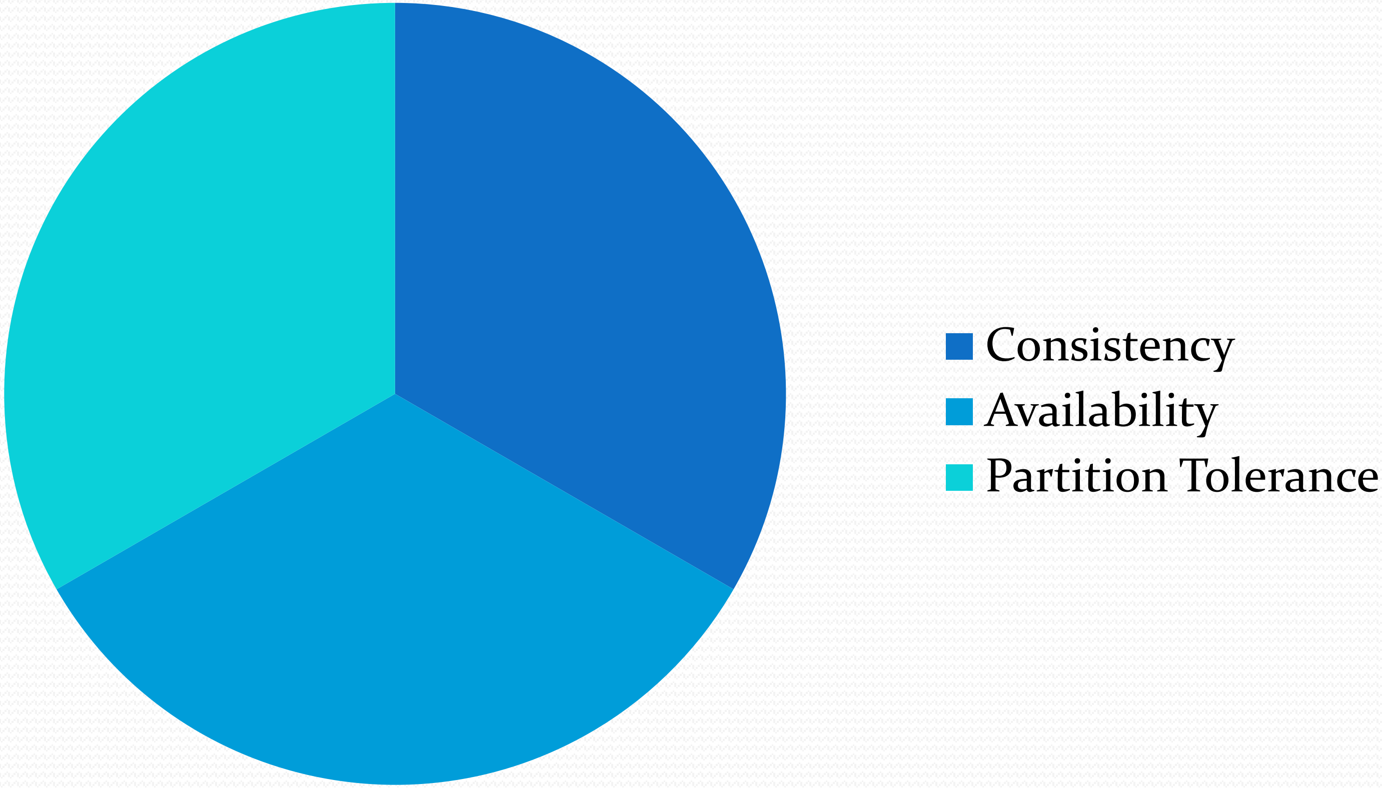
MySQL分布式集群高可用设计及应用

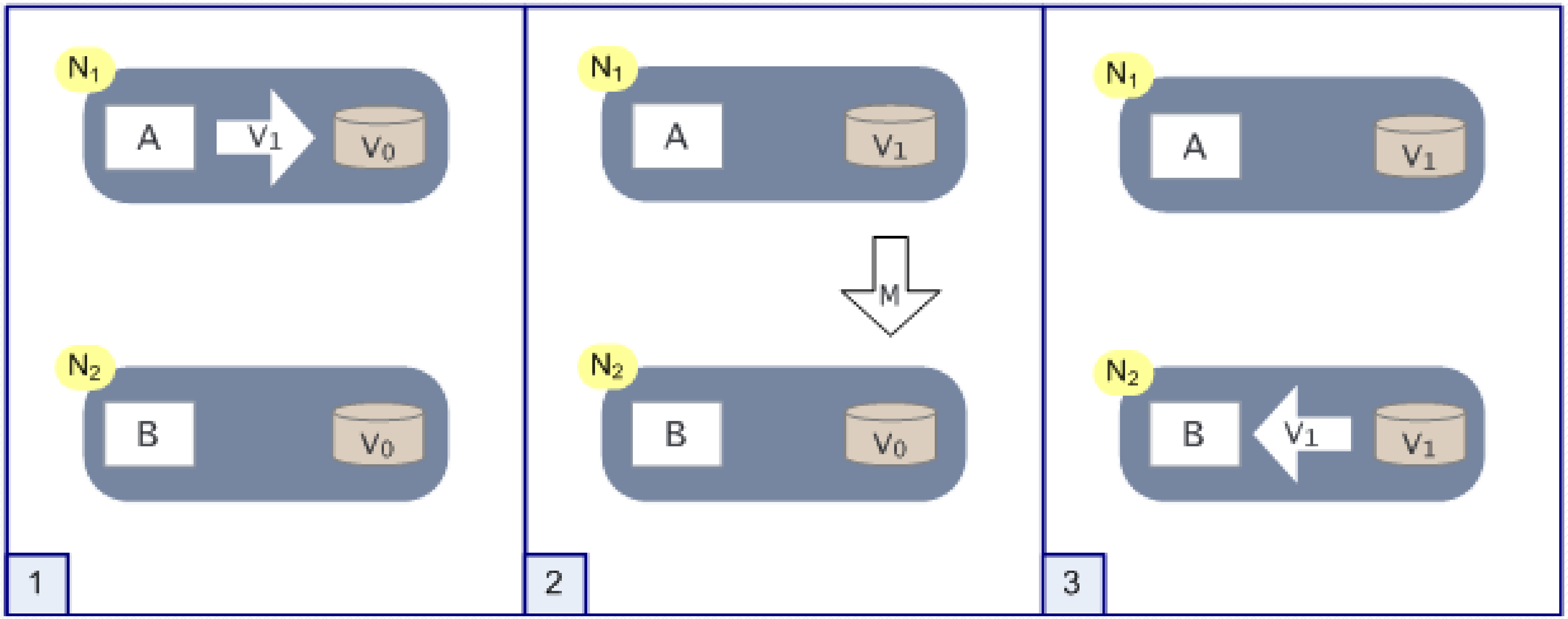
谭俊青@MySQL实验室
(http://www.mysqlab.net)
上海爱可生信息技术有限公司

2010-04-03

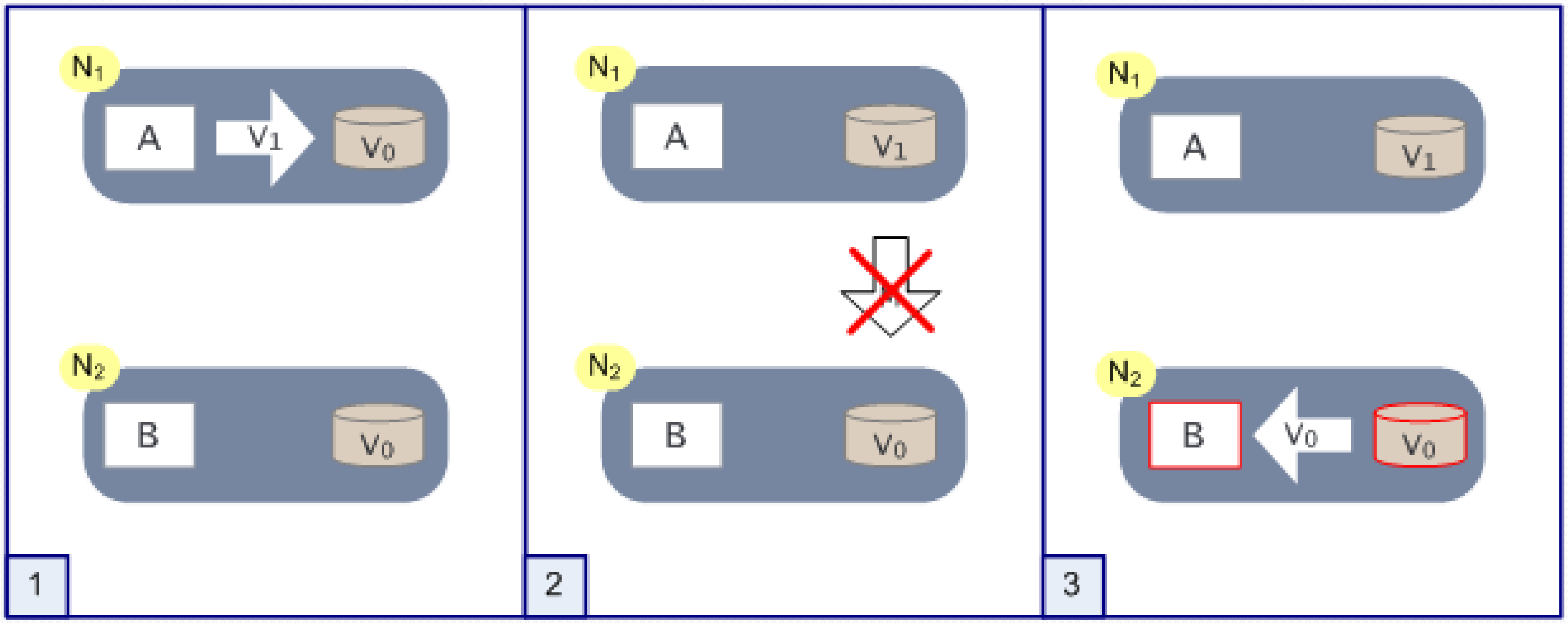
CAP Theorem and NoSQL



CAP Theorem (2/3)



CAP Theorem (3/3)

