Erasure Code in Ceph

Loic Dachary @ Red Hat

ceph

Save Space

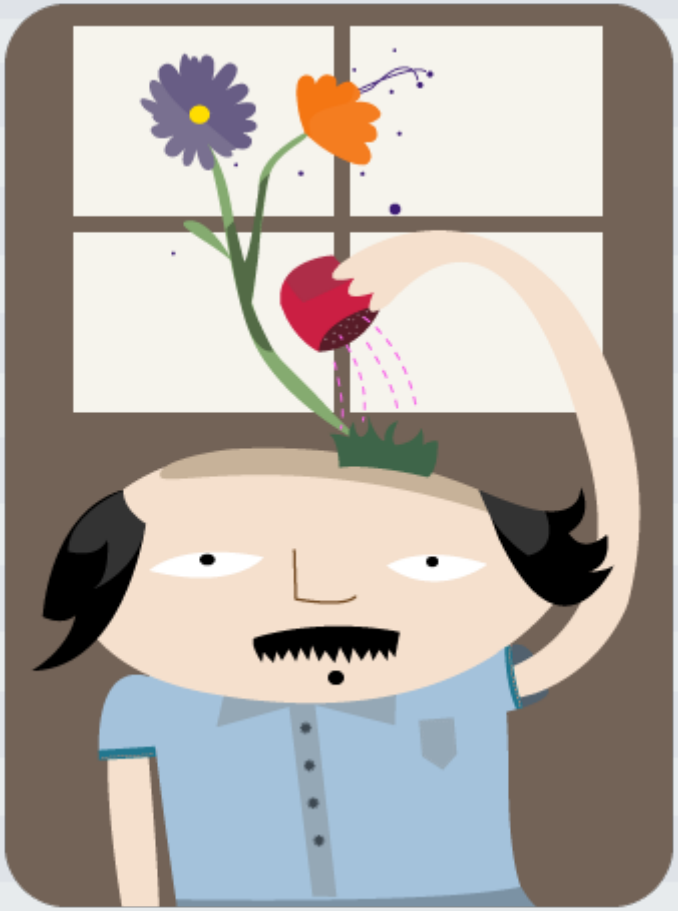
5 minutes role playing game

XOR

Input		Output
A	B	
0	0	0
0	1	1
1	0	1
1	1	0

3 peta => 1.3 peta

Harder object mutations / recovery



Simple operations and tiering

Replicated

