facebook

facebook

f4: Facebook's Warm BLOB Storage System

Subramanian Muralidhar*, Wyatt Lloyd*ᴪ, **Sabyasachi Roy***, Cory Hill*, Ernest Lin*, Weiwen Liu*, Satadru Pan*, Shiva Shankar*, Viswanath Sivakumar*, Linpeng Tang*⁺, Sanjeev Kumar*

*Facebook Inc., ᴪUniversity of Southern California, ⁺Princeton University

BLOBs@FB

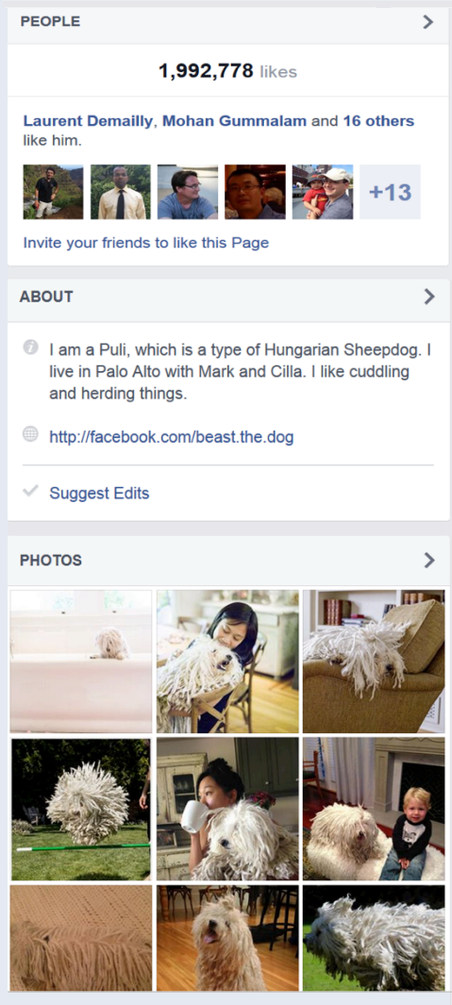
Immutable
&
Unstructured

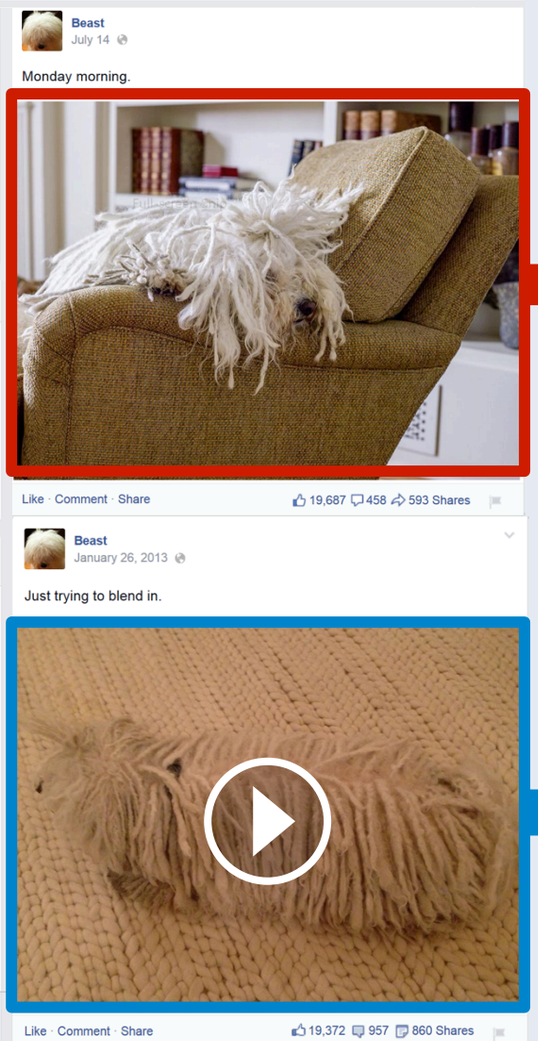
Diverse

A LOT of them!!

