

# **An Introduction to MySQL Cluster Architecture and Use**

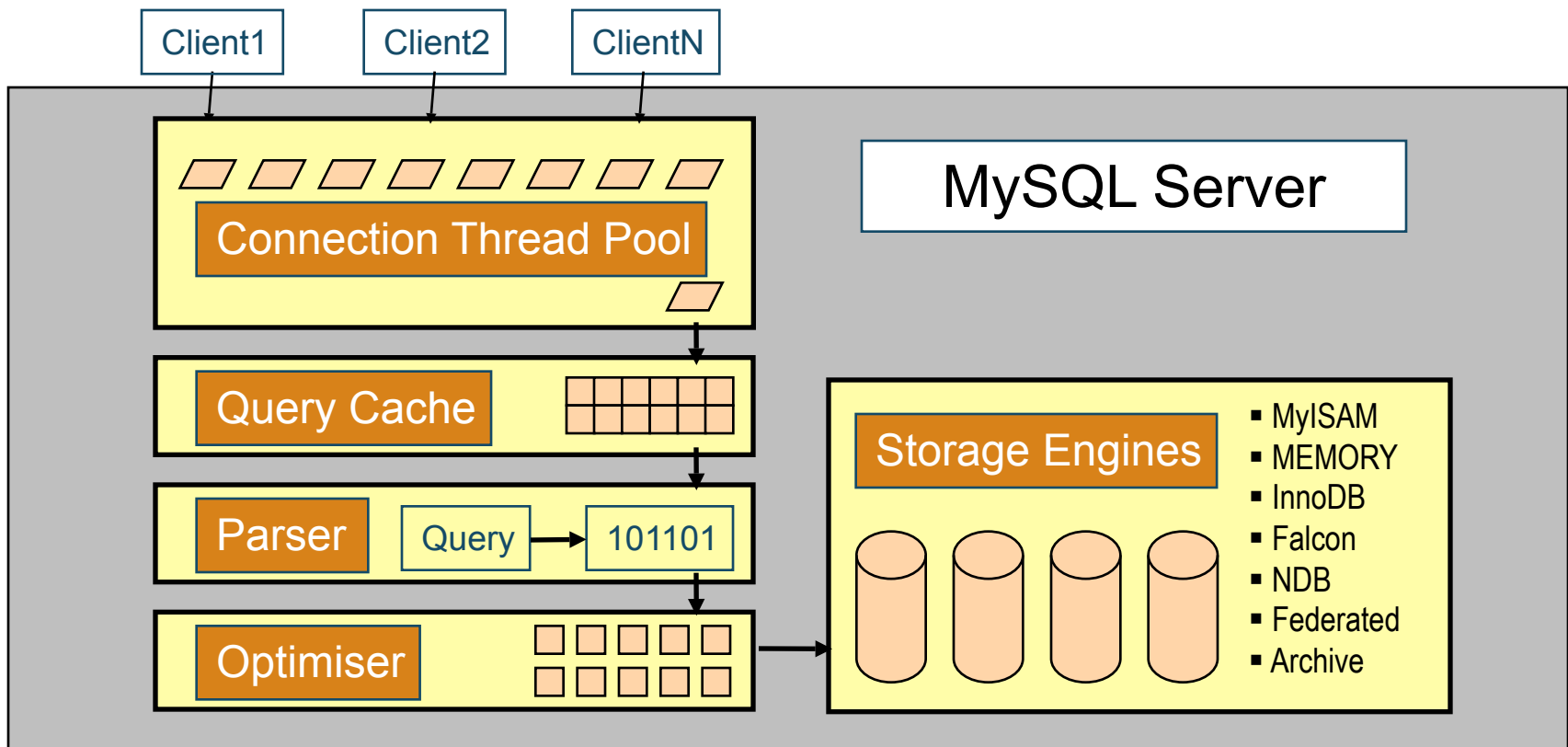
**November 2006**

**Arjen Lentz  
Support Engineer & Trainer**

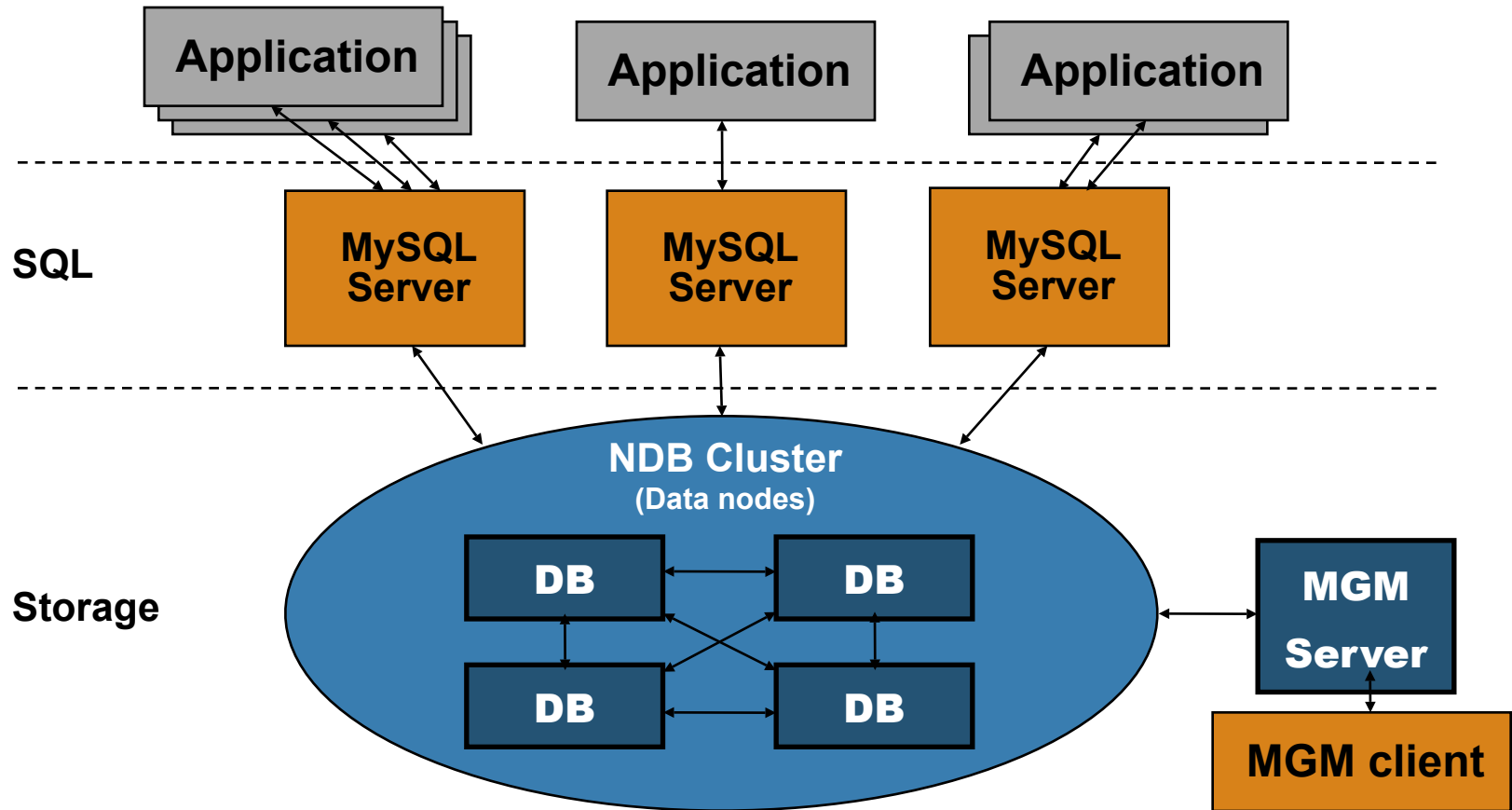
## Topics

- Learn about the architecture of MySQL Cluster
- Current and upcoming features
- What tasks MySQL Cluster is currently good at solving
  - and those where it isn't

# MySQL Server Architecture



# MySQL Cluster Architecture



# Physical Requirements

- A node is a process, not a computer
- Need at least three machines in a cluster
  - This avoids the split brain problem