Erasure Coded

Promoted to replica on read

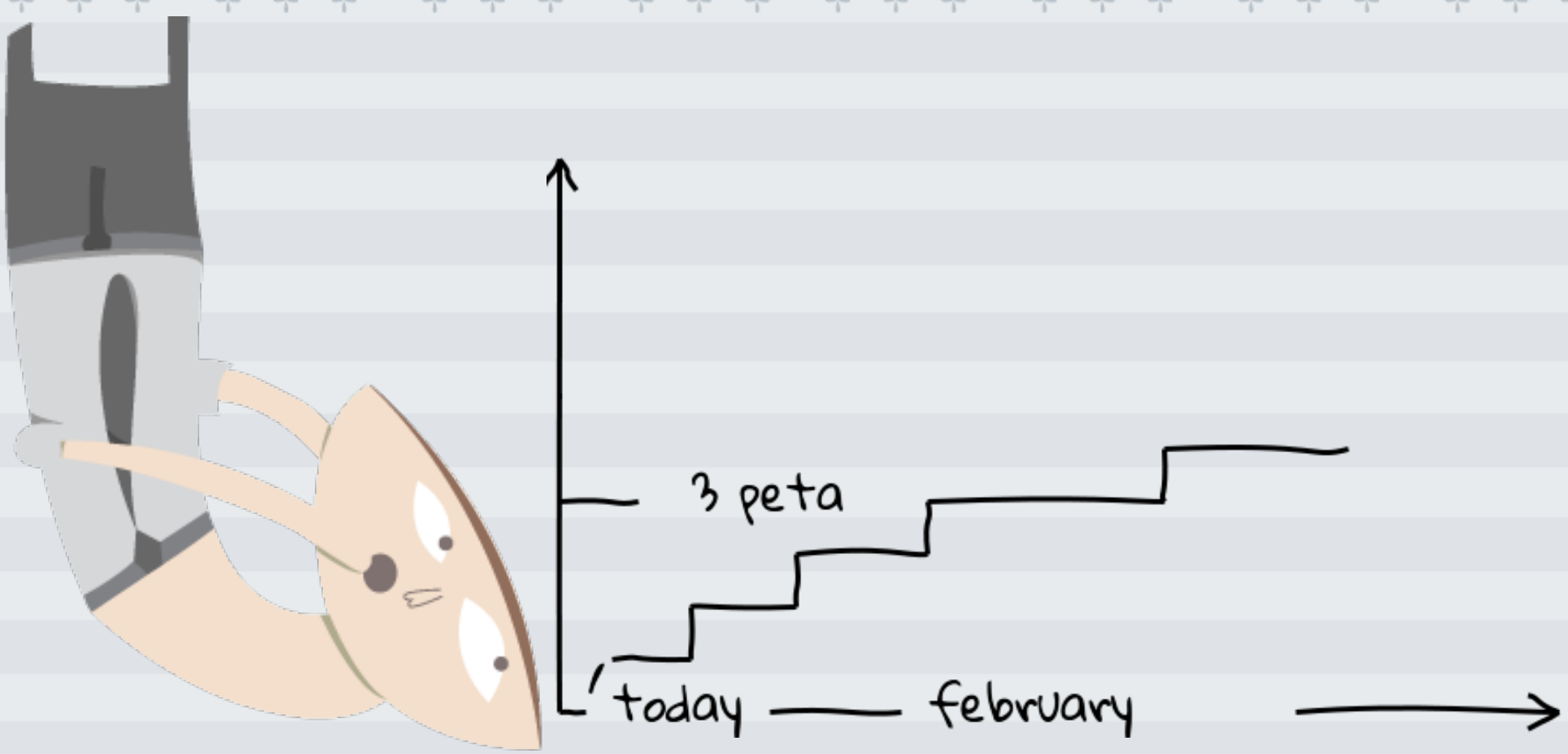
Replicated

Erasure Coded

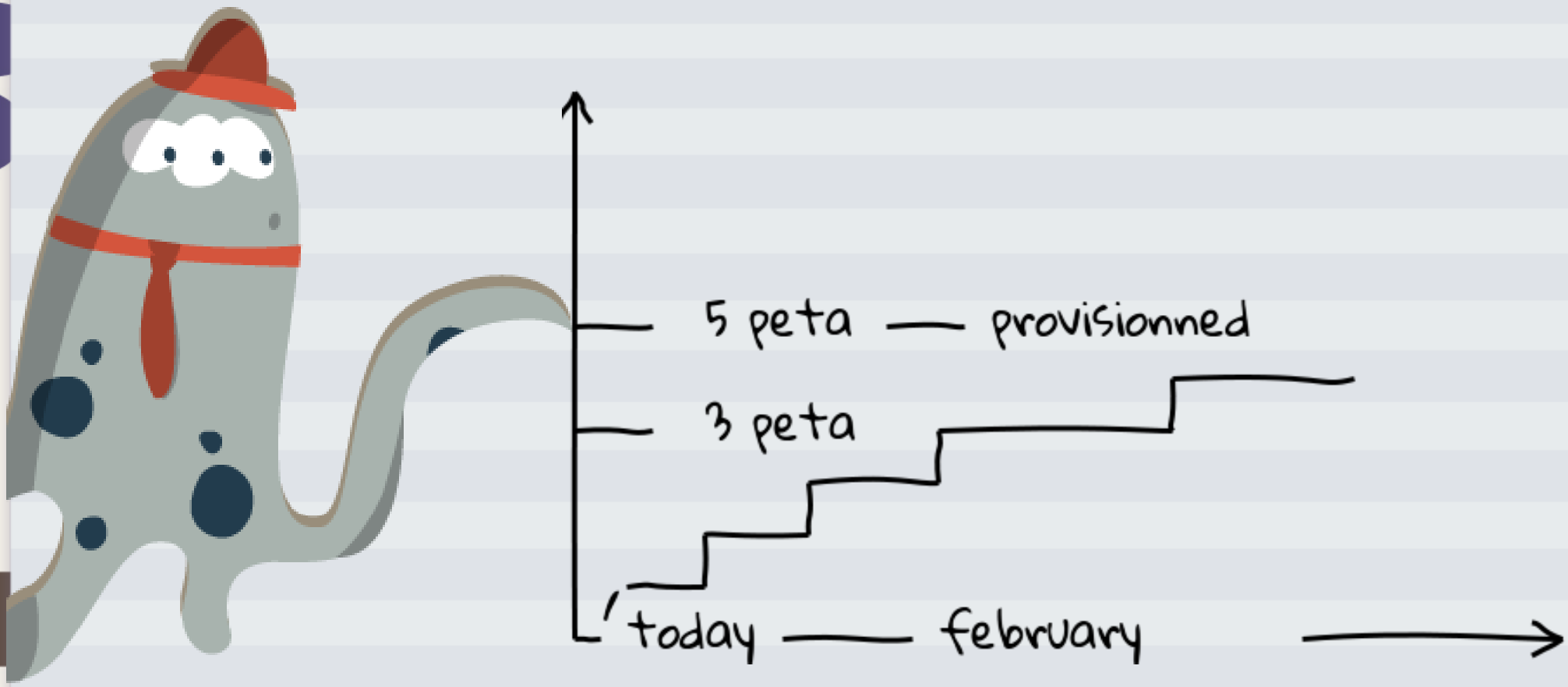
Sam & David : internals

Janne & Andreas & Loic & Takeshi erasure code

Released May 2014 : Firefly



Why save space before shortage ?