Profile Photo

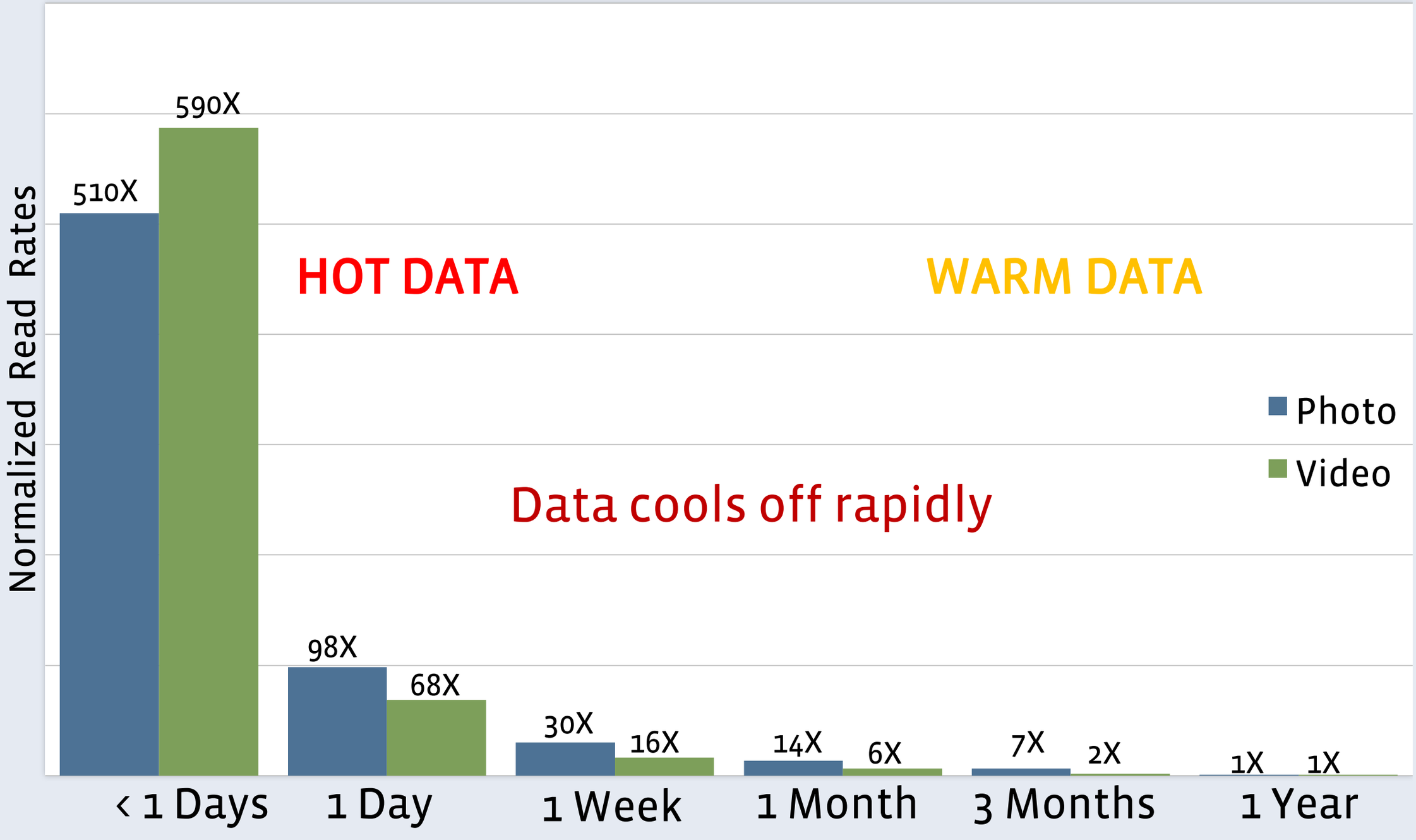
Cover Photo



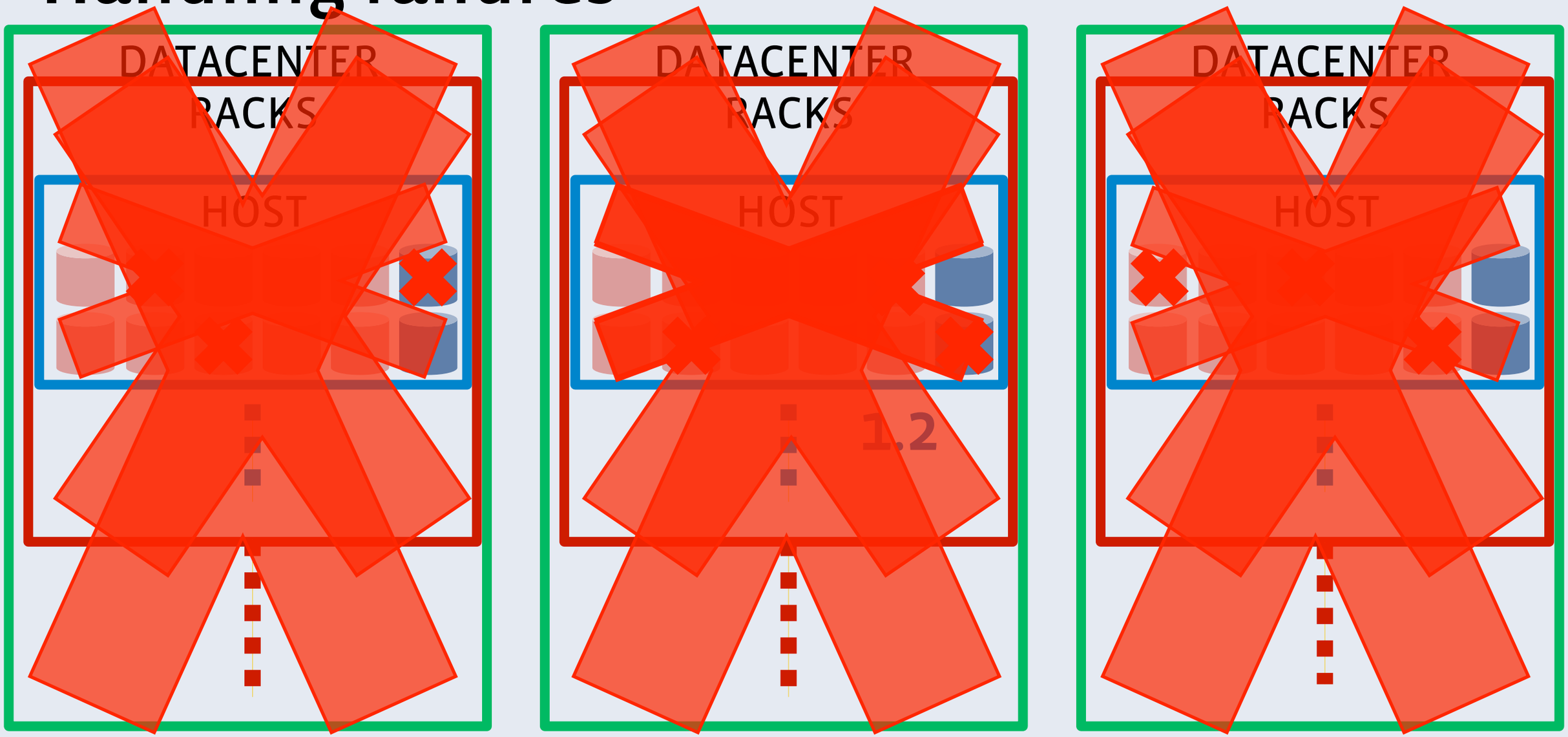


Feed Photo

Feed Video

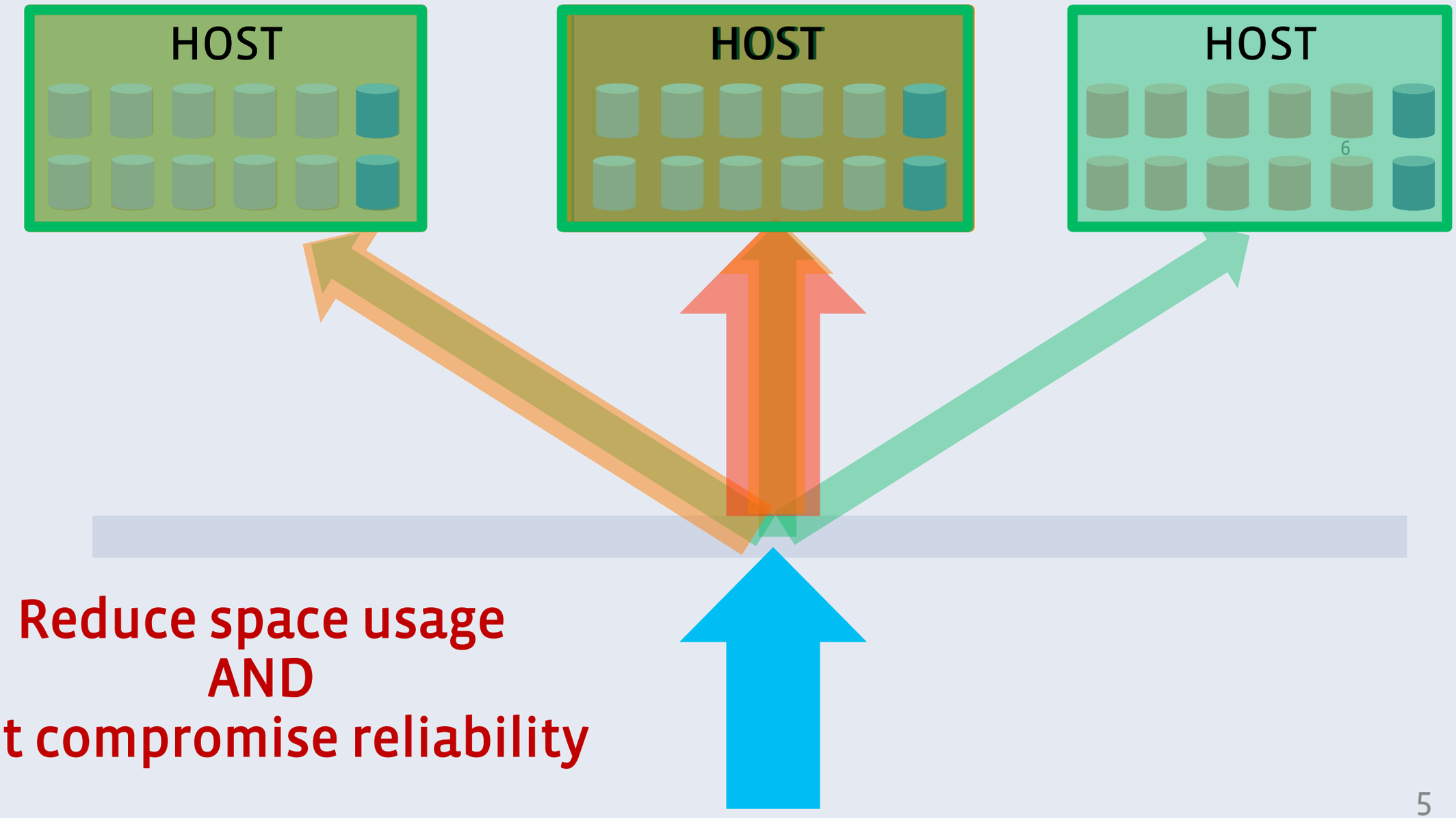

Normalized Read Rates
510X
590X
98X
68X
30X
16X
14X
6X
7X
2X
1X
1X
HOT DATA
WARM DATA
Photo
Video
Data cools off rapidly
< 1 Days
1 Day
1 Week
1 Month
3 Months
1 Year

