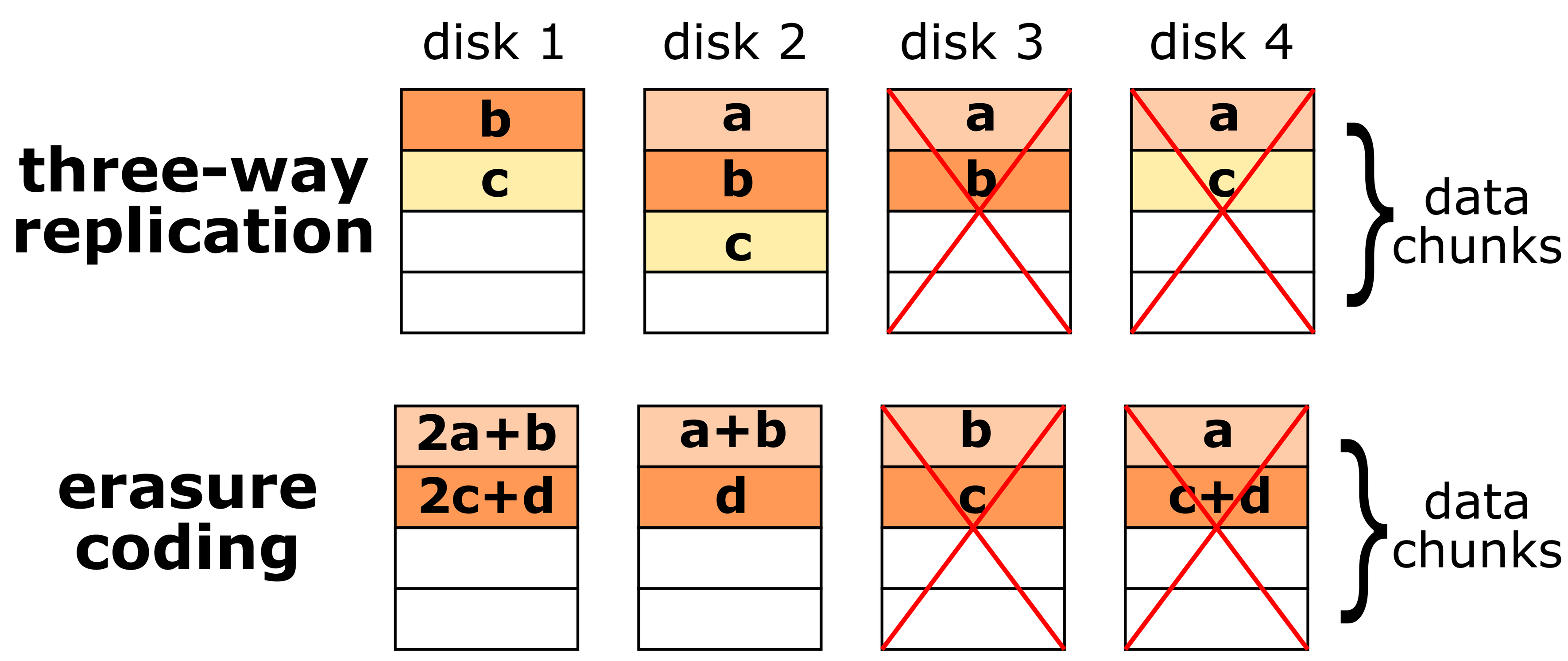
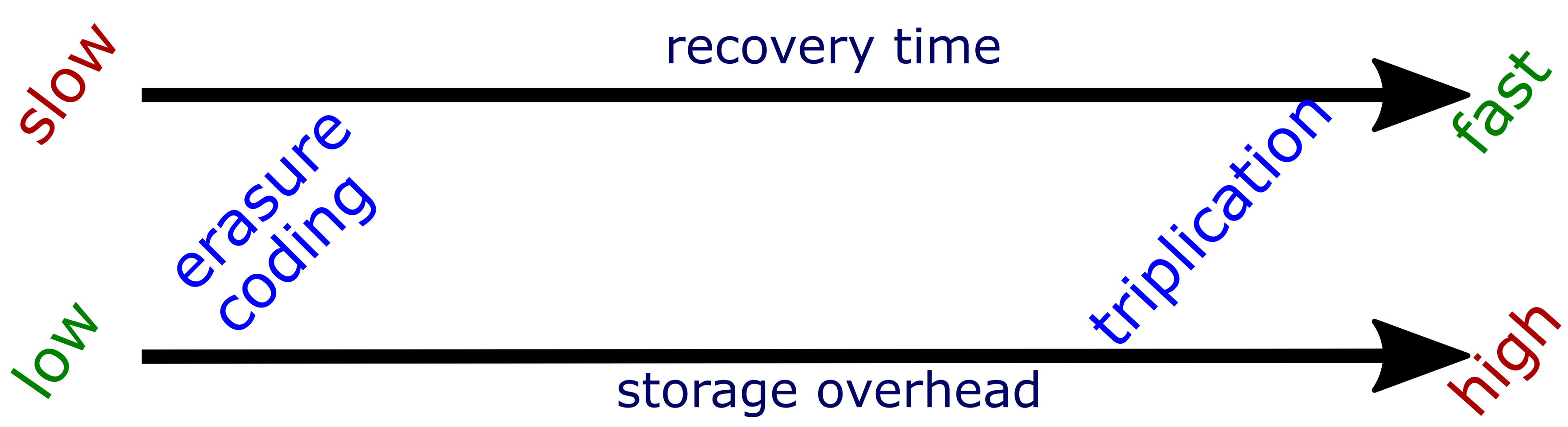
RAIDP: ReplicAtion with Intra-Disk Parity