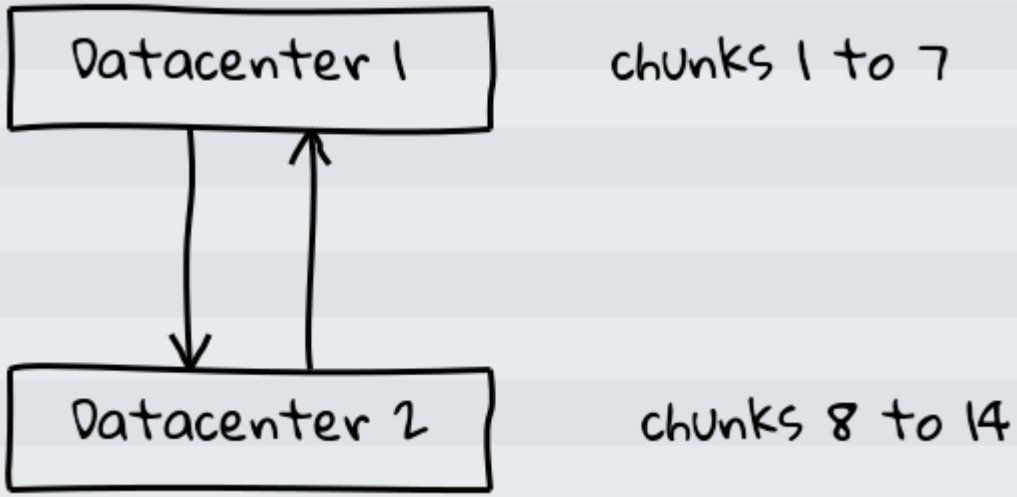


Reliability Model

- After an OSD is lost
- Recovery
- Backfilling a new OSD

April 2015 : Hammer

Repair K=10, M=4



Locally Recoverable Codes
LRC @ Red Hat

local chunk