Handling failures



Replication: * 3 = 3.6

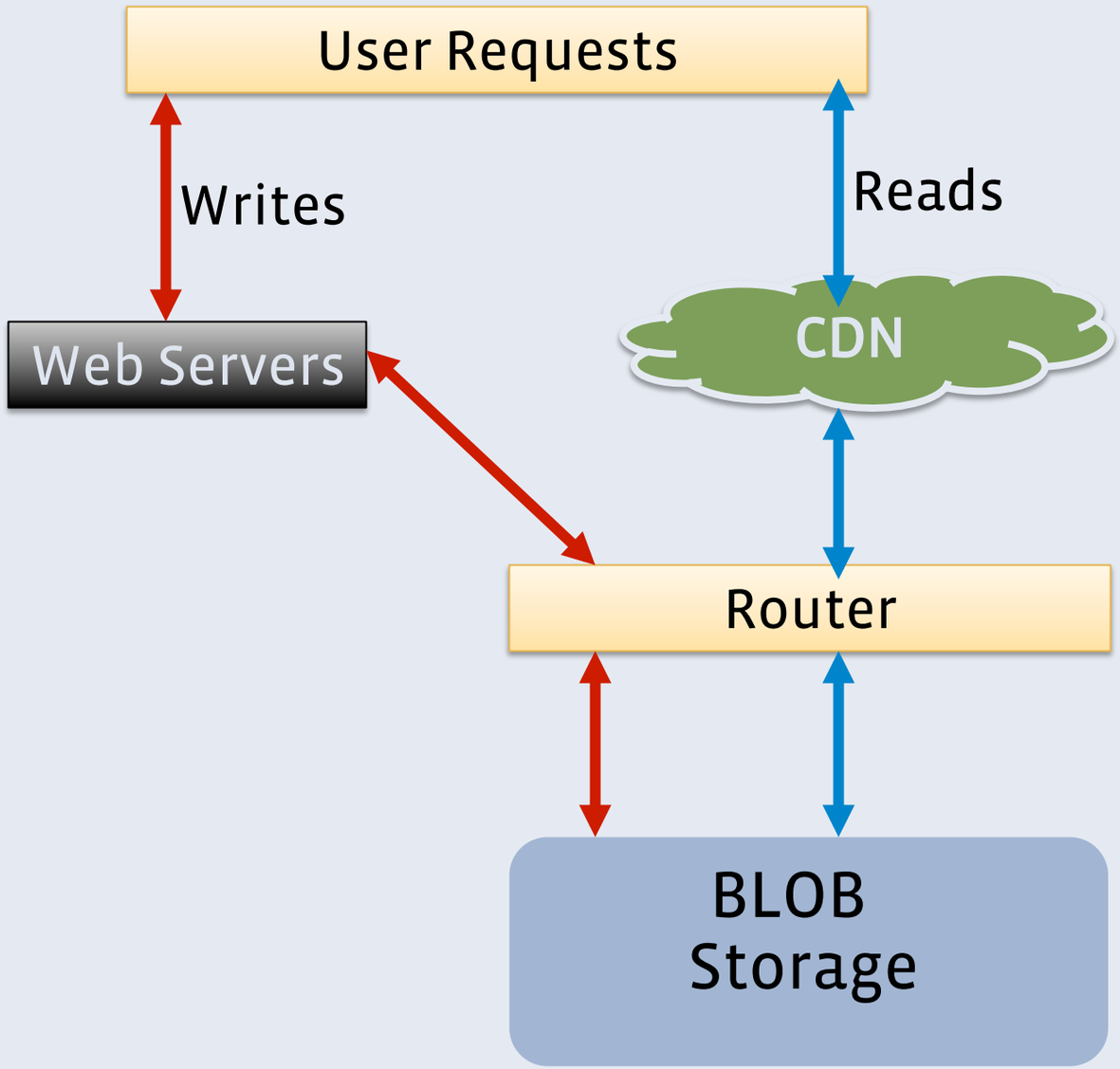
Handling load



Reduce space usage
AND
Not compromise reliability

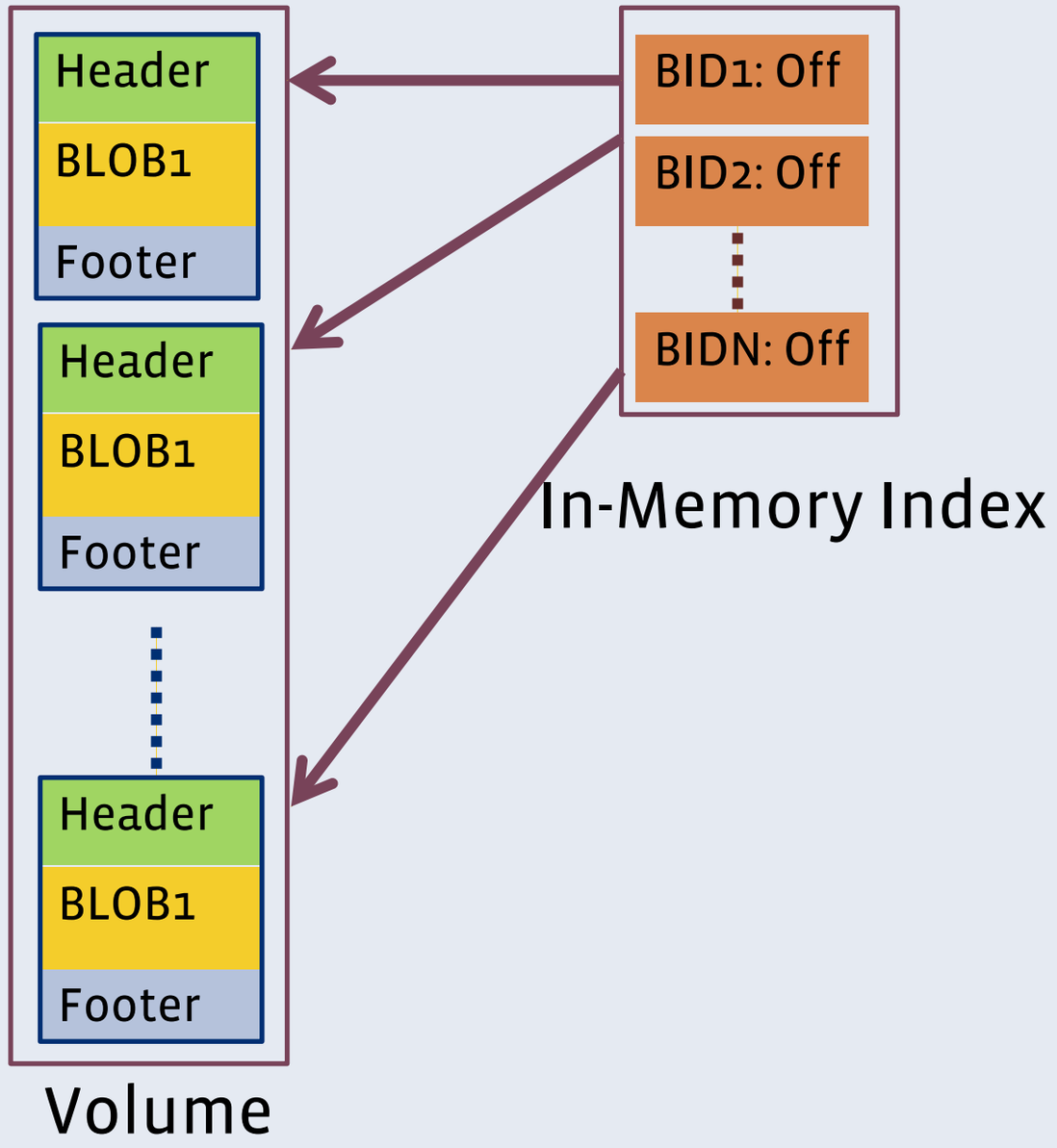
Background: Data serving

- CDN protects storage
- Router abstracts storage
- Web tier adds business logic



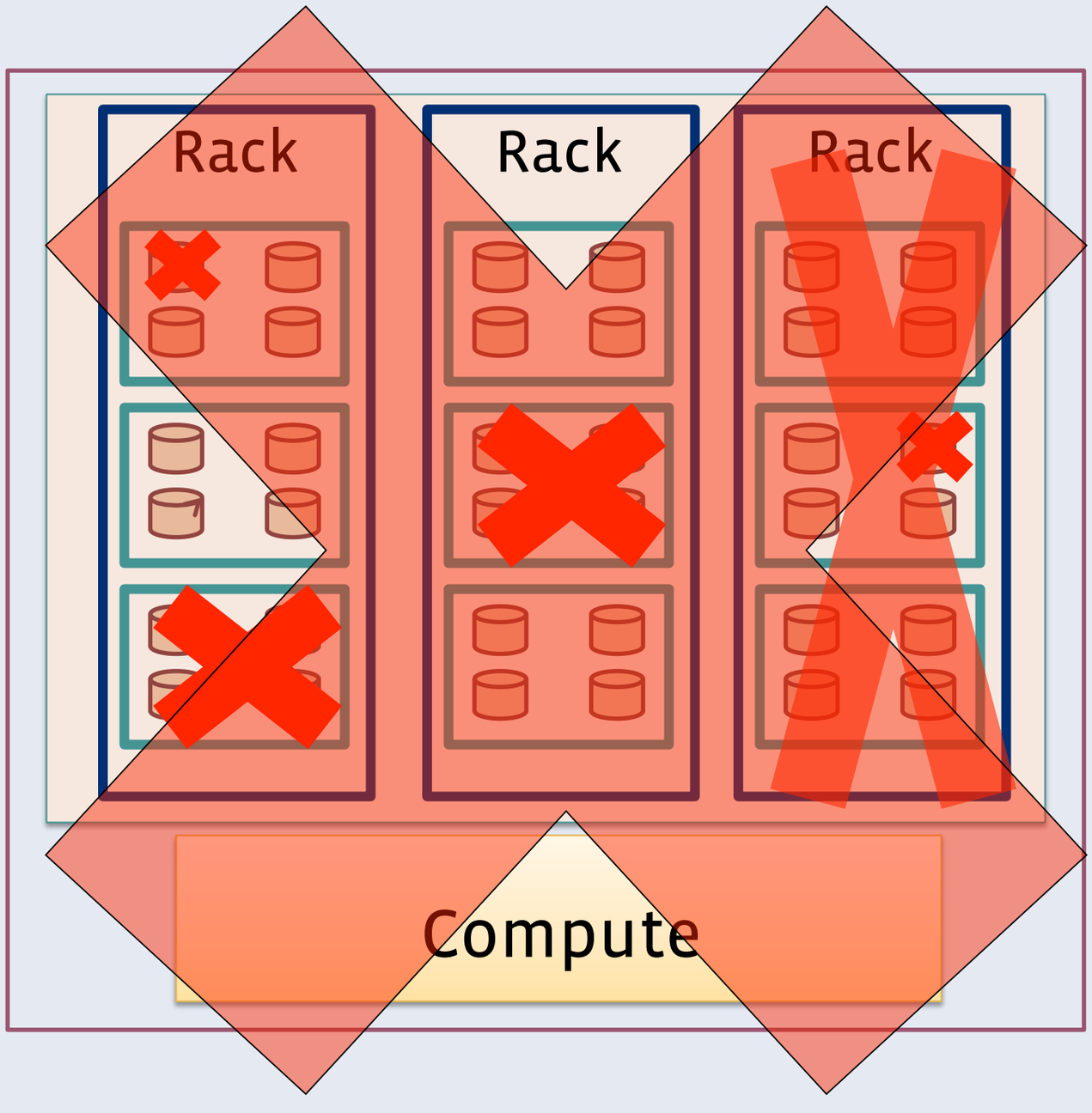
Background: Haystack [OSDI2010]

