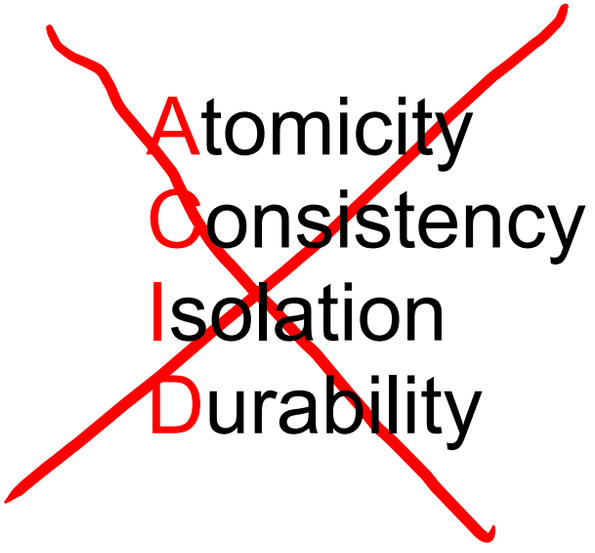
CS244b

Dynamo (2007)

- Eventual Consistency
- Quorum Systems
- Tail Latency
- Trade-offs in messy real-world systems

key-value put/get

NoSQL



Availability & response time

TAIL LATENCY

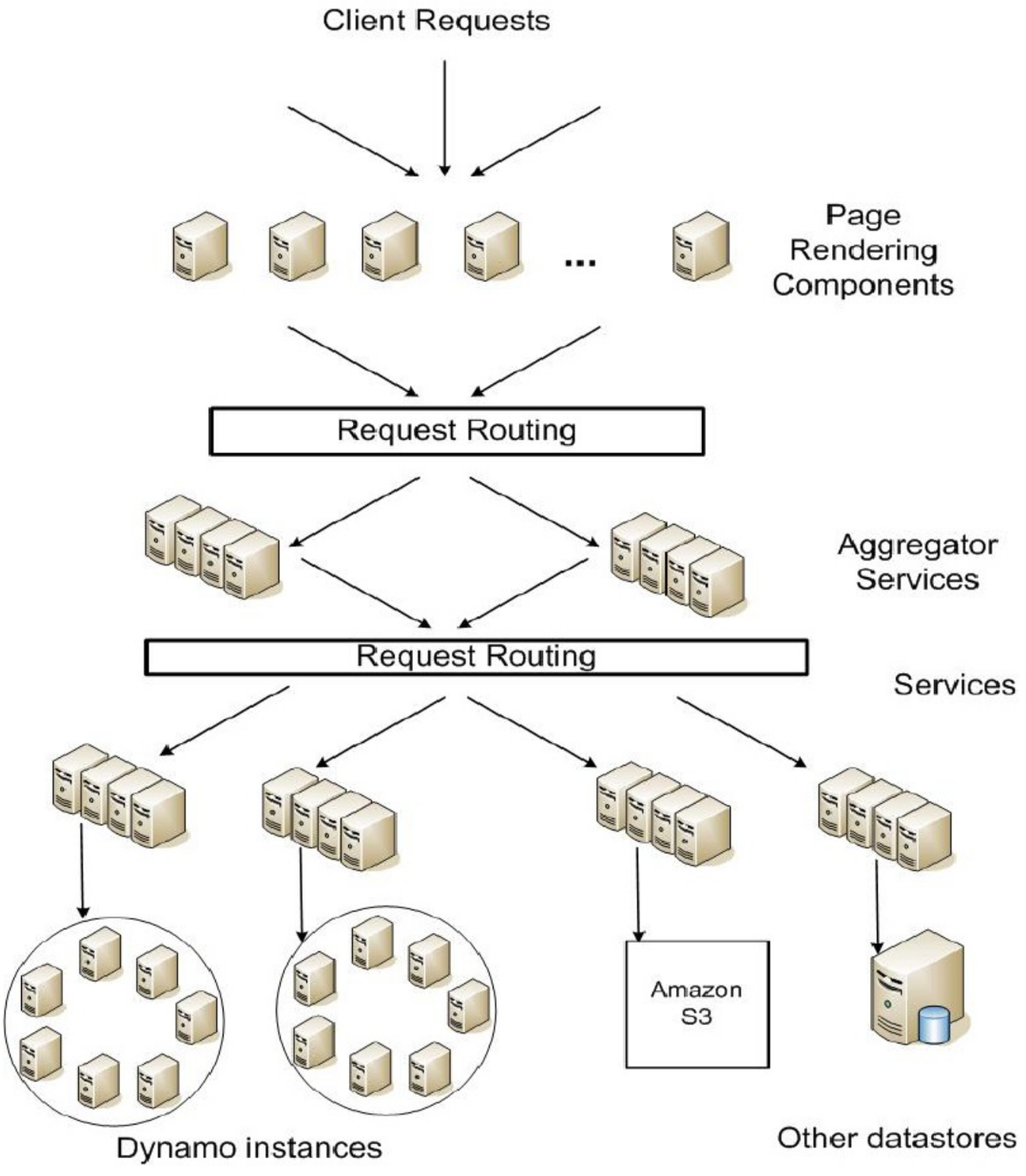
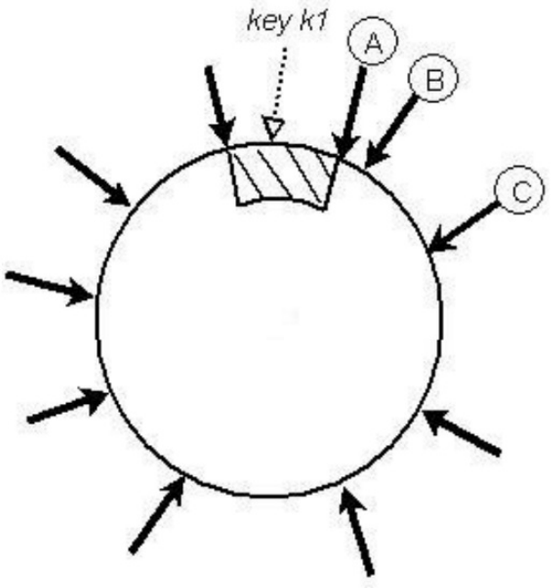


