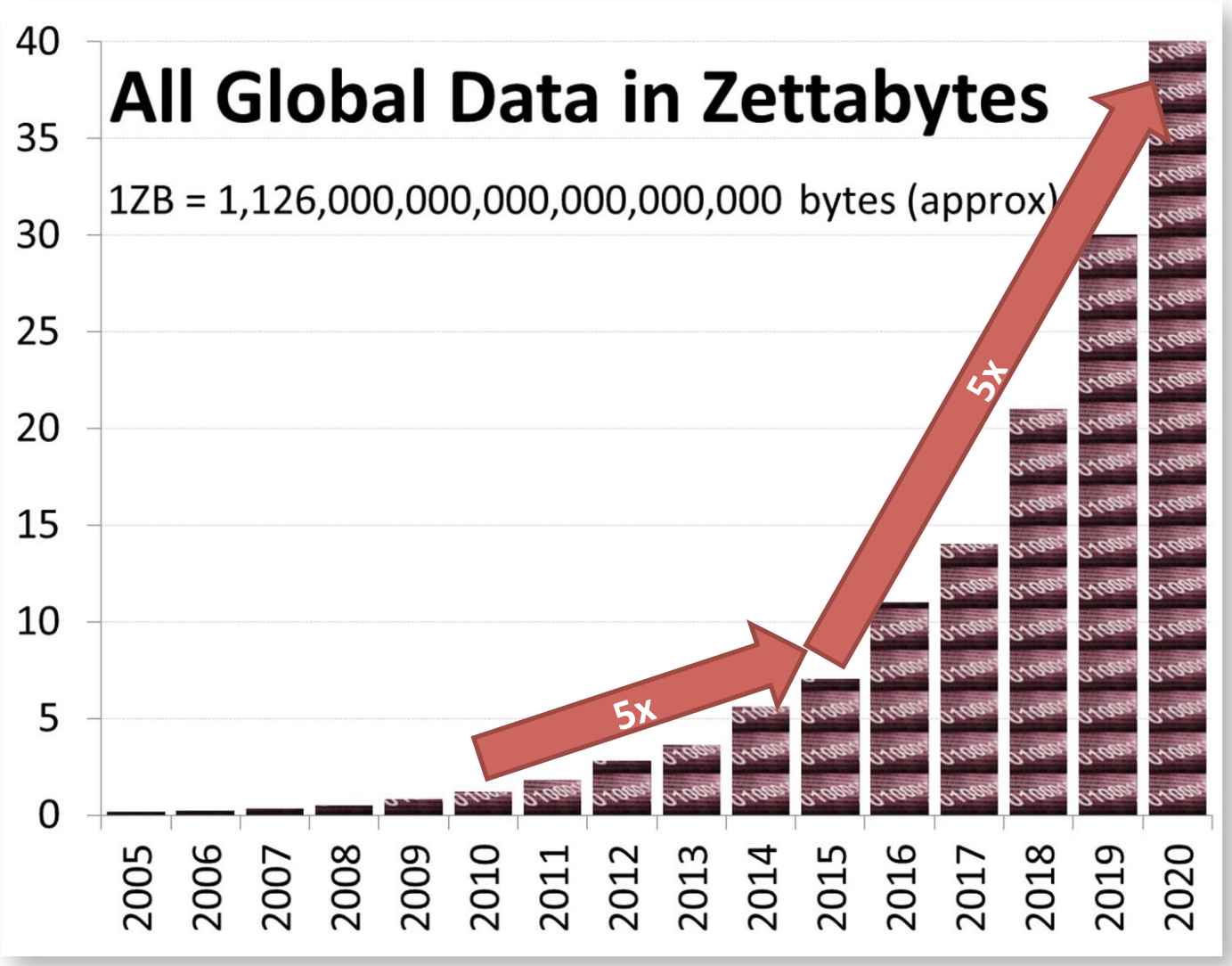
FAST'15

A Tale of Two Erasure Codes in HDFS

Mingyuan Xia, Mohit Saxena
Mario Blaum, David Pease

IBM Research Almaden & McGill University

Really Big Data Today

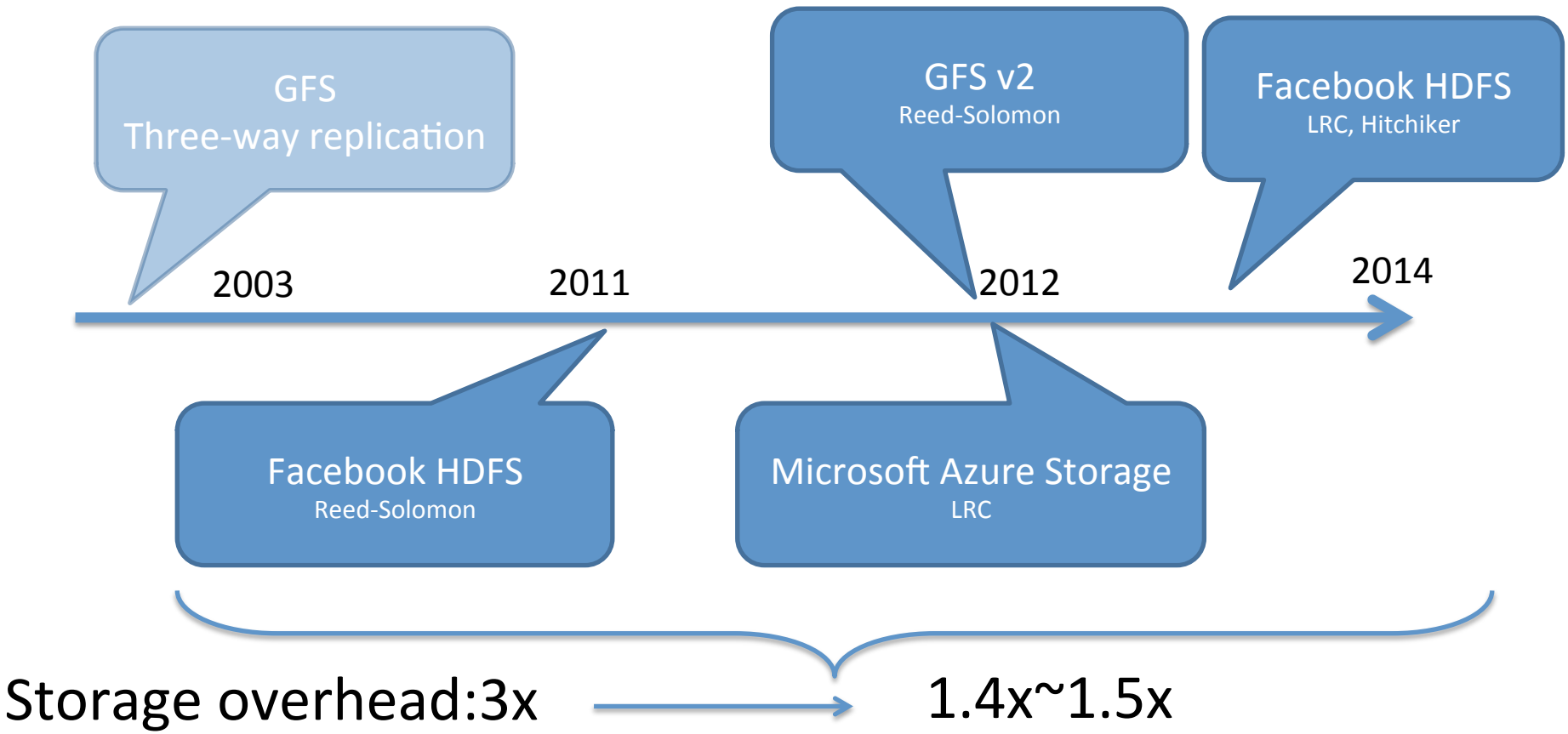


UNECE data

Big Data Storage

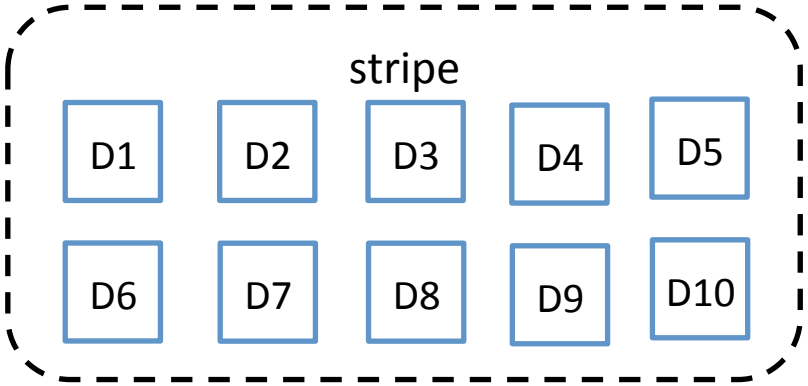
Storage overhead:3x

Big Data Storage



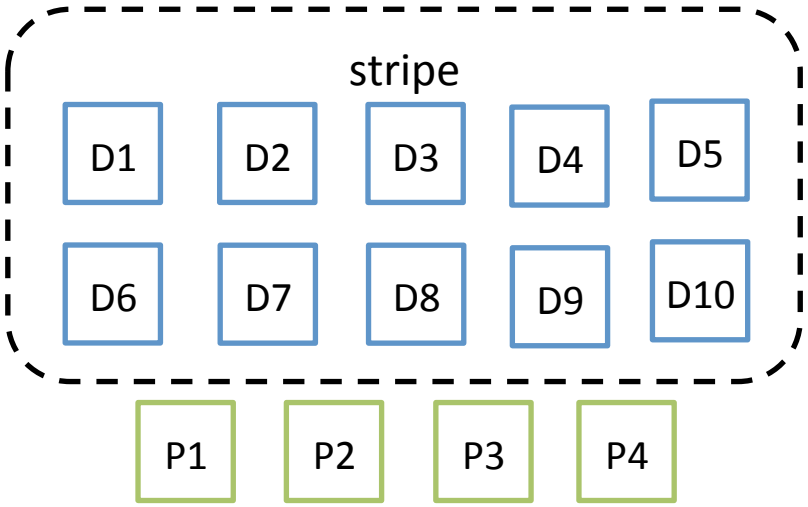
Erasure Coded Storage saves millions of $ for capital cost

Erasure Coding 101



Facebook HDFS: Reed Solomon (14,10)

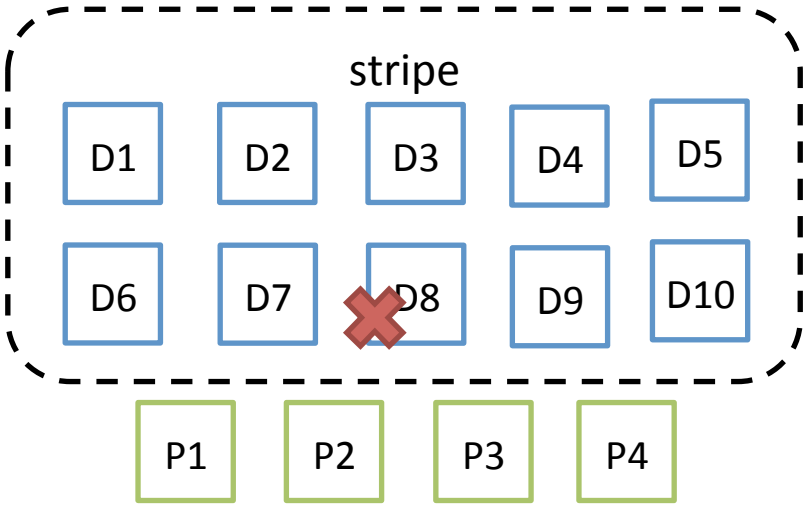
Erasure Coding 101



Facebook HDFS: Reed Solomon (14,10)

- Compute 4 parities per 10 data blocks
- Storage overhead: 1.4x=14/10

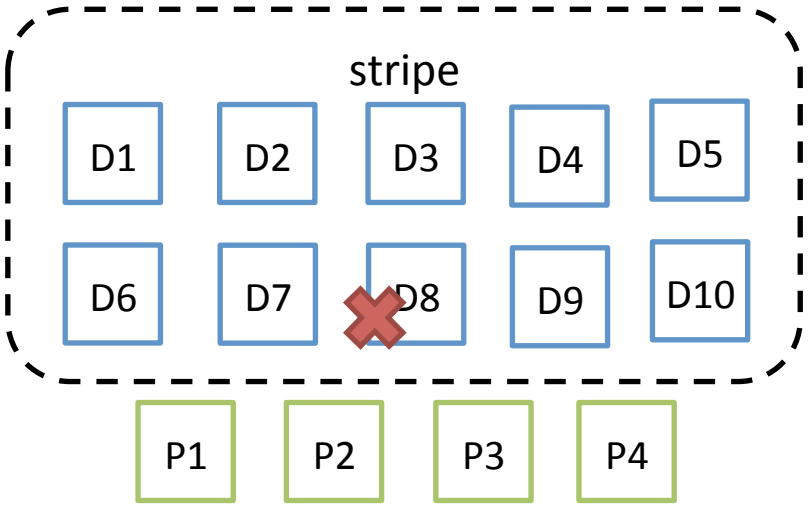
Erasure Coding 101



Facebook HDFS: Reed Solomon (14,10)

- Compute 4 parities per 10 data blocks
- Storage overhead: 1.4x=14/10
- For a single failure, RS needs any 10 blocks over network from other nodes to recover