- Management server doesn't need a powerful machine
- Data nodes
  - generally have a lot of memory
  - Disk IO requirements can be calculated from configuration parameters
  - not CPU bound
  - ndbd is single threaded
- SQL nodes need the most CPU
  - have more of these than data nodes
  - mysqld is multithreaded

# A Minimalistic Configuration

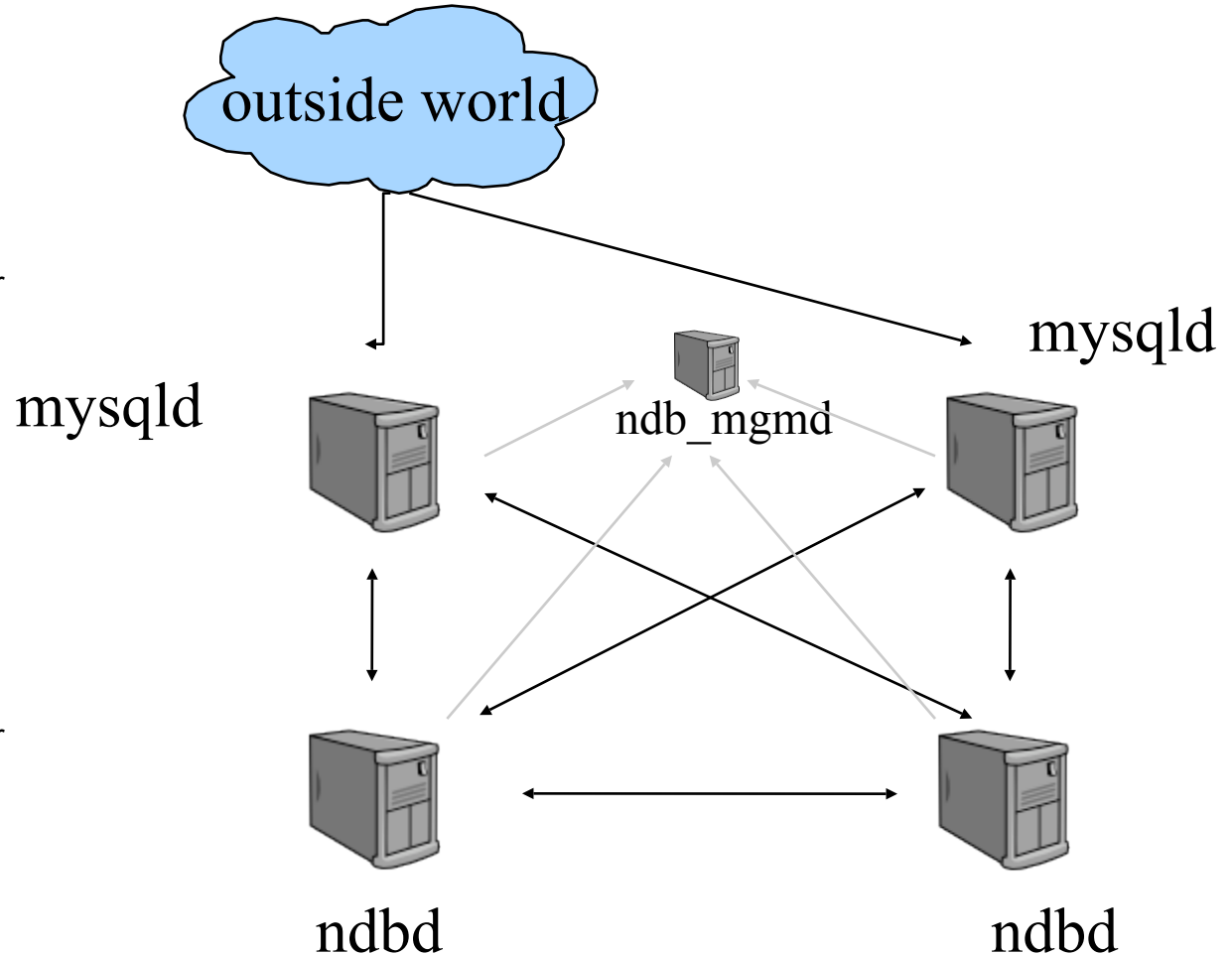
```
[ndbd default]
NoOfReplicas = 2
DataMemory = 400M
IndexMemory = 32M
DataDir = /usr/local/mysql/cluster
```

```
[ndbd]
HostName = 192.168.0.40
```

```
[ndbd]
HostName = 192.168.0.41
```