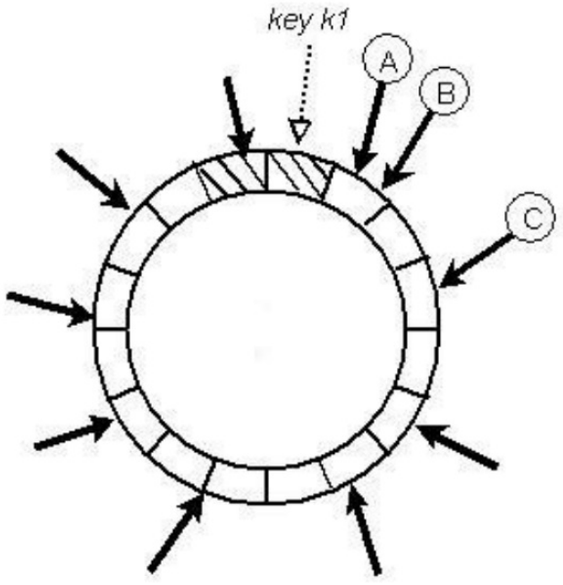
Figure 1: Service-oriented architecture of Amazon's

get (key) -> (context, list of values)
put (key, context, value) -> void

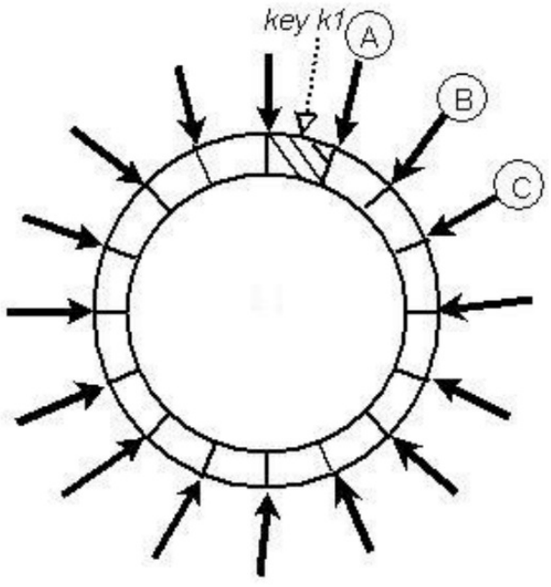
- Replicate on N
- Heterogeneous node
- Minimize Churn



Strategy 1



Strategy 2



Strategy 3

S = #nodes, T = tokens/node, Q = #partitions (#2-3)