Erasure Coding 101

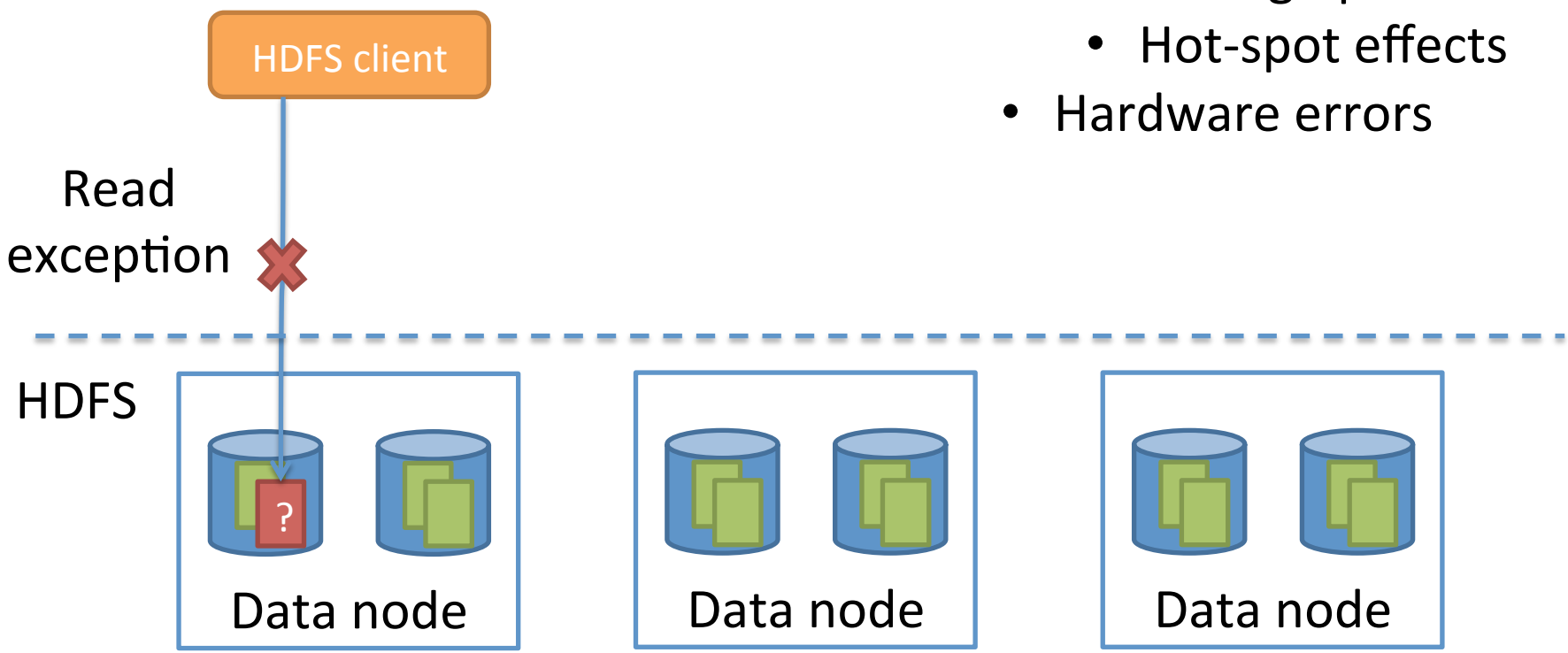


Facebook HDFS: Reed Solomon (14,10)

- Compute 4 parities per 10 data blocks
- Storage overhead: 1.4x
- For a single failure, RS needs any 10 blocks over network from other nodes to recover

Problems:
- Degraded read
- Data reconstruction

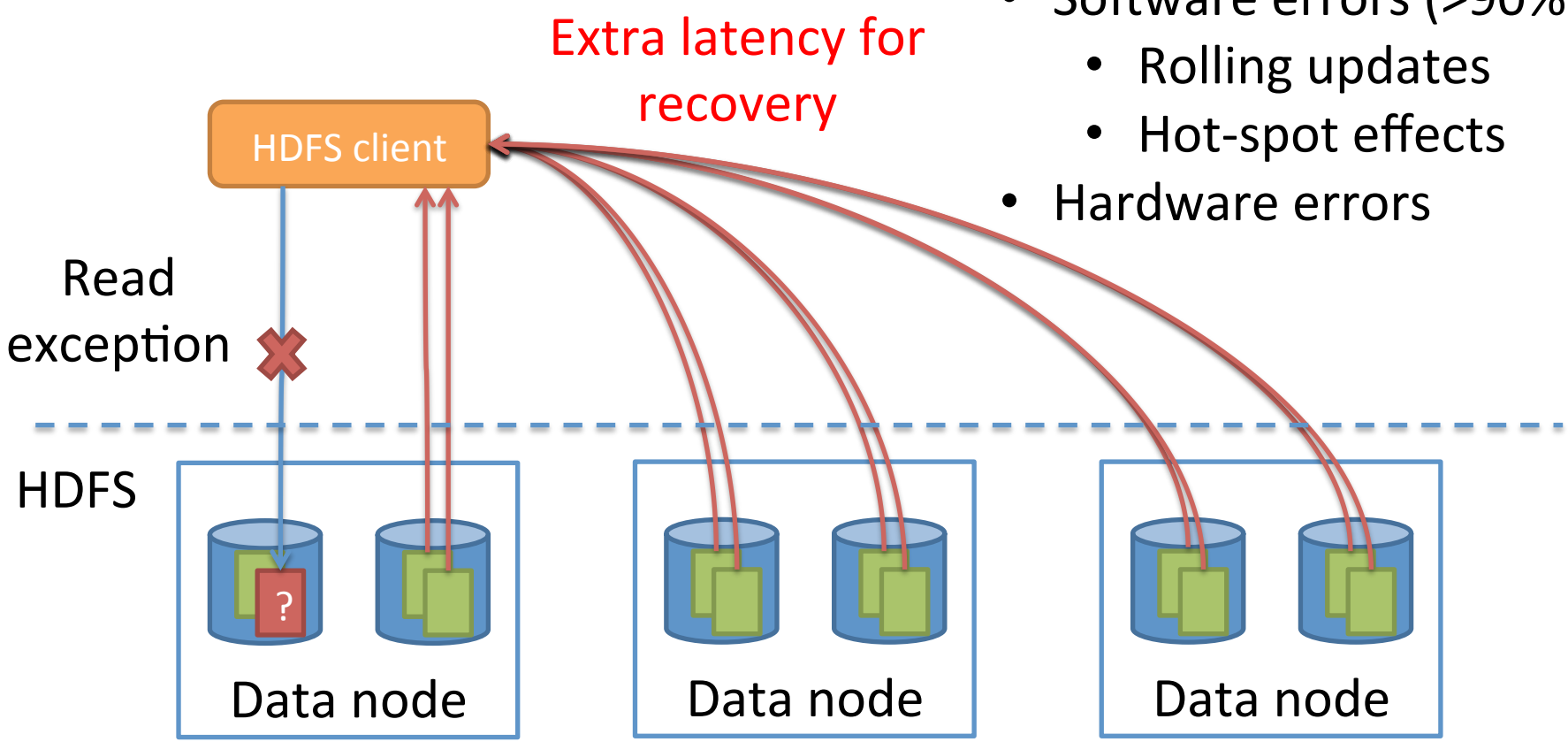
Problem 1: Degraded Read



Causes

- Software errors (>90%)
  - Rolling updates
  - Hot-spot effects
- Hardware errors

Problem 1: Degraded Read



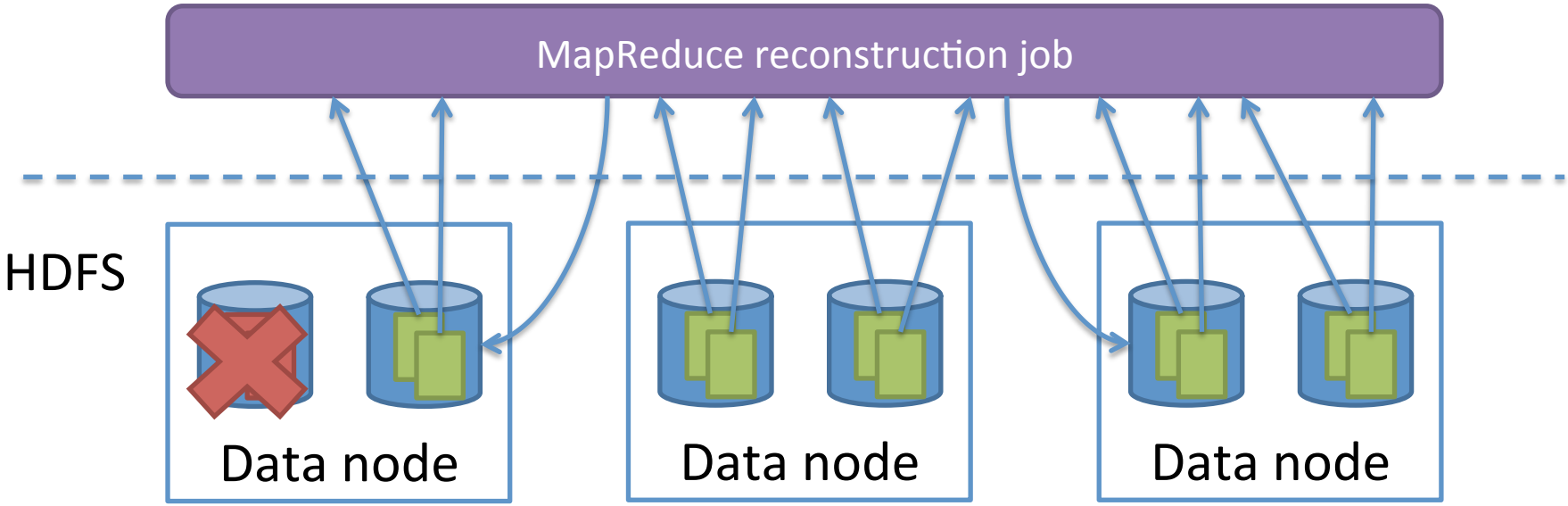
Causes

- Software errors (>90%)
  - Rolling updates
  - Hot-spot effects
- Hardware errors

Problem 2: Data Reconstruction

Causes

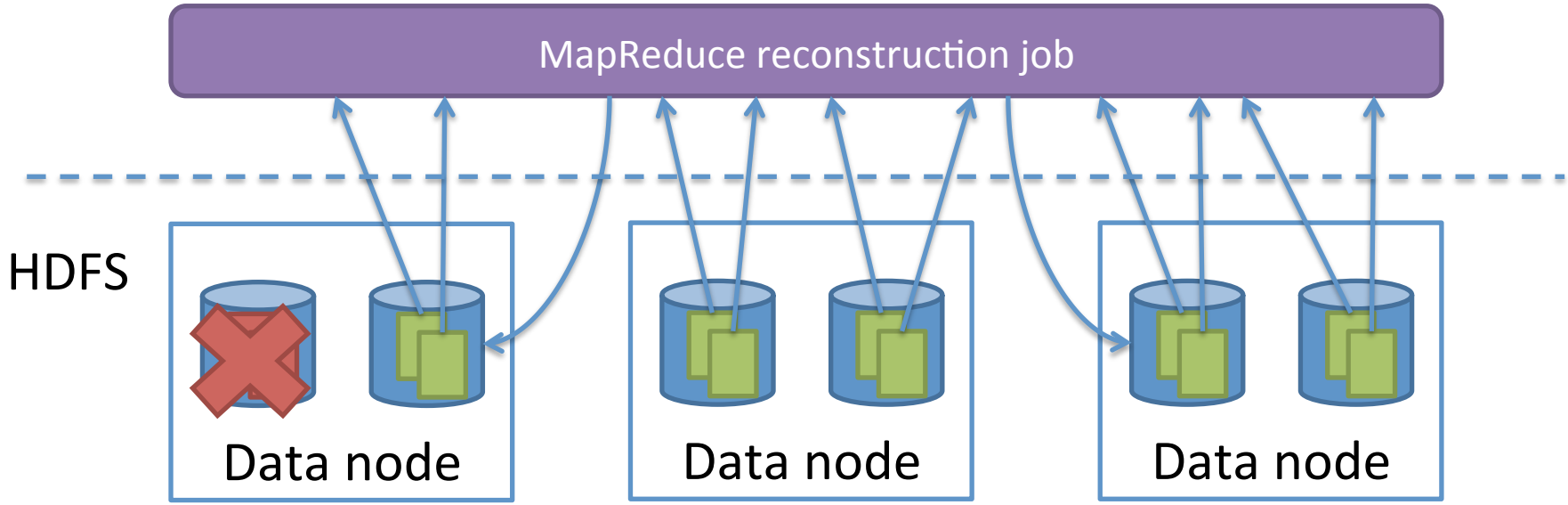
- Disk or node failure
- Decommissioned nodes



Problem 2: Disk/node Reconstruction

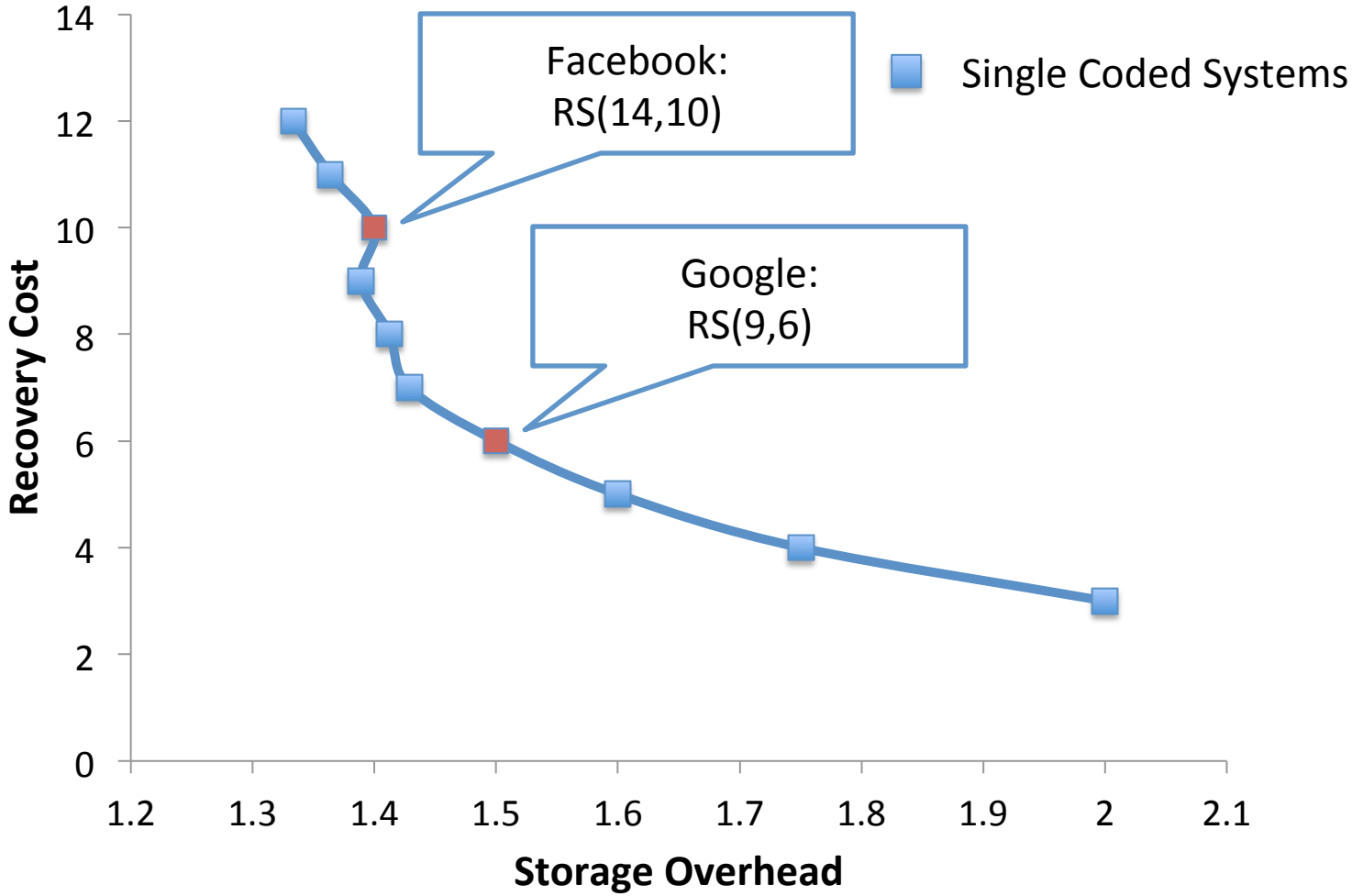
<u>Production Clusters</u>

- New data: 500TB~900TB/day
- Failure: lose ~100k blocks/day
- Reconstruction traffic: 180TB/day