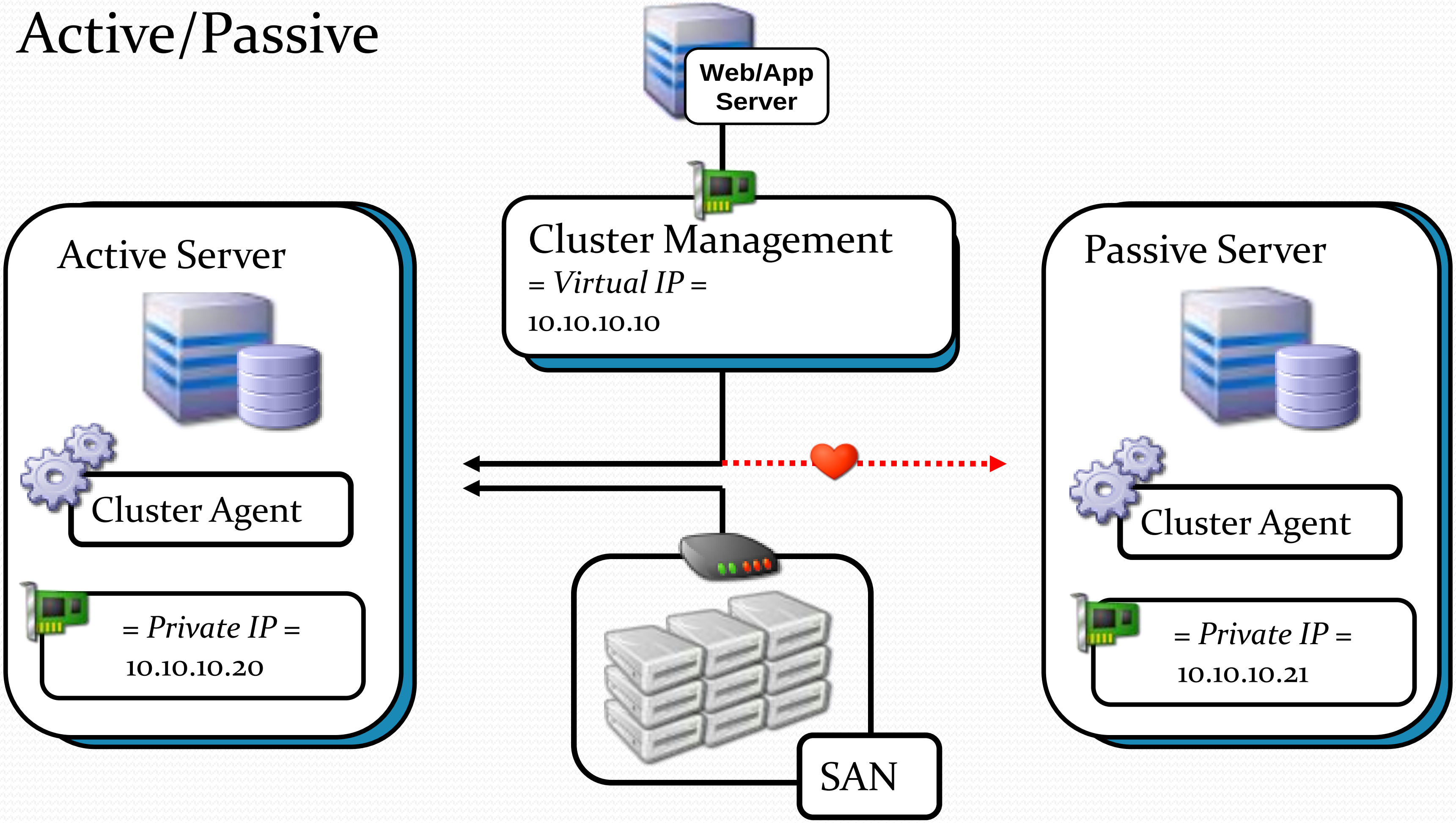


MySQL HA Solution

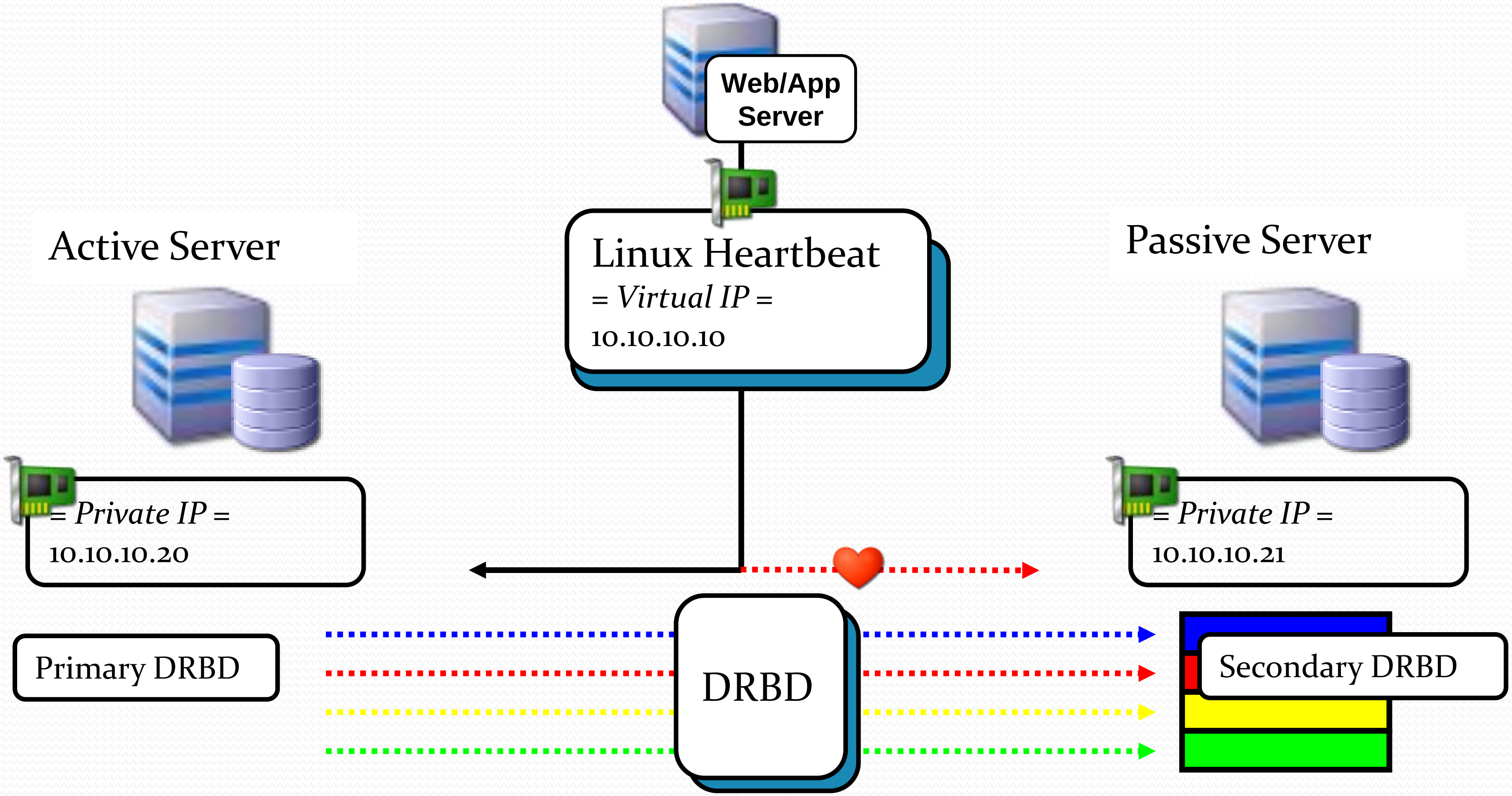
- MySQL + Shared-Storage
- MySQL + DRBD (CP)
- Master + Slave (AP)
- Master + Slave(SemiSyncReplication) (CP/AP)
- Multi-Master (AP)
- MySQL Cluster (CAP? CP/AP)