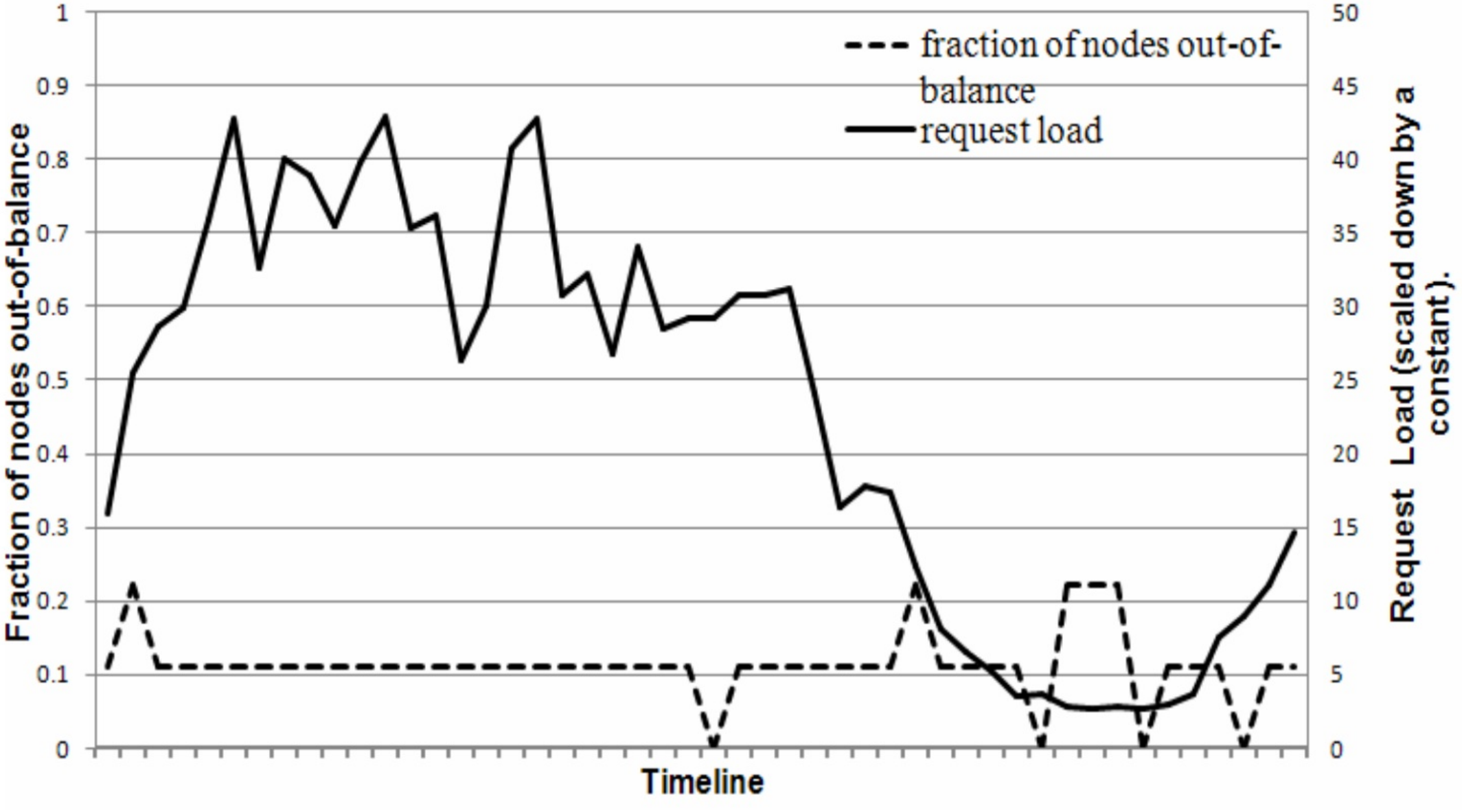


Figure 6: Fraction of nodes that are out-of-balance (i.e., nodes