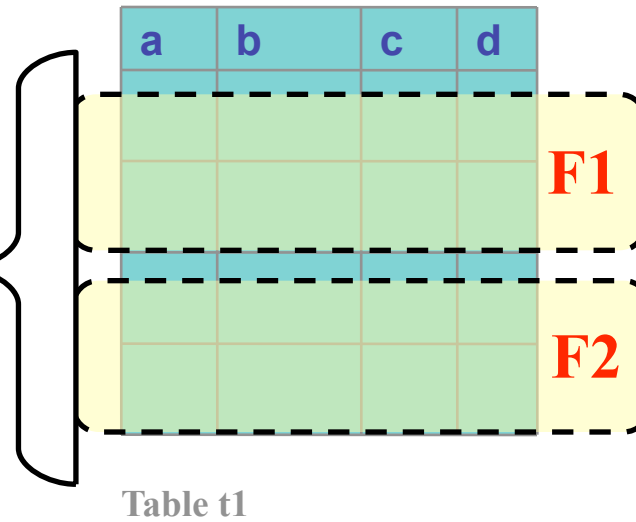
```
[ndb_mgmd]
DataDir = /usr/local/mysql/cluster
HostName = 192.168.0.42
```

```
[mysqld]
[mysqld]
[mysqld]
```

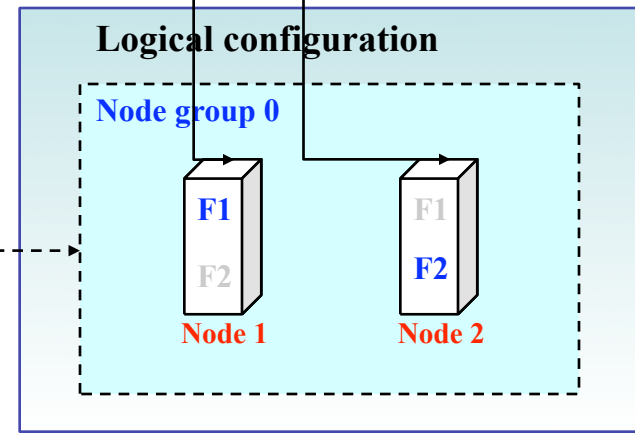
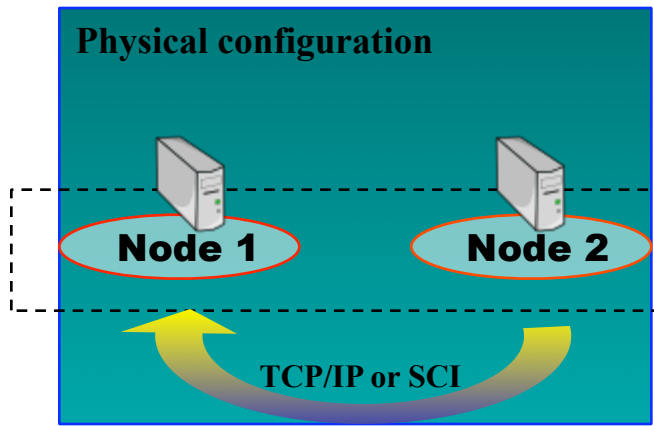


# Data Distribution on 2 Nodes

- Horizontal fragmentation of table t1 (2 fragments)
- Fragments distributed on nodes



