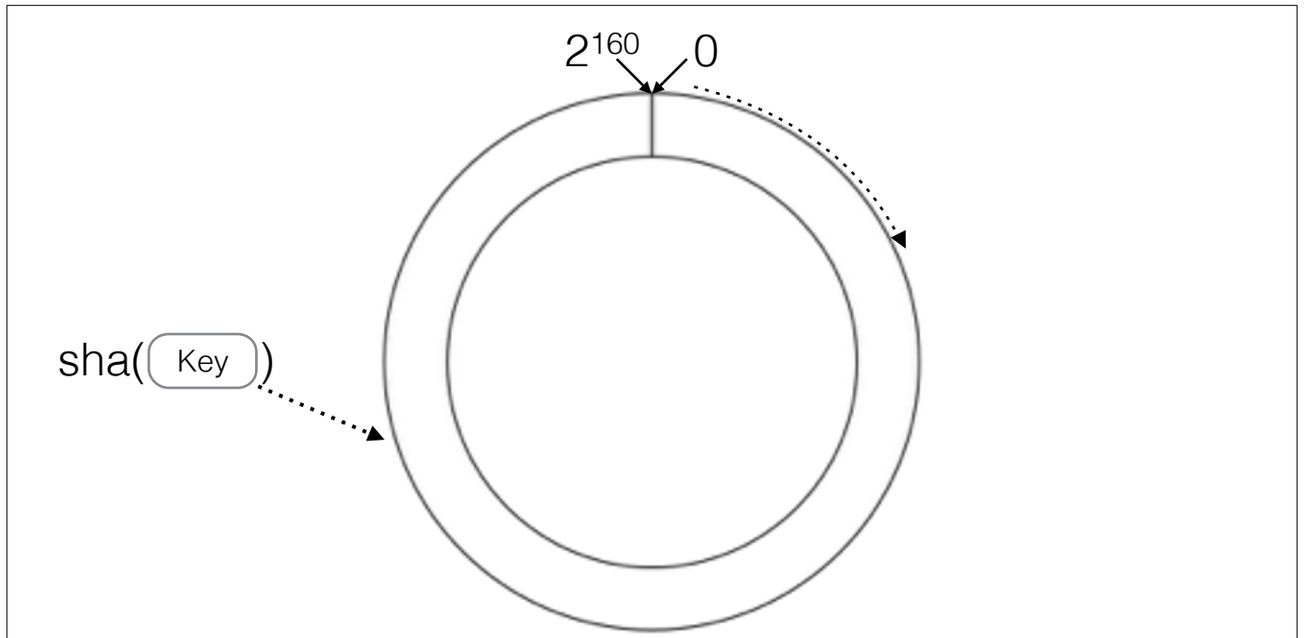
