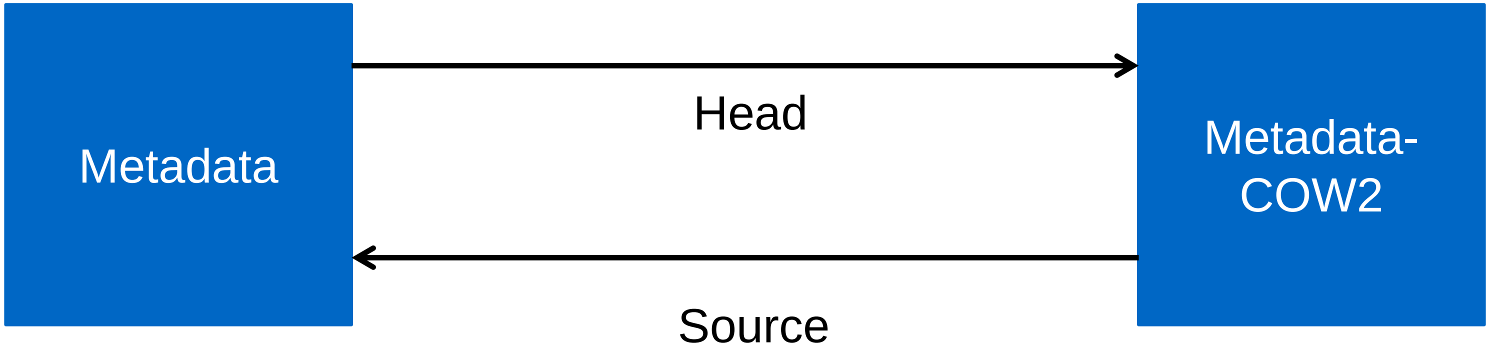
© 2014 NetApp, Inc. All rights reserved.

Deletion in a Snapshot Chain



© 2014 NetApp, Inc. All rights reserved.

Deletion in a Snapshot Chain



2 extra writes

© 2014 NetApp, Inc. All rights reserved.

What about B+ trees?

Tree ops can move pointers

Need additional pointers: next and previous

Can increase deletion overhead

© 2014 NetApp, Inc. All rights reserved.

GCTree summary

One extra write per metadata block when writing

One to two extra writes for most deletions

Avoid deletion overhead with background scans

© 2014 NetApp, Inc. All rights reserved.

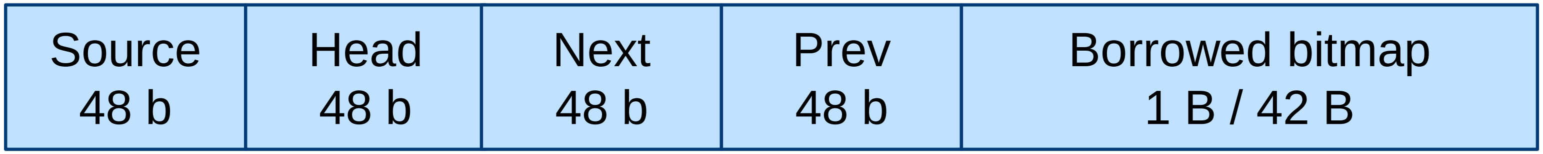
Implementing gcext4

Ext4: Extent-based file system

Two metadata types: inodes, extent blocks

© 2014 NetApp, Inc. All rights reserved.

GCTree Metadata Layout

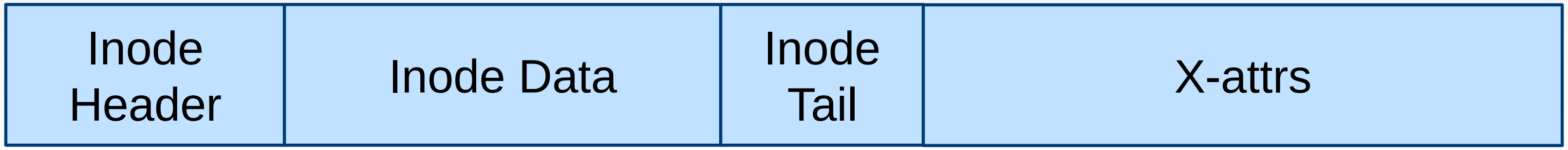


Source	Head	Next	Prev	Borrowed bitmap
48 b	48 b	48 b	48 b	1 B / 42 B

© 2014 NetApp, Inc. All rights reserved.

