Lotus knows.

Smarter software for a Smarter Planet.

Domino Cluster 的資料同步和抄寫技術

沈偉 | IBM 高級軟體工程師

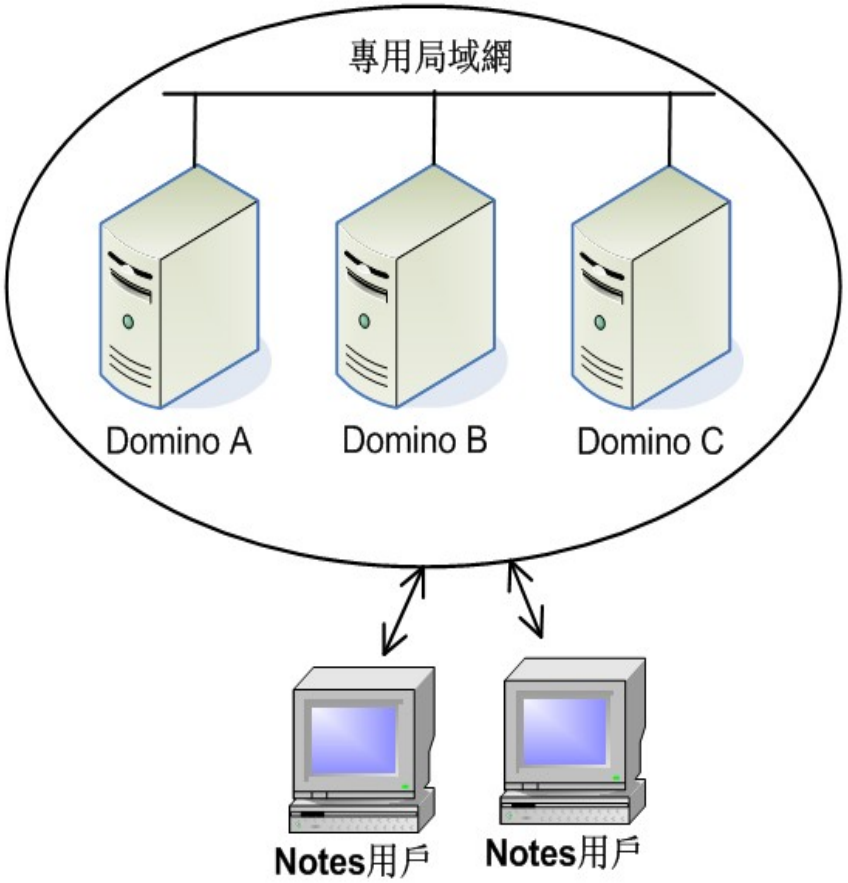
lotusknows.com

綱要

- Domino Cluster
- 容錯移轉和負載均衡
- Cluster 內的資料庫抄寫
- Cluster 的規劃、創建和監控

Domino Cluster

- 兩個或以上 Domino
- 屬於同一 Domino 域
- 磁片，CPU 資源較多
- 高速局域網
- 推薦同一版本



Domino Cluster 的好處

- 提高資料庫可用性
  - 容錯移轉
  - 重要資料庫
- 負載均衡
  - 提高存取速度，優化性能
- 利於資料備份、災備
- 利於硬體、作業系統和 Domino 伺服器升級
- 提高系統可擴展性
- .....