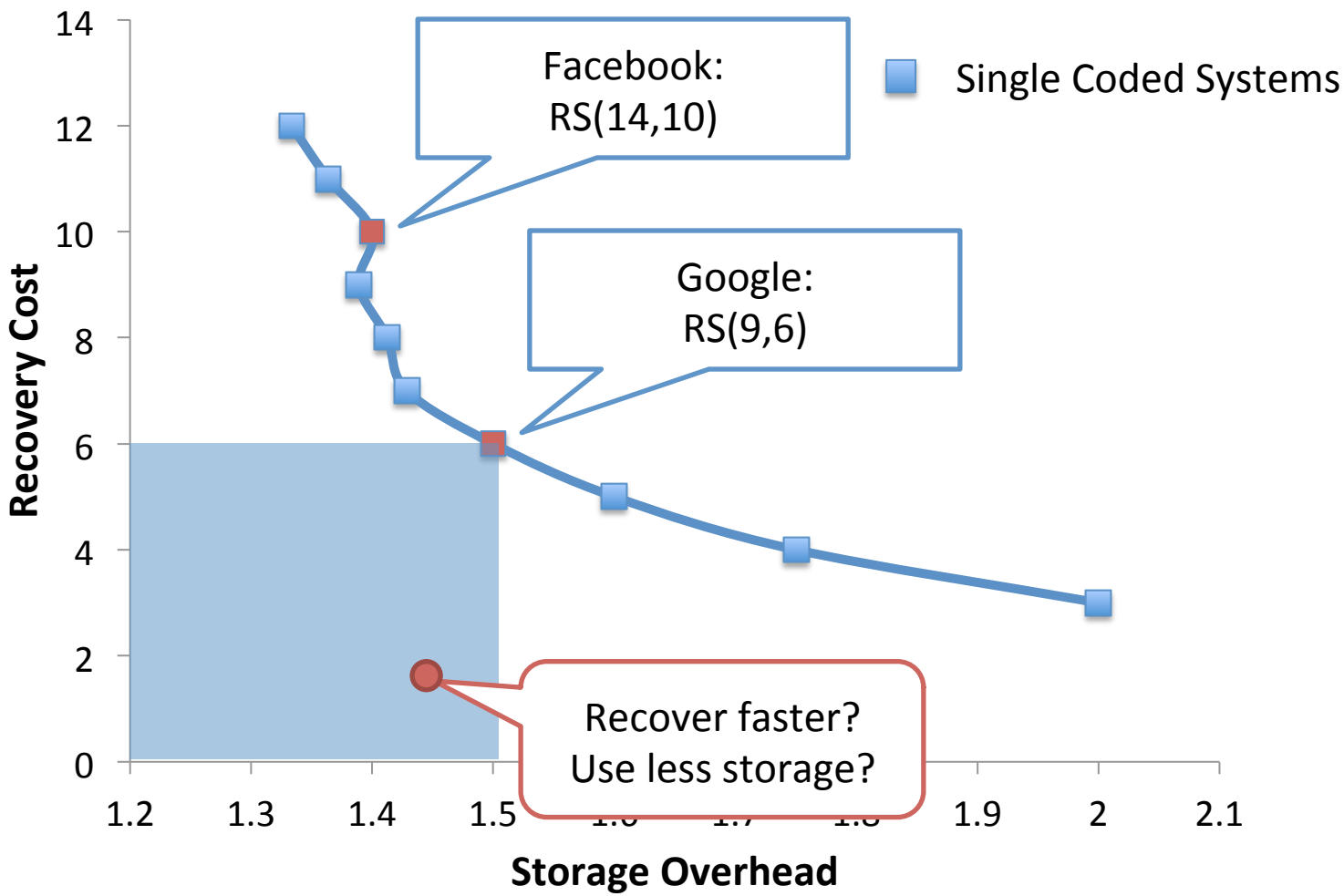


[SIGCOMM'14] Rashmi et al. A "hitchhiker's" guide to fast and efficient data reconstruction in erasure-coded data centers

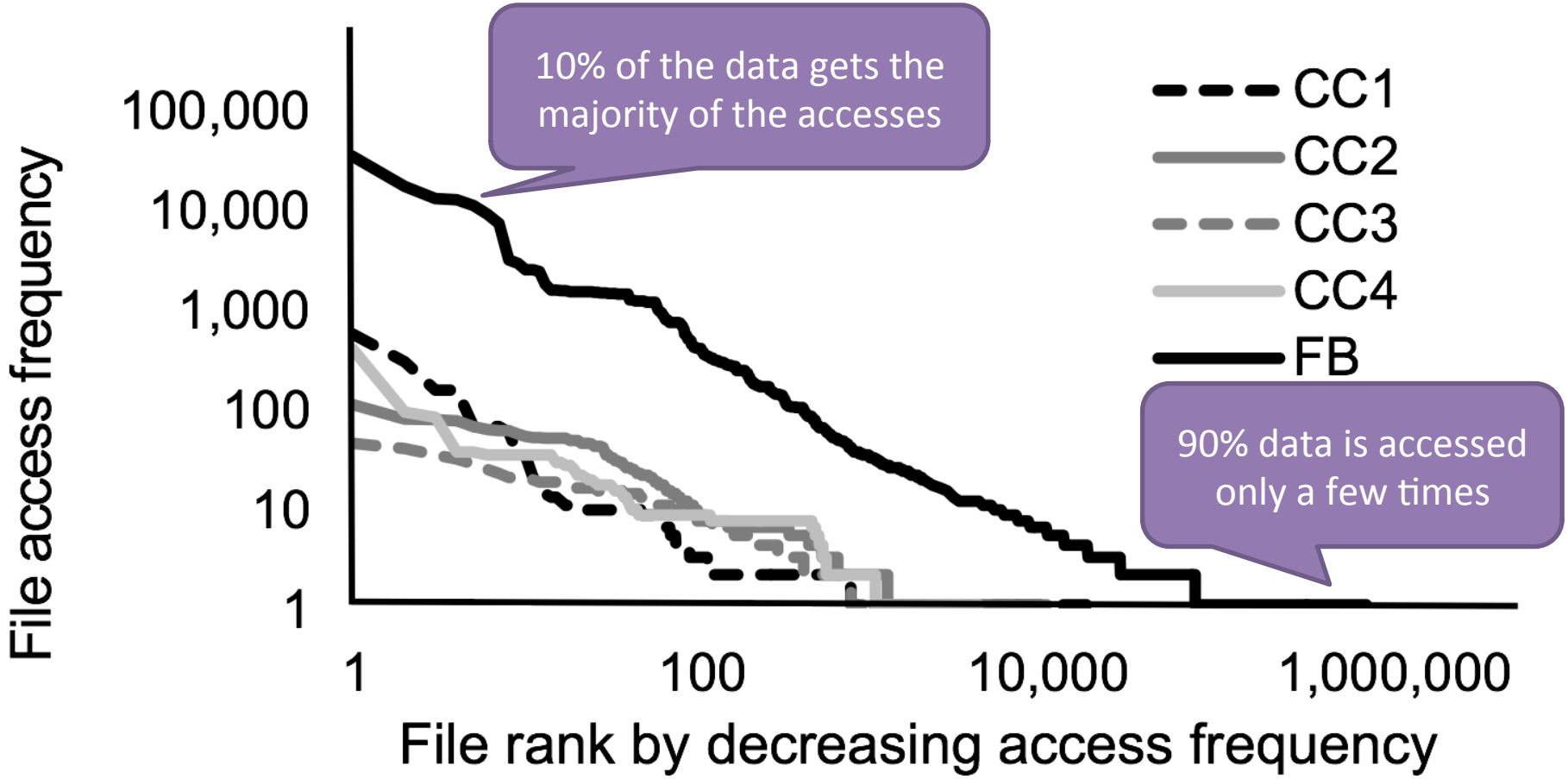
Recovery Cost vs. Storage Overhead



Recovery Cost vs. Storage Overhead



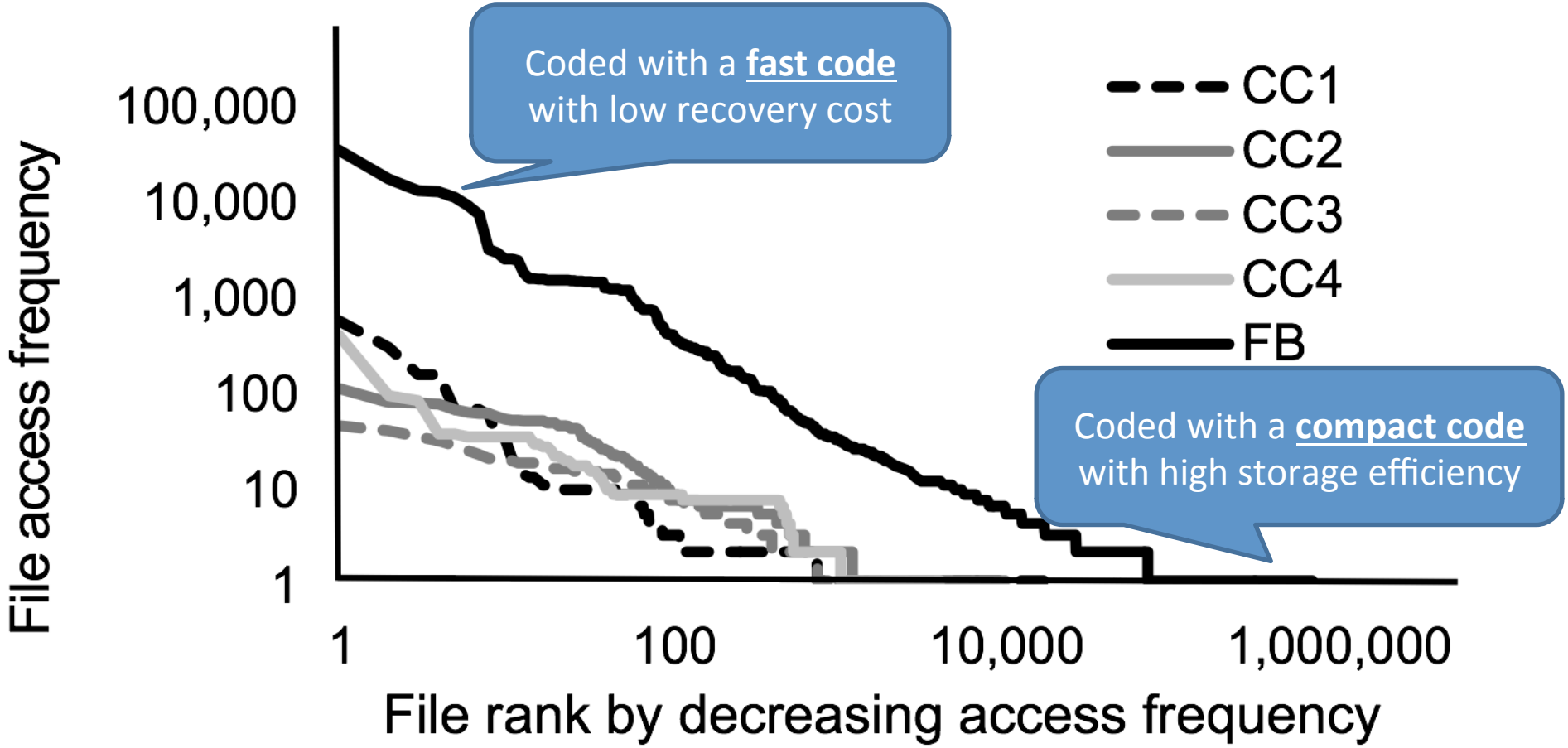
HDFS Data Access Skew



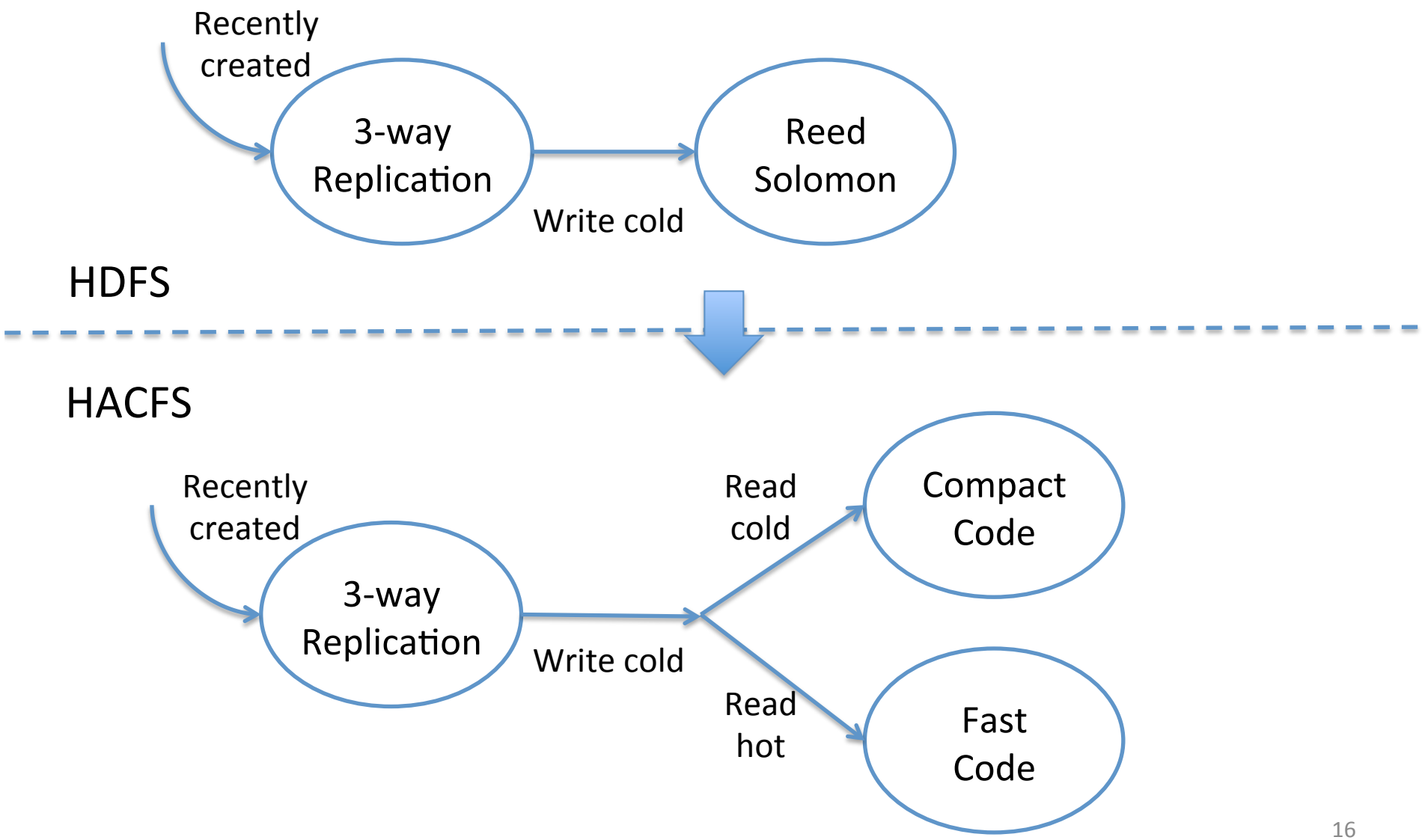
Four Cloudera customers and one Facebook workload