Datacenter 1

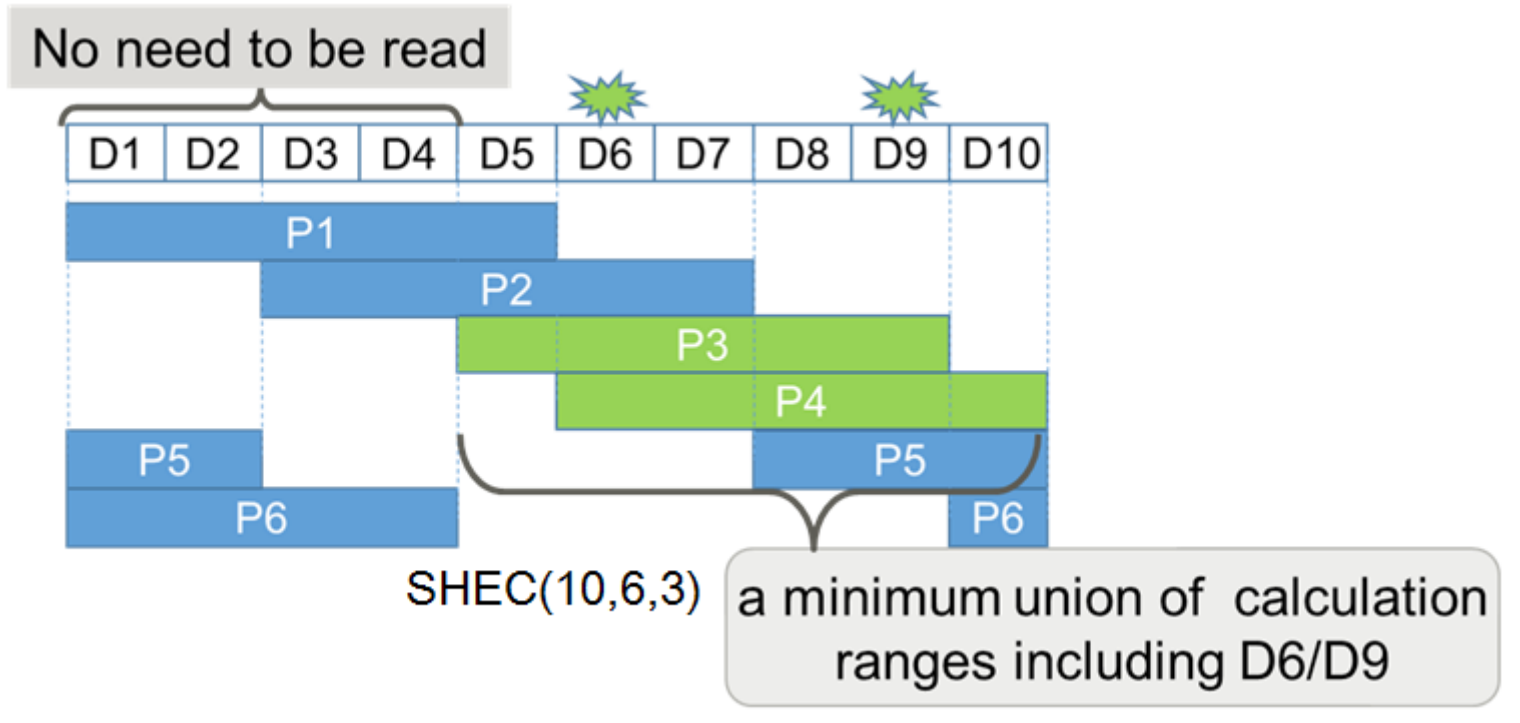
chunks 1 to 7

Datacenter 2

chunks 8 to 14

local chunk

SHEC Takeshi @ Fujitsu



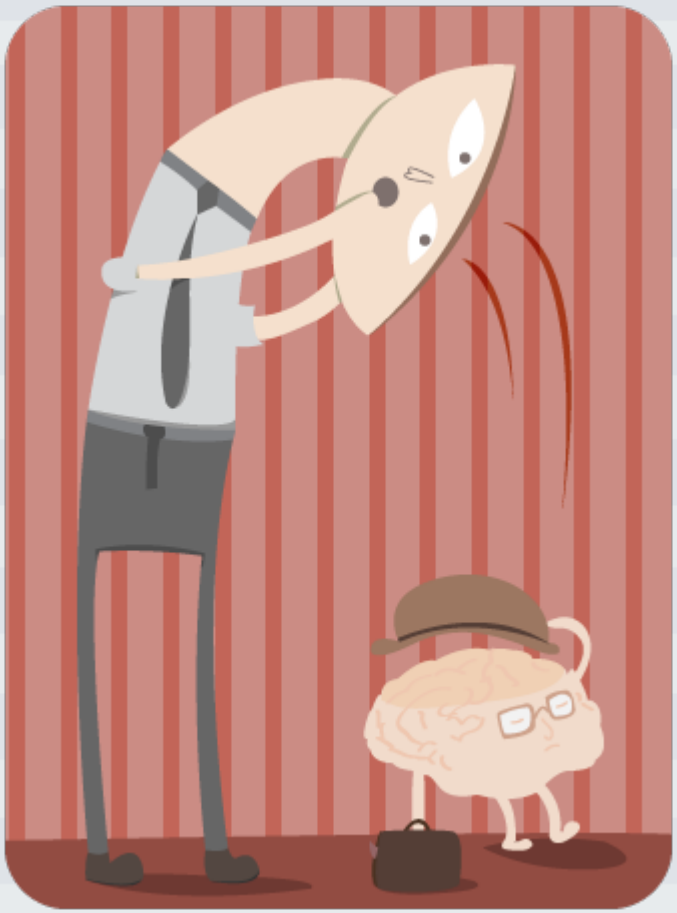
ISA plugin Yuan @ Intel

SIMD aka SSE2, SSE3, SSE4