2 copies of data  
 Fx – primary replica  
 Fx – secondary replica



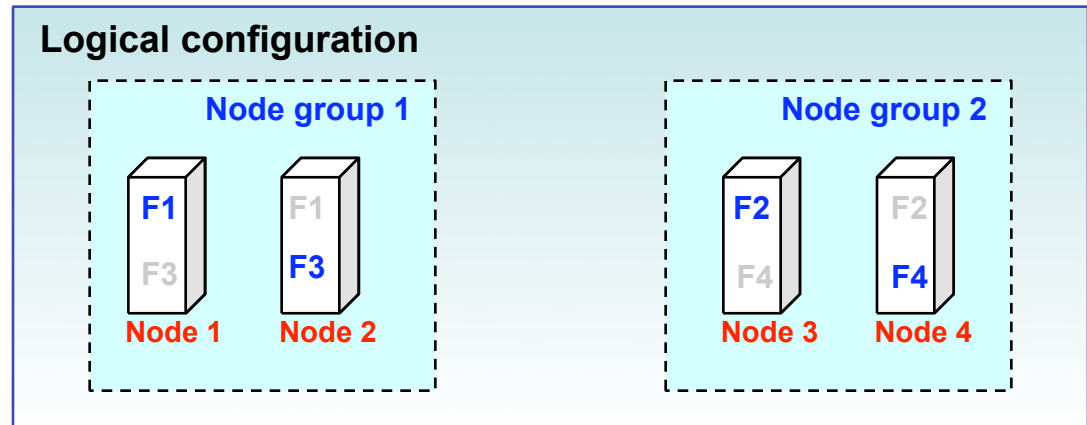
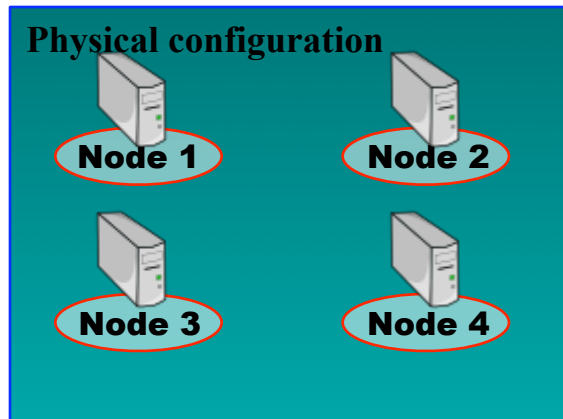
# Data Distribution on 4 Nodes

- Horizontal fragmentation of Table 1 (4 fragments)
- Fragments distributed on nodes

Pnr	AccNo	Val	\$\$	
				<b>F1</b>
				<b>F2</b>
				<b>F3</b>
				<b>F4</b>

Table 1

2 copies of data  
**F<sub>x</sub>** – primary replica  
 F<sub>x</sub> – secondary replica



# Data Distribution on 4 Nodes

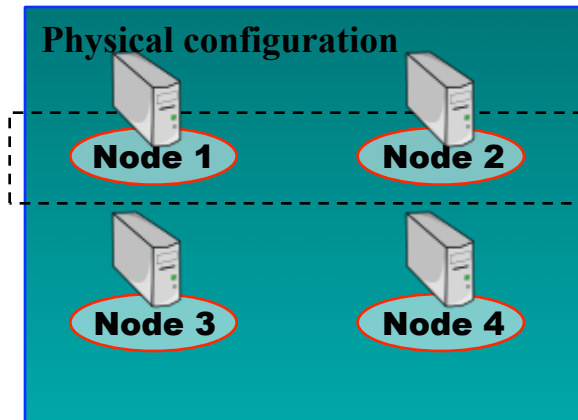
- Horizontal fragmentation of Table 1 (4 fragments)
- Fragments distributed on nodes

Pnr	AccNo	Val	\$\$	
				F1
				F2
				F3
				F4

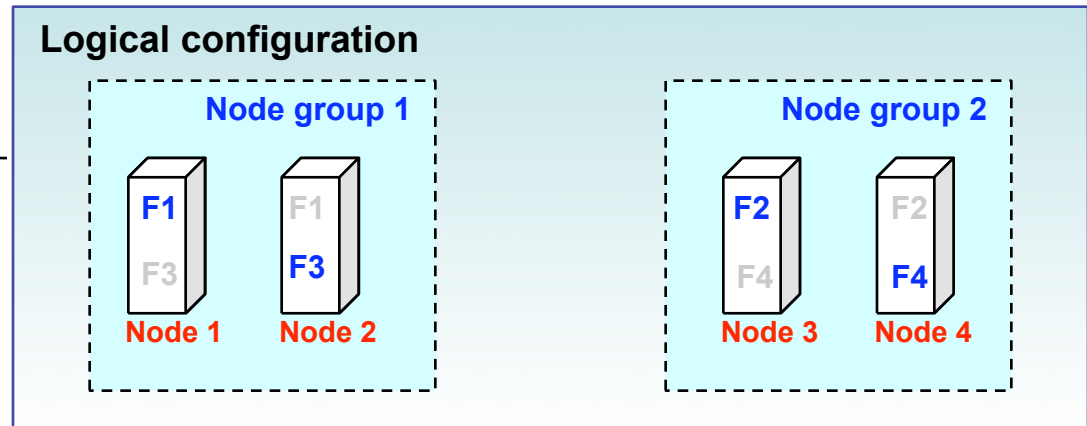
Table 1

2 copies of data  
 Fx – primary replica  
 Fx – secondary replica

## Physical configuration



## Logical configuration



# Data Distribution on 4 Nodes

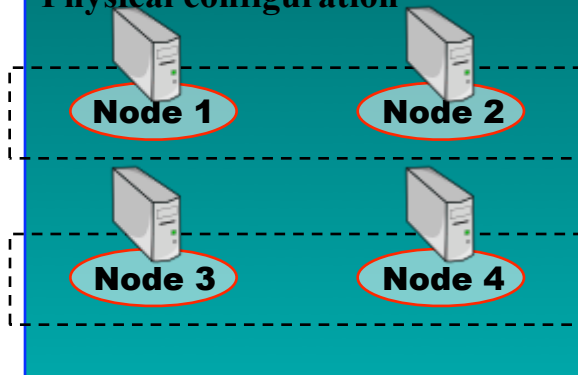
- Horizontal fragmentation of Table 1 (4 fragments)
- Fragments distributed on nodes