[VLDB'12] Chen et al., Interactive Query Processing in Big Data Systems: A Cross-Industry Study of MapReduce Workloads

Two Erasure Codes



Adaptive Coding in HDFS



Popular code families

- **Product Code**
- **Local Reconstruction Code**
- Reed-Solomon Code (MDS code)
- Partial MDS Code
- HoVer Code

Low recovery cost codes

Popular code families

- **Product Code**
- **Local Reconstruction Code**
- Reed-Solomon Code (MDS code)
- Partial MDS Code
- HoVer Code

Low recovery cost codes

Fast and Compact Product Codes

D1	D2	D3	D4	D5	P1
D6	D7	D8	D9	D10	P2
P3	P4	P5	P6	P7	P8

<u>Fast code</u>
<u>(Product Code 2x5)</u>

Storage overhead: 1.8x (18/10)

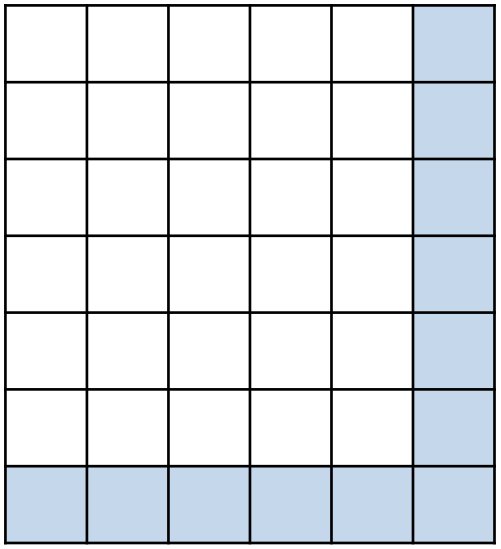
Fast and Compact Product Codes

D1	D2	D3	D4	D5	P1
D6	D7	D8	D9	D10	P2
P3	P4	P5	P6	P7	P8

<u>Fast code</u>
<u>(Product Code 2x5)</u>
Recovery cost: 2
Storage overhead: 1.8x (18/10)

Fast and Compact Product Codes

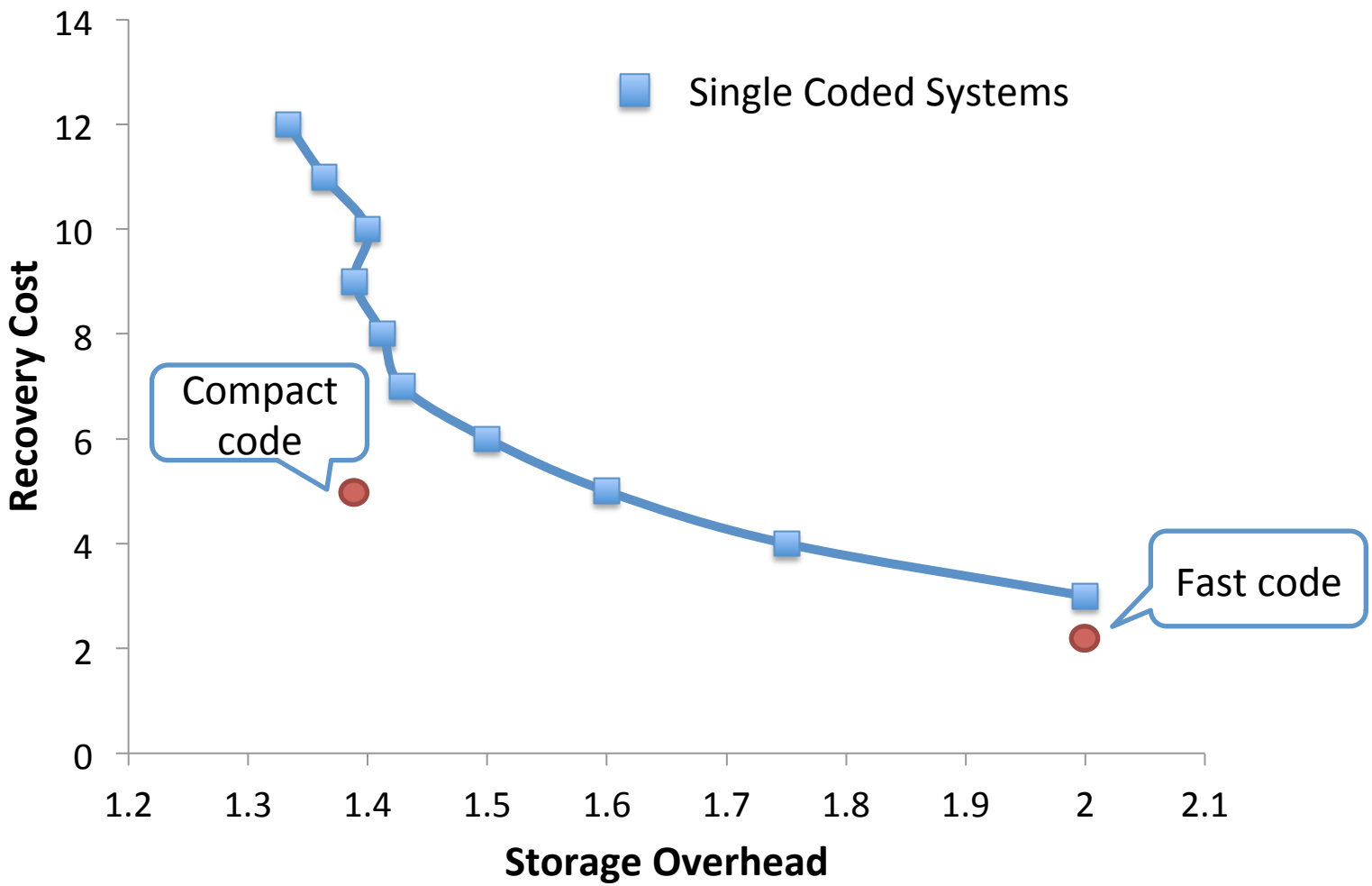
D1	D2	D3	D4	D5	P1
D6	D7	D8	D9	D10	P2
P3	P4	P5	P6	P7	P8



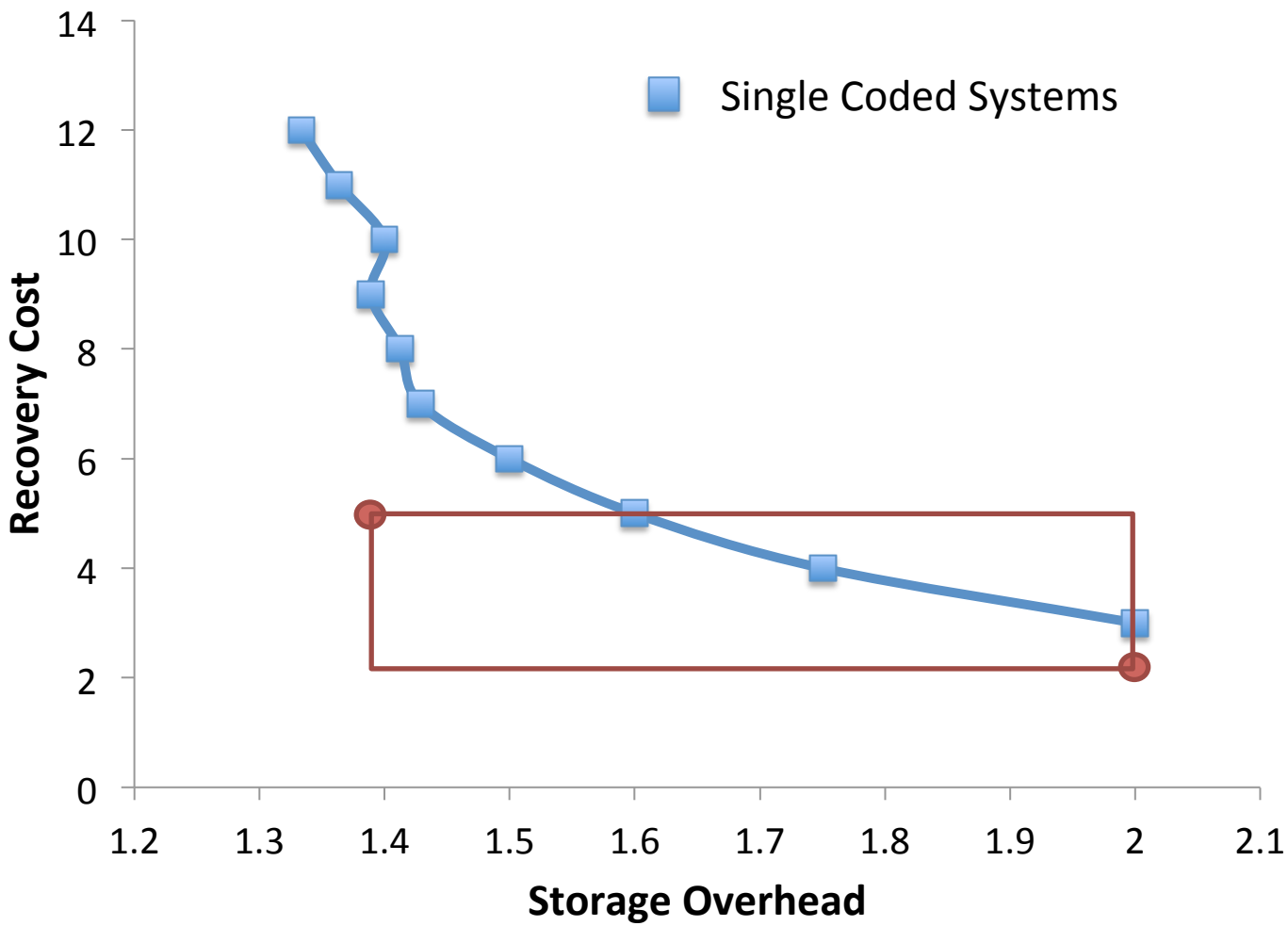
<u>Fast code</u>
<u>(Product Code 2x5)</u>
Recovery cost: 2
Storage overhead: 1.8x

<u>Compact code</u>
<u>(Product Code 6x5)</u>
Recovery cost: 5
Storage overhead: 1.4x