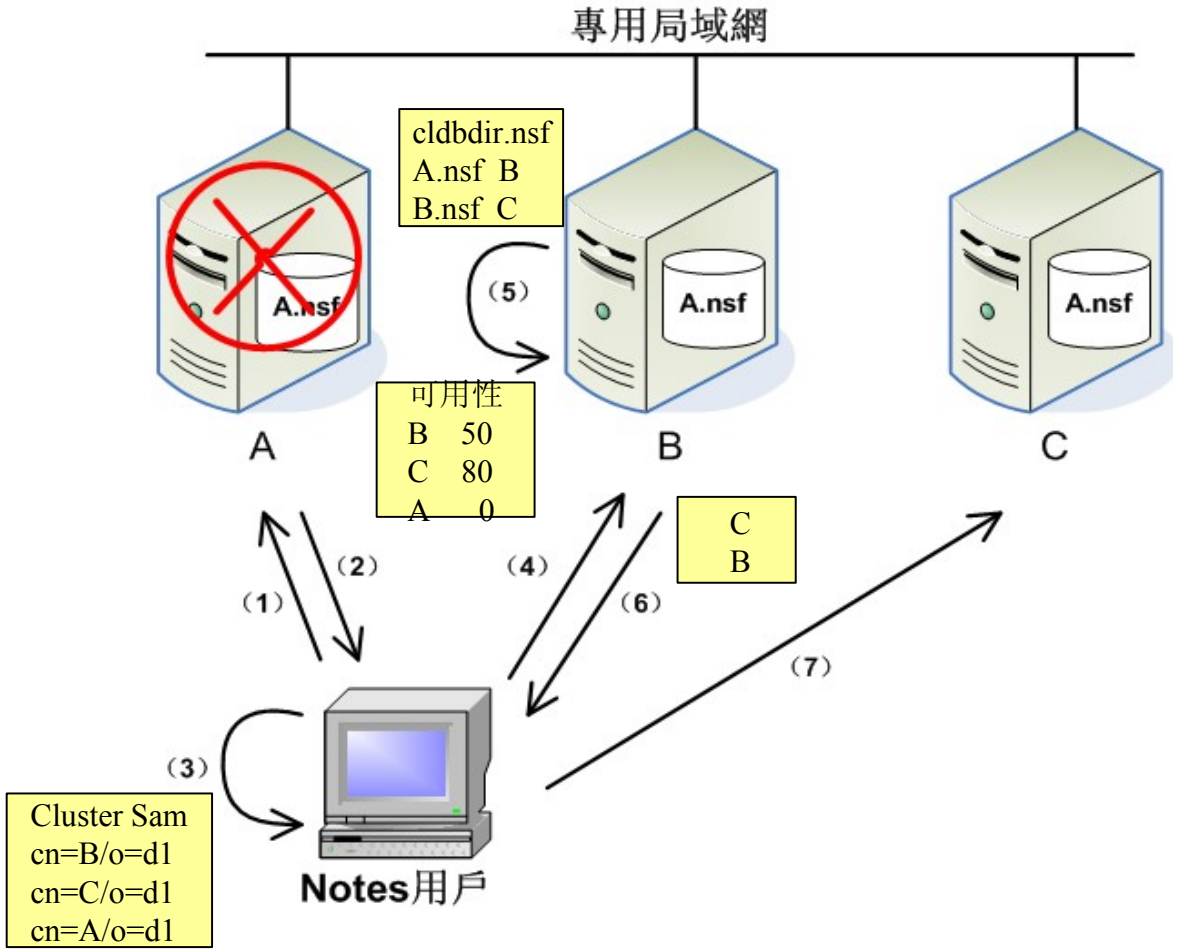
綱要

- Domino Cluster
- 容錯移轉和負載均衡
- Cluster 內的資料庫抄寫
- Cluster 的規劃、創建和監控

觸發容錯移轉

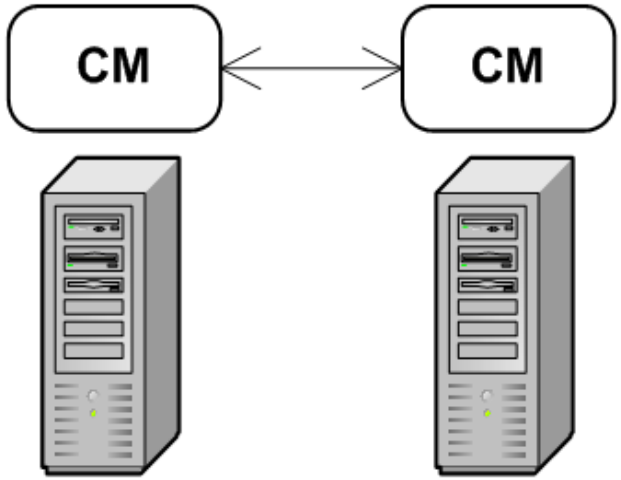
類別	觸發容錯移轉的動作
打開資料庫操作	從書簽打開一個資料庫
	點擊文檔連結、視圖連結或者資料庫連結
	索引伺服器無法工作時使用域搜索
	漫遊伺服器無法工作時訪問漫遊文件
	啟動包含 @command([FileOpenDatabase]) 的域，動作或按鈕
	LS 中調用 NotesDatabase 類的 OpenWithFailover 方法
	Java 中調用 DbDirectory 類的 OpenDatabase 方法
	複製資料庫時目標伺服器不可用
郵件伺服器相關操作	發送郵件時
	查找名字時
	Type-ahead
	路由郵件時
	郵件預發送代理
	會議邀請
	閒置時間查找
	伺服器查找
Web 伺服器操作	選中打開 URL 的圖示
	點擊一個 URL 熱點
	Web 流覽器中訪問 URL