Adding GCTree metadata

ext4 inode (256 B):



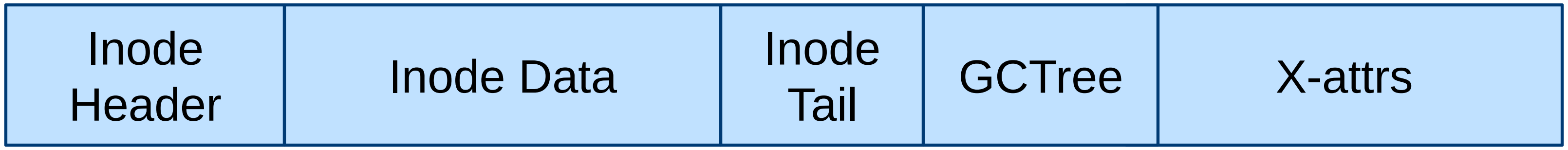
Inode Header

Inode Data

Inode Tail

X-attrs

gcext4 inode (256 B):



Inode Header

Inode Data

Inode Tail

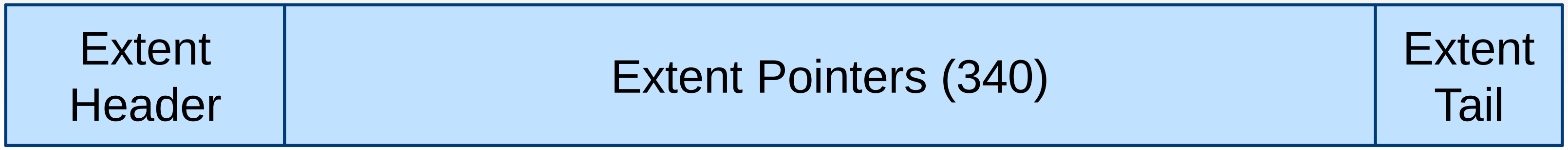
GCTree

X-attrs

© 2014 NetApp, Inc. All rights reserved.

Adding GCTree metadata

ext4 extent (4096 KB):

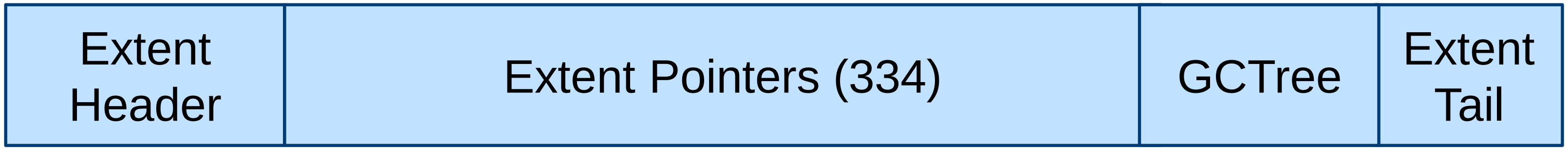


Extent Header

Extent Pointers (340)

Extent Tail

gcext4 extent (4096 KB):



Extent Header

Extent Pointers (334)

GCTree

Extent Tail

© 2014 NetApp, Inc. All rights reserved.

Adapting ext4 to COW:

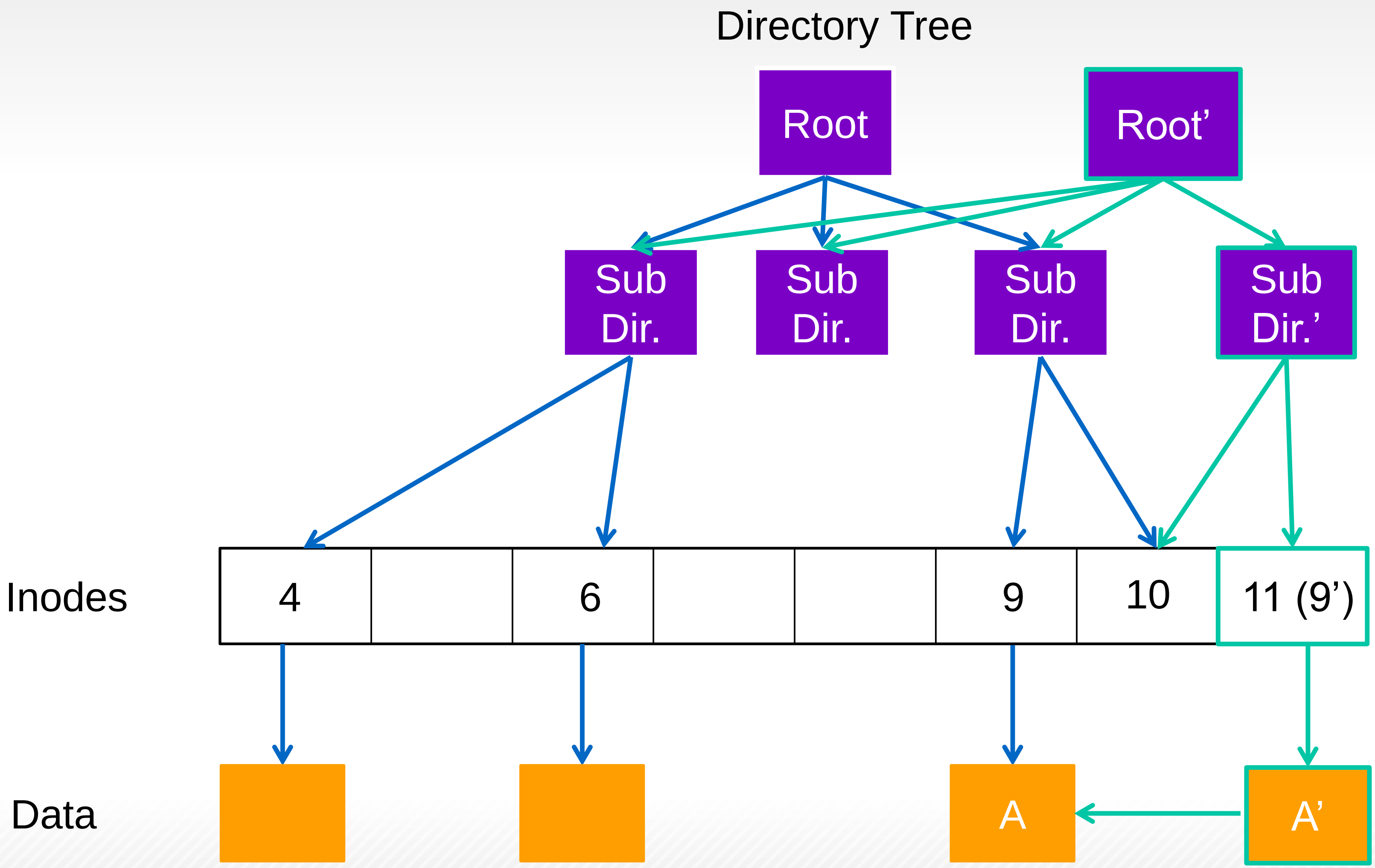
Straightforward, but fiddly

COW once per snapshot

Inodes proved problematic

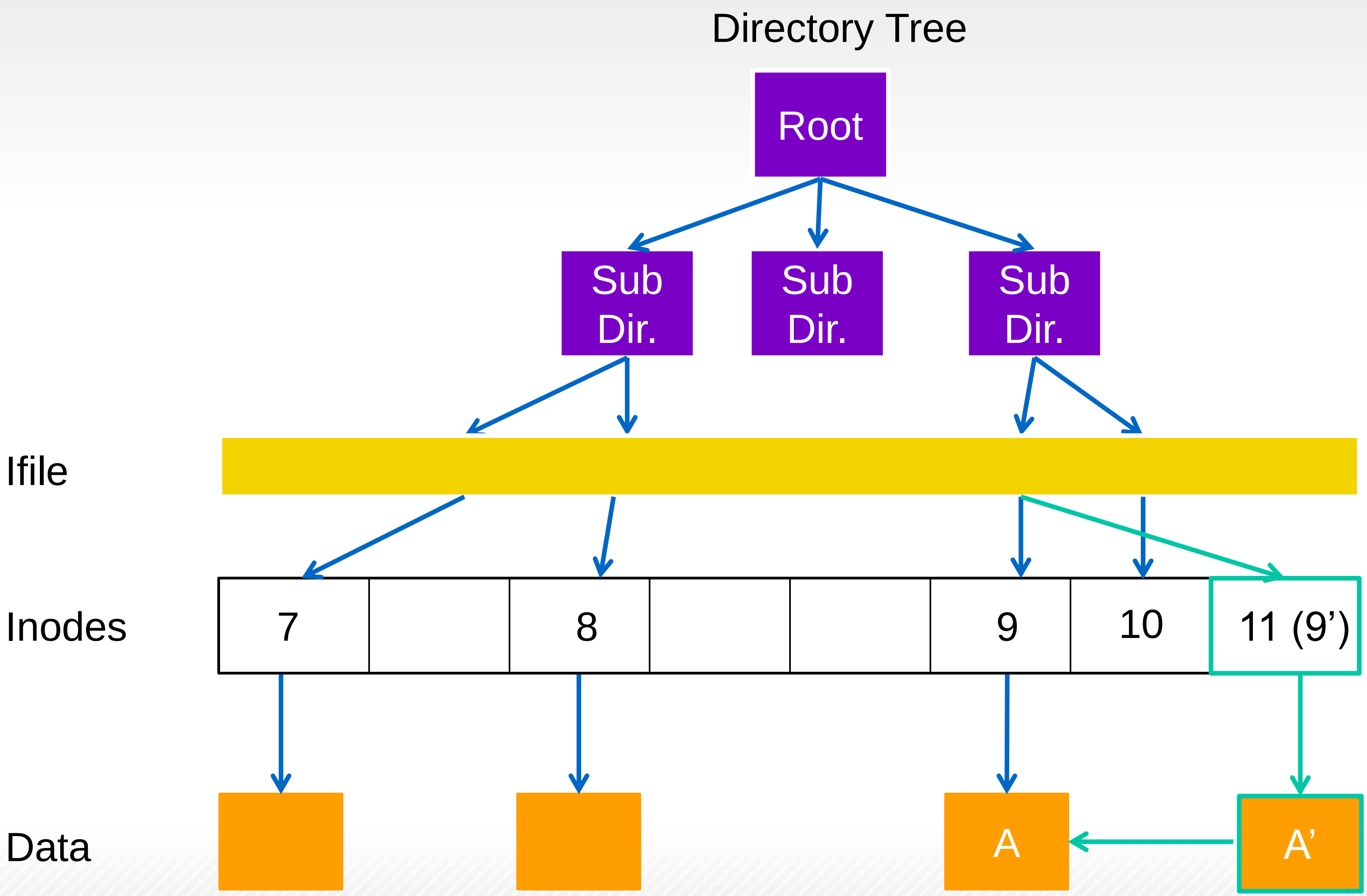
© 2014 NetApp, Inc. All rights reserved.

Fixed Location Inodes



© 2014 NetApp, Inc. All rights reserved.

Solution: ifile



© 2014 NetApp, Inc. All rights reserved.

Implementing Deletion

Separate kernel threads act as scanners

Deletion enqueues a message

Scanners process message, do actual deletion

Removing a snapshot deletes an ifile

© 2014 NetApp, Inc. All rights reserved.

Evaluation

Do they work?

How do they compare?

© 2014 NetApp, Inc. All rights reserved.

Experimental Set-up

3GHz Core 2 Duo, 6GB RAM

7200 RPM, 160GB hard drive

© 2014 NetApp, Inc. All rights reserved.

Basic Benchmarks

Fileserver

OLTP

VM: fileserver in VirtualBox

© 2014 NetApp, Inc. All rights reserved.