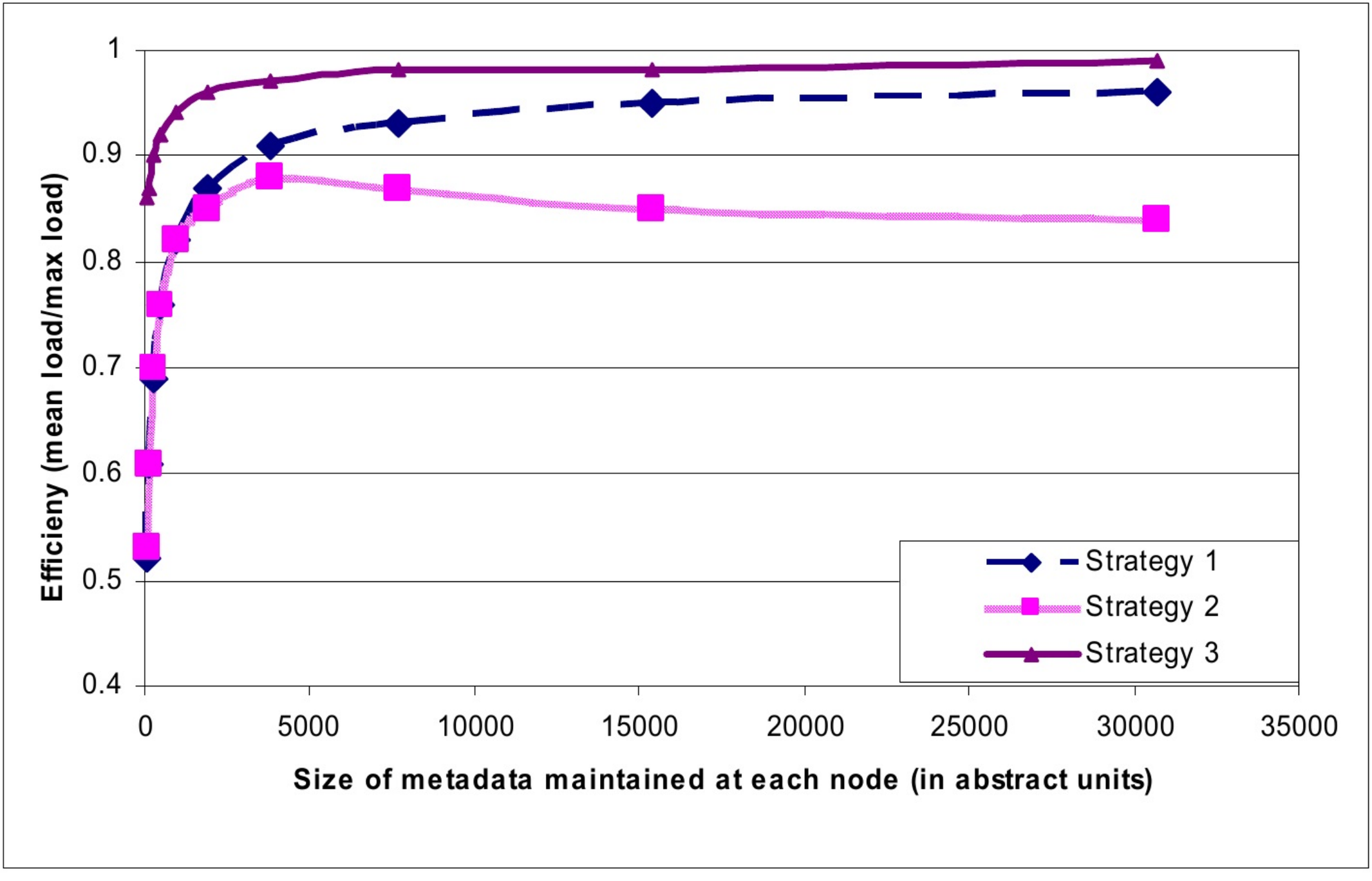
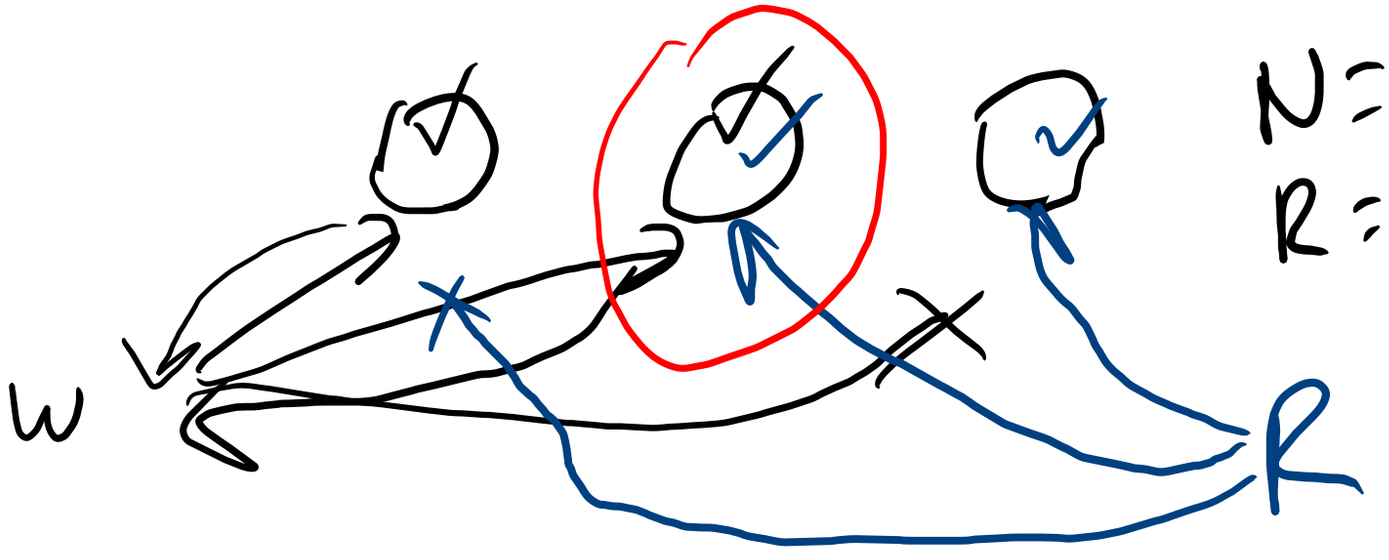