MySQL + Shared-Storage

- Active/Passive

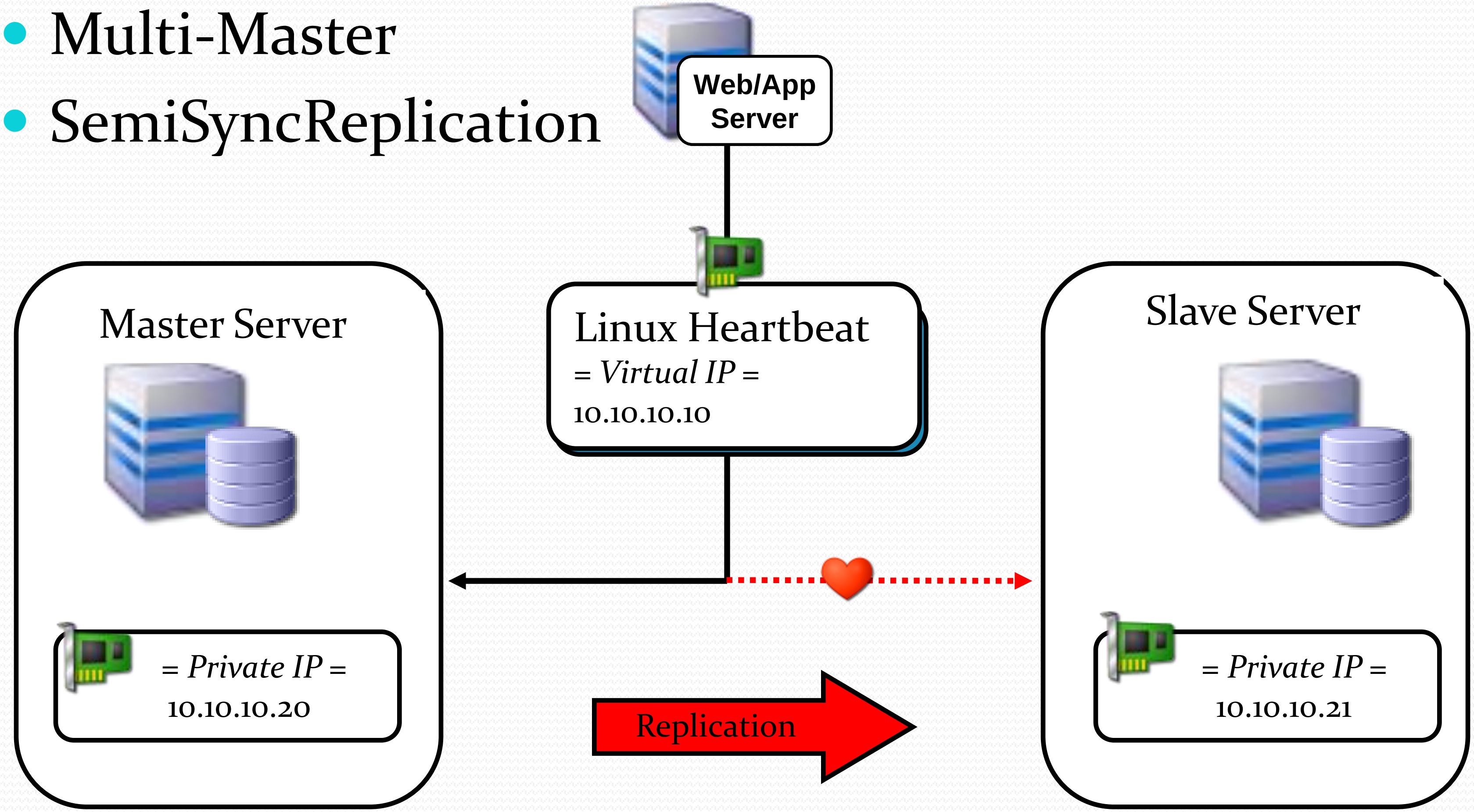


MySQL + DRBD

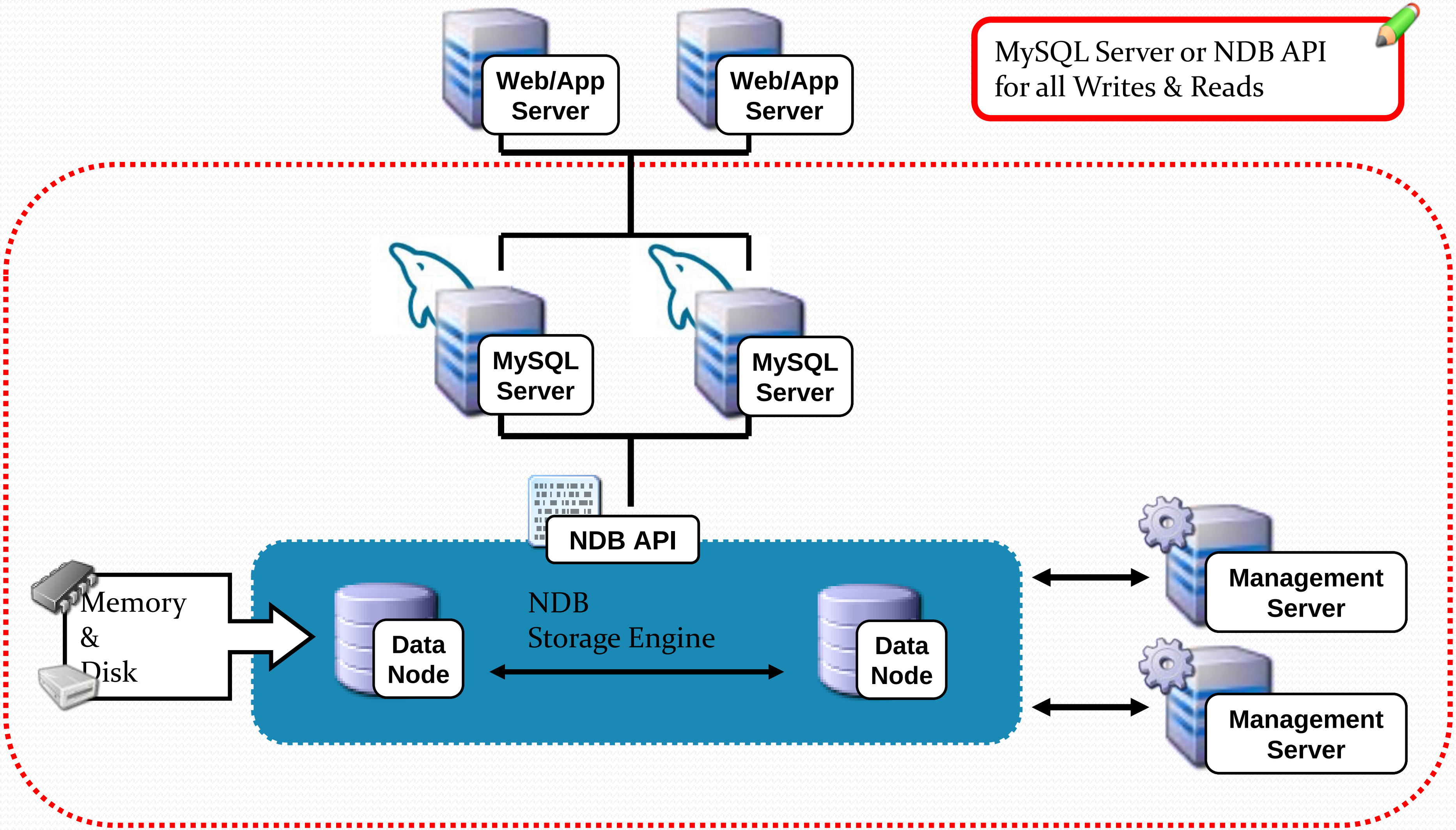

Web/App Server
Linux Heartbeat
= Virtual IP =
10.10.10.10
Active Server
= Private IP =
10.10.10.20
Passive Server
= Private IP =
10.10.10.21
Primary DRBD
DRBD
Secondary DRBD

Master + Slave

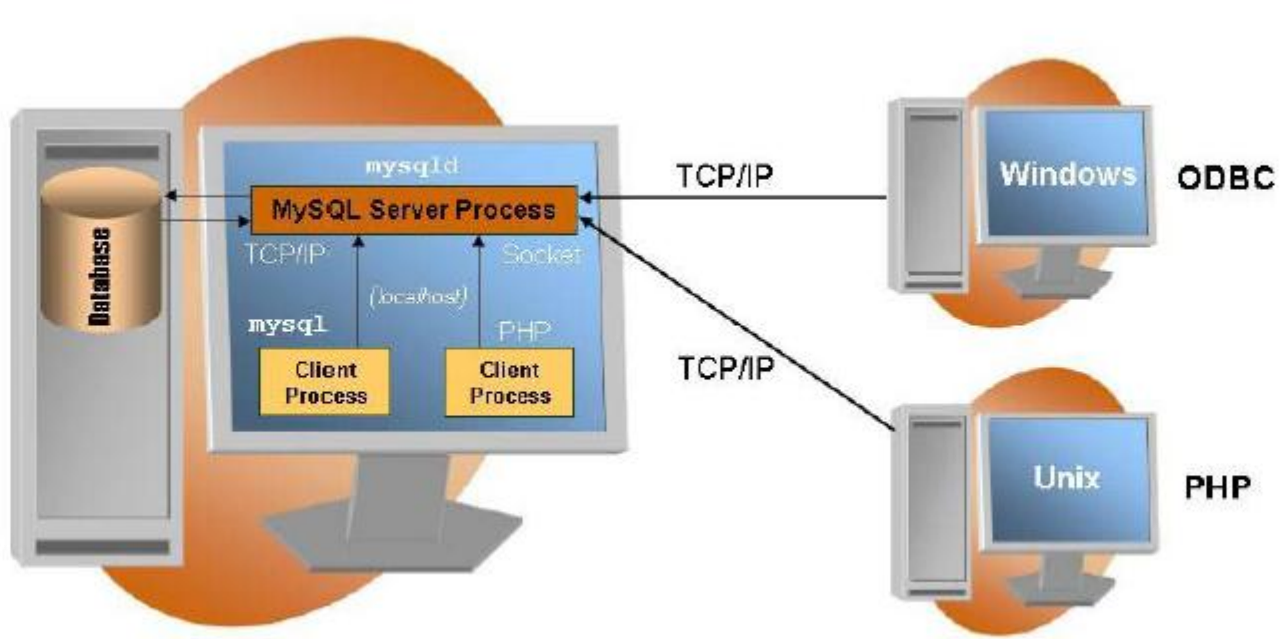
- Multi-Master
- SemiSyncReplication



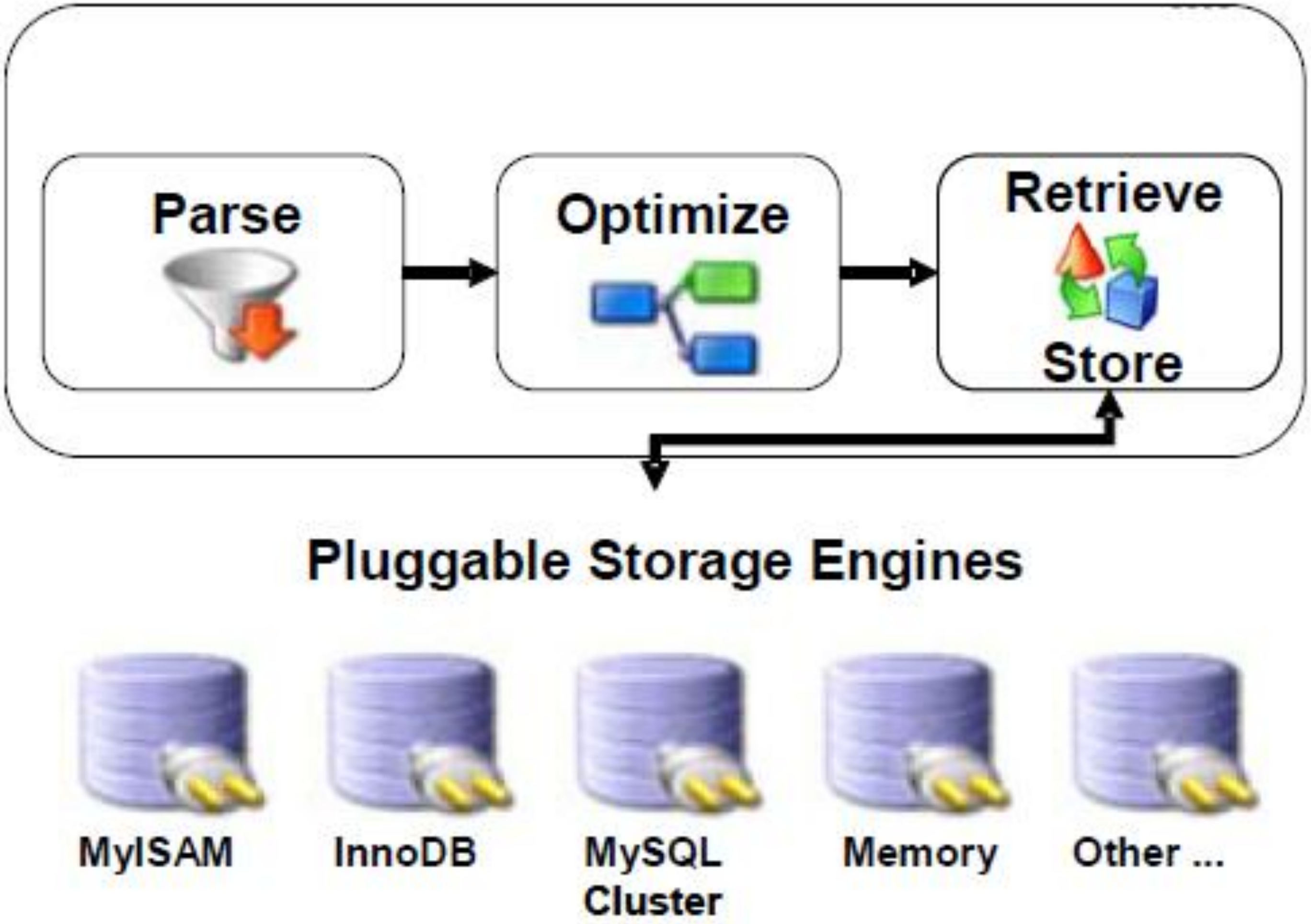
MySQL Cluster



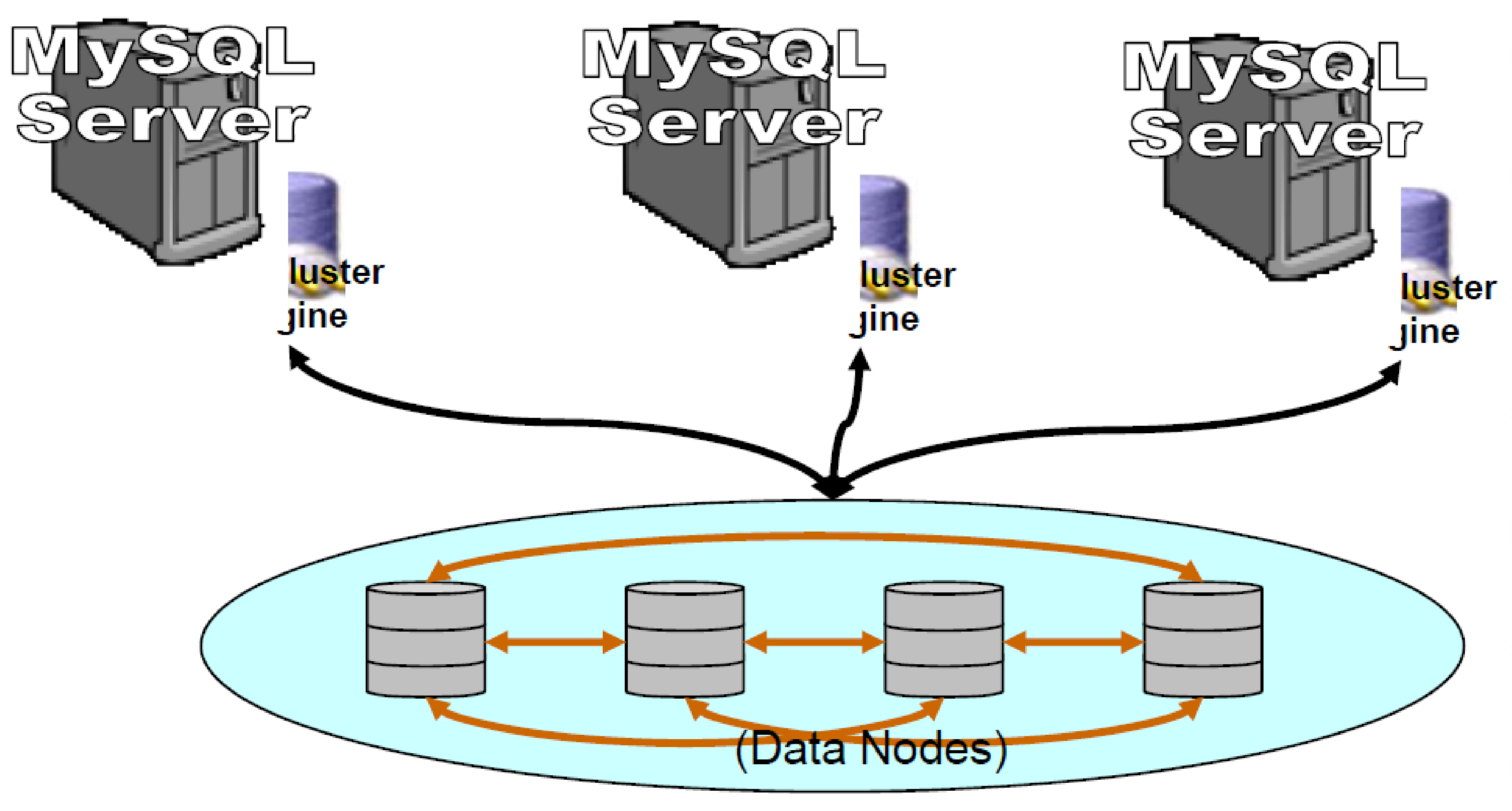
MySQL General Architecture (1/3)