Cluster 中的容錯移轉



(1) Notes 打開 Domino A 上 A.nsf
(2) 沒有回應
(3) Notes 查找本地 cluster.ncf，讀取與 A 同一群集的第一個 Domino 伺服器 (B)
(4) Notes 訪問 B
(5) B 找出群集中含 A.nsf 的 Domino 伺服器清單 (按可用性高低排序)
(6) B 將列表發送給 Notes
(7) Notes 打開列表上第一個 Domino 伺服器 (C) 上的 A.nsf

Cluster Manager (CM)



- 向 cluster 其他伺服器發送心跳
  - 伺服器的可用性
- 監聽其他 CM 發送的心跳
  - 在緩存中更新
- 監控其他伺服器的可用性
- 向 Notes 發送可用伺服器列表
- 定期監控伺服器文檔，維護 cluster 列表
- 往伺服器日誌中寫入容錯移轉、負載均衡事件

Cluster 資料庫目錄 (Cluster Database Directory)

- CLDBDIR.NSF
  - 每個文檔包含一個資料庫資訊，包括資料庫名字，伺服器名字，路徑， replica ID 等
  - 存在于 cluster 中每個伺服器上，互為 replica
- 維護
  - 由 cluster 資料庫目錄管理器 (Cluster Database Directory Manager) 創建
  - 資料庫更新及時
  - replicate 及時 (cluster replicator)

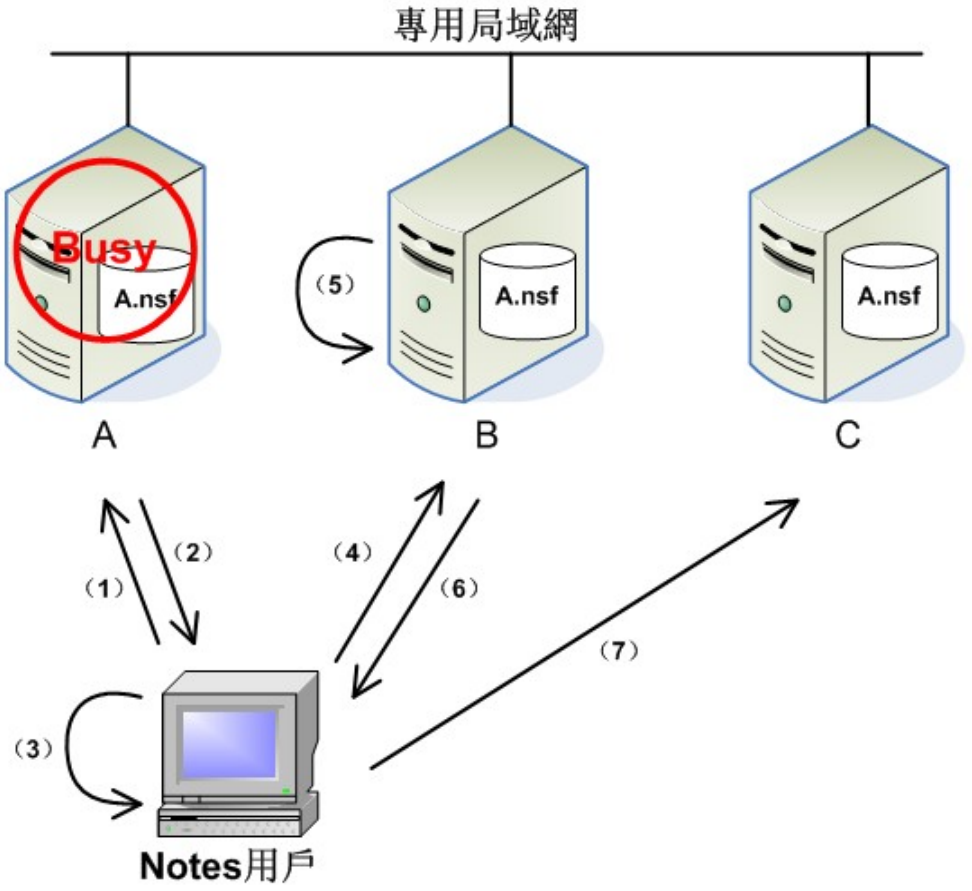
其他組件

- Cluster Replicator
  - 即時 replicate
  - 一個或多個
- Cluster Administrator
  - 啟動 / 關閉 cluster 資料庫目錄， cluster replicator
  - 清除 cluster 資料庫目錄

負載均衡

- 可用性閥值
- 最大用戶數

(1) Notes 打開 Domino A 上 A.nsf
(2) A 回應“伺服器忙”
(3) Notes 查找本地 cluster.ncf，讀取與 A 同一群集的第一個 Domino 伺服器 (B)
(4) Notes 訪問 B
(5) B 找出群集中含 A.nsf 的 Domino 伺服器清單 (按可用性高低排序)
(6) B 將列表發送給 Notes
(7) Notes 打開列表上第一個 Domino 伺服器 (C) 上的 A.nsf