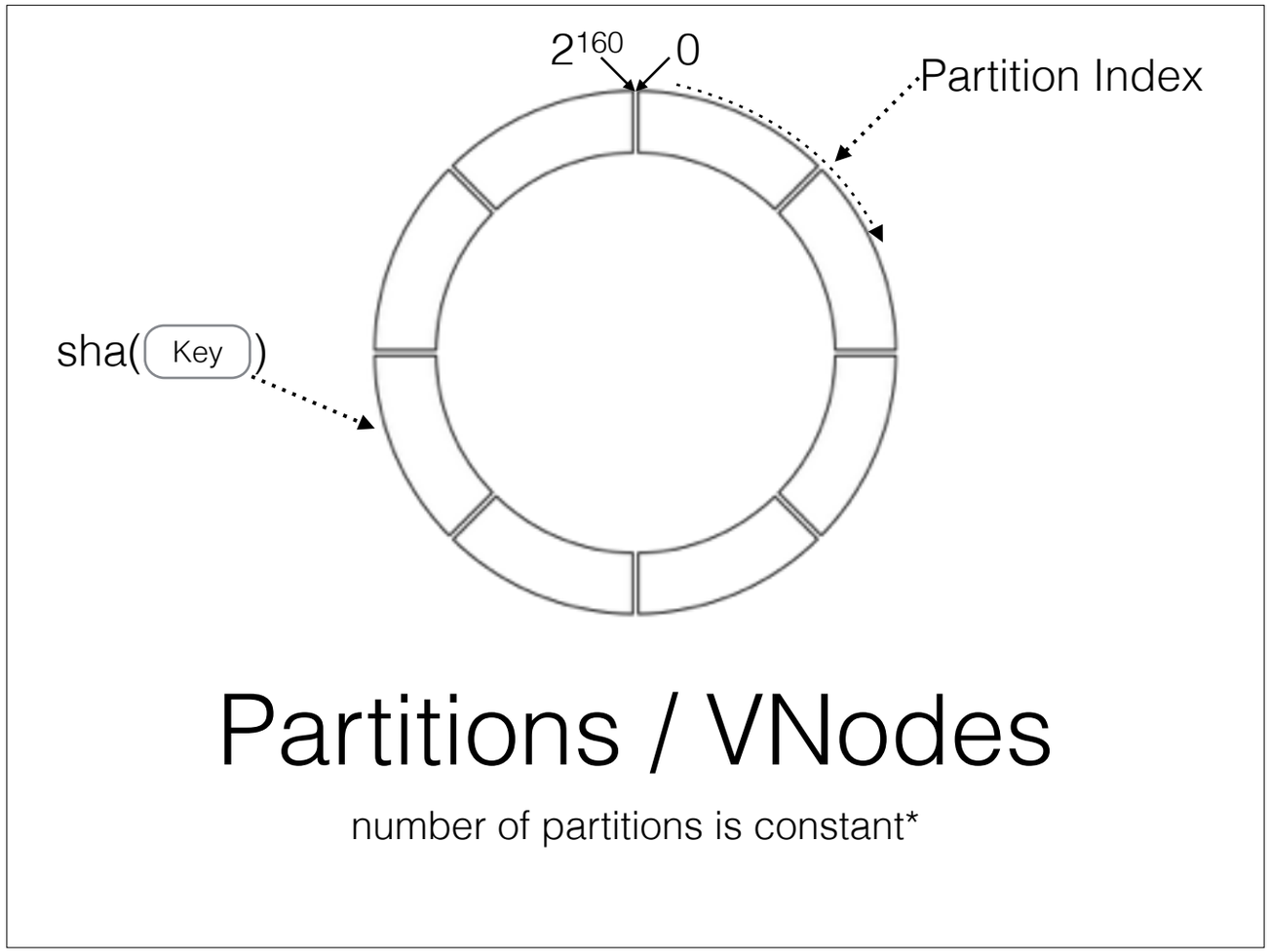
# riak_core