- Volume is a series of BLOBs
- In-memory index



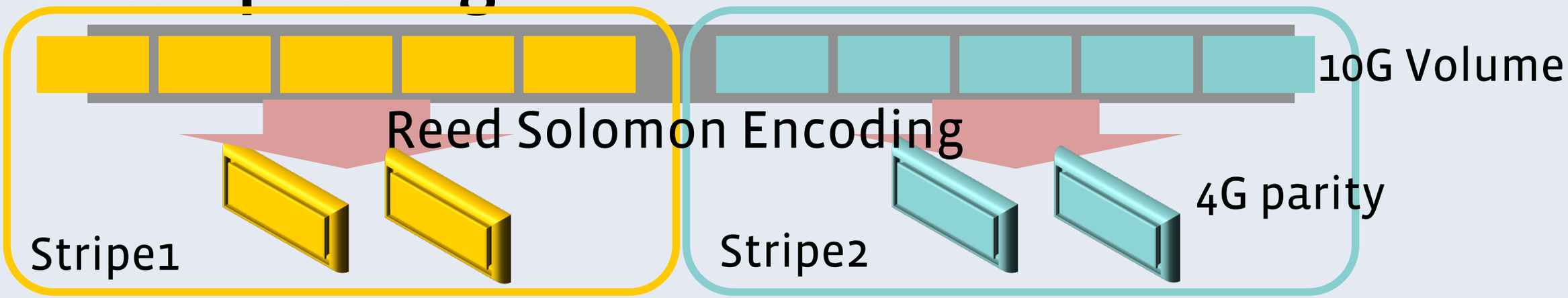
Introducing f4: Haystack on cells

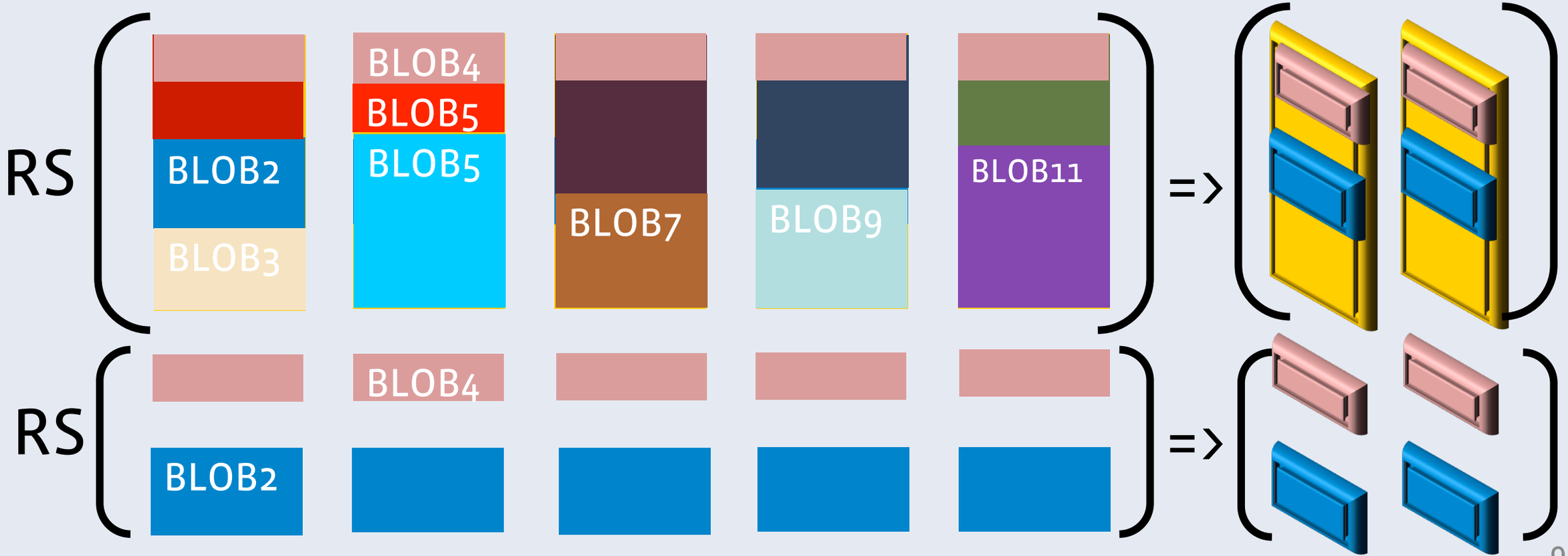
Data+Index



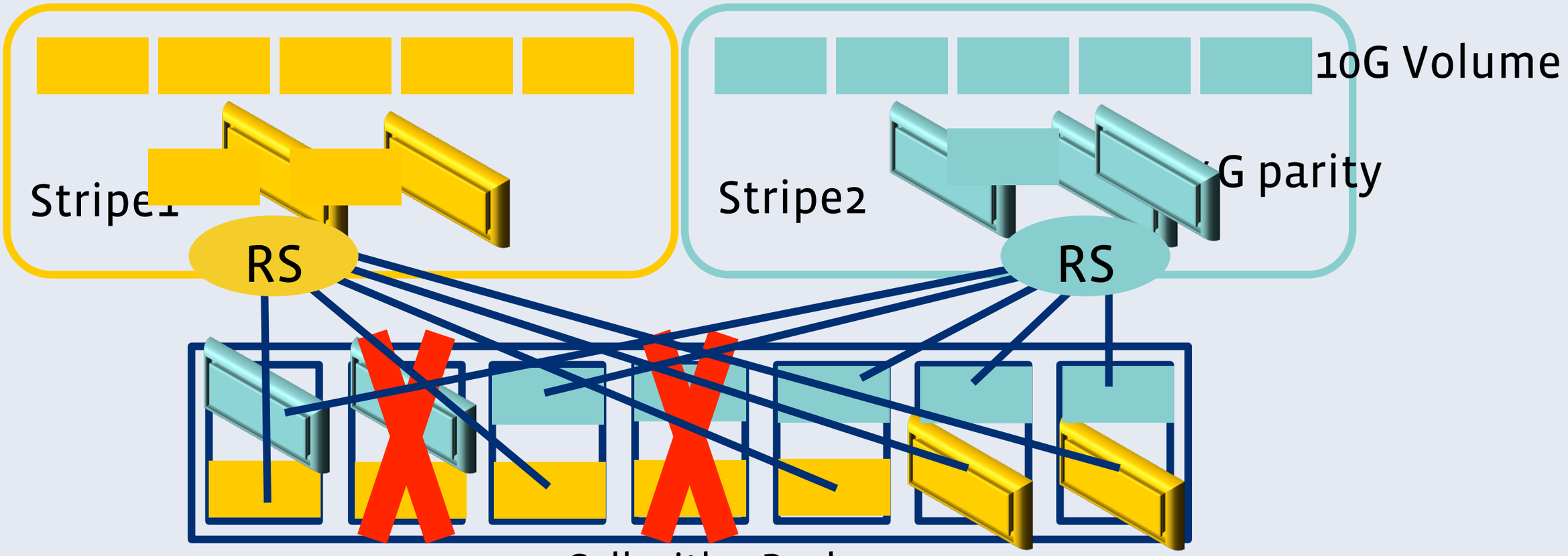
Cell

Data splitting





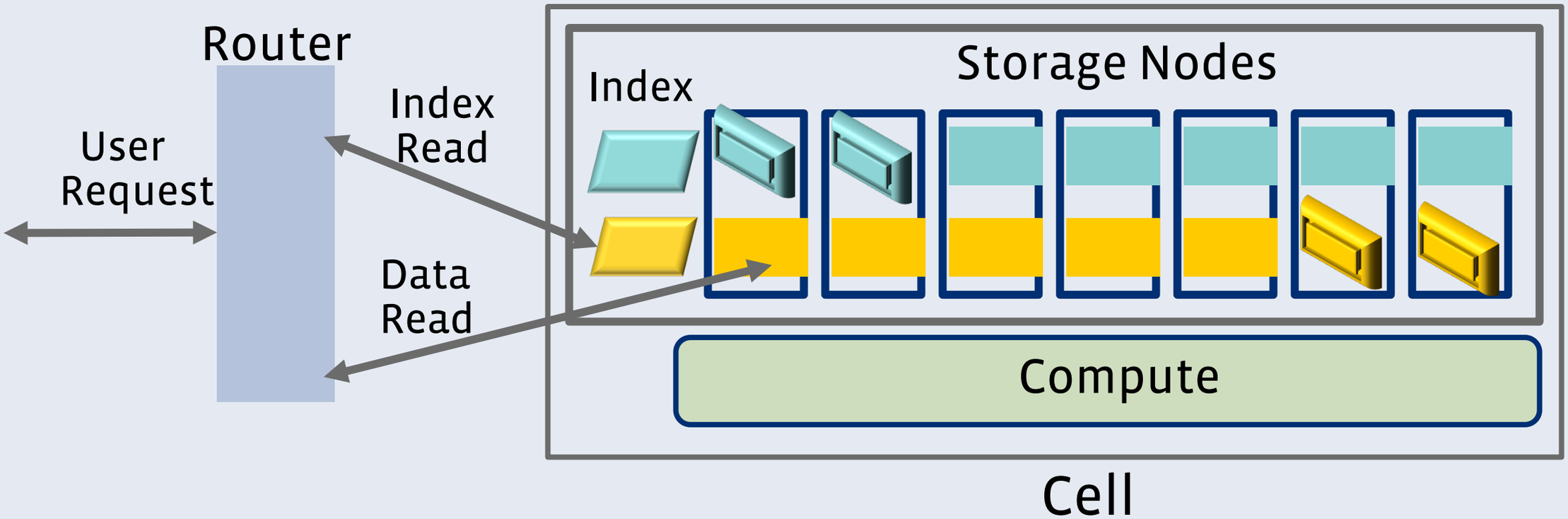