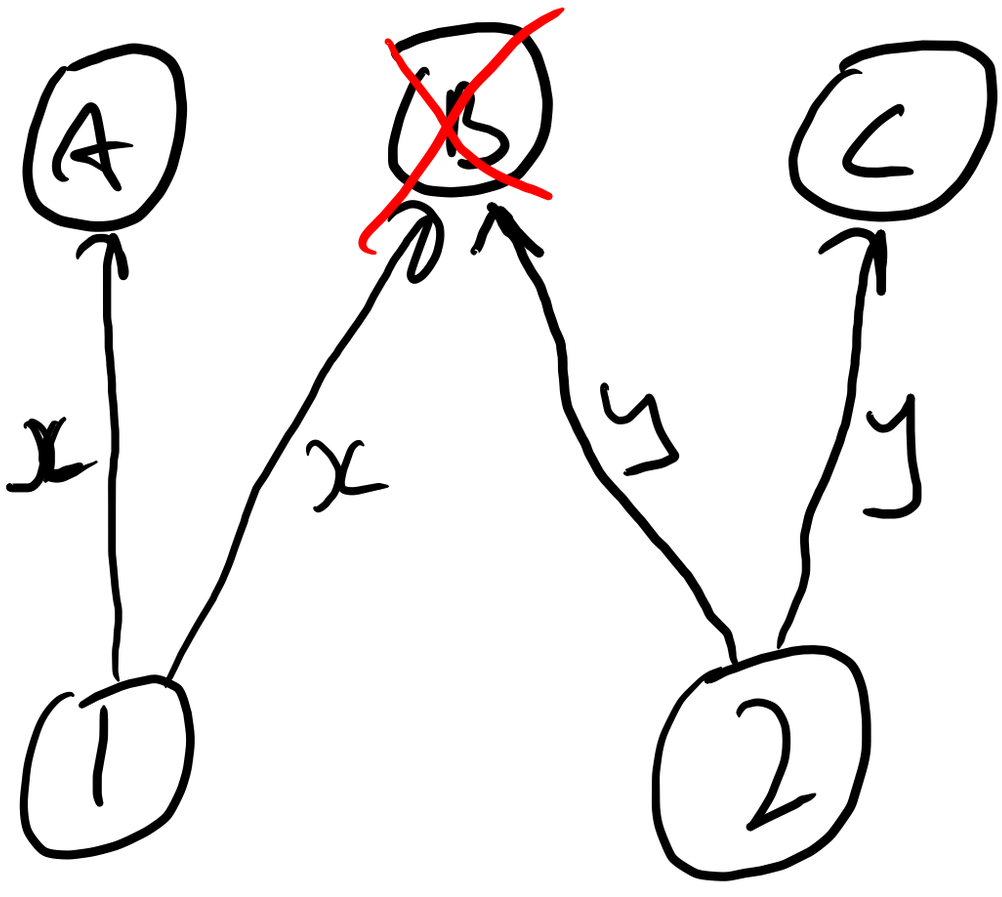
Figure 8: Comparison of the load distribution efficiency of