Benchmark Configuration

Storage footprint: 2x memory

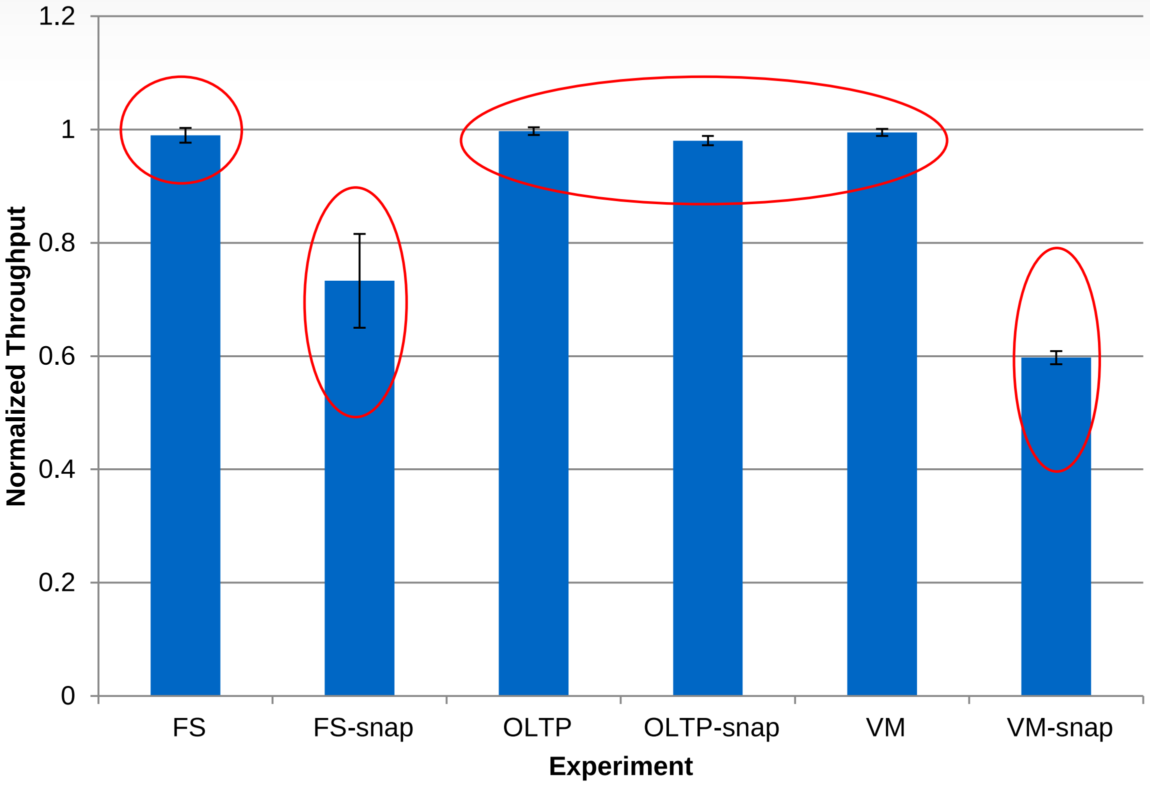
5 repetitions, 3 hours each

Snapshot per hour for Fileserver, VM

Snapshot per five minutes for OLTP

© 2014 NetApp, Inc. All rights reserved. NetApp Confidential – Limited Use

Comparison to ext4



© 2014 NetApp, Inc. All rights reserved.

Hierarchical Refcount Comparison

Direct performance comparison infeasible

Look at block-write overhead

© 2014 NetApp, Inc. All rights reserved.

File Systems for Comparison

Btrfs: refcounting file system

Simulation in gcext4: accounts for differences in btrfs

© 2014 NetApp, Inc. All rights reserved. NetApp Confidential – Limited Use

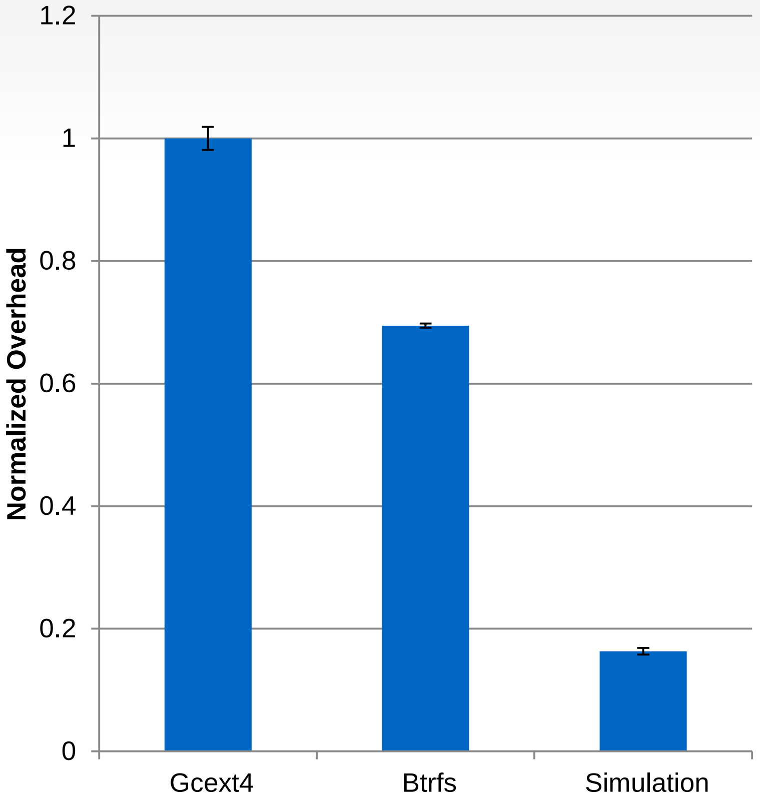
Simulation Methodology

Assumes refcounts stored in a contiguous map

Use a 15MB durable log

© 2014 NetApp, Inc. All rights reserved.