MySQL General Architecture (2/3)



MySQL General Architecture (3/3)



What is a Cluster

- Shared-nothing vs Shared-disk

  - In a shared-nothing , which MySQL Cluster is, each node has its own complete set of hardware

  - In a shared-disk architecture, there is a central storage location that all of the nodes will access and make use of

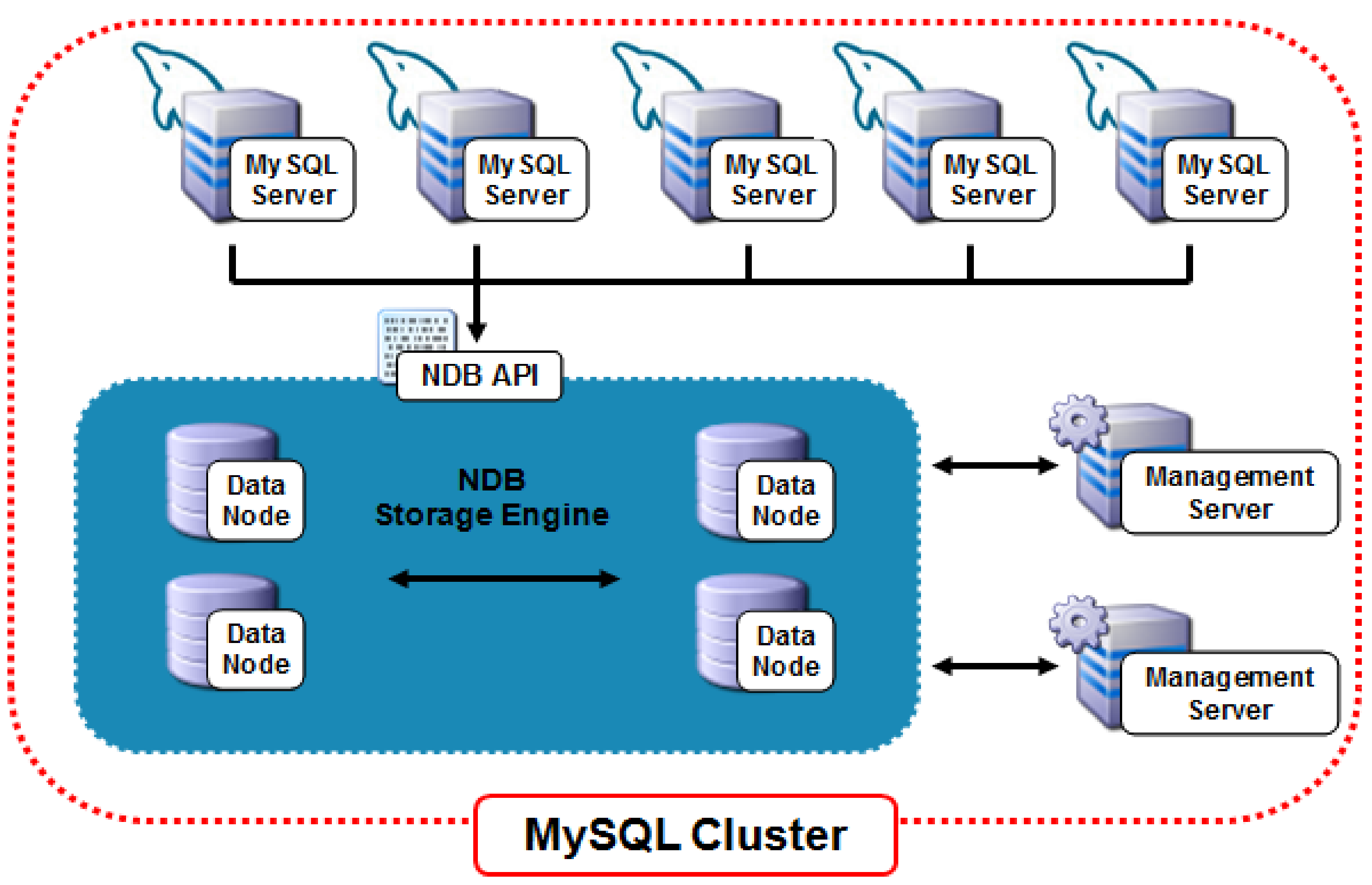
- MySQL Cluster Hardware

  - MySQL Cluster does not require and special hardware, such as SAN or NAS

  - Each node can run on commodity type hardware

  - Designed to allow many maintenance operations to be completed in a online fashion

MySQL Cluster Architecture



Node Types

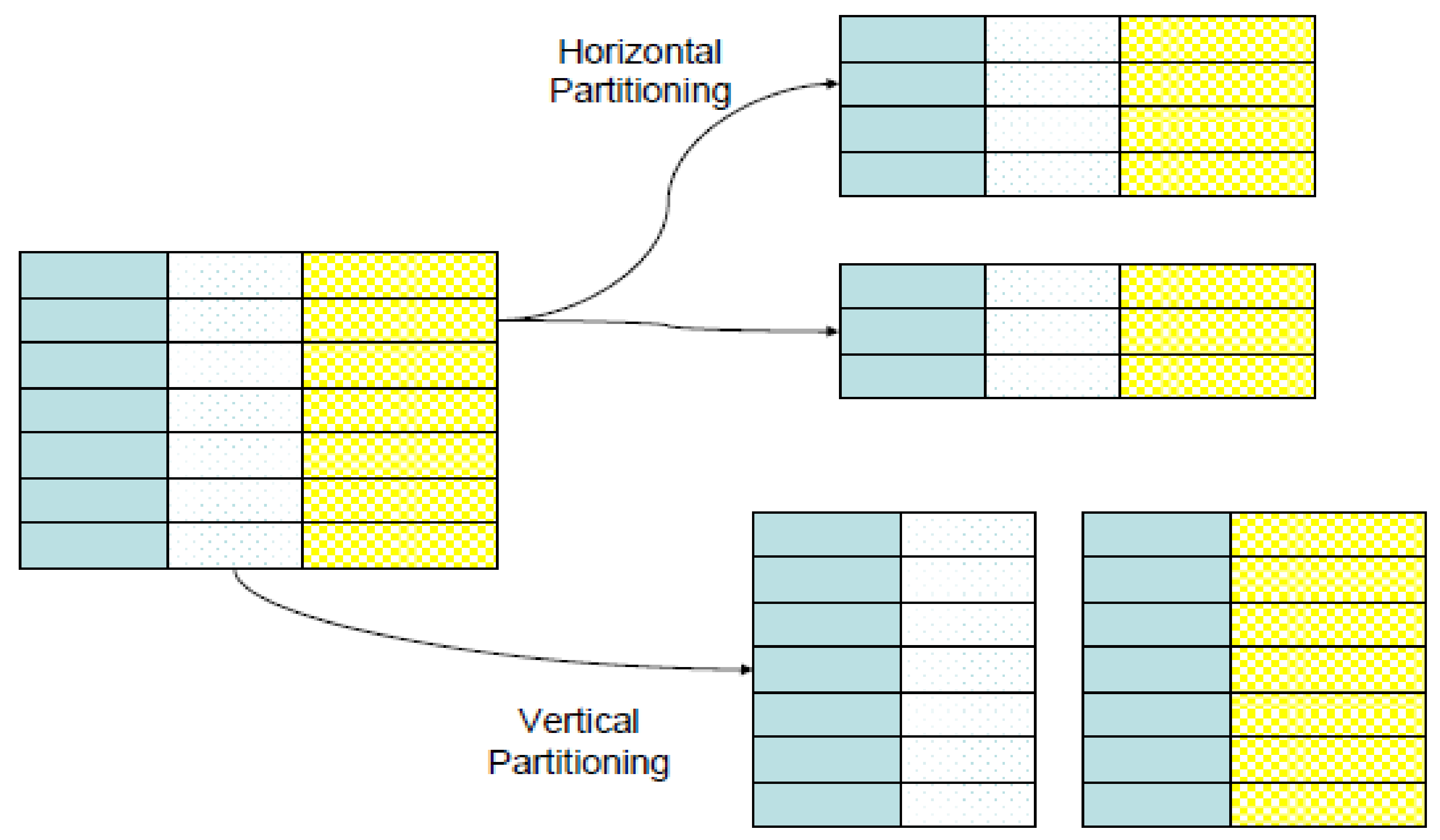
- Data Nodes

  -- Storage Nodes
- API Nodes

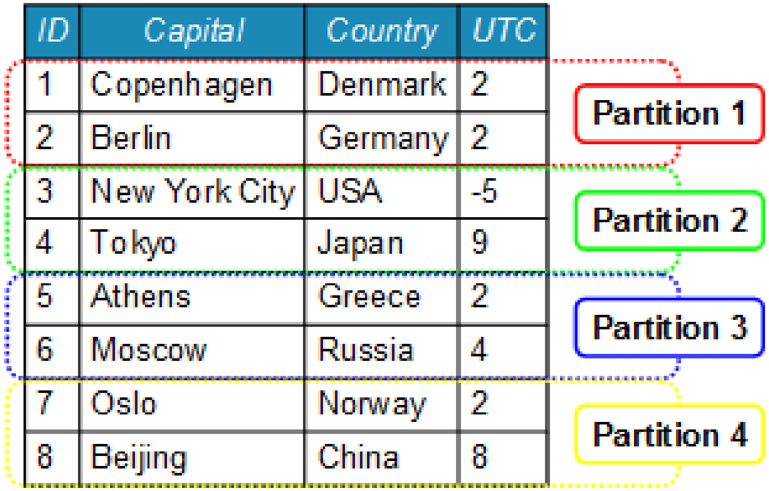
  -- Mediators between the end process
  and the data nodes
- Management Node