Data placement



Cell with 7 Racks

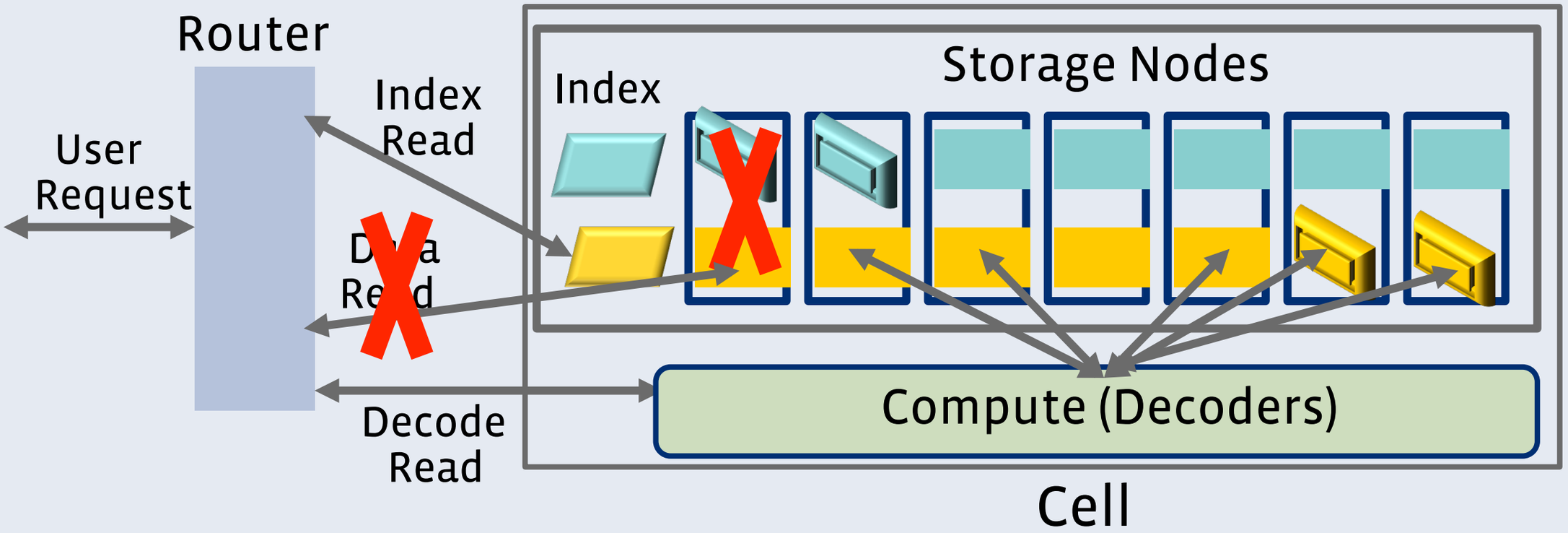
- Reed Solomon (10, 4) is used in practice (1.4X)
- Tolerates 4 racks (→ 4 disk/host) failures

Reads



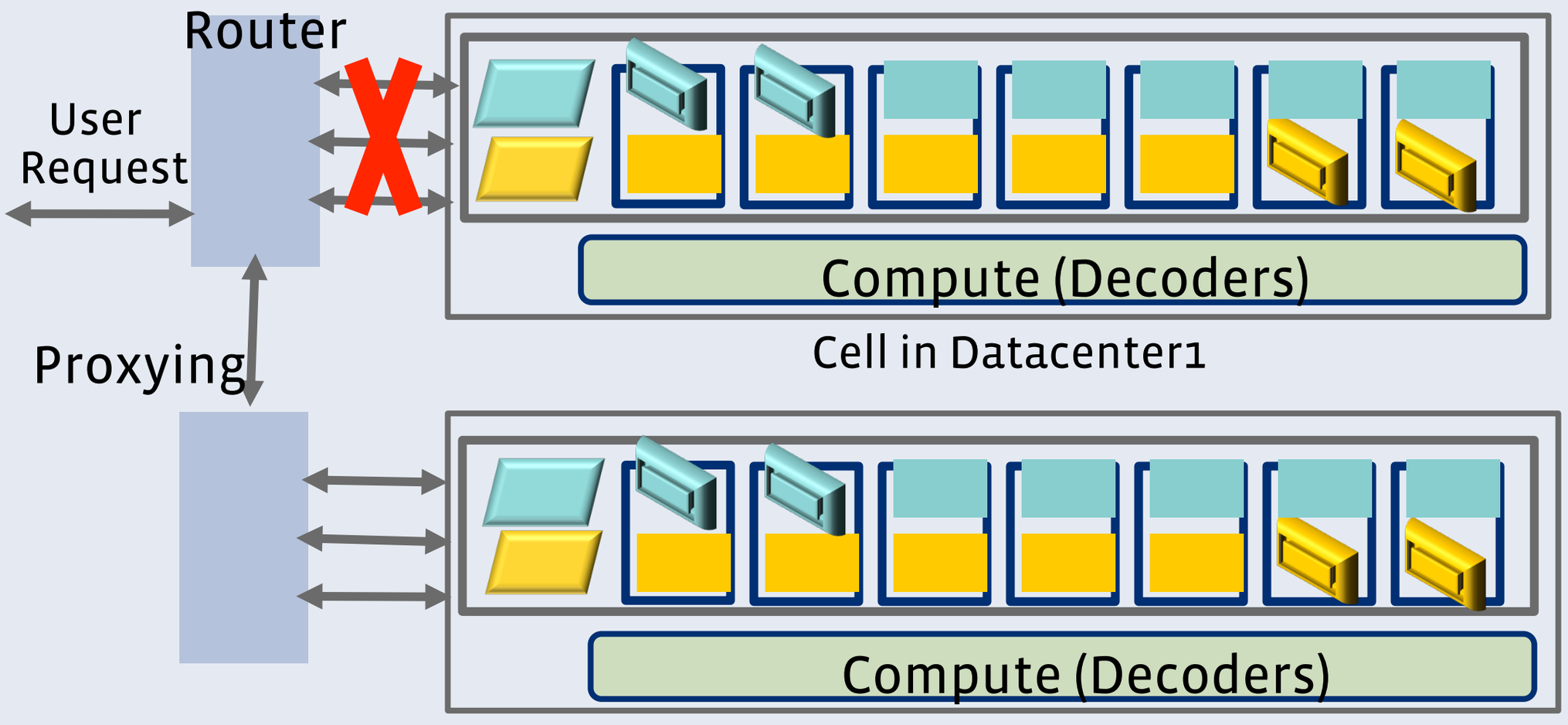
- 2-phase: Index read returns the exact physical location of the BLOB