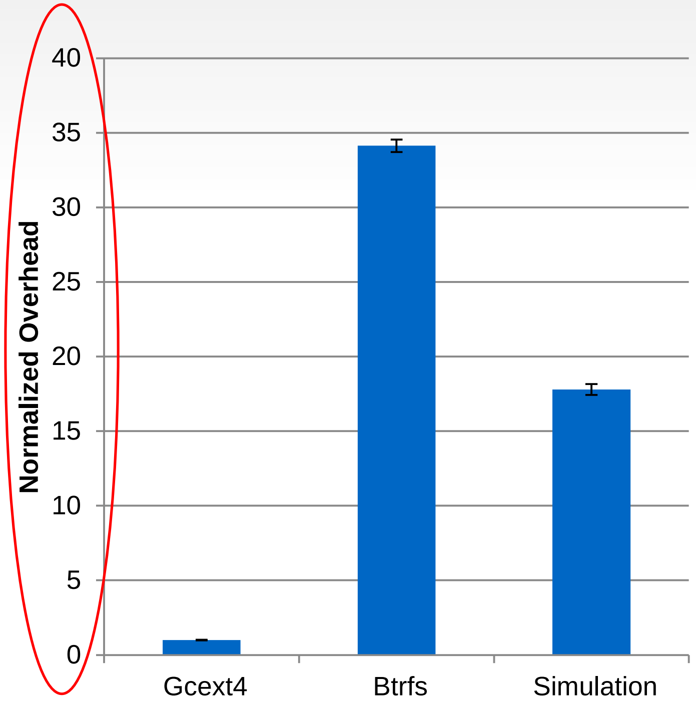
Results: Traditional Workloads



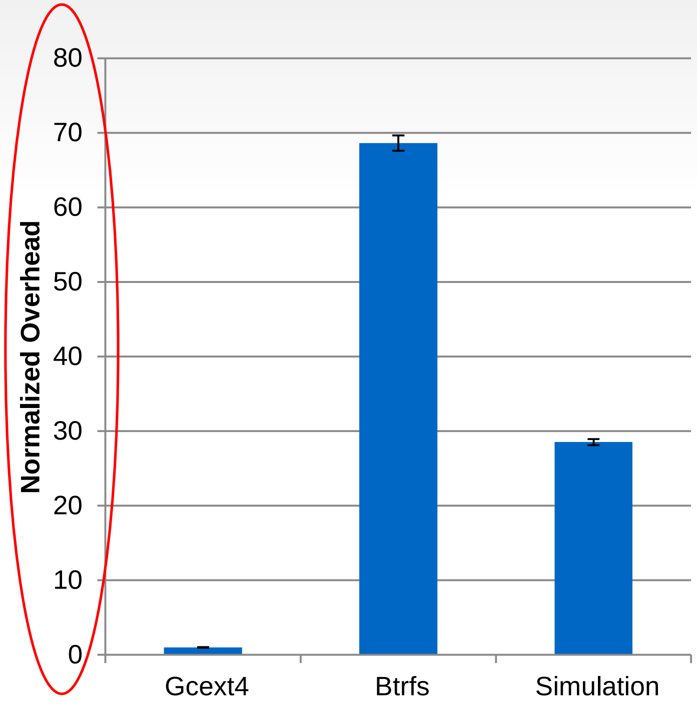
Fileserver

© 2014 NetApp, Inc. All rights reserved. NetApp Confidential – Limited Use

Results: Enterprise Workloads



OLTP



VM

© 2014 NetApp, Inc. All rights reserved. NetApp Confidential – Limited Use

Conclusion

GCTrees: snapshots for enterprise workloads

Substantial gulf between refcounts and GCTrees

Optimal choice depends on workload

© 2014 NetApp, Inc. All rights reserved.

Thank you!