  -- Manages the configuration and control
  of the MySQL Cluster

Partition

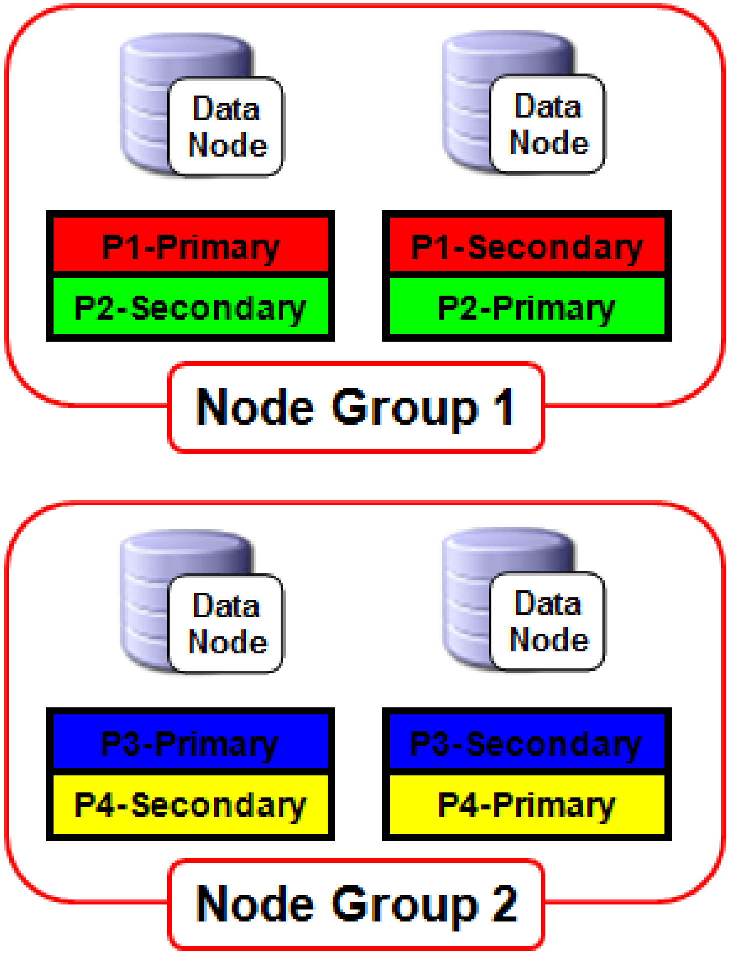


MySQL Cluster Architecture



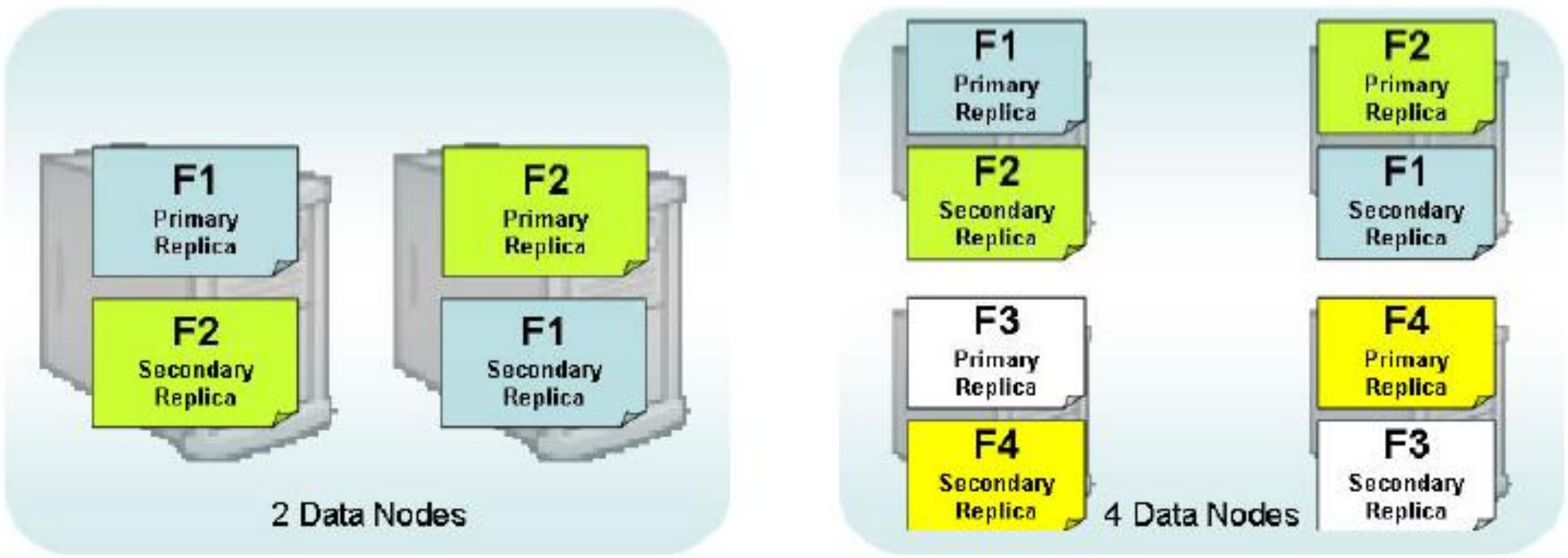
ID	Capital	Country	UTC
1	Copenhagen	Denmark	2
2	Berlin	Germany	2
3	New York City	USA	-5
4	Tokyo	Japan	9
5	Athens	Greece	2
6	Moscow	Russia	4
7	Oslo	Norway	2
8	Beijing	China	8

Partition 1
Partition 2
Partition 3
Partition 4

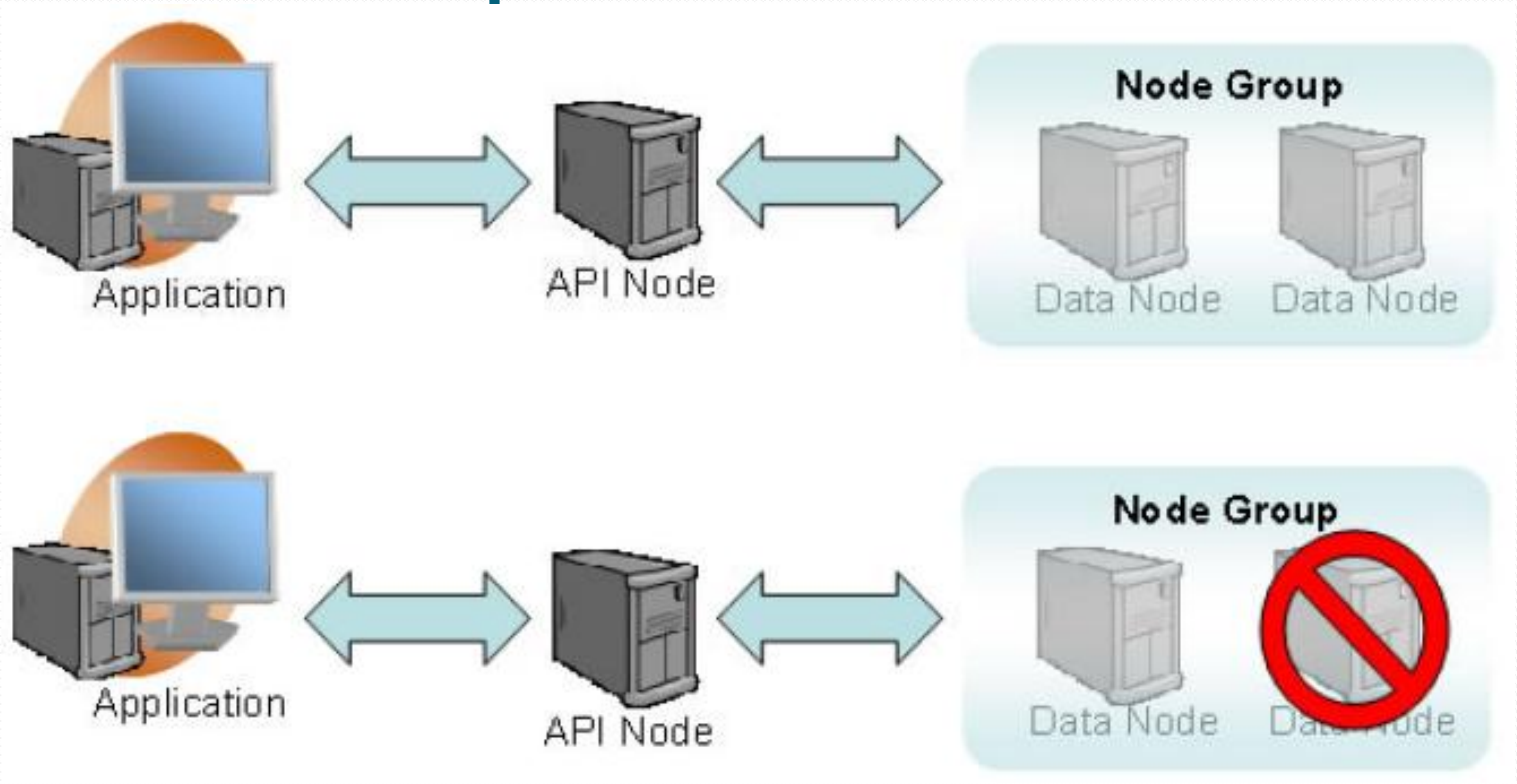


- Four Data Nodes
- Two Replicas
- Two Node Groups

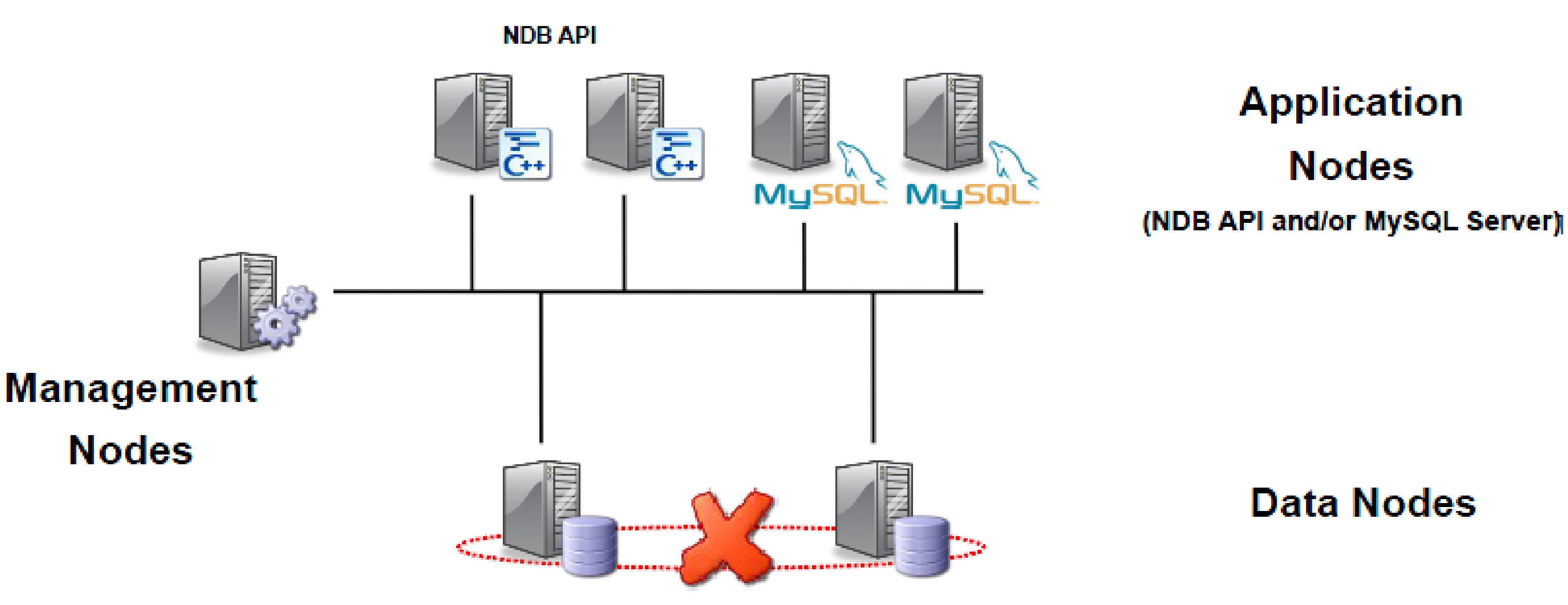
Fragments


F1
Primary
Replica
F2
Secondary
Replica
F2
Primary
Replica
F1
Secondary
Replica
2 Data Nodes
F1
Primary
Replica
F2
Secondary
Replica
F2
Primary
Replica
F1
Secondary
Replica
F3
Primary
Replica
F4
Secondary
Replica
F4
Primary
Replica
F3
Secondary
Replica
4 Data Nodes