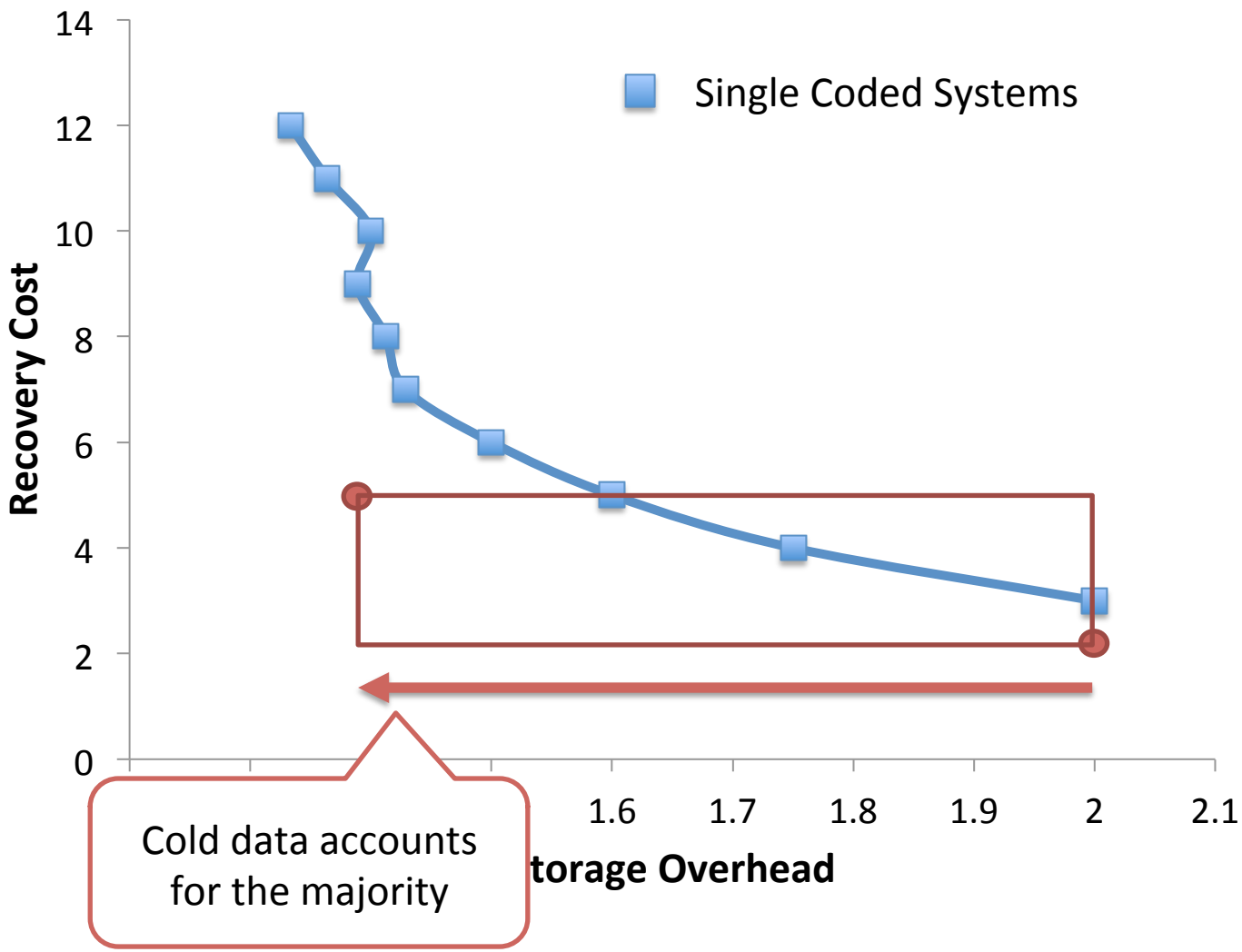
HACFS with Product Codes



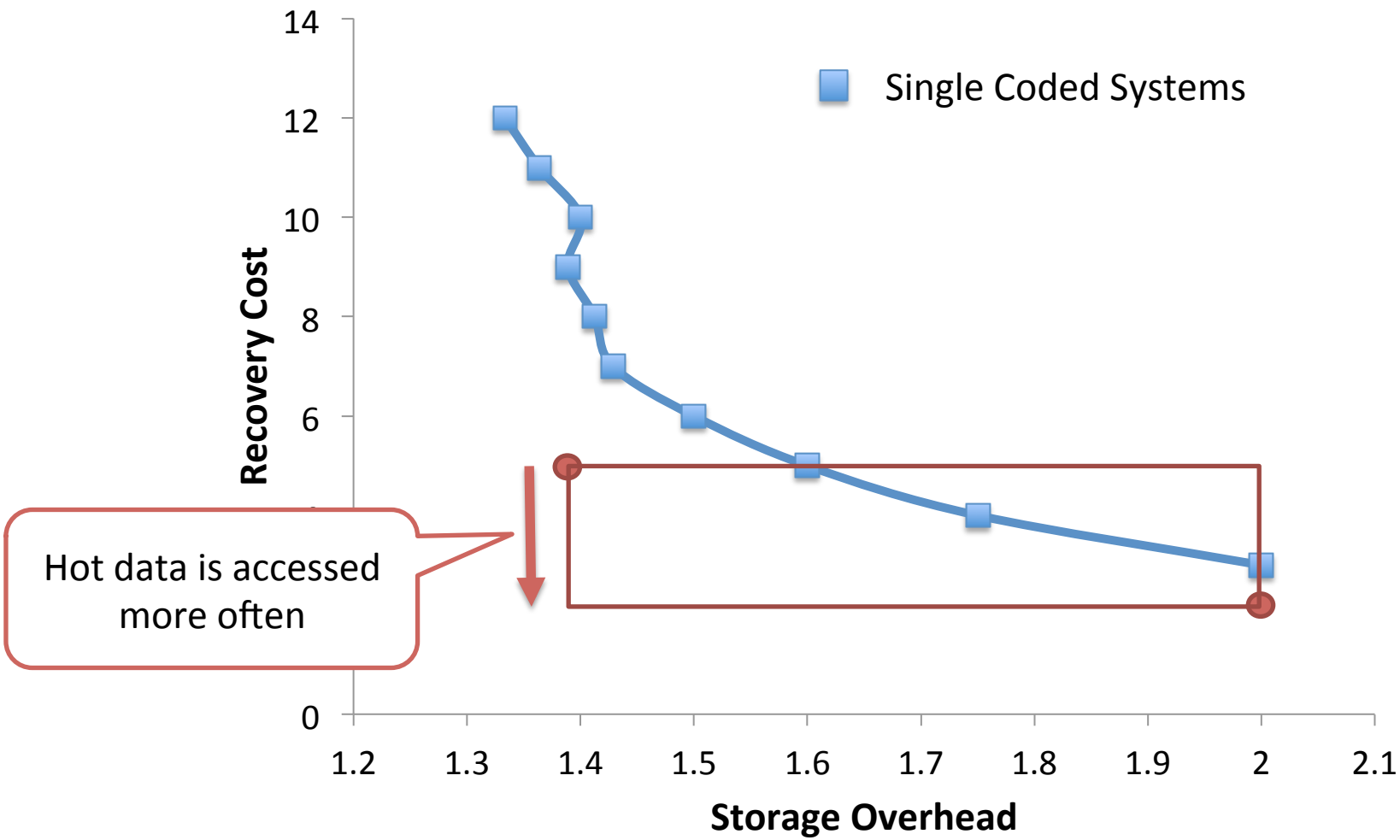
HACFS with Product Codes



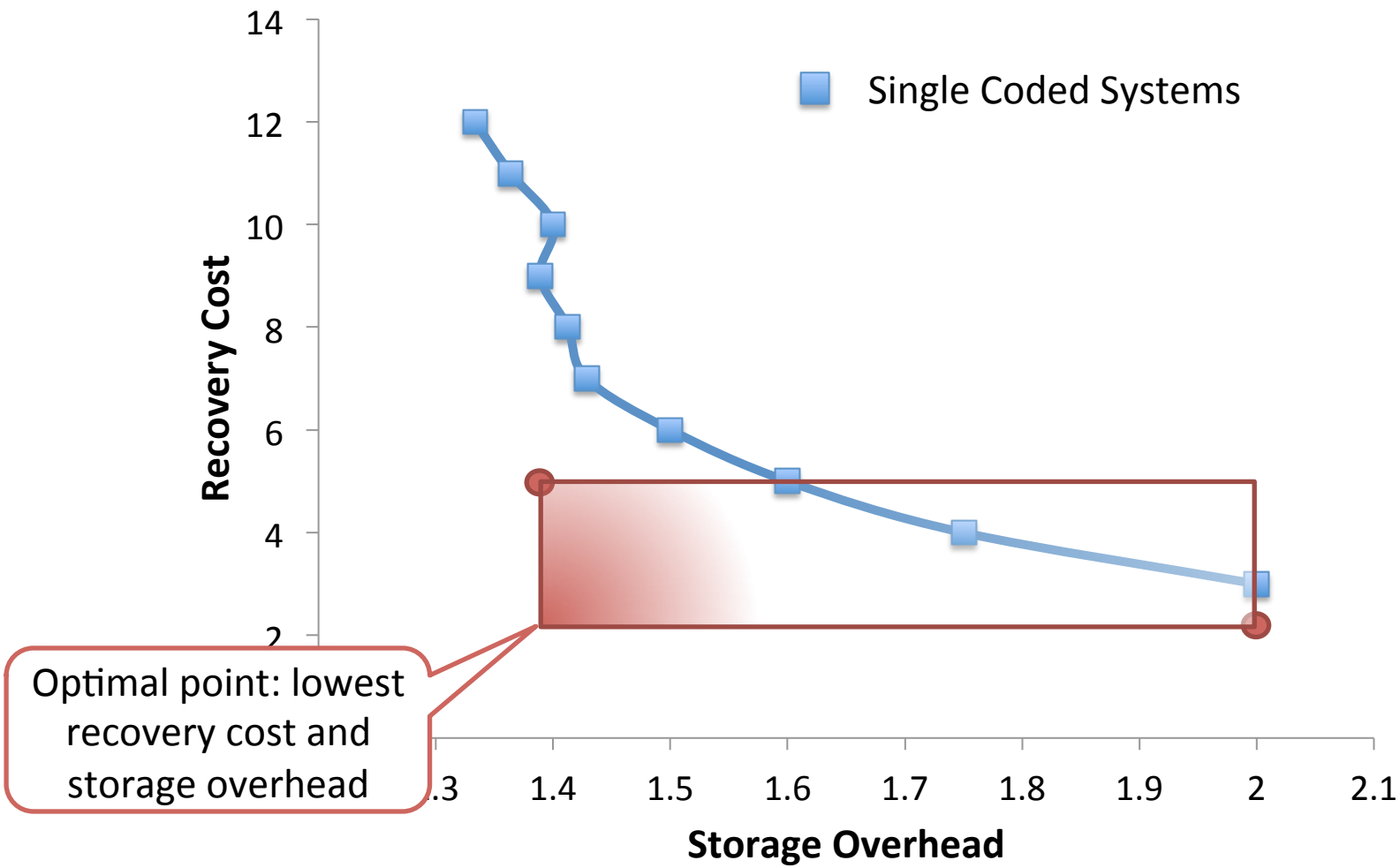
HACFS with Product Codes



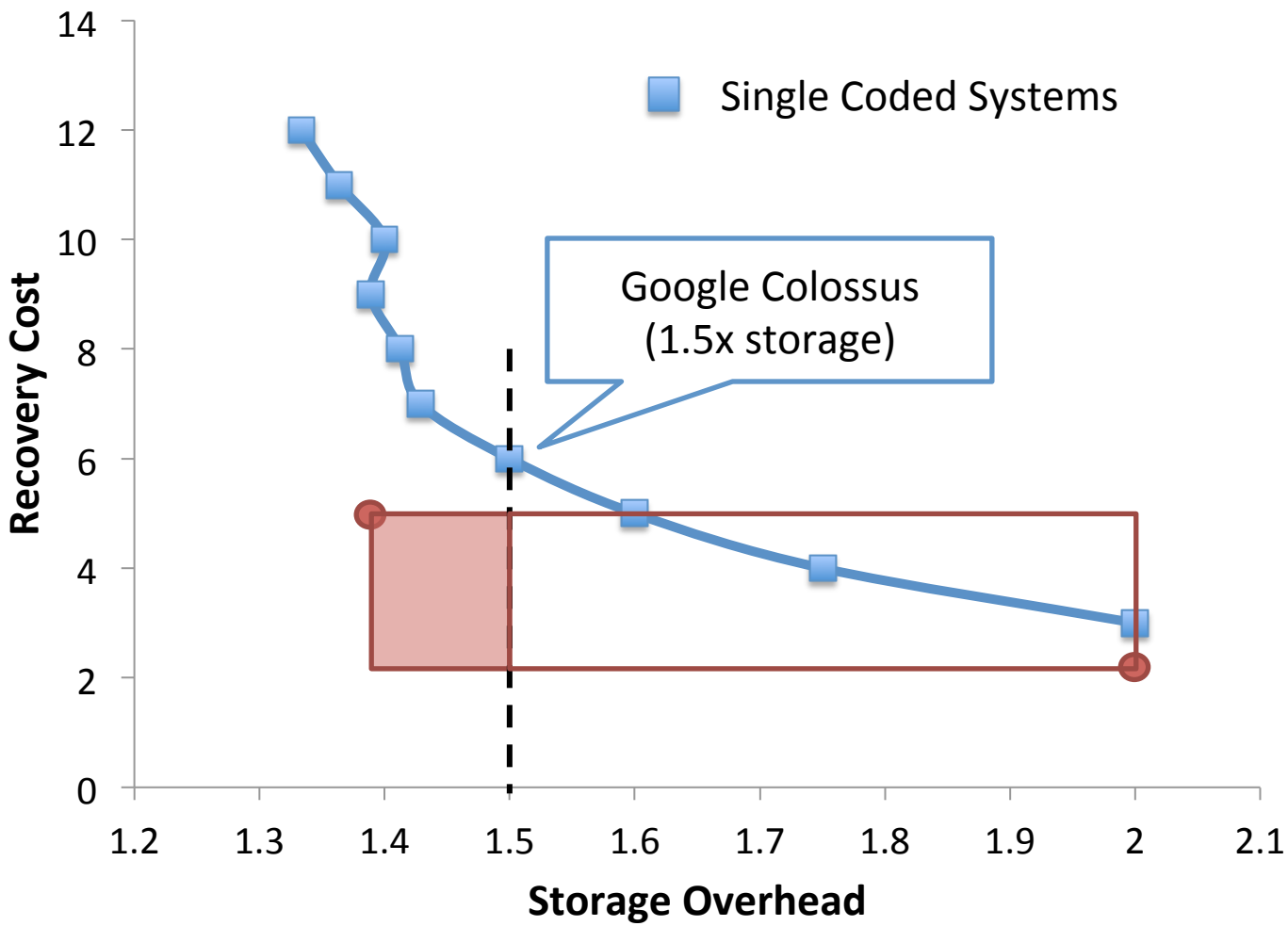
HACFS with Product Codes



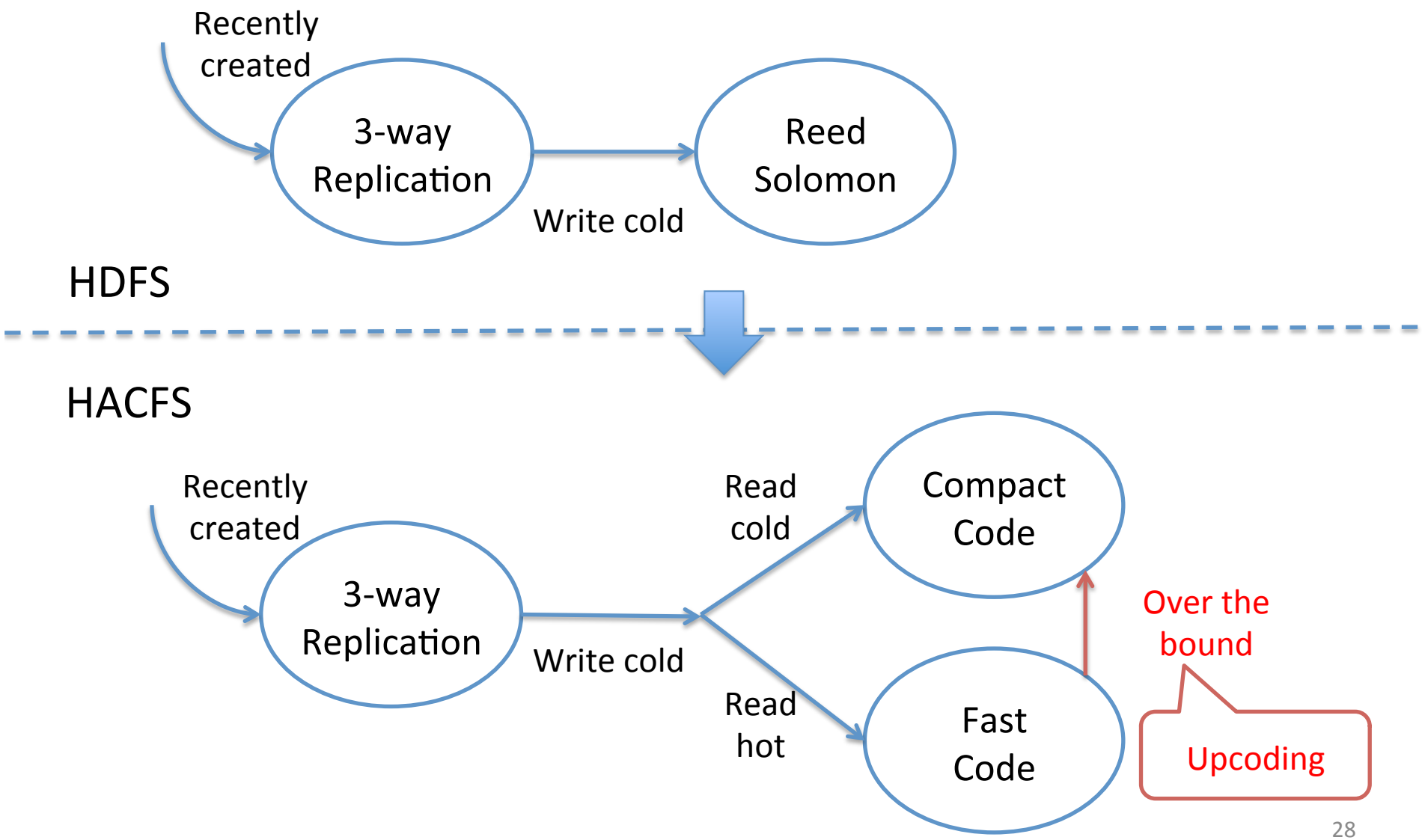
HACFS with Product Codes



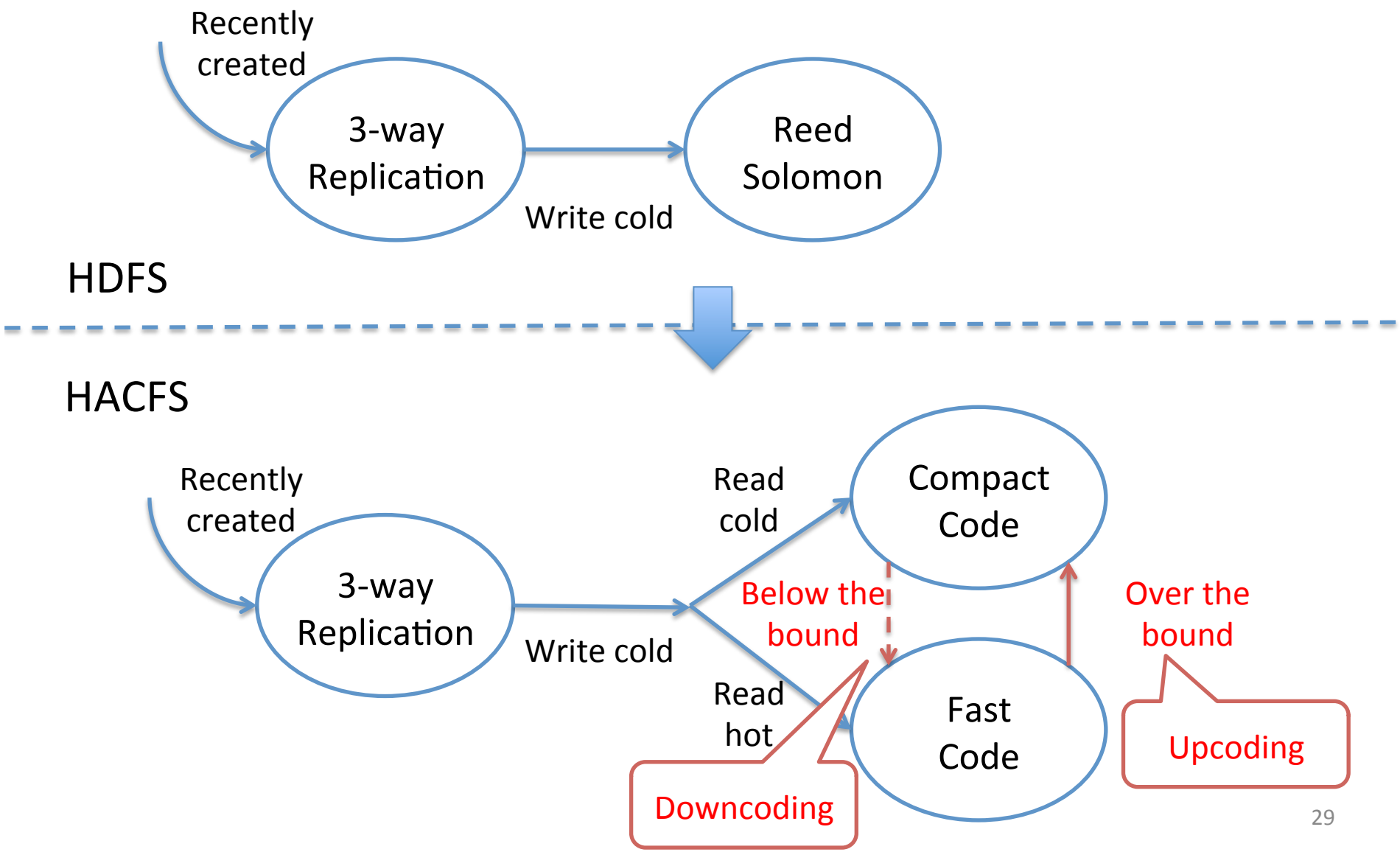
Storage Bound



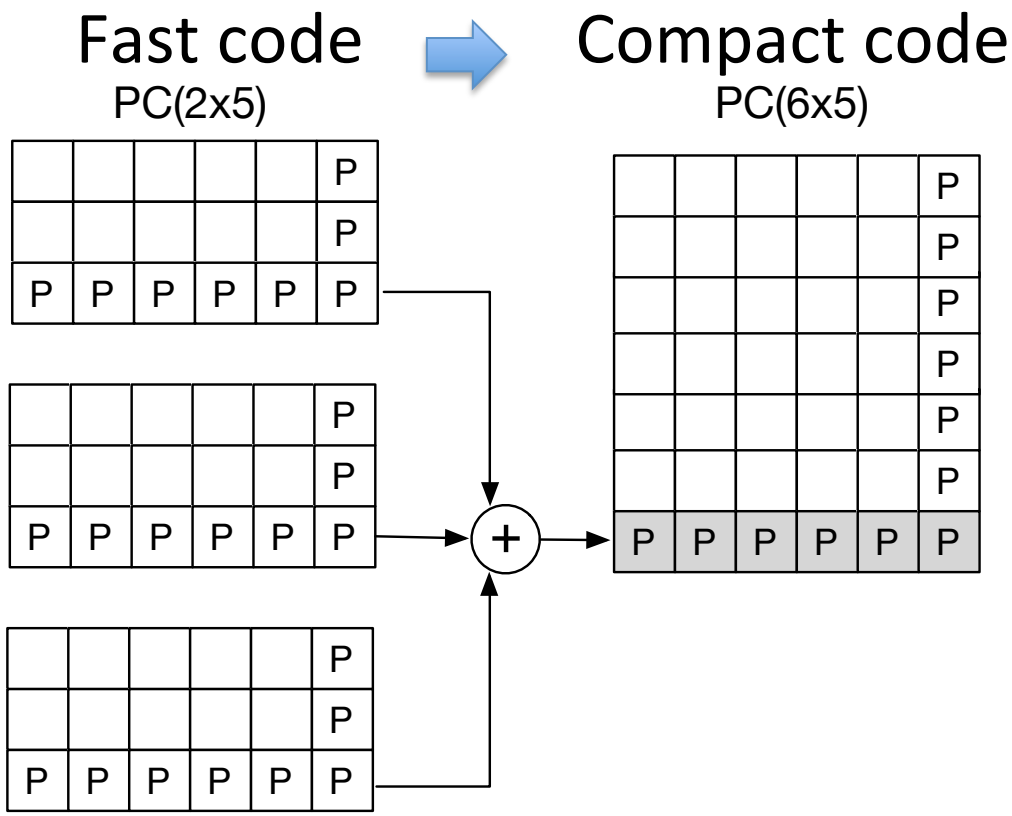
Upcoding/Downcoding



Upcoding/Downcoding



Upcoding for Product Codes



Parity-only Conversion

- Horizontal parties require no re-computation
- Vertical parities require no data block transfer