Concepts and Misconceptions




# The Hash Ring

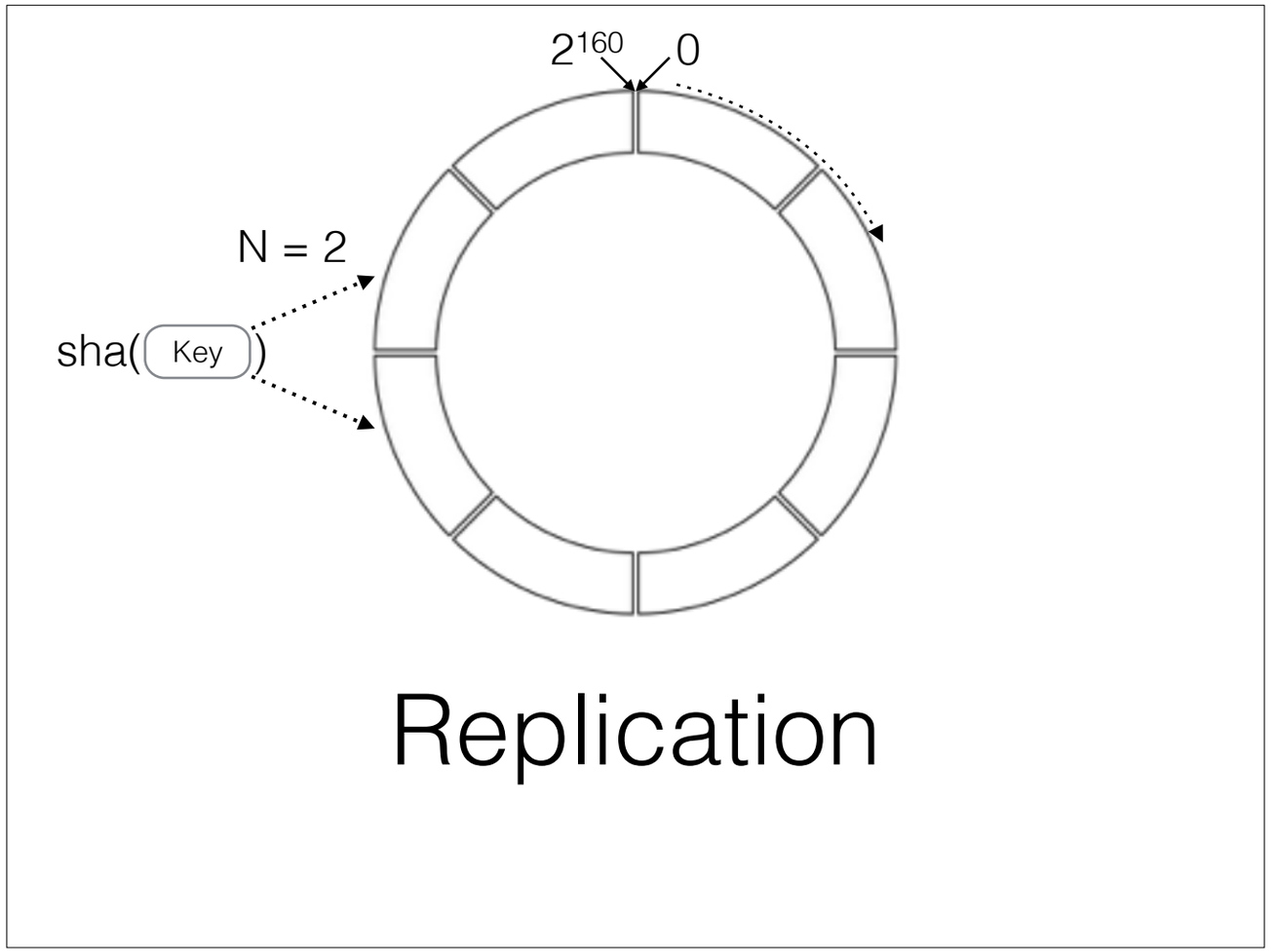
to rule them all




# Partitions / VNodes

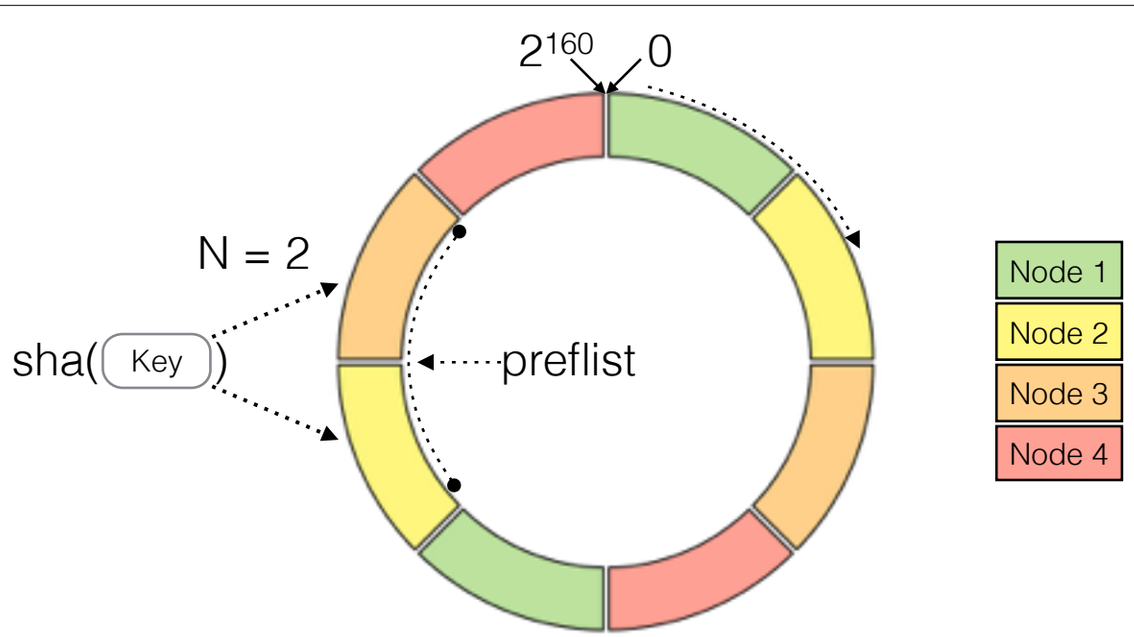
number of partitions is constant*