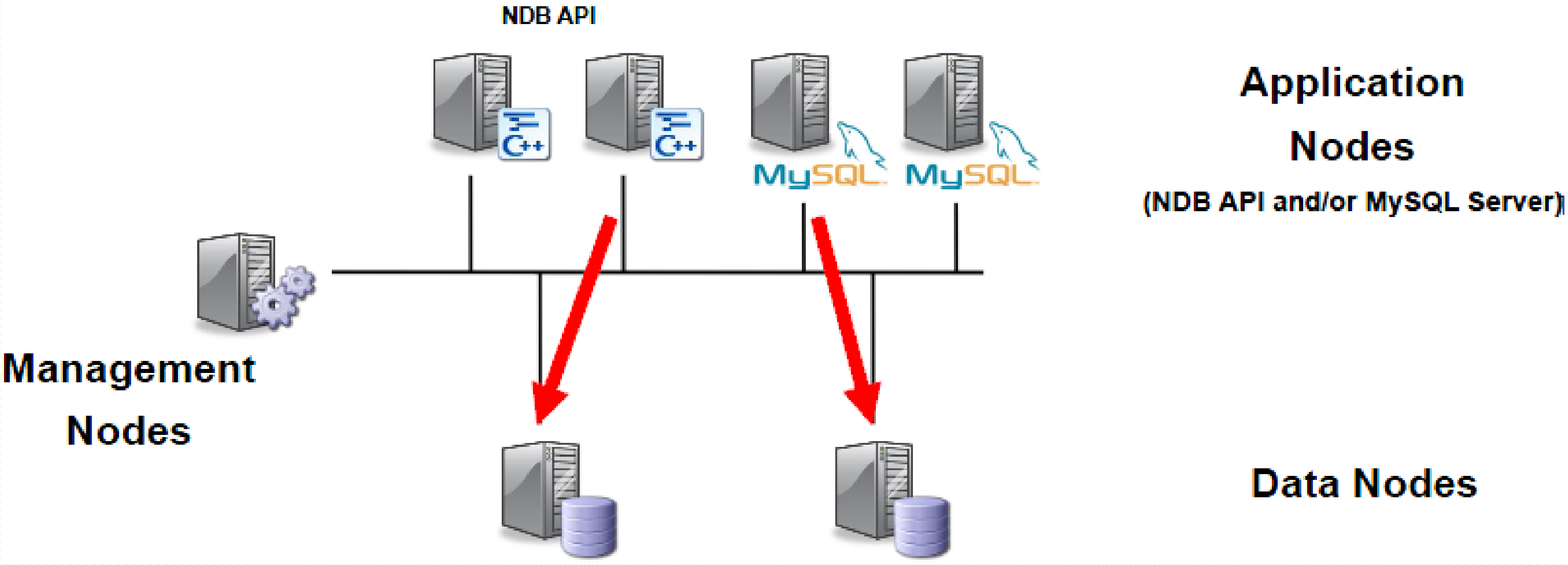
Node Groups


Application
API Node
Node Group
Data Node
Data Node
Application
API Node
Node Group
Data Node
Data Node

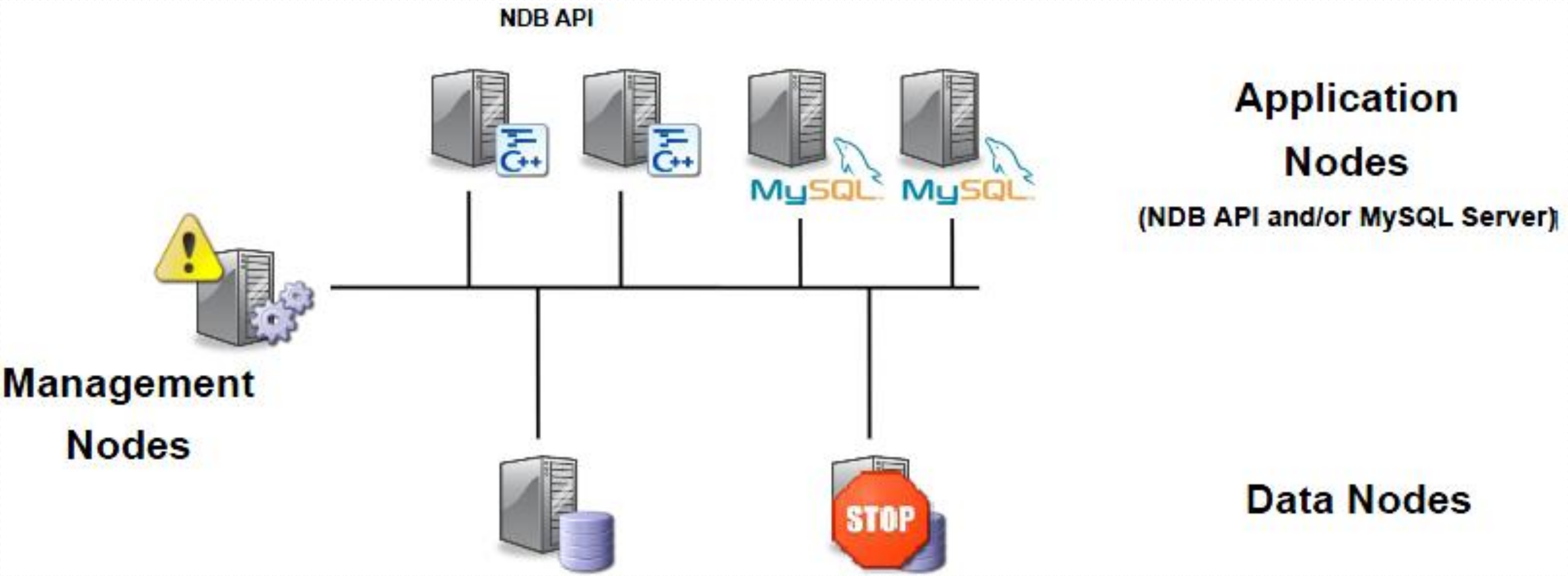
Split Brain (1/3)



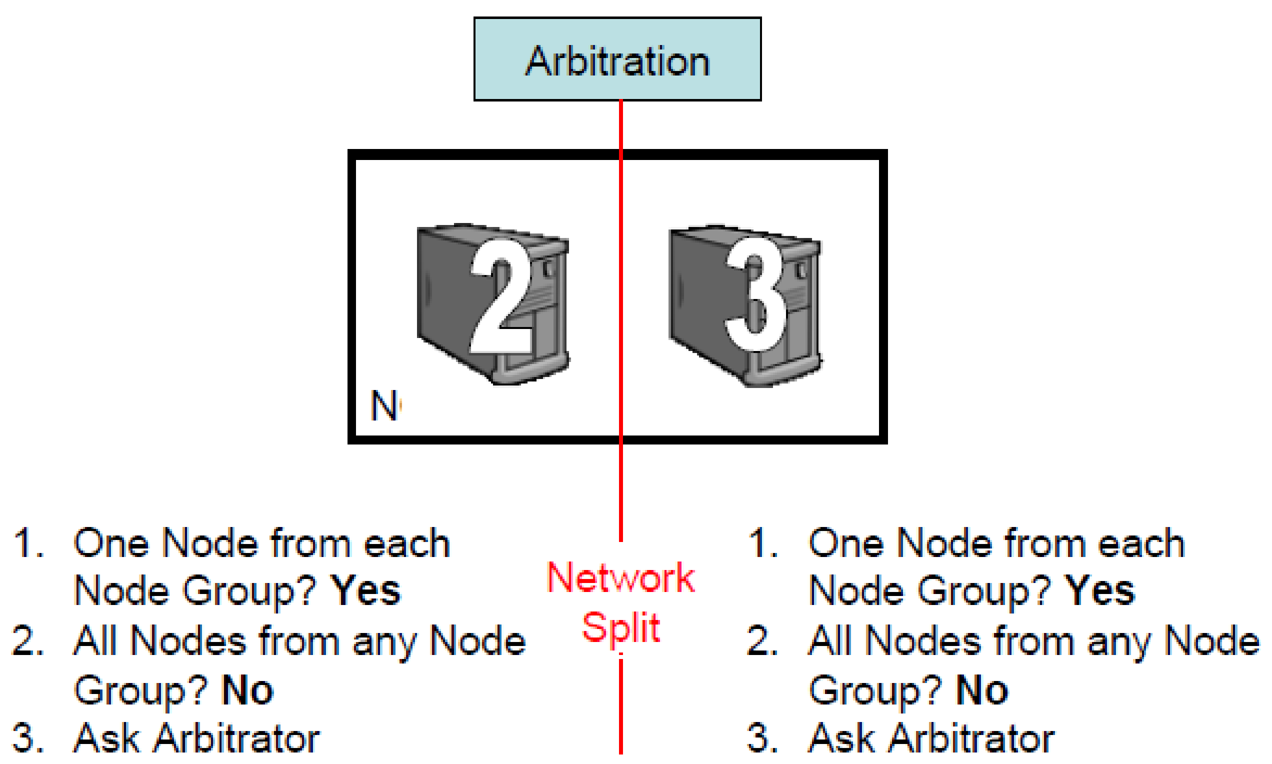
Split Brain (2/3)



Split Brain (2/3)