Reads under cell-local failures



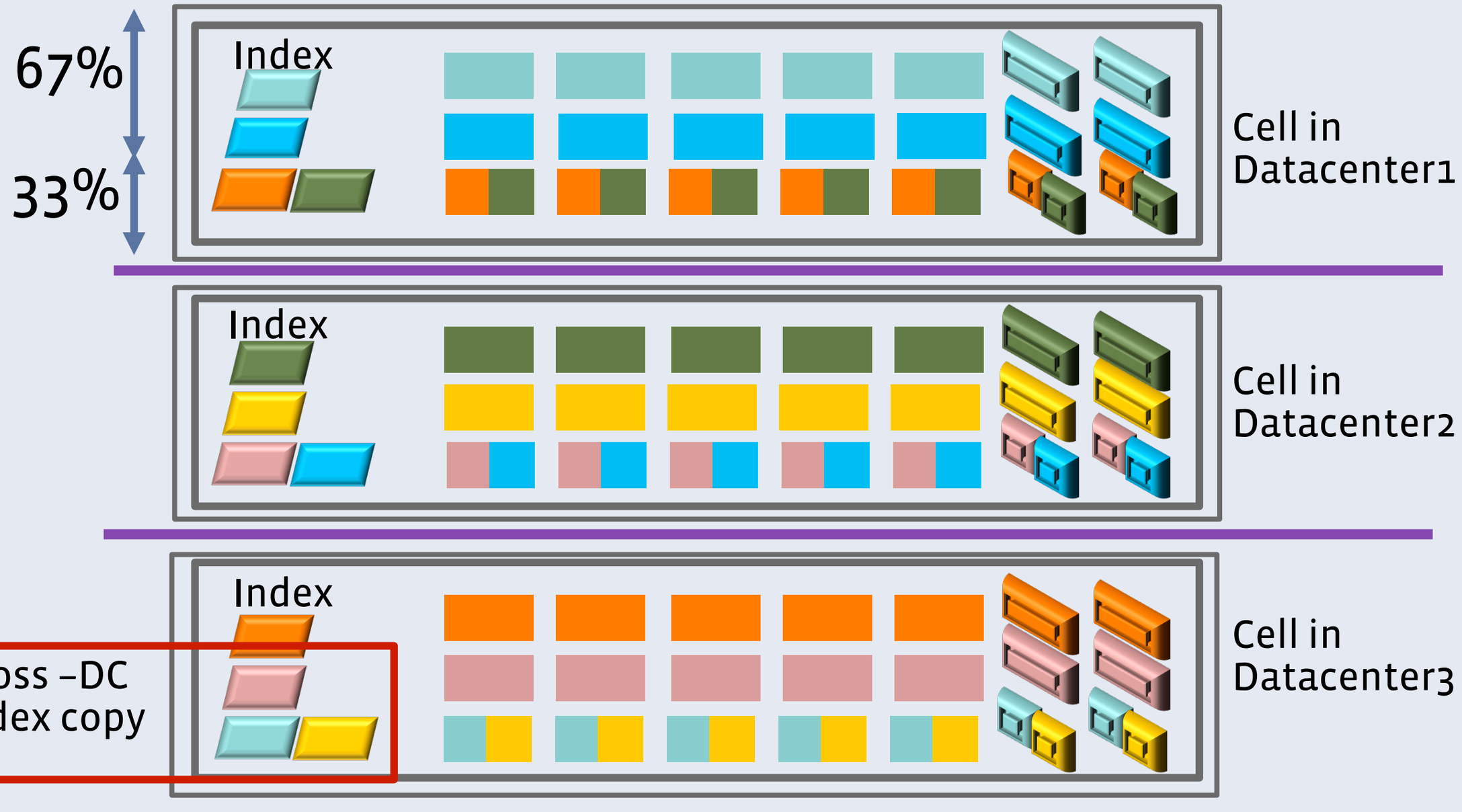
- Cell-Local failures (disks/hosts/racks) handled locally

Reads under datacenter failures (2.8X)

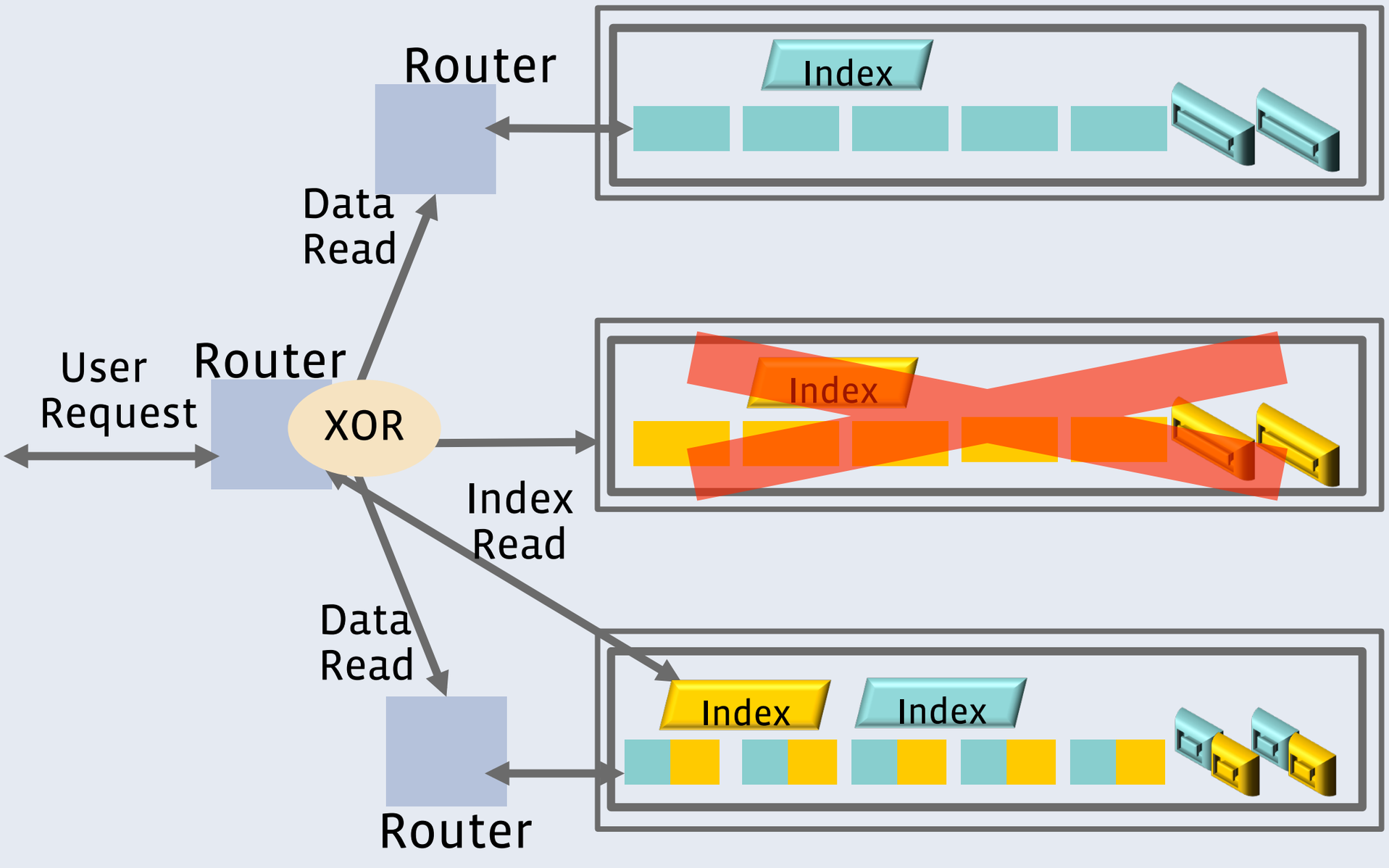


2 * 1.4X = 2.8X

Cross datacenter XOR (1.5 * 1.4 = 2.1X)