Pnr	AccNo	Val	\$\$	
				F1
				F2
				F3
				F4

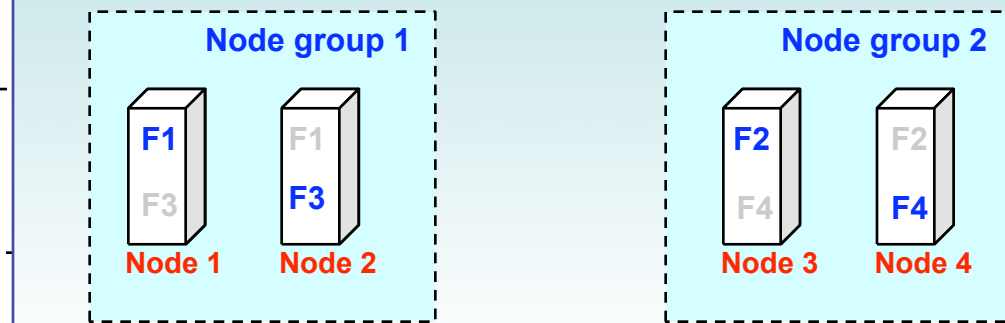
Table 1

2 copies of data  
 Fx – primary replica  
 Fx – secondary replica

## Physical configuration

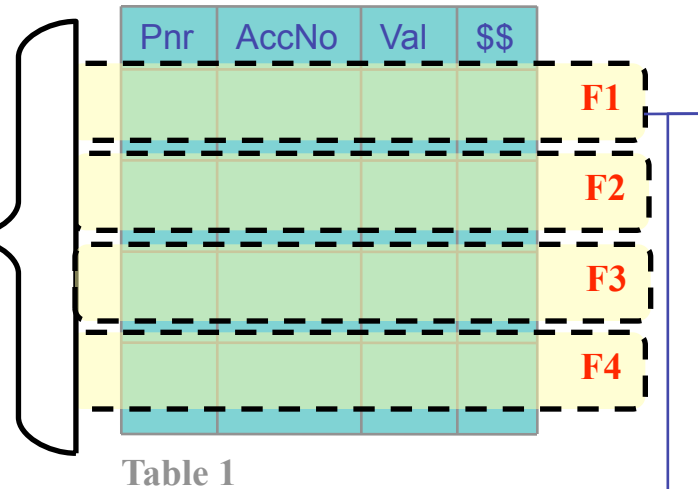


## Logical configuration

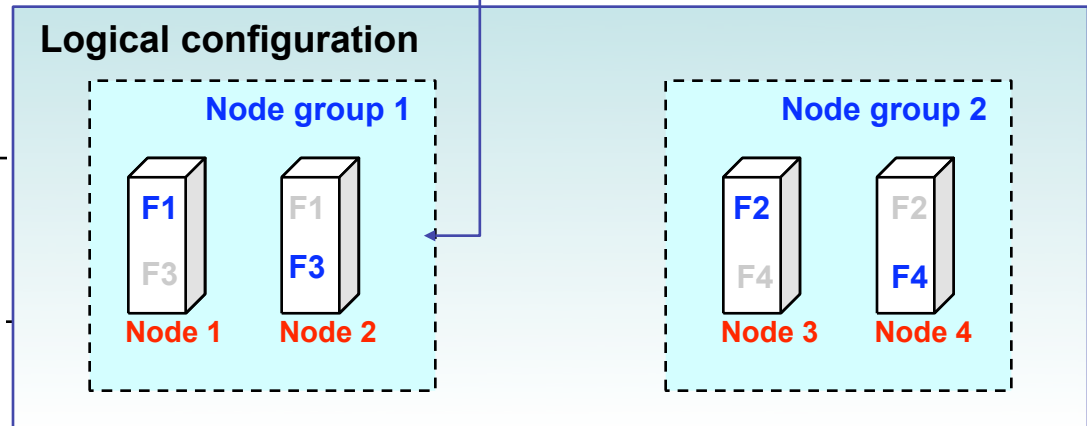
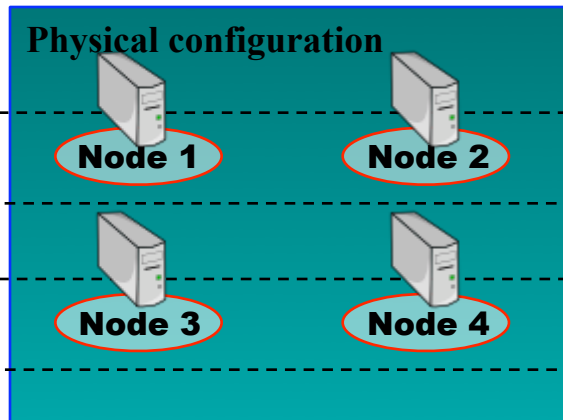


# Data Distribution on 4 Nodes

- Horizontal fragmentation of Table 1 (4 fragments)
- Fragments distributed on nodes