Arbitration

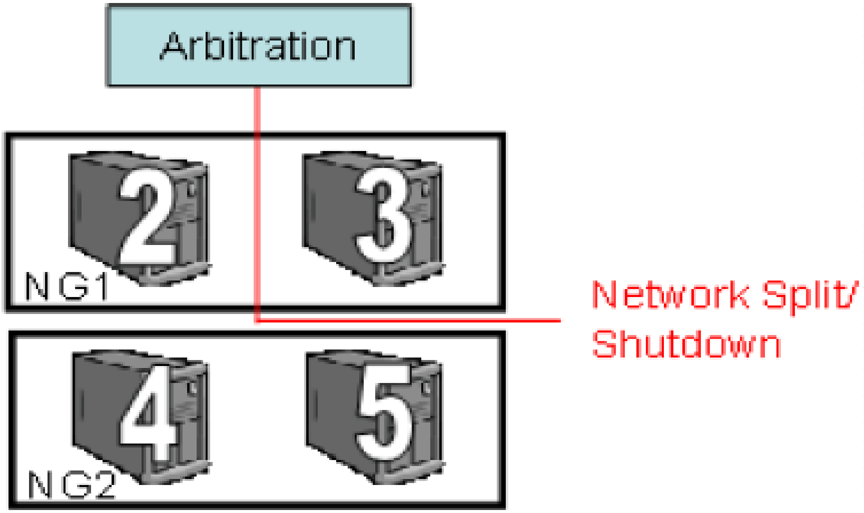


1. One Node from each Node Group? **Yes**
2. All Nodes from any Node Group? **No**
3. Ask Arbitrator

1. One Node from each Node Group? **Yes**
2. All Nodes from any Node Group? **No**
3. Ask Arbitrator

First Node to ask will continue while the other will be shut down

More Data Nodes



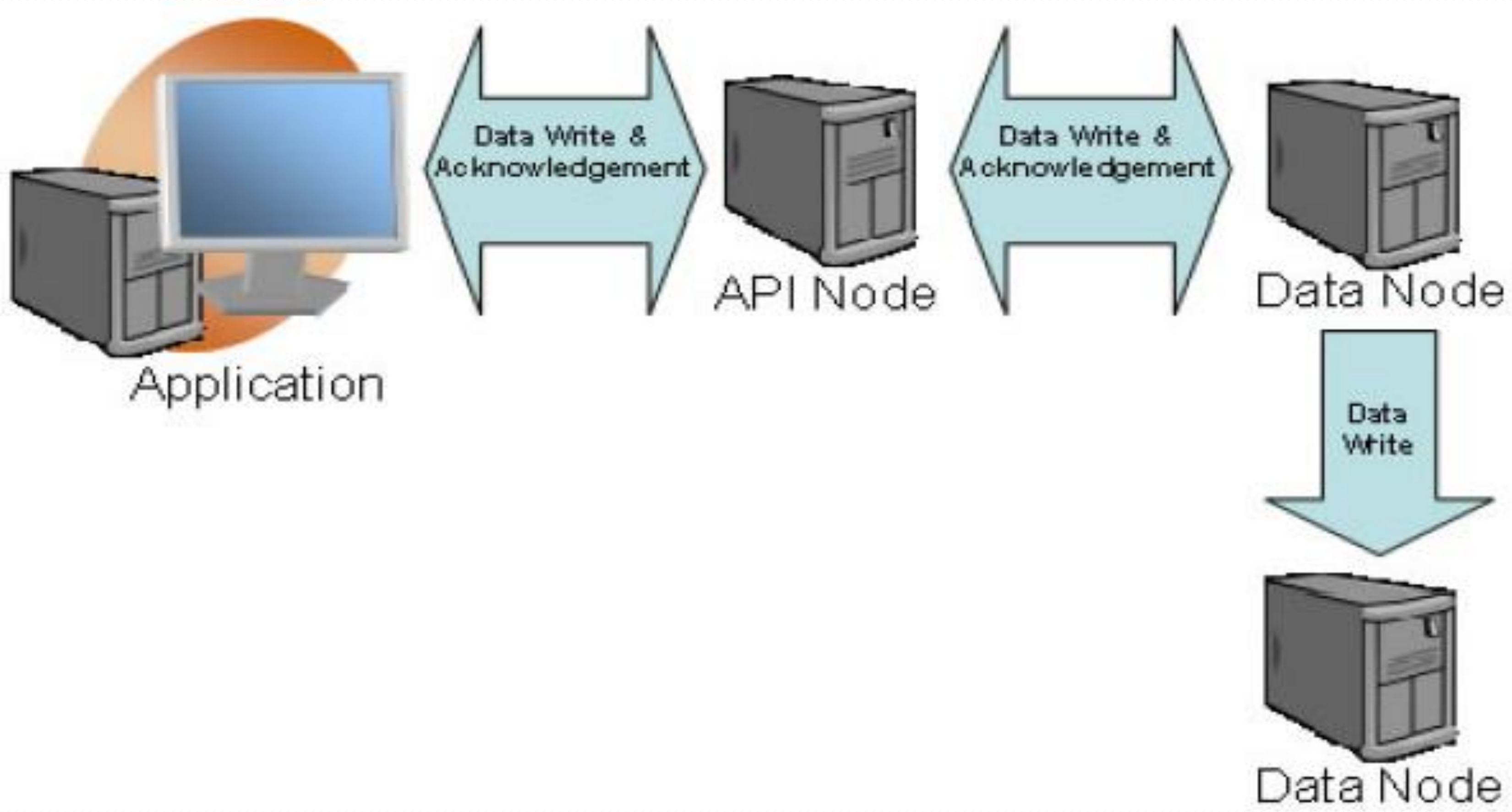
Nodes 2,4&5

1. One Node from each Node Group? **Yes**
2. All Nodes from any Node Group? **Yes**
3. Continue as the Cluster, no Arbitrator is needed

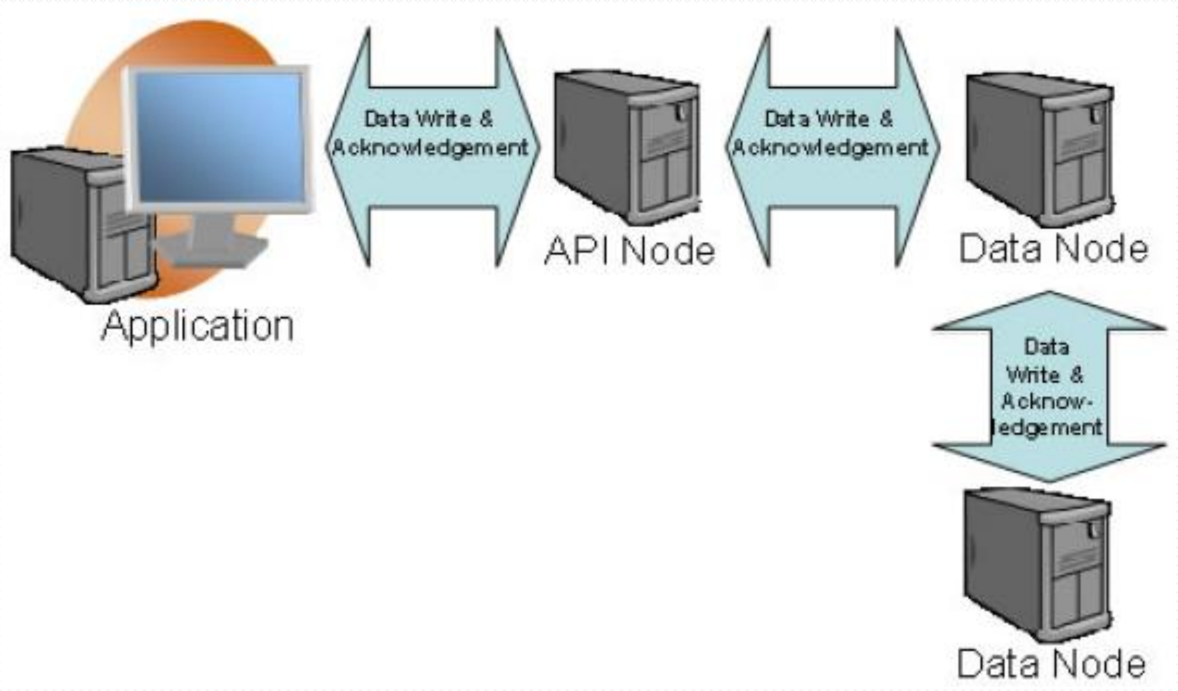
Node 3

1. One Node from each Node Group? **No**
2. Shut Down

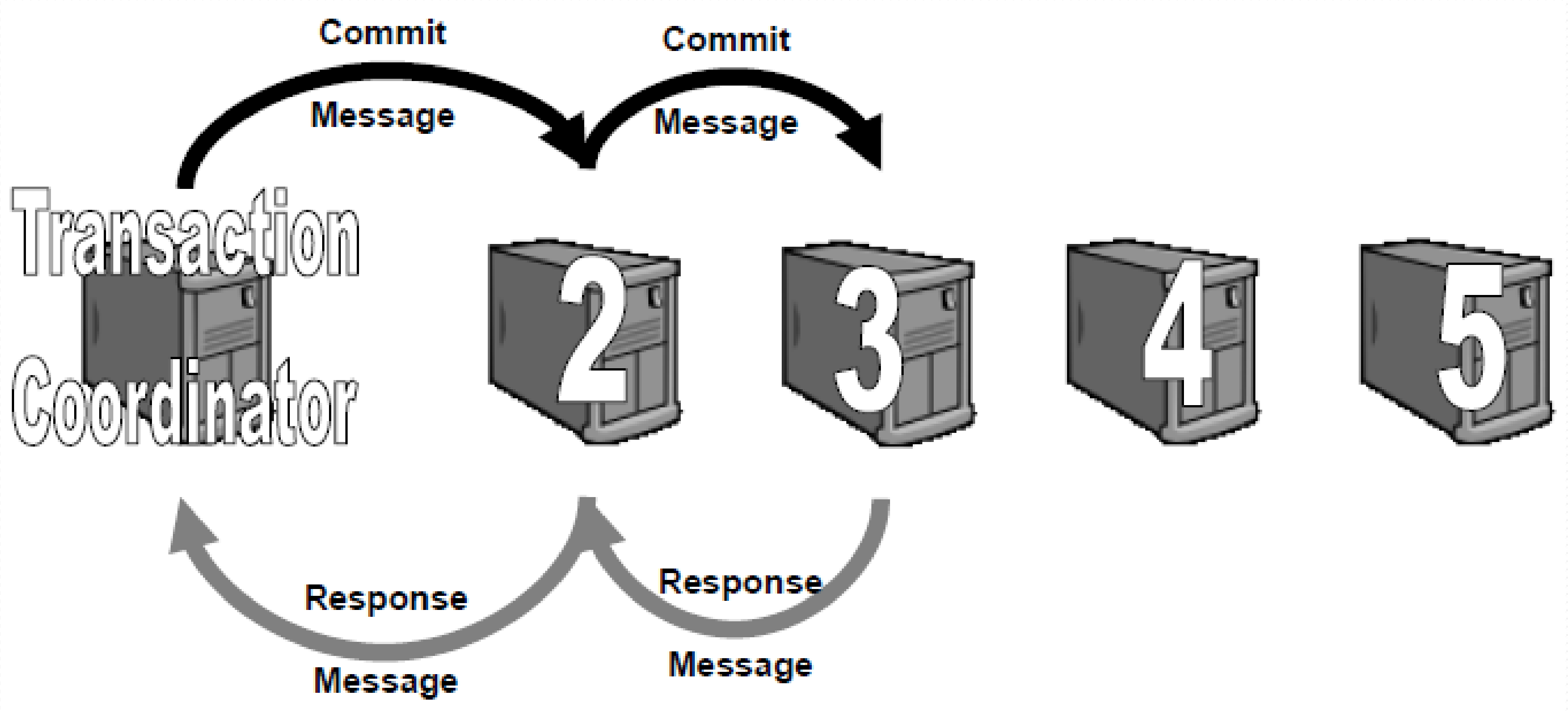
Asynchronous (AP of CAP)



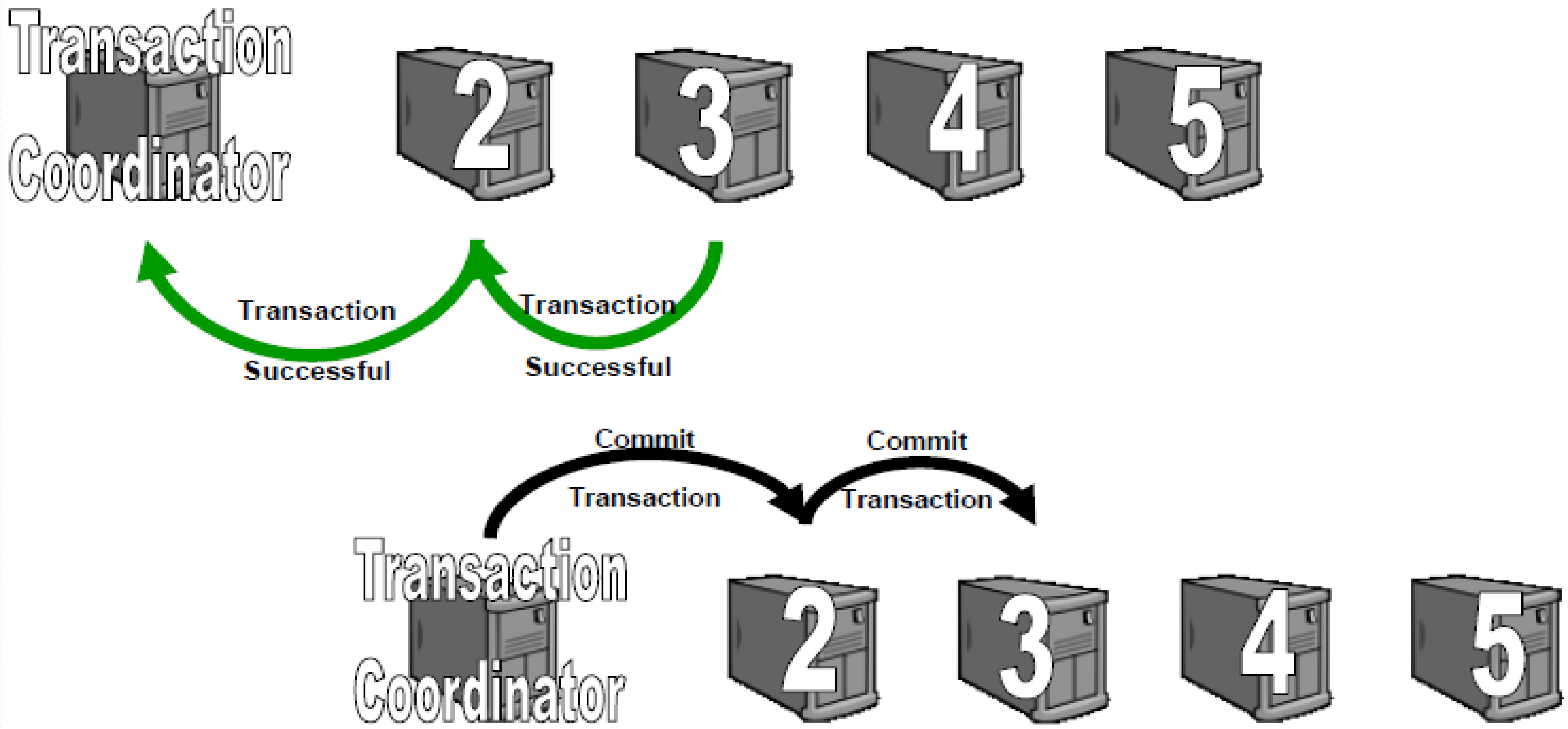
Synchronous (CP of CAP)