Only for Intel processors

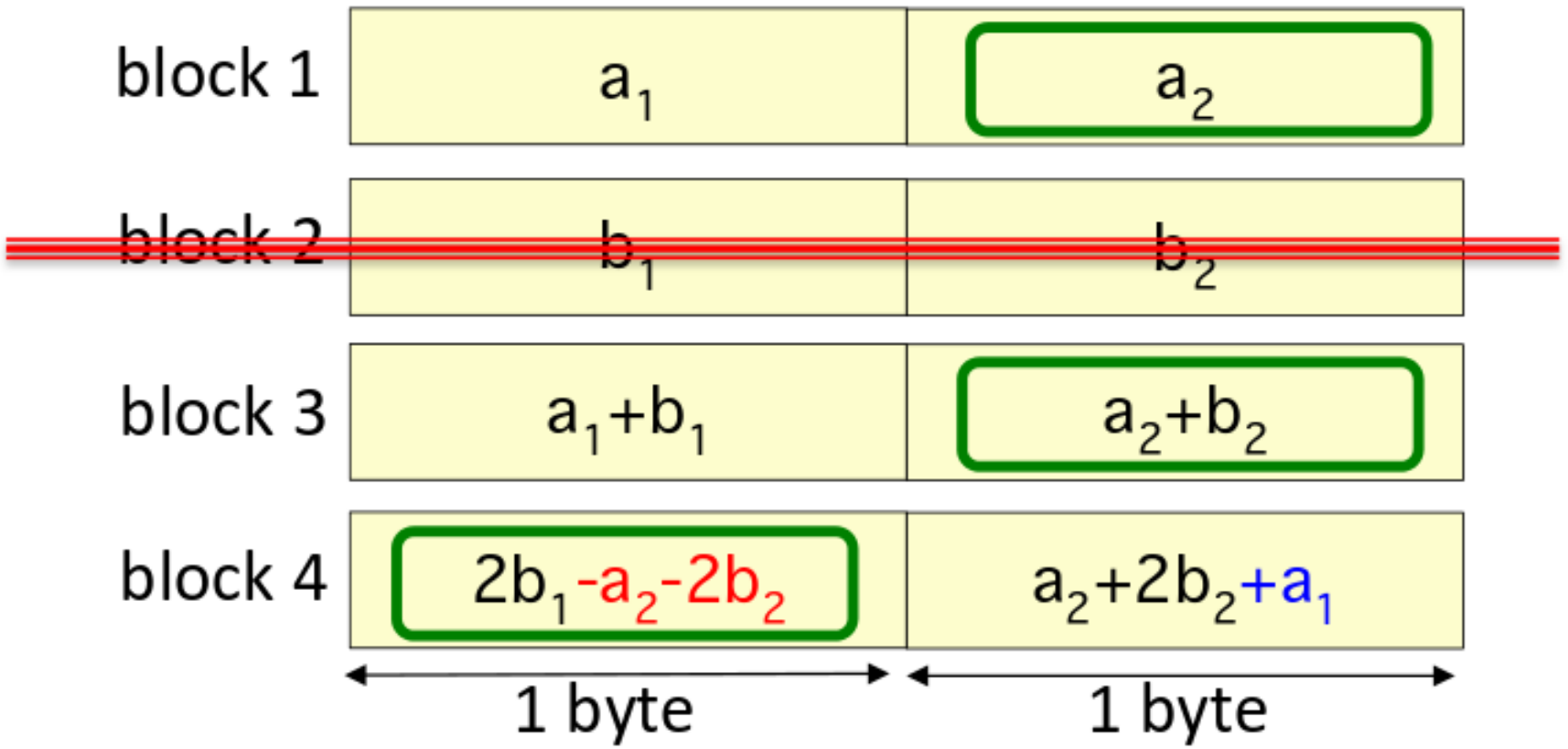
~50% Faster

jerasure ARM Janne @ ARM



Infernalis

Hitchhiker Rashmi @ U.C. Berkeley



ldachary@redhat.com

Artwork GPLv3+ Tartaruga Feliz