vnode is a unit of migration/replication/sharding



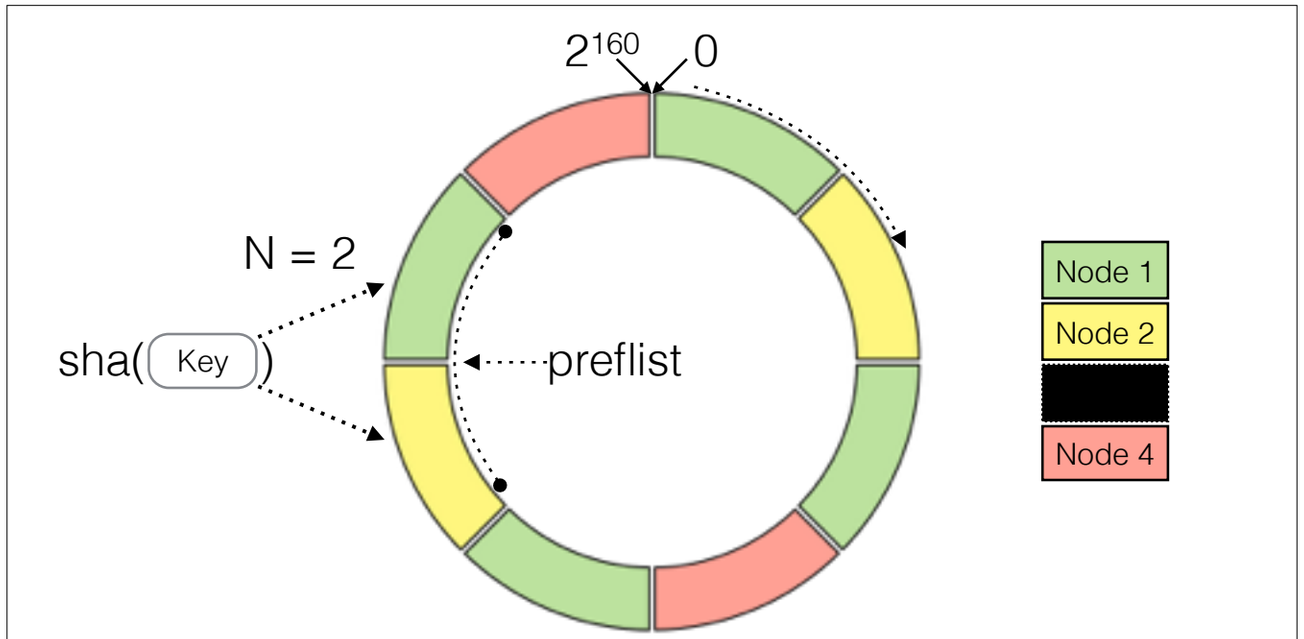

# Replication

N is usually 3, but for the sake of example…




# VNode / Node

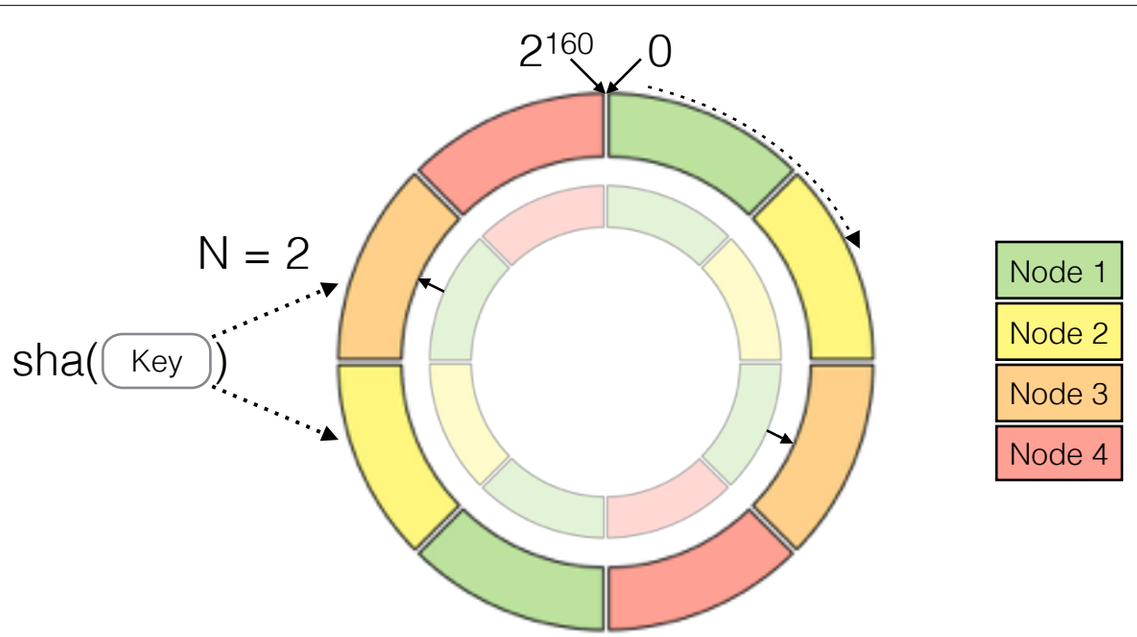
ownership, preflists




# Fallbacks

`{Index, OwnerNode}` ⇒ `{Index, FbNode}`

Misconception #1 fallback is never performed by another partition, it goes to another node but keeps the index.