Reads with datacenter failures (2.1X)



Haystack v/s f4 2.8 v/s f4 2.1

	Haystack with 3 copies	f4 2.8	f4 2.1
Replication	3.6X	*2.8X*	*2.1X*
Irrecoverable Disk Failures	9	*10*	*10*
Irrecoverable Host Failures	3	*10*	*10*
Irrecoverable Rack failures	3	*10*	*10*
Irrecoverable Datacenter failures	*3*	2	2
Load split	*3X*	2X	1X

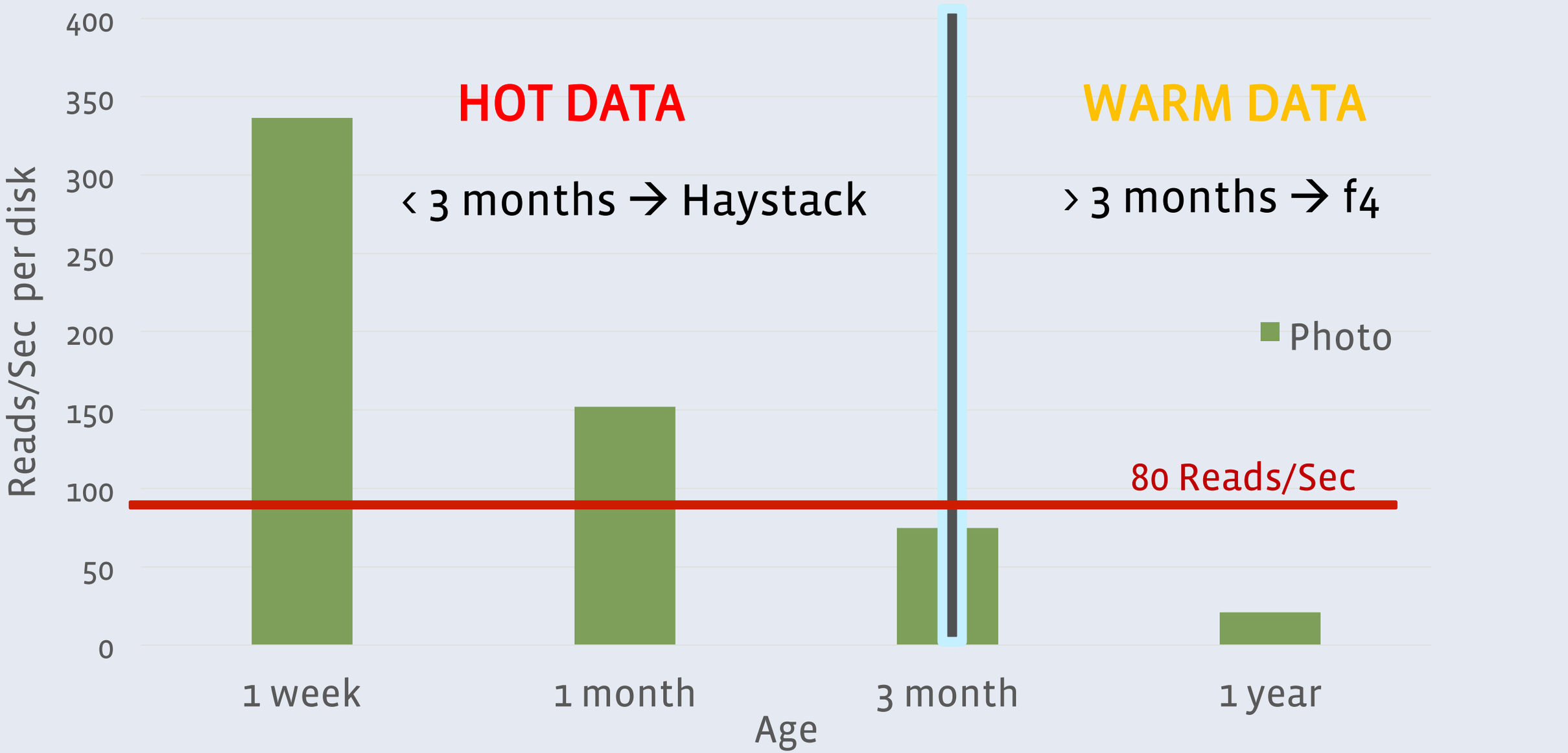
Evaluation

- What and how much data is “warm”?
- Can f4 satisfy throughput and latency requirements?
- How much space does f4 save
- f4 failure resilience

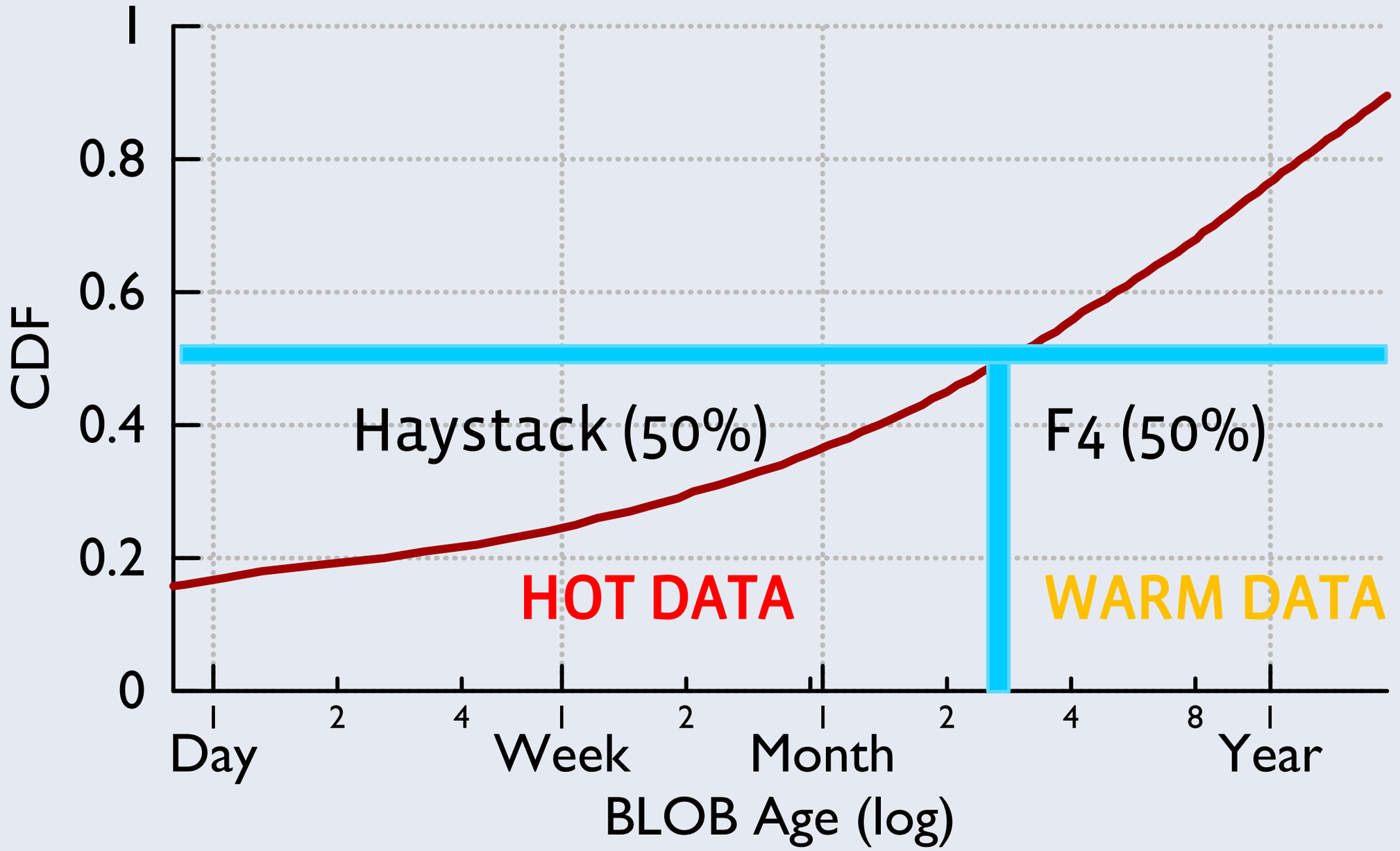
Methodology

- CDN data: 1 day, 0.5% sampling
- BLOB store data: 2 week, 0.1%
- Random distribution of BLOBs assumed
- The worst case rates reported