2 copies of data  
 Fx – primary replica  
 Fx – secondary replica



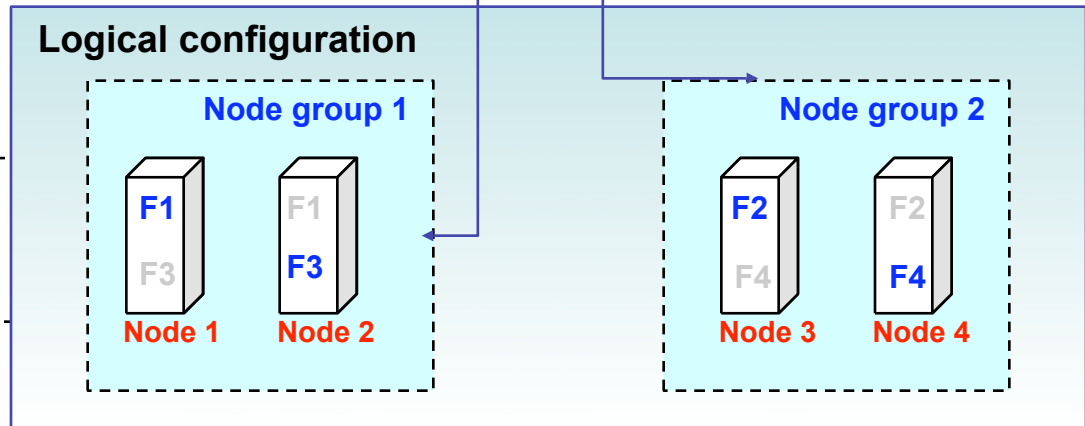
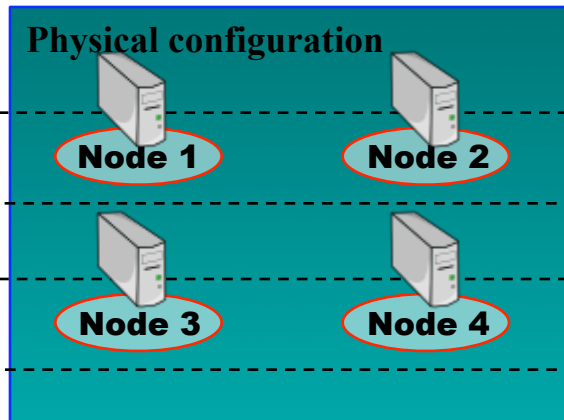
# Data Distribution on 4 Nodes

- Horizontal fragmentation of Table 1 (4 fragments)
- Fragments distributed on nodes

Pnr	AccNo	Val	\$\$	
				F1
				F2
				F3
				F4

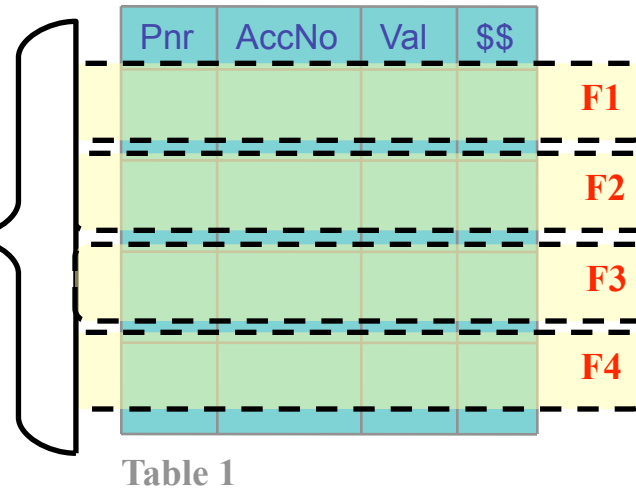
Table 1

2 copies of data  
 Fx – primary replica  
 Fx – secondary replica

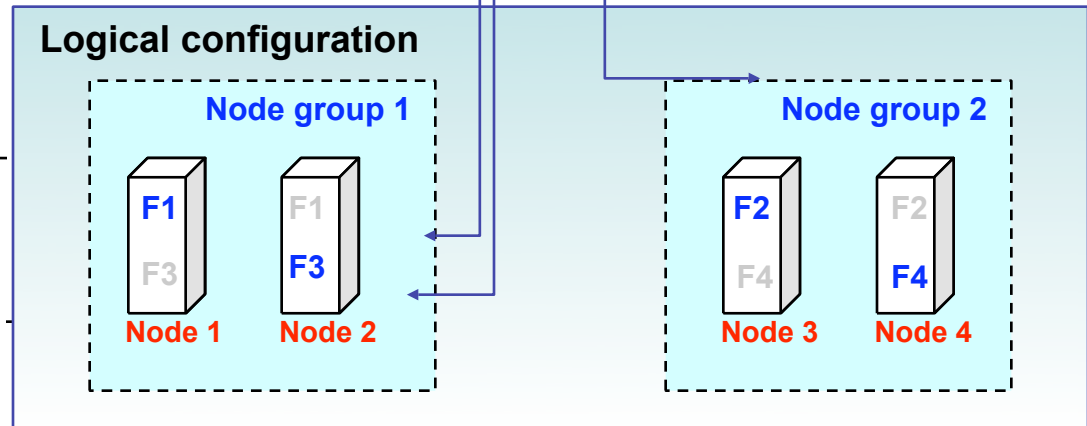
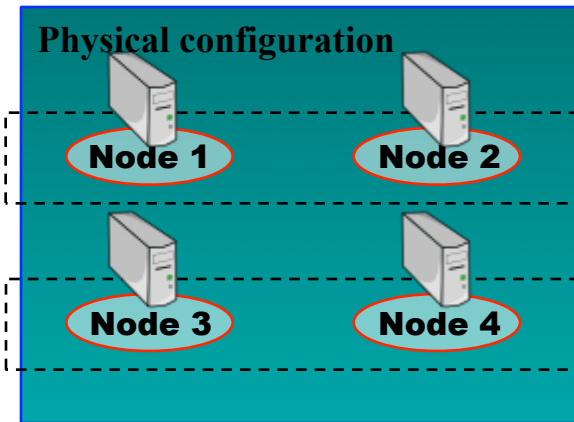


# Data Distribution on 4 Nodes

- Horizontal fragmentation of Table 1 (4 fragments)
- Fragments distributed on nodes