- All parity updates can be done in parallel and in a distributed manner

Efficient Up/Down-coding

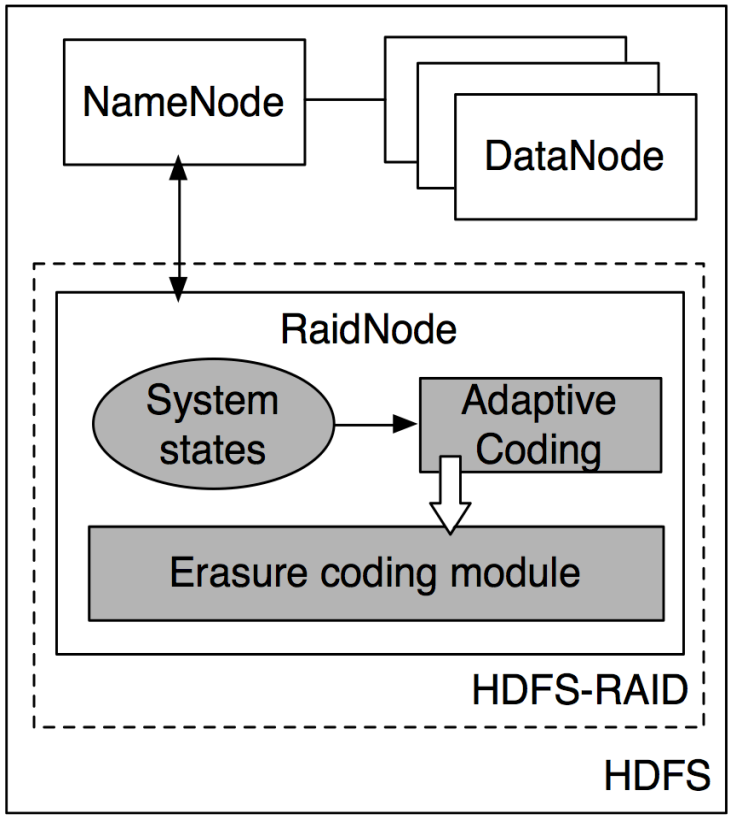
- Popular code families with efficient up/down-coding
  - ✓ **Product Code**
  - ✓ **Local Reconstruction Code**
  - ✓ Reed-Solomon Code (MDS code)
  - ✓ Partial MDS Code
  - ✓ HoVer Code

HACFS implementation

Applicable to other codes as well

Evaluation

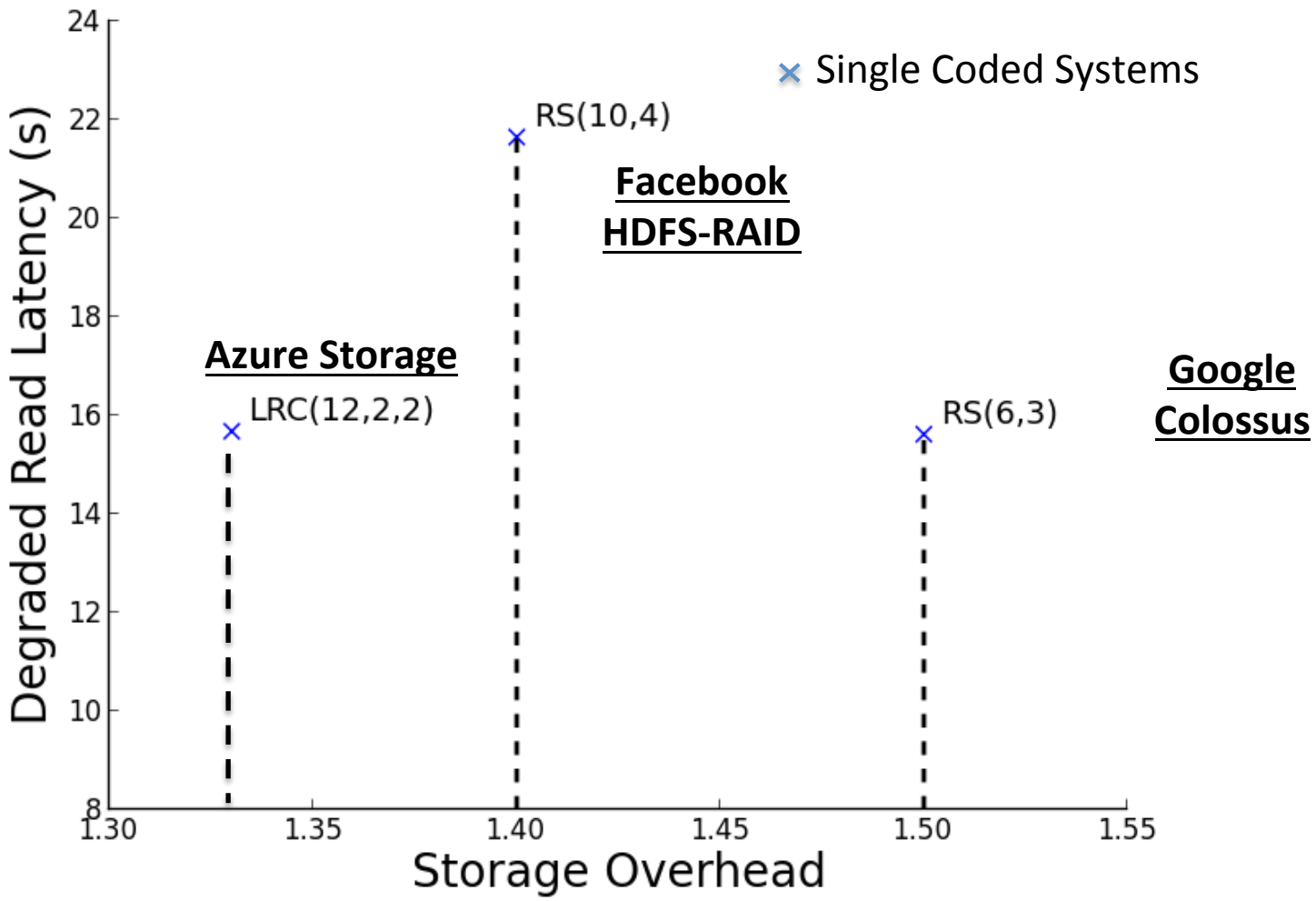
- HACFS Implementation
  - Extension to Facebook's HDFS
  - 3k LOC: three new modules
- Methodology
  - Five workloads: four Cloudera customers, one Facebook [VLDB'2012]
  - HDFS cluster: 11 nodes
  - Each node: 24 cores, 6 disks, 1 Gbps network