Application
Data Write & Acknowledgement
API Node
Data Write & Acknowledgement
Data Node
Data Write & Acknow-ledgement
Data Node

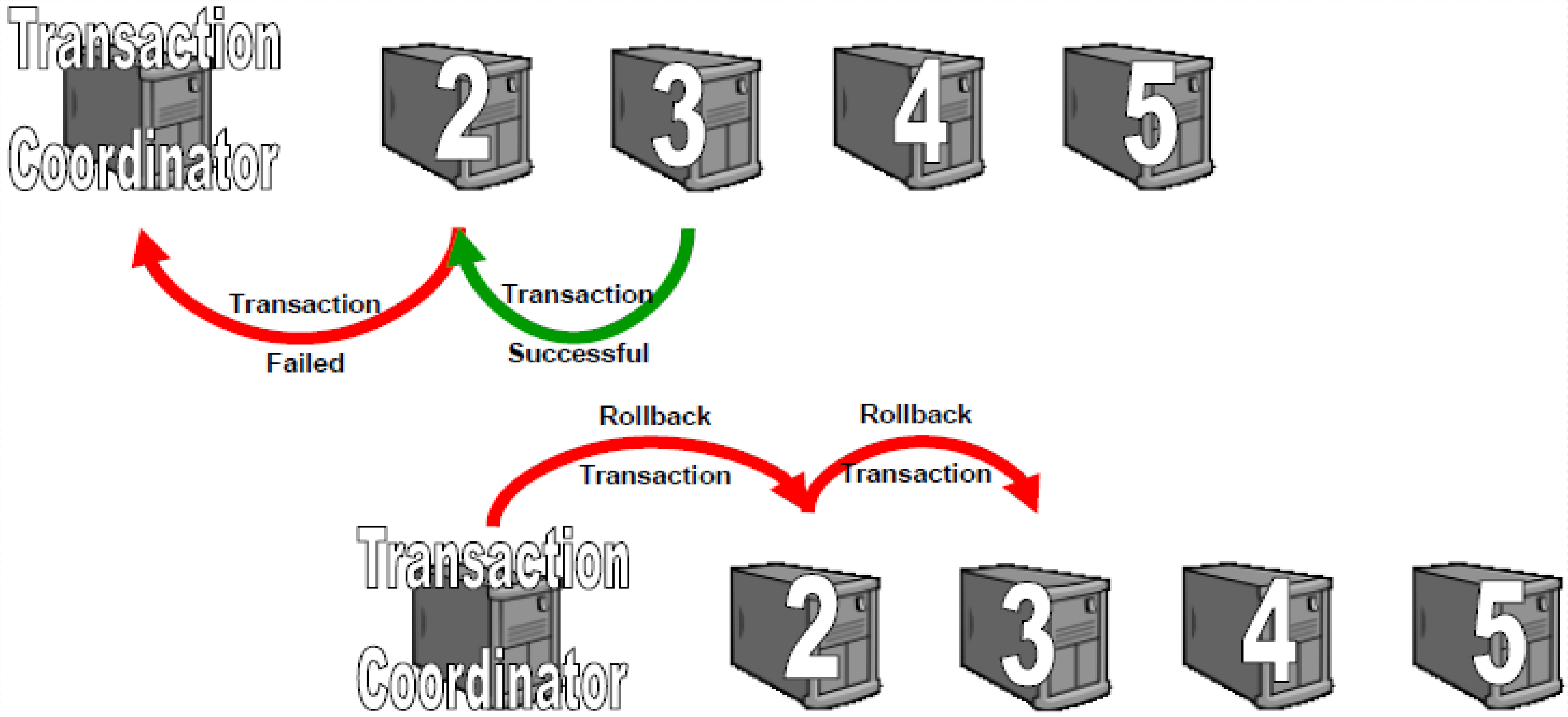
Phase One: Commit-Request

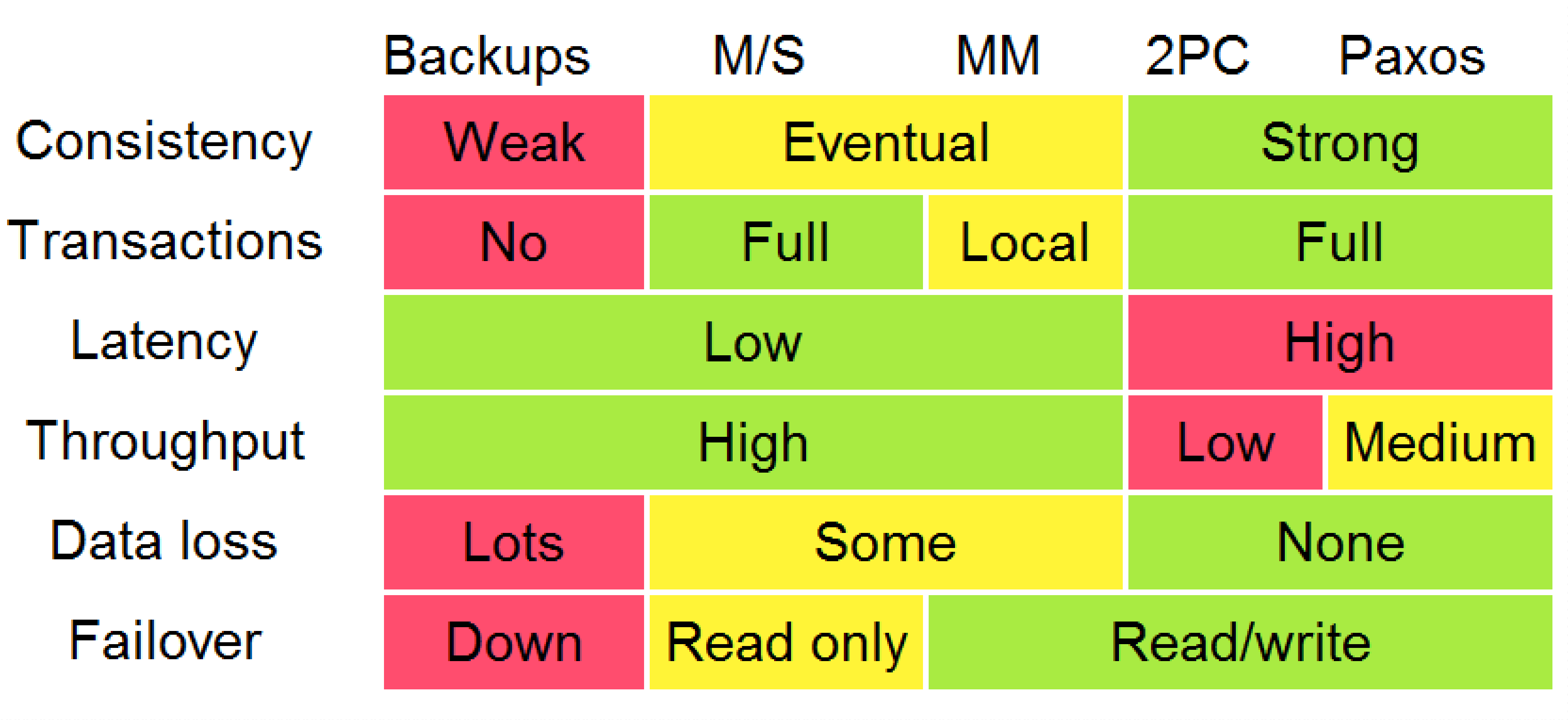


Phase Two: Successful Commit



Phase Two: Failure(Abort, Rollback)



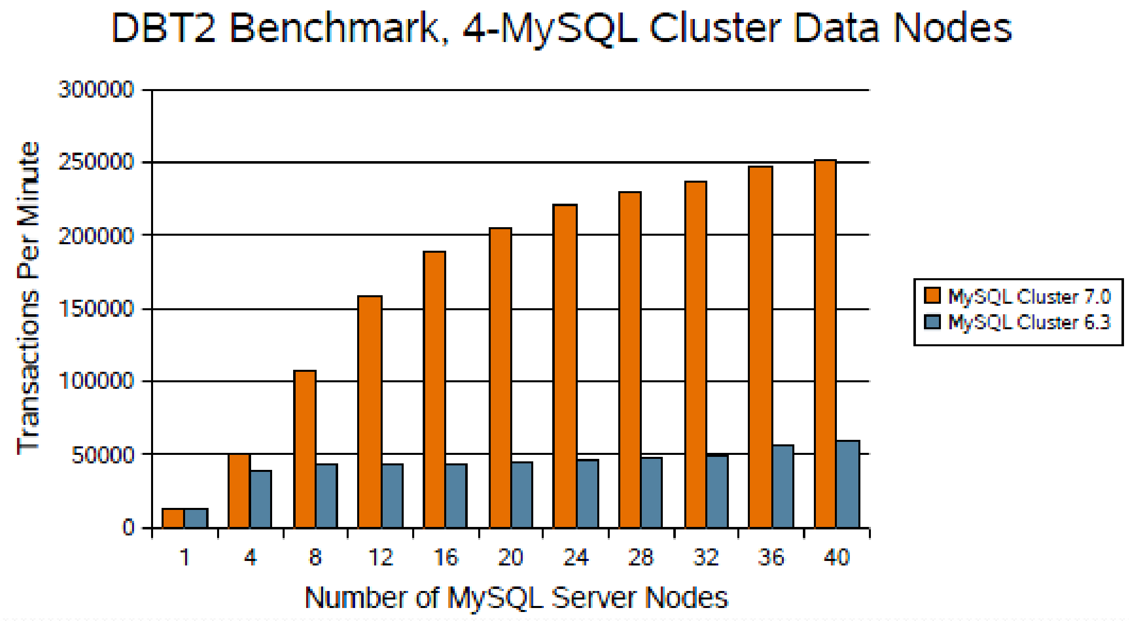


	Backups	M/S	MM	2PC	Paxos
Consistency	Weak	Eventual		Strong	
Transactions	No	Full	Local	Full	
Latency	Low			High	
Throughput	High			Low	Medium
Data loss	Lots	Some		None	
Failover	Down	Read only	Read/write		

MySQL Cluster Benchmarks (1/2)

- For 4 Node Cluster, MySQL Cluster 7 achieved 251,000 Transactions per minute which is more than 4X improvement over the MySQL Cluster 6.3 release.
- For 2 Node Cluster, MySQL Cluster 7 achieved 143,000 Transactions per minute which is more than 4X improvement over the MySQL Cluster 6.3 release.

MySQL Cluster Benchmarks (2/2)



- Data Nodes
  - Sun Fire x4450s
- SQL Nodes
  - Sun Fire x4600s & x4450s
- OpenSolaris
- Gigabit Ethernet

Questions / Discussion