2 copies of data  
 Fx – primary replica  
 Fx – secondary replica



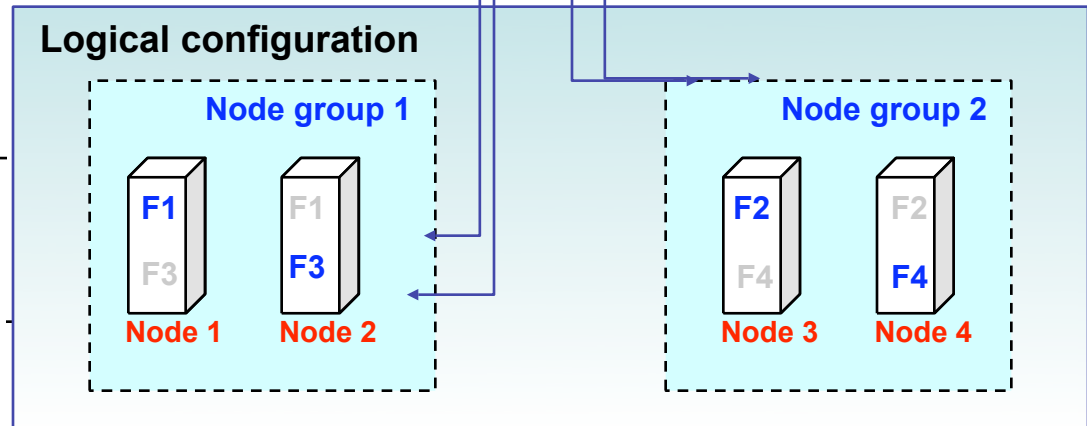
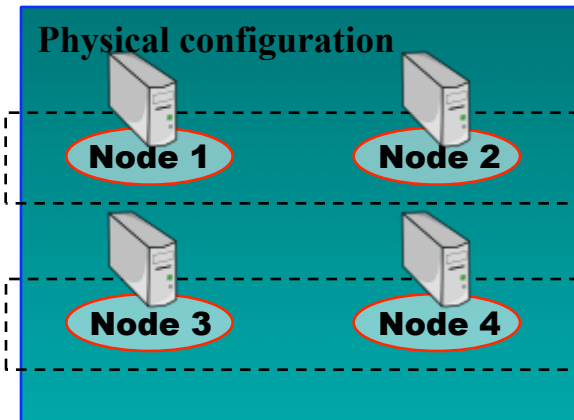
# Data Distribution on 4 Nodes

- Horizontal fragmentation of Table 1 (4 fragments)
- Fragments distributed on nodes

Pnr	AccNo	Val	\$\$	
				F1
				F2
				F3
				F4

Table 1

2 copies of data  
 Fx – primary replica  
 Fx – secondary replica



## Failure Scenarios

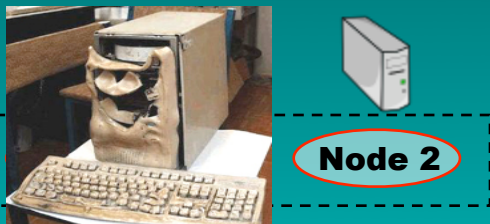
- MySQL Server Node
  - Can be restarted and reconnected to the cluster.
  - Applications can connect to other MySQL Server nodes
- Storage Node
  - All other storage nodes are informed about failure
  - Data is replicated, so there is another storage node to service transaction requests
  - Currently, transactions using a failed node are aborted, and have to be restarted by the application
- Management Server Node
  - Continued operation not dependent on management server
  - may be restarted, or have redundant nodes

# Single Node Failure

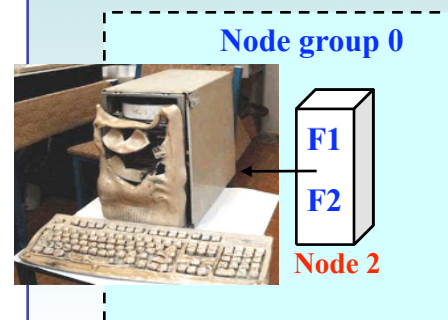
What happens if *node 1* fails?

- Detection of node failure. Primary fragment F1 handled by node 2
- Automatic restart and recovery of node 1
  - Node 1 recovers F1 and F2 from node 2 and rejoins the cluster.

## Physical configuration



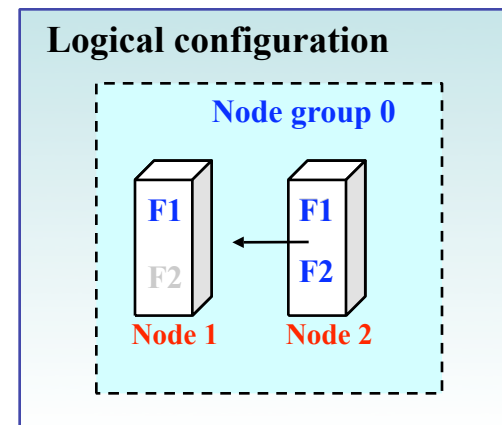
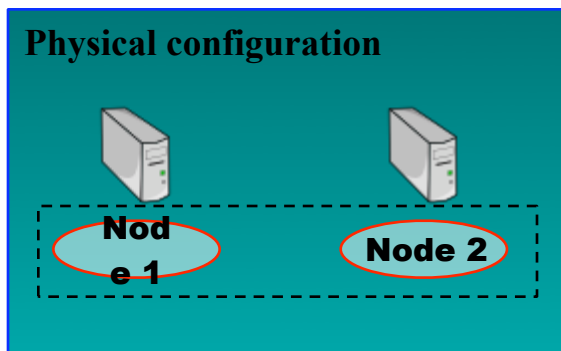
## Logical configuration



# Single Node Failure

What happens if *node 1* fails?

- Detection of node failure. Primary fragment F1 handled by node 2
- Automatic restart and recovery of node 1
  - Node 1 recovers F1 and F2 from node 2 and rejoins the cluster.