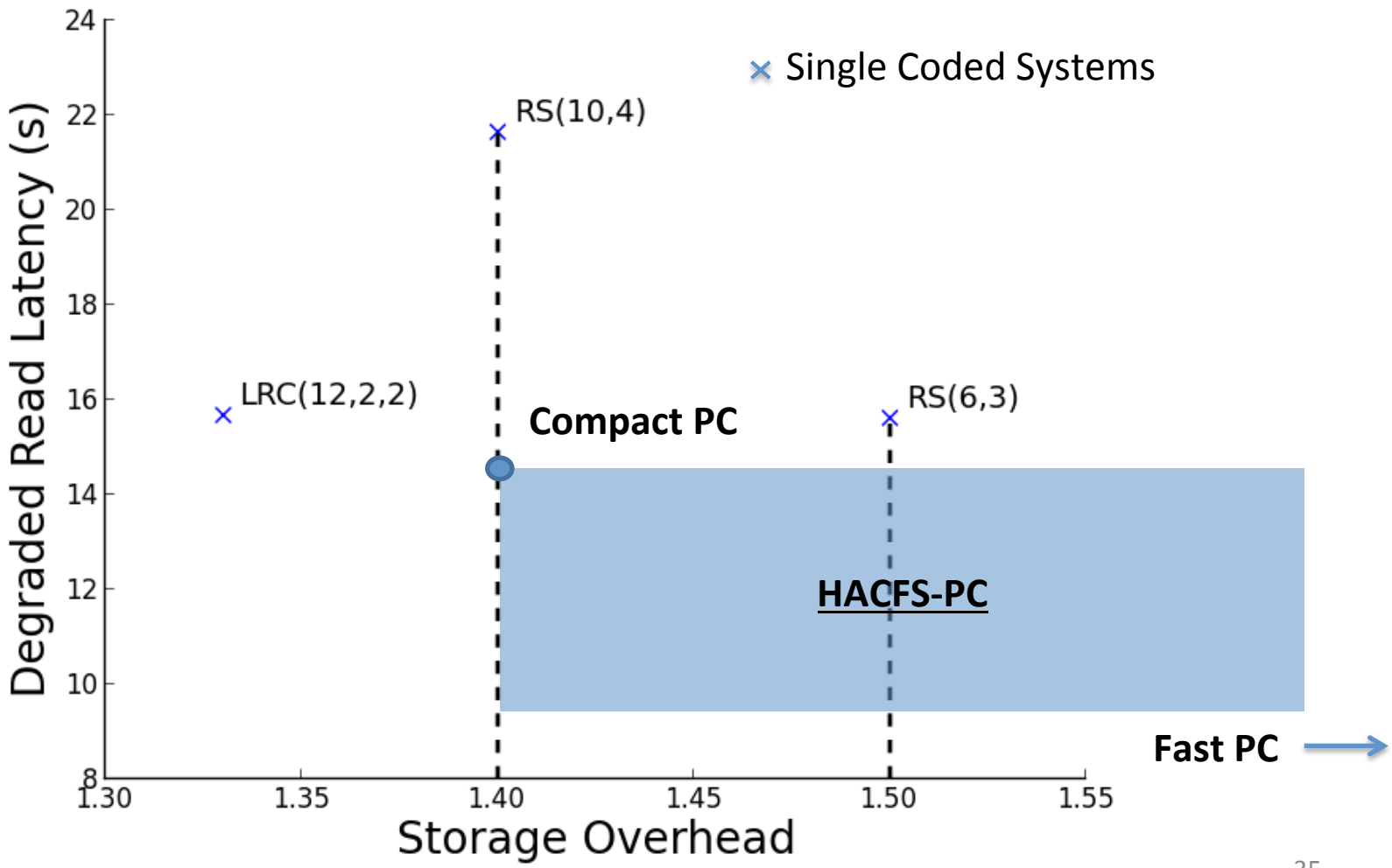
Experiment metrics

- Degraded read latency
  - Foreground read request delay
  - Caused mostly by software issues
- Reconstruction time
  - Background recovery for failures
  - Caused mostly by hardware failures
- Storage overhead (bounded)