Quorum Systems



N = 3
R = W = 2

- write to W servers
- read from R servers
- $R + W > N$


A
B
C
x
x
y
y
1
2

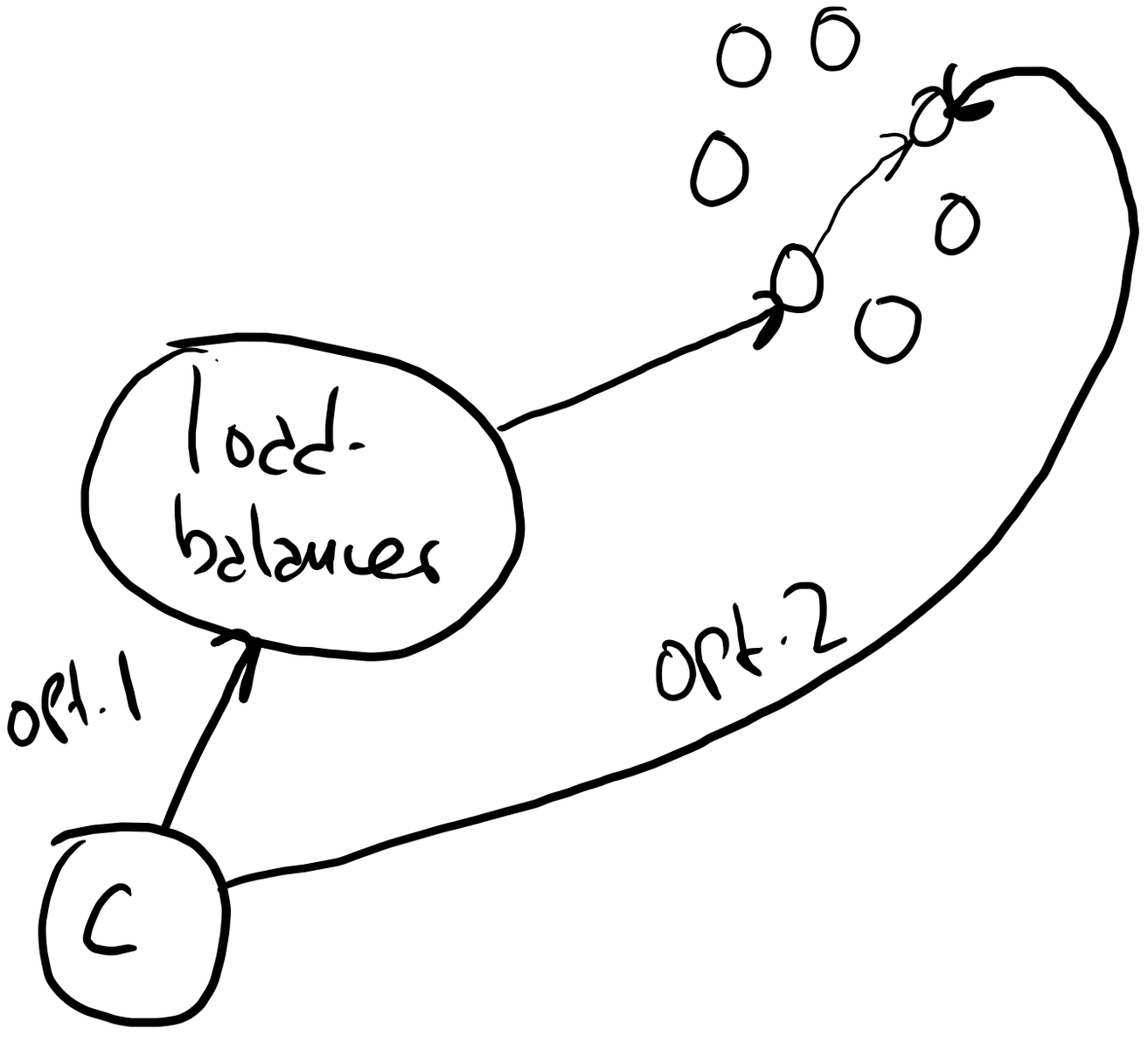
- Write
  - Ask R replicas for vers#
  - Pick new vers# higher
  - Send new version w. # to servers
  - return if/when hear from W
- Read
  - ask for latest vers. (value, vers)
  - received R matching replies → done
  - Re-broadcast latest (val, vers)