Eitan Rosenfeld, Aviad Zuck, Nadav Amit,
Michael Factor, Dan Tsafrir

triplication vs. erasure coding

background: modern cloud storage systems use redundancy to withstand simulataneous disk failures





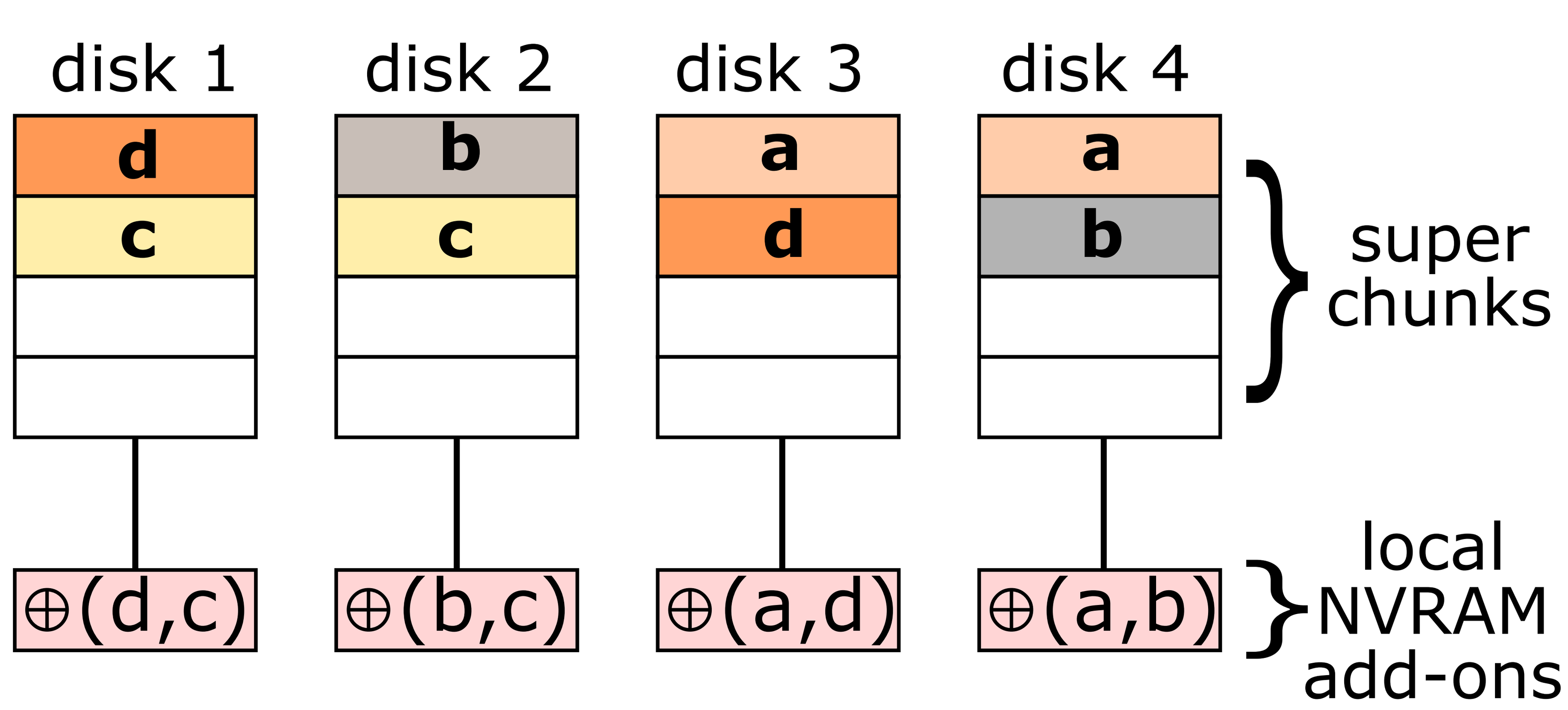
RAIDP hybrid layout

problem: triplication is expensive and used typically for warm data only

idea: two independent NVRAM disk add-ons are cheaper than a third replica

proposed solution:

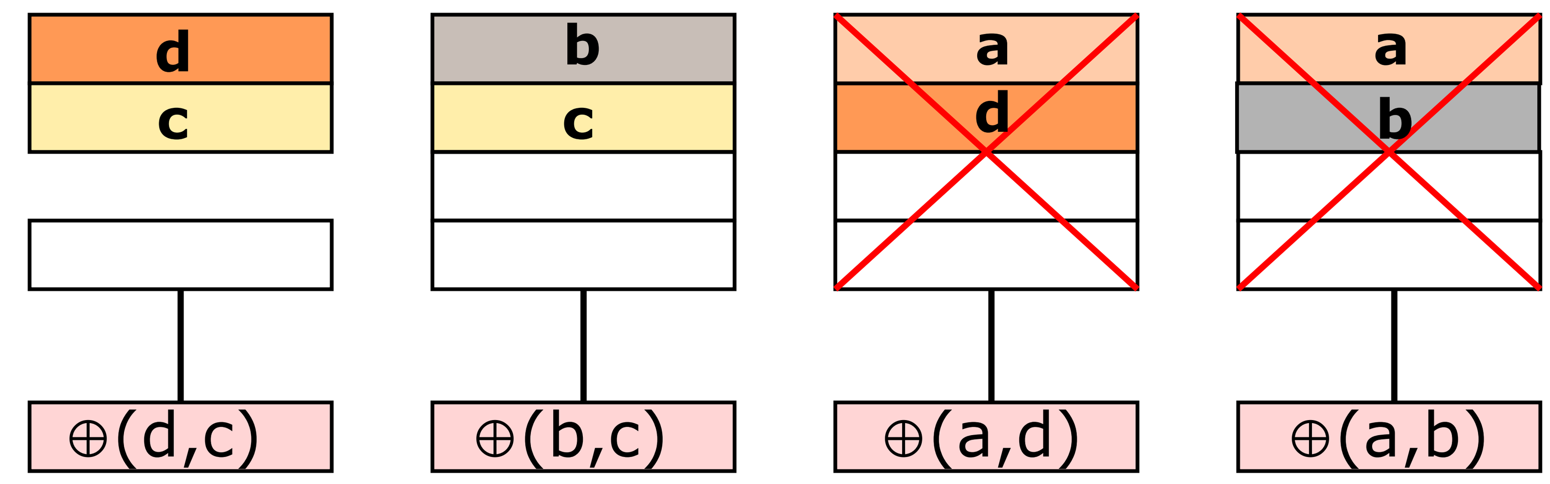
- disks partitioned to superchunks (e.g., 4GB) with *two* replicas only
- bitwise mirror superchunk writes
- disks share ⩽ 1 superchunk
- disk add-ons use NVRAM to store *local chunk parity*
- add-ons fail separately from disk