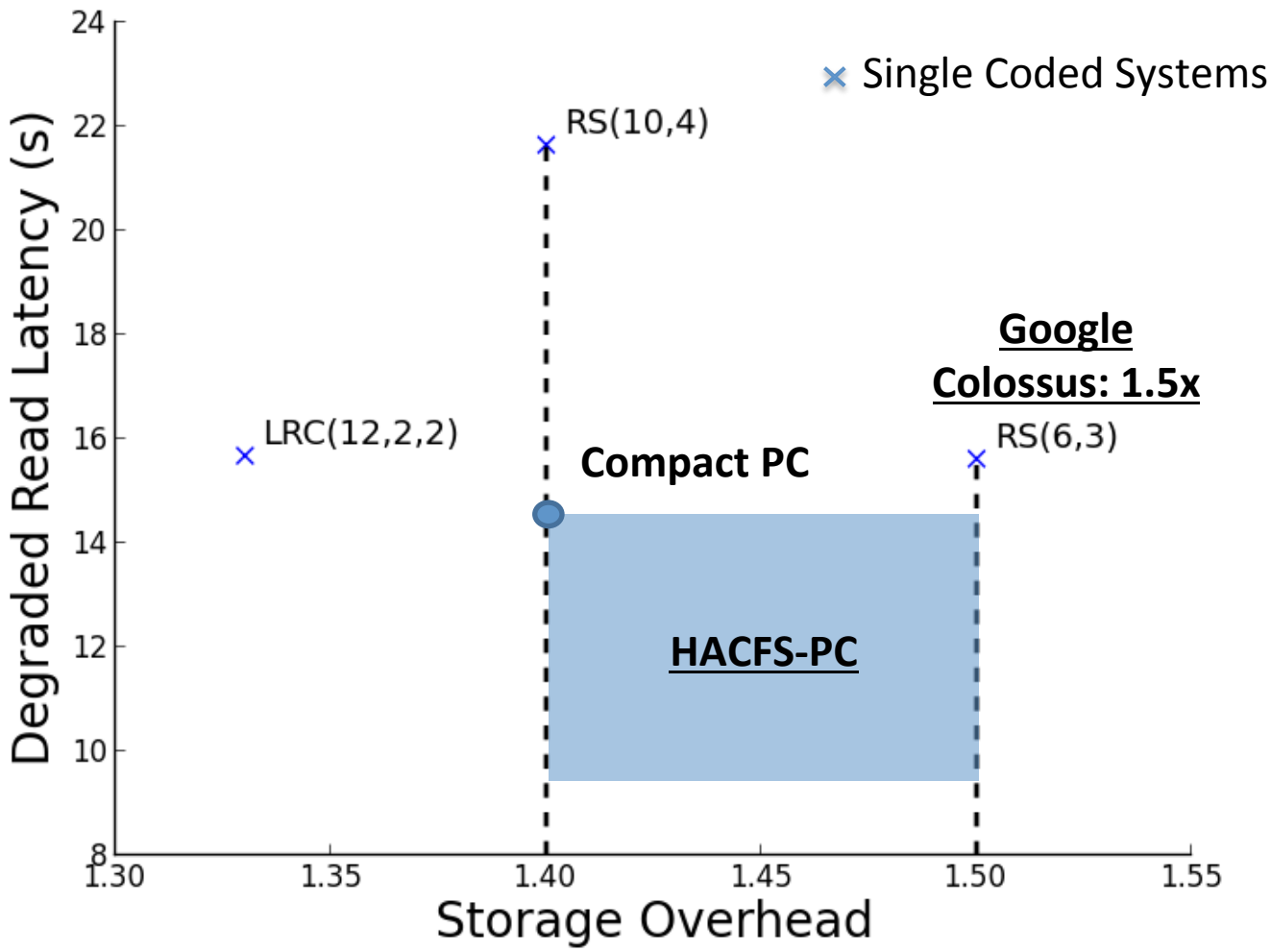
Degraded Read Latency



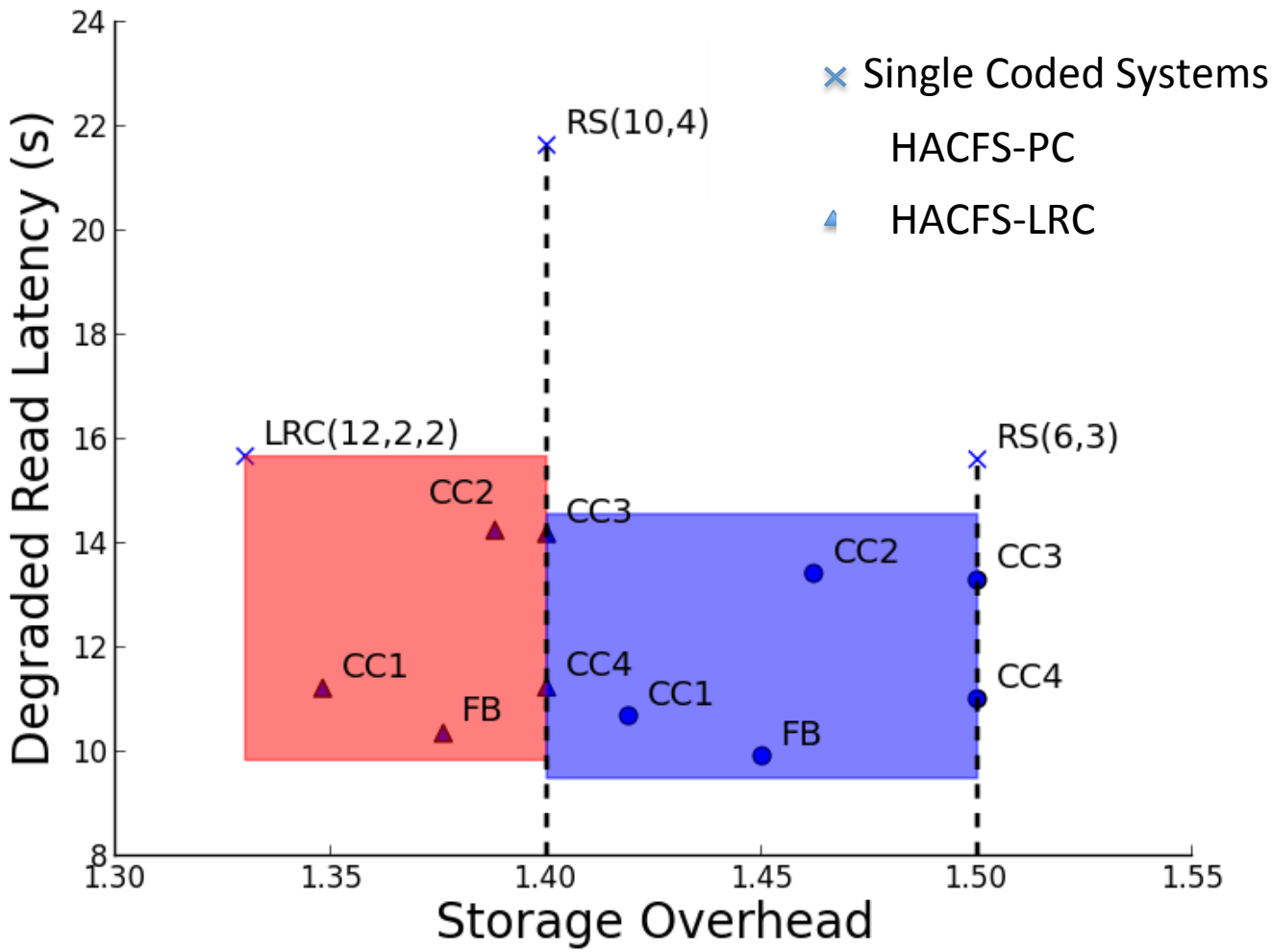
Degraded Read Latency



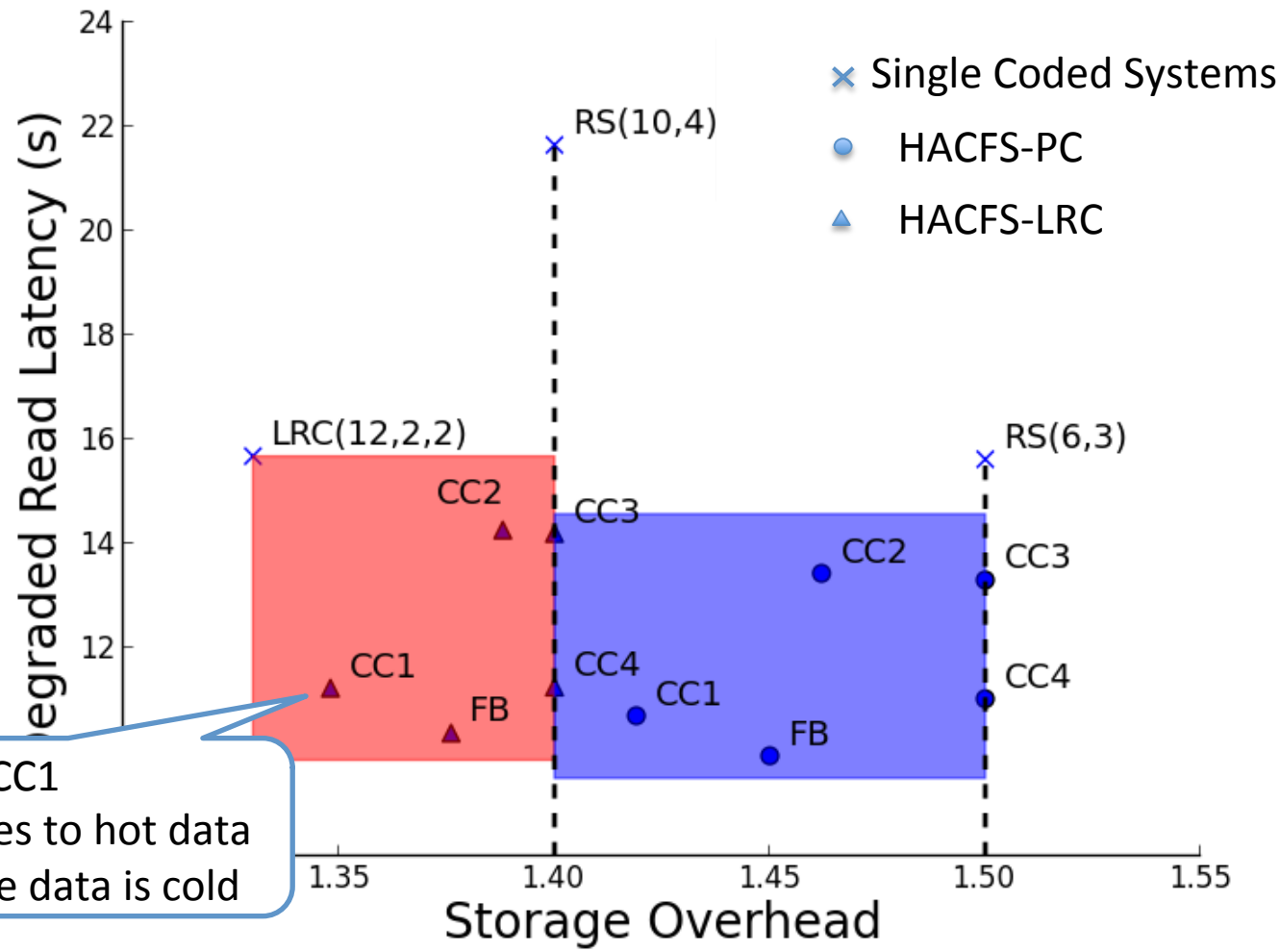
Degraded Read Latency



Degraded Read Latency

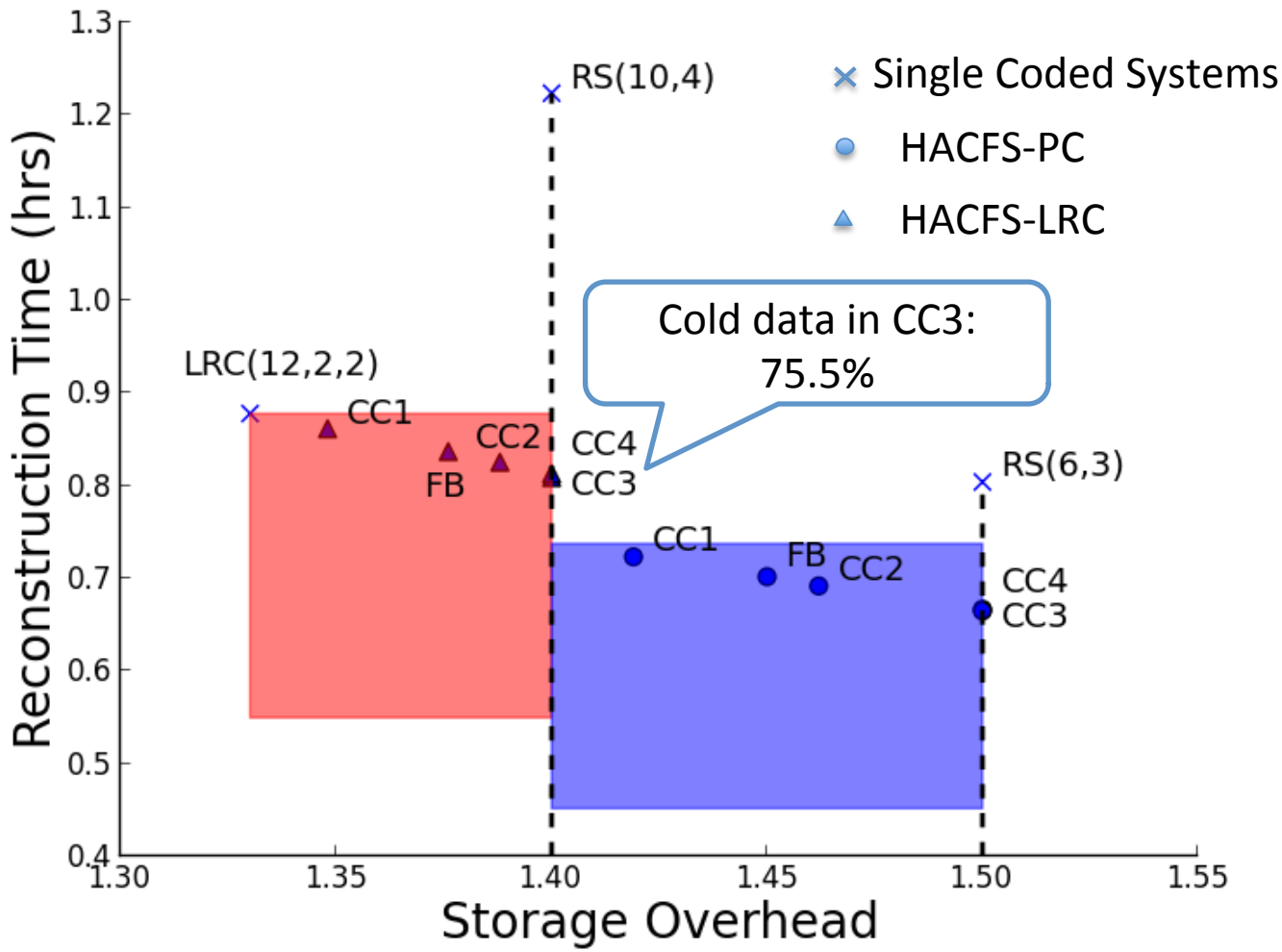


Degraded Read Latency

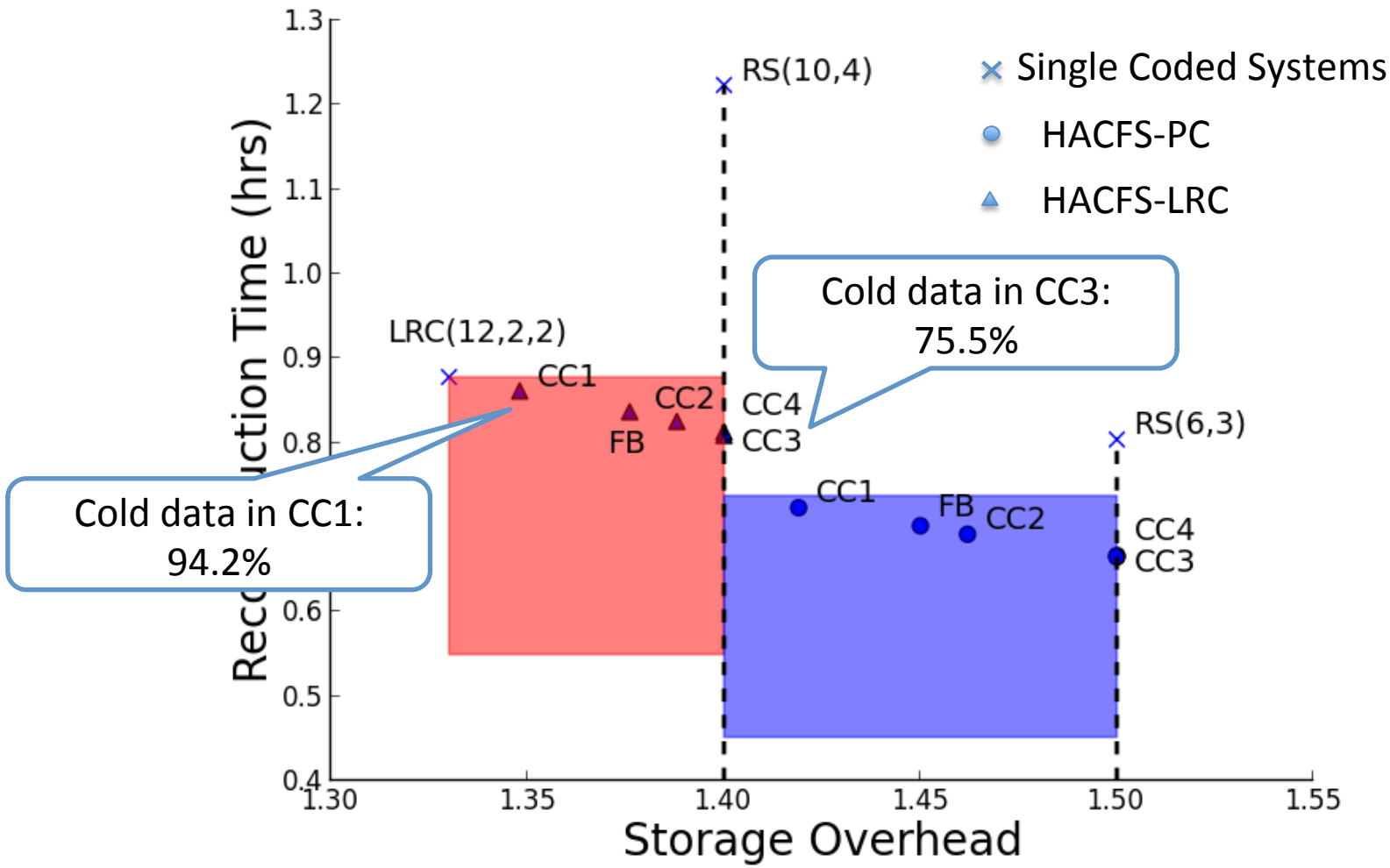


CC1
86% accesses to hot data
94.2% of the data is cold

Reconstruction Time



Reconstruction Time



System Comparisons

	HACFS using Product Codes		
	Colossus FS	FB HDFS	Azure
Degraded Read Latency	25.2%	46.1%	25.4%
Reconstruction Time	14.3%	43.7%	21.4%
Storage Overhead	2.3%	-4.7%	-10.2%

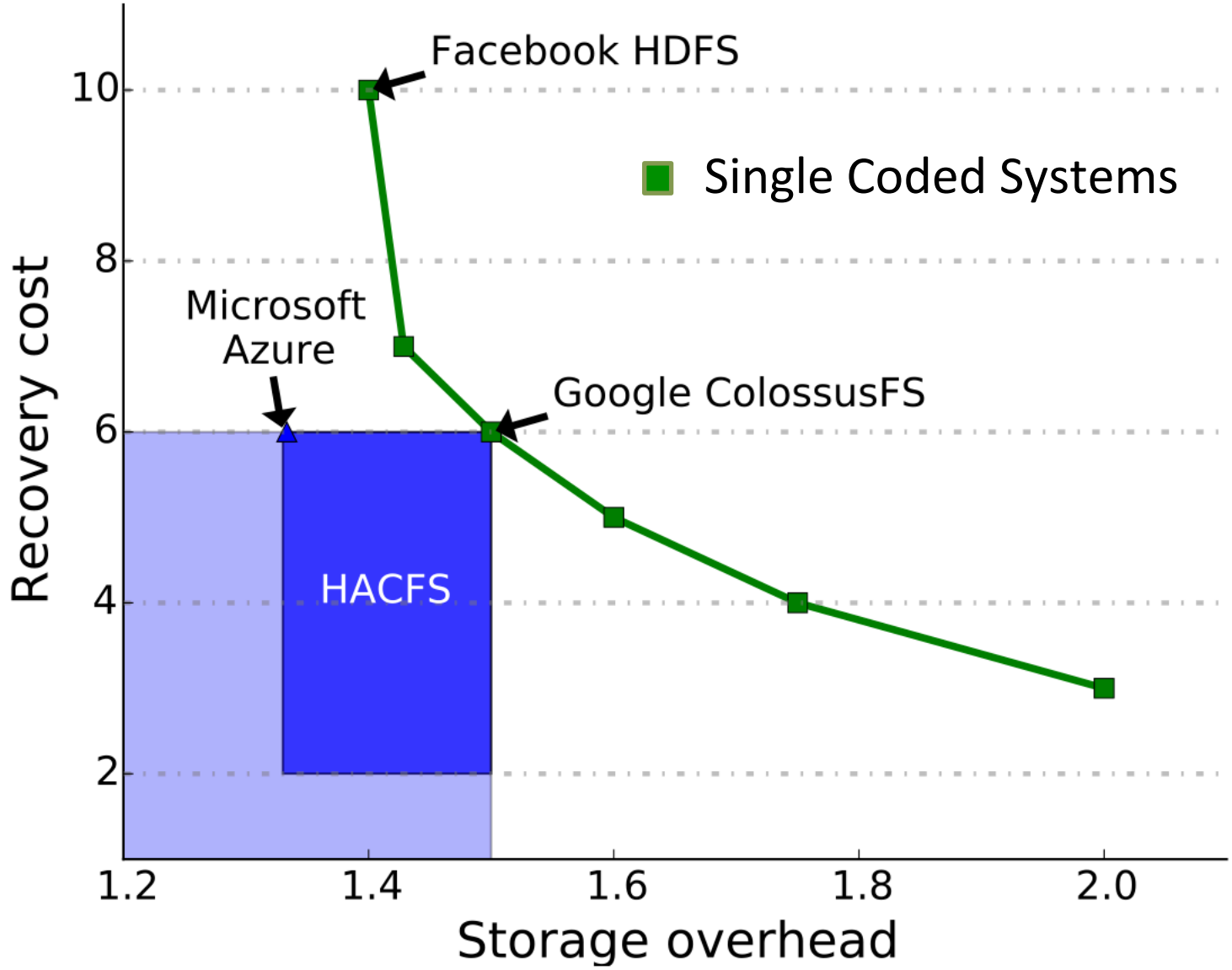
	HACFS using LRCs		
	Colossus FS	FB HDFS	Azure
Degraded Read Latency	21.5%	43.3%	21.2%
Reconstruction Time	-3.1%	32.2%	5.6%
Storage Overhead	7.7%	1.1%	-4.2%

System Comparisons

	HACFS using Product Codes		
	Colossus FS	FB HDFS	Azure
Degraded Read Latency	25.2%	46.1%	25.4%
Reconstruction Time	14.3%	43.7%	21.4%
Storage Overhead	2.3%	-4.7%	-10.2%

	HACFS using LRCs		
	Colossus FS	FB HDFS	Azure
Degraded Read Latency	21.5%	43.3%	21.2%
Reconstruction Time	-3.1%	32.2%	5.6%
Storage Overhead	7.7%	1.1%	-4.2%

Conclusions



Thanks Q/A

FAST'15

A Tale of Two Erasure Codes in HDFS

Mingyuan Xia, Mohit Saxena
Mario Blaum, David Pease

IBM Research Almaden & McGill University