# Handoff

hinted / ownership / resize

for some time fallback and primary coexist

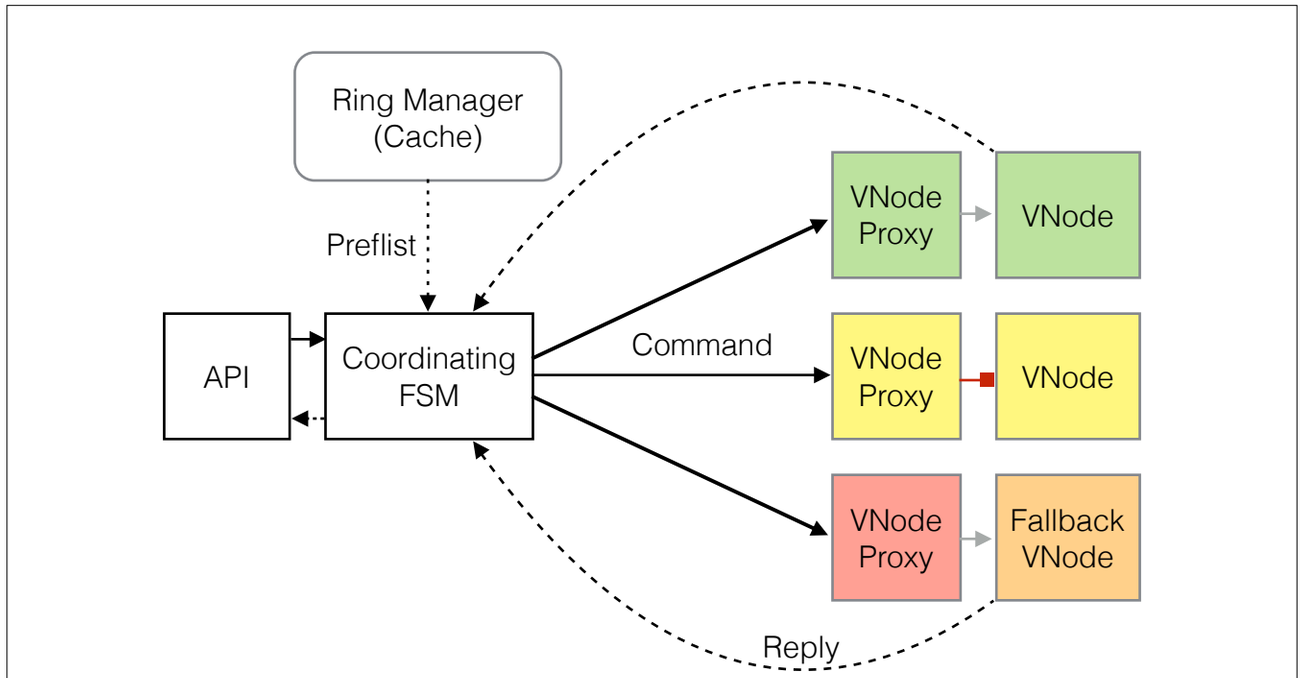
# Handoff

- Hinted
  - A fallback vnode is idle for a period of time and the primary, is up again.
- Ownership
  - A ring update event for a ring that all other nodes have already seen (by vclock).
- Resize

# Gossip & The Claimant

- Every node sends known cluster state to some other node once in a while
  - Cluster state eventually converges
- One node is assigned the task of coordinating the addition and removal of nodes
  - If this node goes down, cluster changes will not be possible until a new one is chosen or the claimant node is marked as down

Misconception #2 new ‘cluster view’ is formed by the claimant, it is not ad-hoc




# The Request

or There and Back Again

Yellow Proxy performs backpressure for the overwhelmed VNode

# How to converge

- On the previous slide the yellow node will never see the command
- Read-repair
- Anti-entropy
  - hashtrees

# Coverage

- Enumerate thru the dataset
- FSM perspective
  - Set coverage over replicas
    - contact as few nodes as possible
    - each possible key gets at least one node
- VNode perspective
  - FOLD command
  - may be used separately

# Cluster metadata

- cluster-global KV storage
- bucket types
- backed by epidemic broadcast

# The Rest

- Other subsystems
  - capabilities
  - epidemic broadcast
  - security
  - cluster info & monitoring facilities
- Algorithms and building blocks
  - vclocks
  - dvv
  - bloom filter
  - hashtrees
  - ...

# Further reading

- http://marianoguerra.github.io/presentations/riak-core-small-bytes-berlin-efl-2014.html
- https://github.com/marianoguerra/flaviodb
- https://github.com/basho/riak_core/wiki
- https://github.com/basho/riak_core/blob/1.4.0/docs/ring-resizing.md
- https://github.com/mrallen1/udon