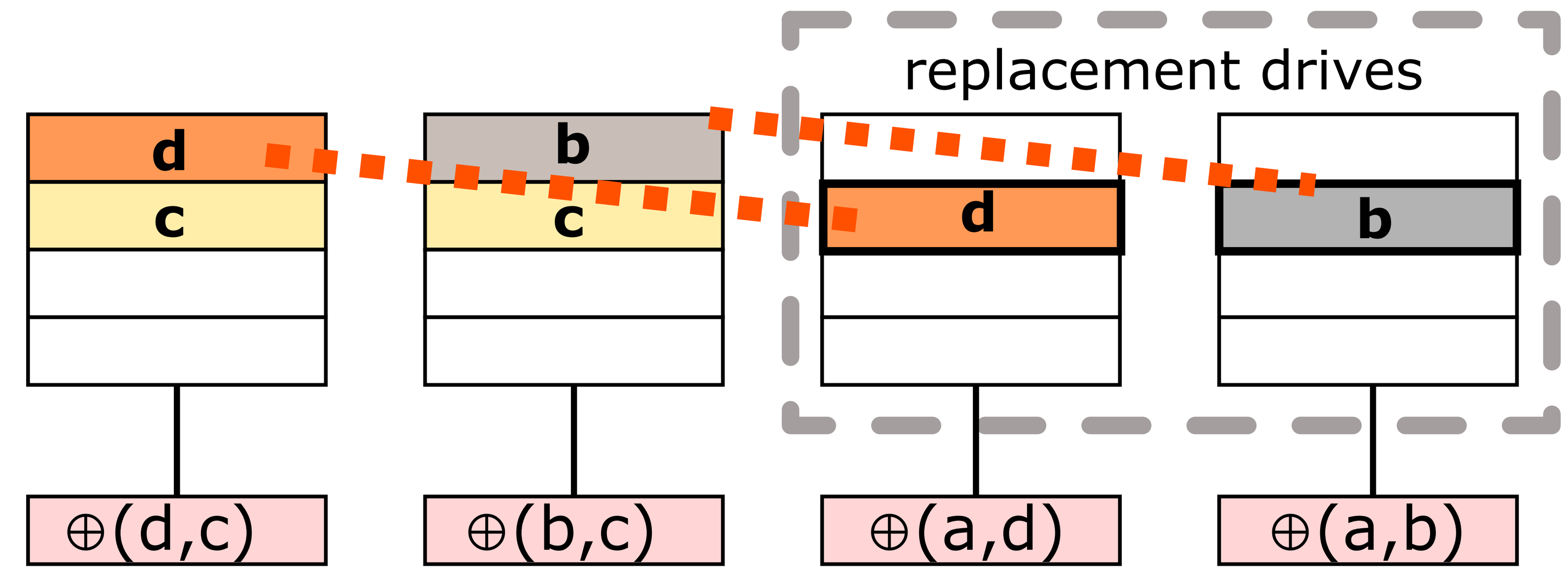
vs. erasure coding	vs. triplication
+ faster recovery	+ lower storage overhead
+ better performance	+ faster appends
- more storage overhead	- slower updates
	- slower recovery