Malkhi & Reiter
Byzantine Quorum Systems

Reader needs R — t+1 identical replies

$R + W - N > 2f$

Dynamo Request Flow



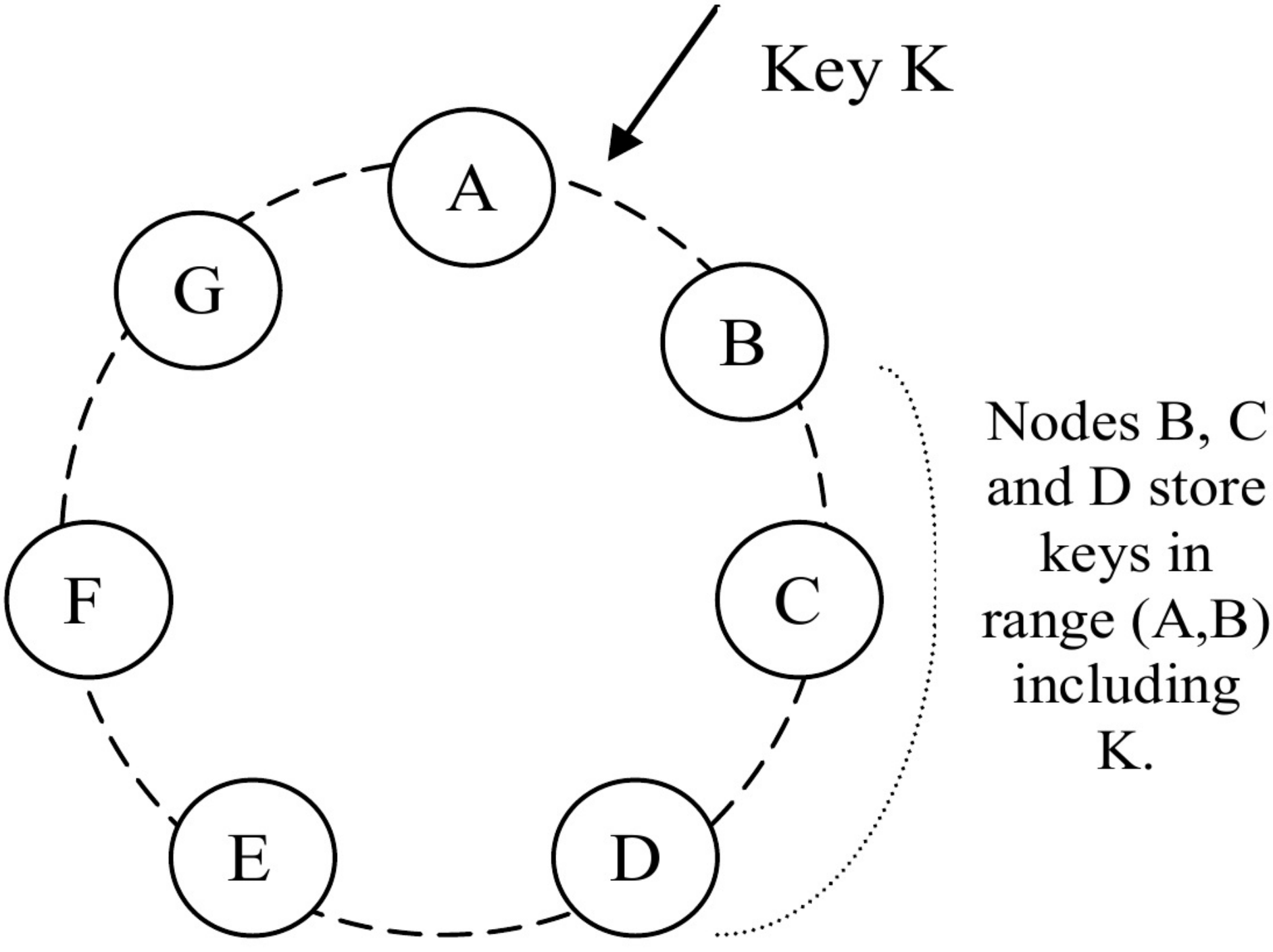
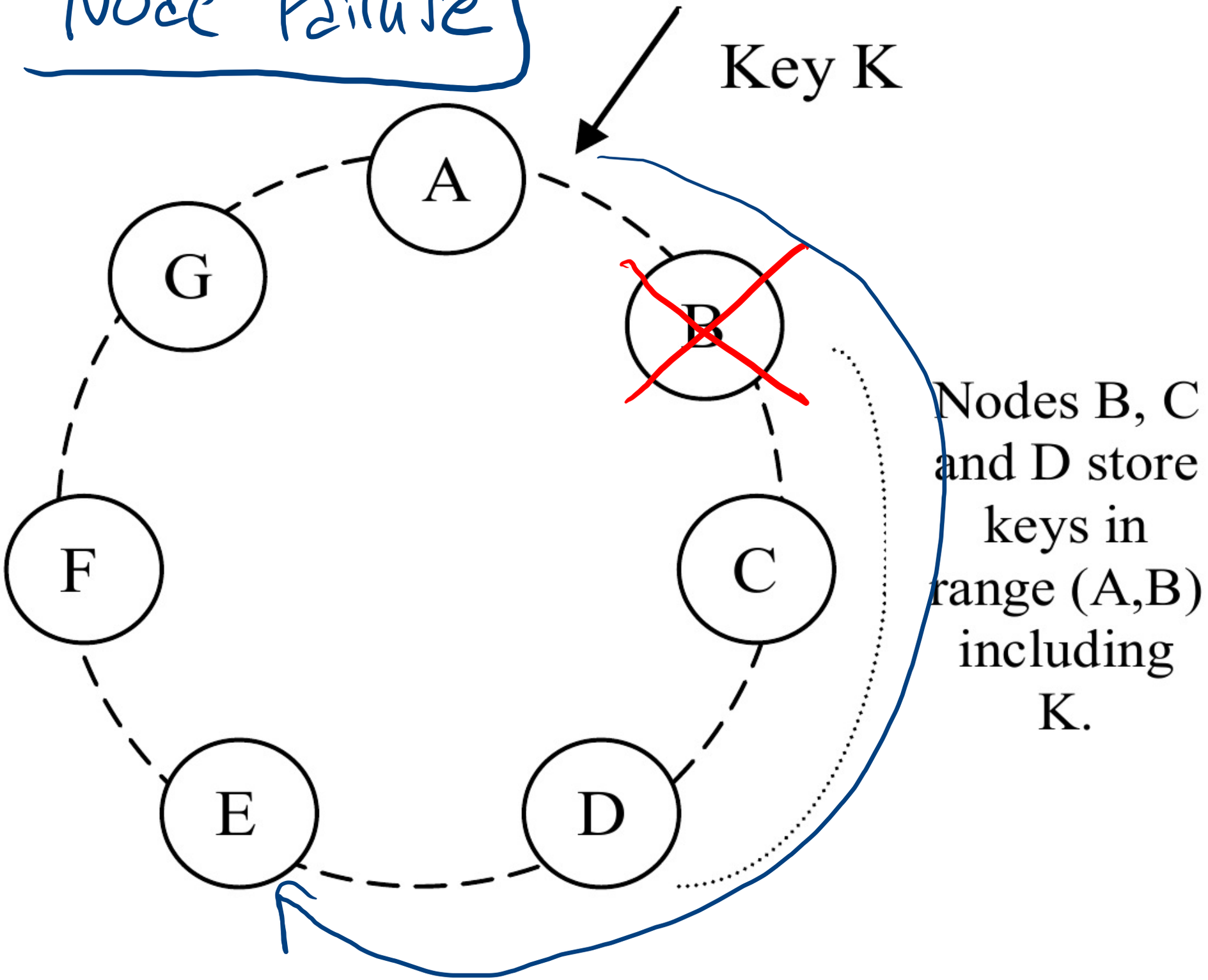