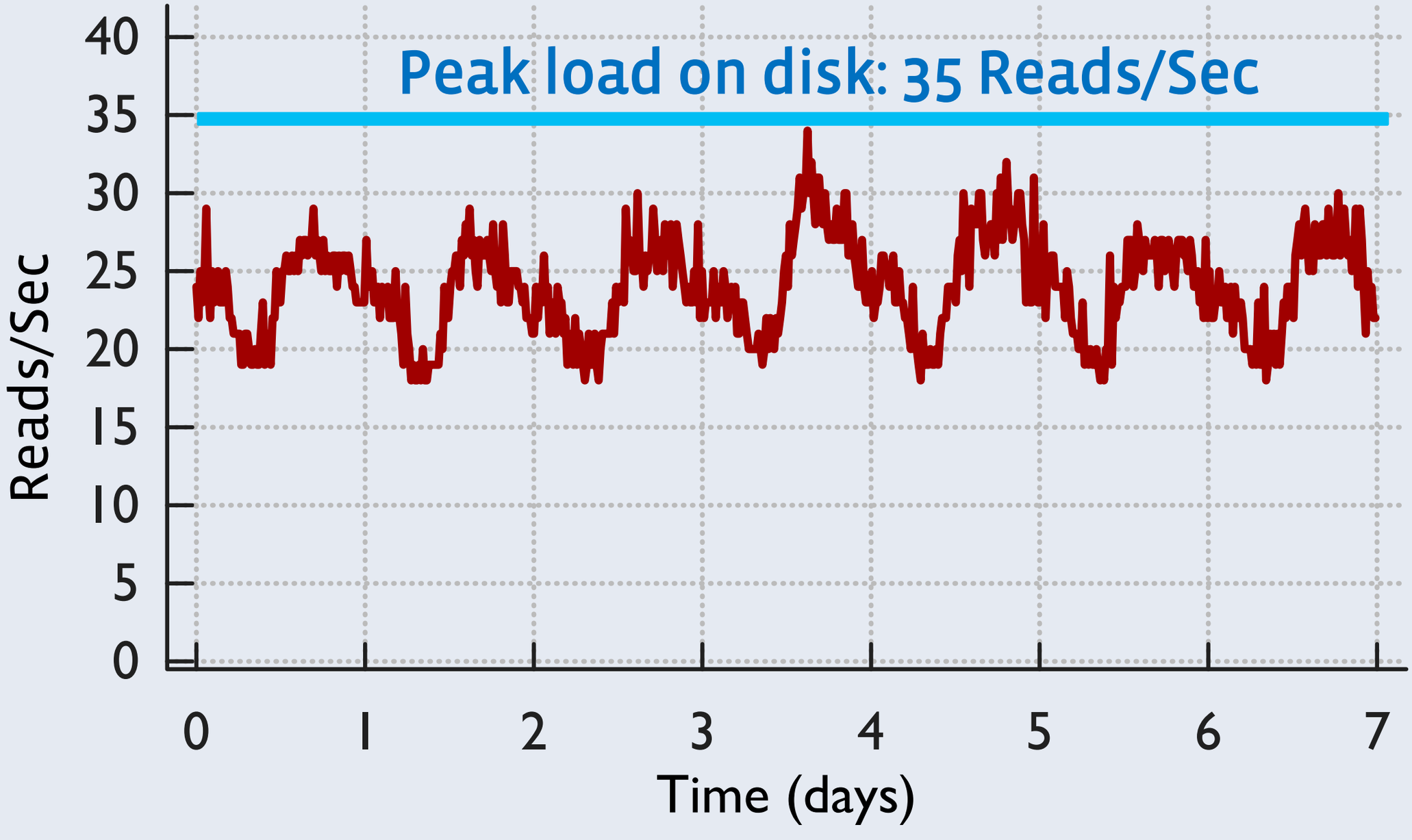
Hot and warm divide



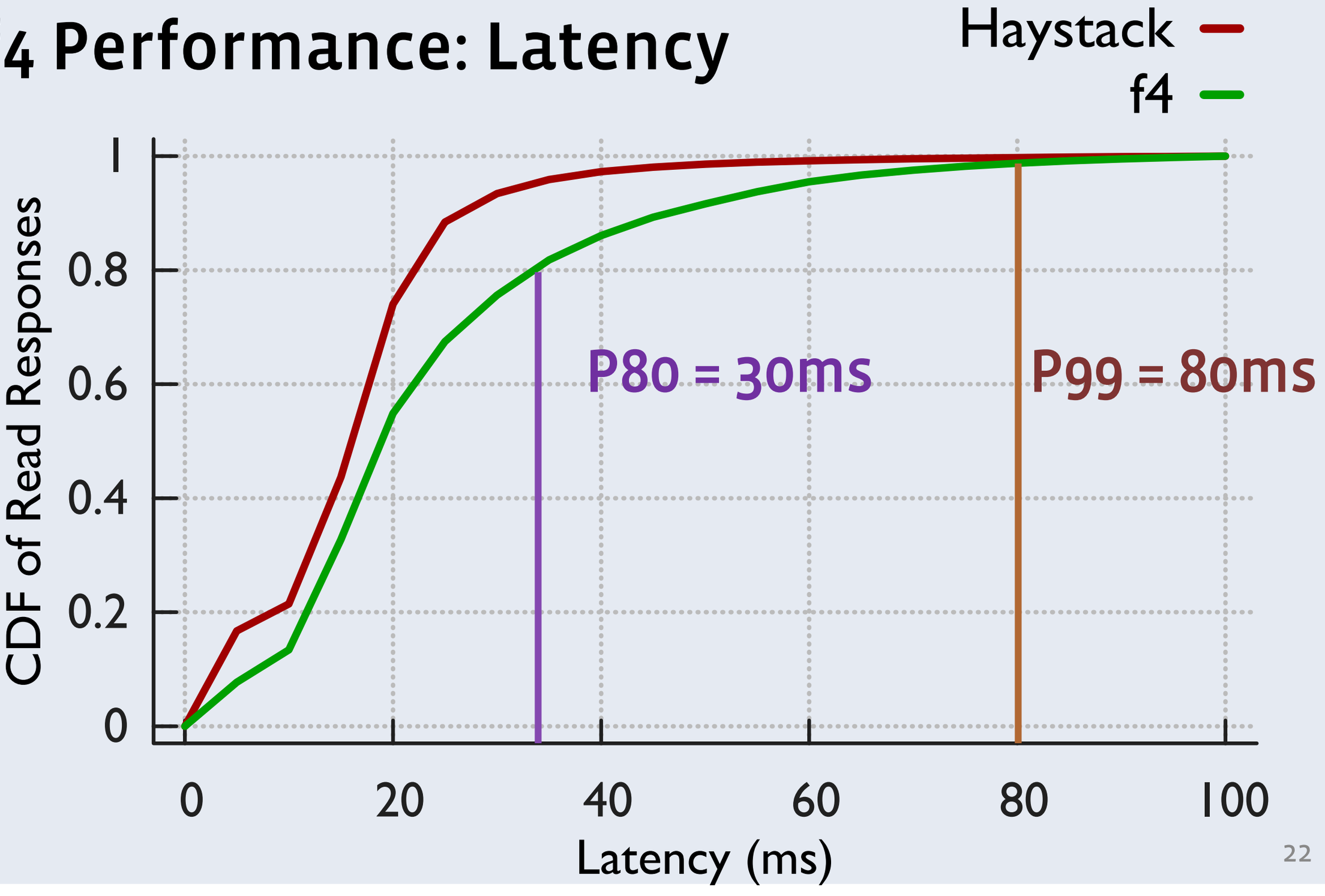
It is warm, not cold



f4 Performance: Most loaded disk in cluster



f4 Performance: Latency



Concluding Remarks

- Facebook's BLOB storage is big and growing
- BLOBs cool down with age
  - ~100X drop in read requests in 60 days
- Haystack's 3.6X replication over provisioning for old, warm data.
- f4 encodes data to lower replication to 2.1X

facebook