recovery in RAIDP

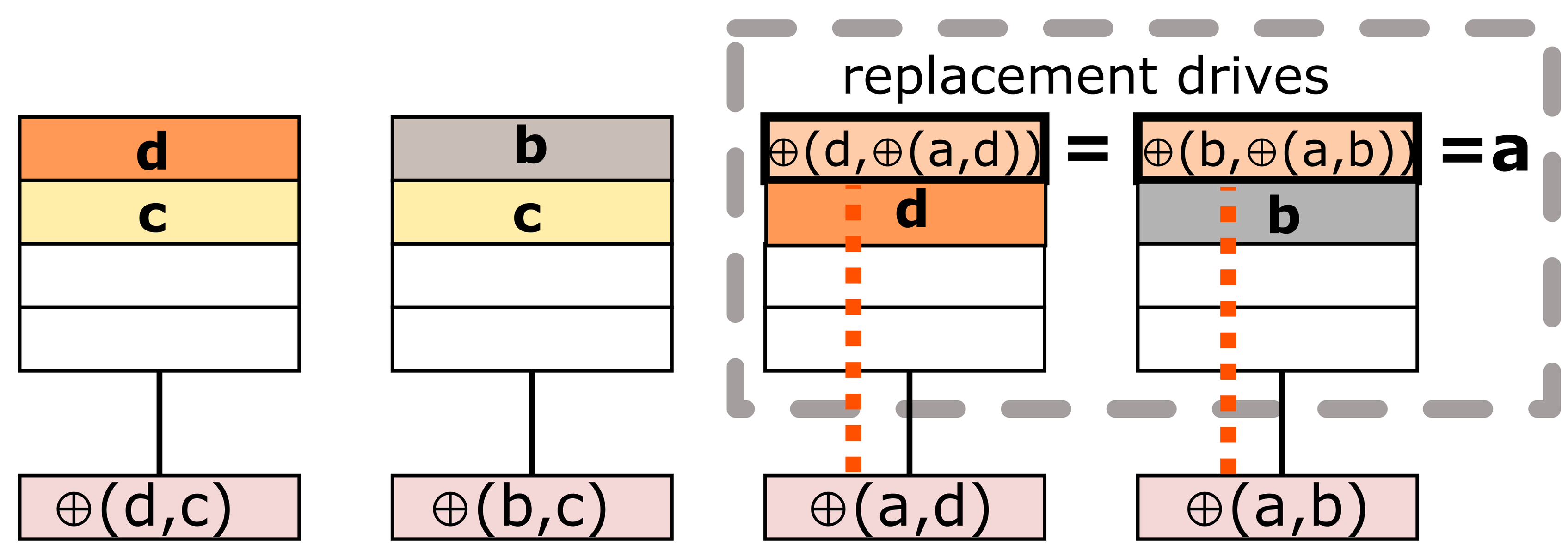
0. two disks fail simultaneously



1. copy superchunk replicas



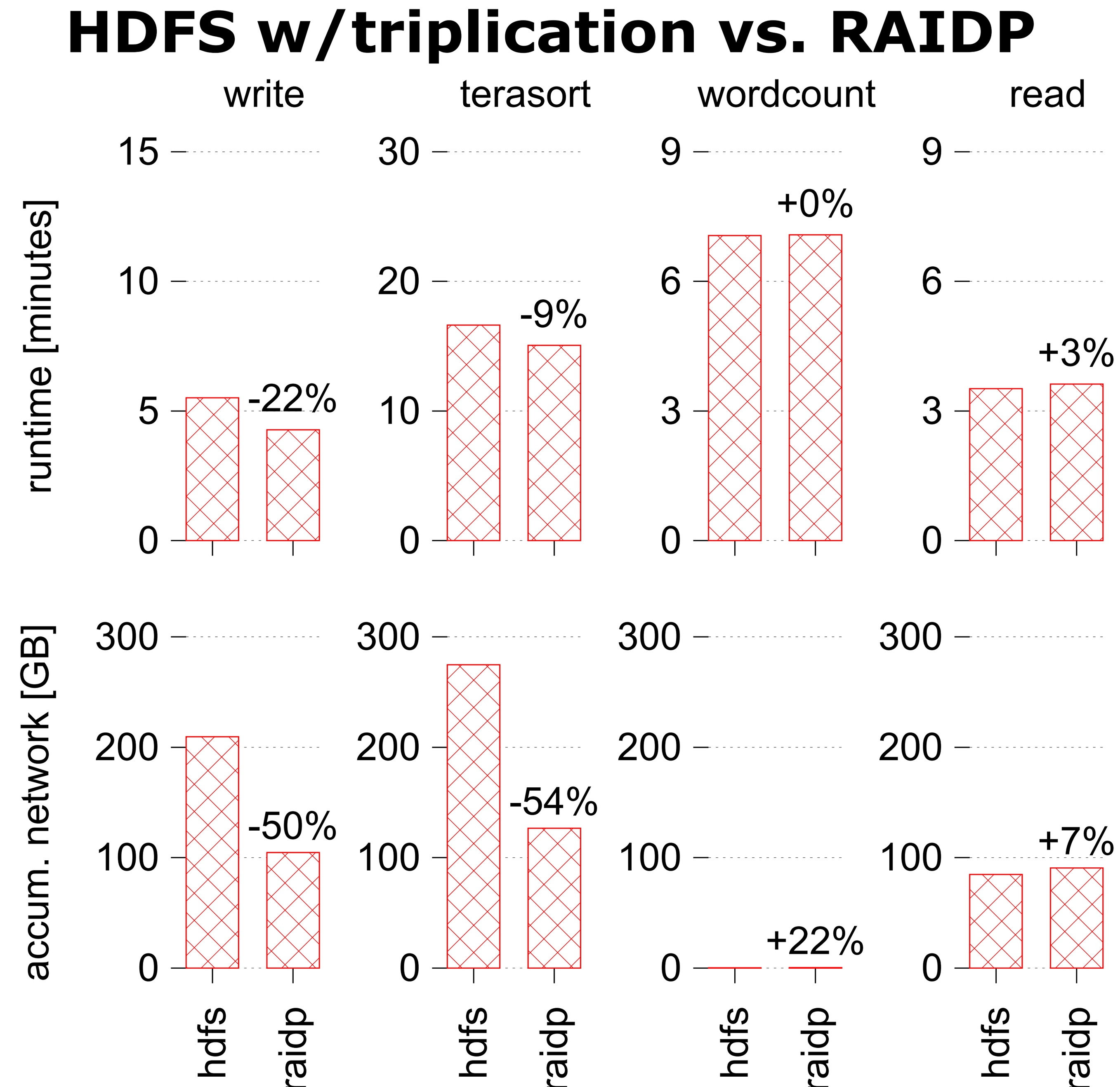
2. reconstruct single missing superchunk



evaluation

- PoC implementation in Hadoop 1.0.4
- cluster of 16 Intel Xeon E3-1220 machines (3.10GHz) Ubuntu 14.04 (kernel 3.13)
- 6GB superchunks, add-ons simulated in memory.
- append-only baseline

HDFS w/triplication vs. RAIDP



Recovery time: 14-15x faster than RAID-6 following double disk failure

TECHNION Israel Institute of Technology
הטכניון מכון טכנולוגי לישראל
vmware®
IBM