Key K
A
B
C
D
E
F
G
Nodes B, C and D store keys in range (A,B) including K.

Table 2: Performance of client-driven and server-driven coordination approaches.

	99.9th percentile read latency (ms)	99.9th percentile write latency (ms)	Average read latency (ms)	Average write latency (ms)
Server-driven	68.9	68.5	3.9	4.02
Client-driven	30.4	30.4	1.55	1.9

Node Failures



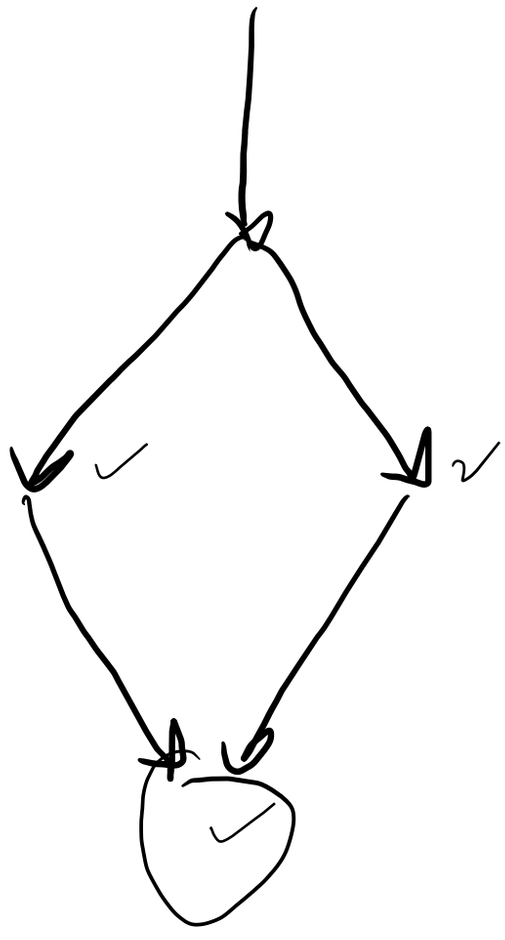
Conflicts

Network partition

Concurrent client updates

Consistency

Syntactic if



Vector clock

$V_1 \langle A\text{-}1, B\text{-}2 \rangle$

$V_2 \langle A\text{-}1, B\text{-}3 \rangle$

$V_1 \leq V_2$ iff $\forall s \; V_1[s] \leq V_2[s]$

⟨A-1, B-1⟩

⟨A-2, B-1⟩ ⟨A-1, B-2⟩

keep timestamps on V.C. entries
truncate at 10 entries (discard oldest)

⟨A-1, B-1, C-1⟩ ⟨~~A-1~~, B-1, C-1, D-1⟩