## Optimisations in 5.0

- engine\_condition\_pushdown option
  - SELECT A from T where unindexed\_field = 42;
  - Condition is evaluated in the data nodes, not in SQL node
  - saving a full table scan
    - which can be rather expensive on large tables
  - Common to see 5-10x speed improvement
  - details in EXPLAIN output
- Integration with query cache
- Batched lookups
  - SELECT \* from t1 where primary\_key IN (1,2,3,4,5,6,7,8,9,10);
  - all 10 key lookups are sent in one batch rather than one at a time
  - 2-3 times performance gain

## New in 5.1

- Variable sized records
  - large space savings for VARCHAR users
- User-defined Partitioning
- Disk based storage option for non-indexed columns
  - Indexes still in memory
- Integration with Replication

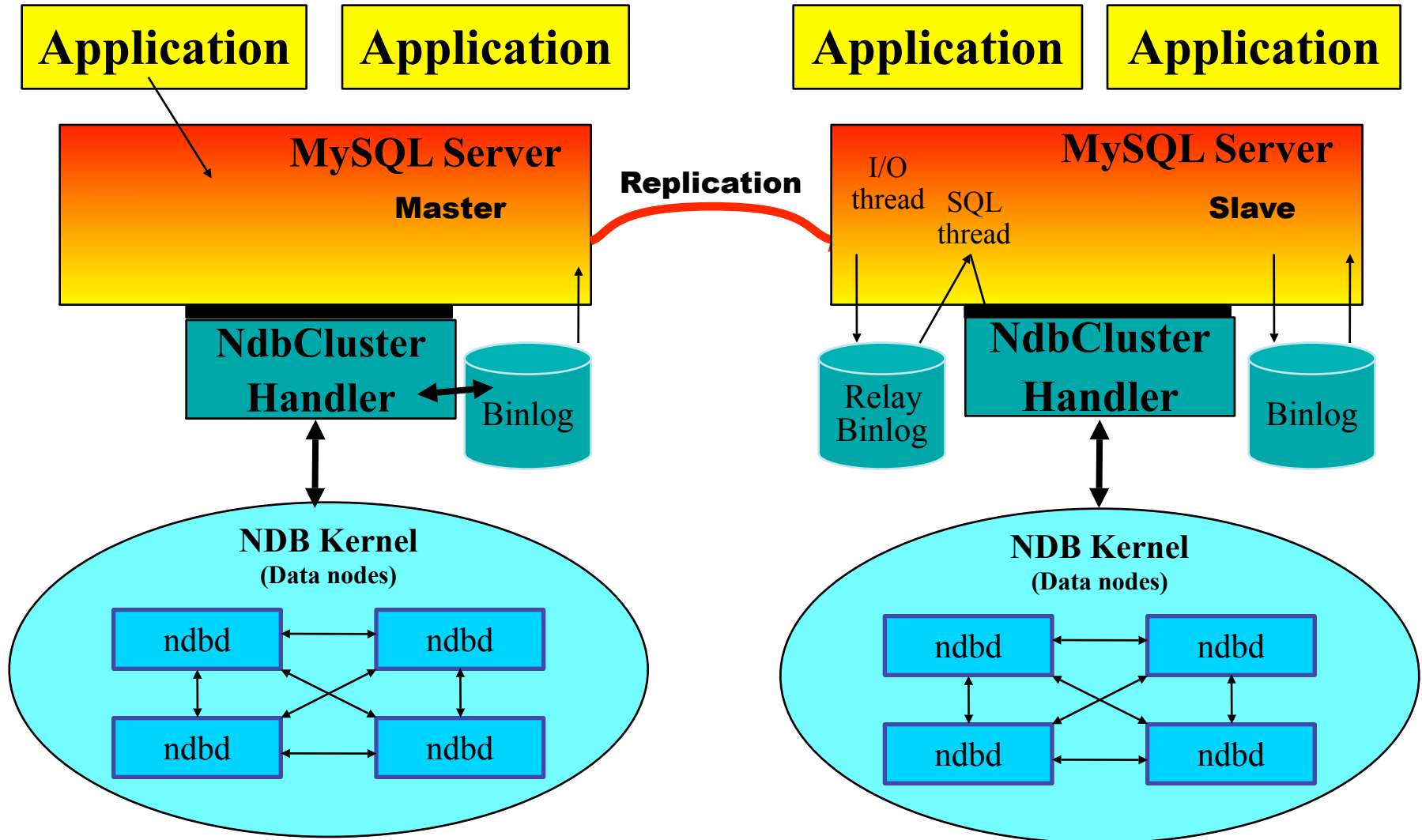
## Custom partitioning

```
CREATE TABLE t1 (  
  a INT NOT NULL,  
  b INT NOT NULL,  
  c INT NOT NULL,  
  PRIMARY KEY(a,b), INDEX (a)  
) ENGINE = NDB  
PARTITION BY RANGE(a) PARTITIONS 3  
(PARTITION x1 VALUES LESS THAN (5),  
PARTITION x2 VALUES LESS THAN (10),  
PARTITION x3 VALUES LESS THAN (20)  
);
```

## Tablespaces and Disk Storage

- CREATE TABLESPACE ts1  
 ADD DATAFILE 'datafile1'  
 USE LOGFILE GROUP lg1  
 INITIAL\_SIZE 32M  
 ENGINE=NDB;
- We can add another datafile too
- ALTER TABLESPACE ts1  
 ADD DATAFILE 'datafile2'  
 INITIAL\_SIZE 48M  
 ENGINE=NDB;
- We currently don't auto-extend
  - So just add another file

# MySQL Replication between Clusters



## Find out More!

- Just download and try!
- MySQL Online Documentation
  - <http://dev.mysql.com/doc/refman/5.1/en/index.html>
- Cluster Forum
  - <http://forums.mysql.com/list.php?25>
- Cluster Mailing List
  - <http://lists.mysql.com/cluster>
- Contact Sales
  - [sales@mysql.com](mailto:sales@mysql.com)
- Contact Me
  - Arjen Lentz
  - [arjen@mysql.com](mailto:arjen@mysql.com)