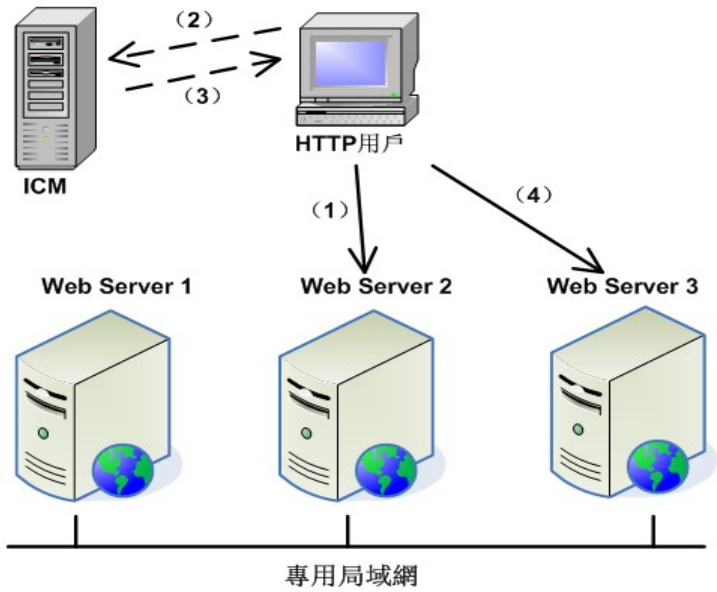
伺服器可用性 (availability)

- 可用性索引值
  - 0 ~ 100， 100 可用性最大
  - 當前交易處理平均時間 / 交易處理最小時間 (loadmon.ncf)
- 可設置閥值
  - SERVER_AVAILABILITY_THRESHOLD

```
OPEN_DB = 291.8
CLOSE_DB = 52.4166666666667
GET_NOTE_INFO = 95.5294117647059
OPEN_NOTE = 112.125
READ_OBJECT = 84.2
NIF_OPEN_NOTE = 3000000
GET_OBJECT_SIZE = 129
DB_INFO_GET = 69.4
DB_REPLINFO_GET = 64.6666666666667
DB_READ_HIST = 3000000
DB_WRITE_HIST = 3000000
GET_SPECIAL_NOTE_ID = 68.3333333333333
```

Web Server Cluster

- Web 伺服器
  - 收到 HTTP 請求後，查看伺服器文檔得到 ICM 位址
  - 生成指向 ICM 的 URL
- Internet Cluster Manager(ICM)
  - 負責重定向 HTTP 請求
  - 週期性向 cluster 內 web 伺服器發送探針，獲取資訊
  - 將最可用的 relica 連結返回給 HTTP 用戶端
  - 可在 cluster 外



綱要

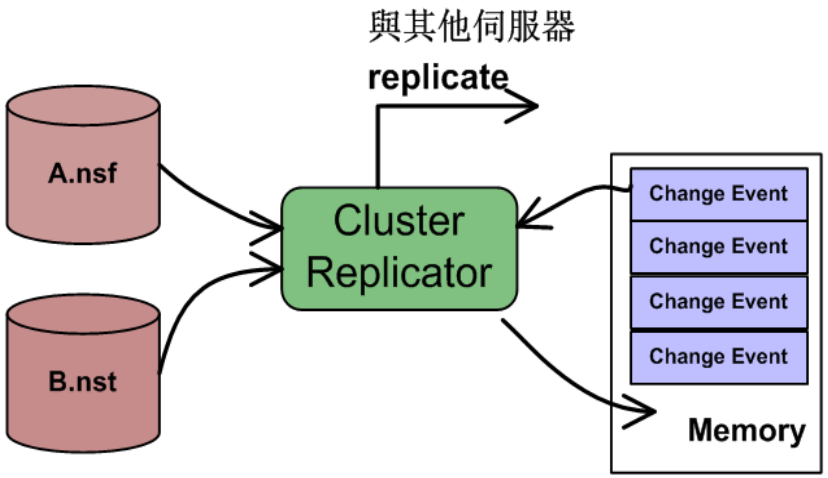
- Domino Cluster
- 容錯移轉和負載均衡
- Cluster 內的資料庫抄寫
- Cluster 的規劃、創建和監控

cluster 資料庫抄寫

- 保持資料庫 replica 一致
- 容錯移轉、負載均衡實現的基礎
- 時延性要求高

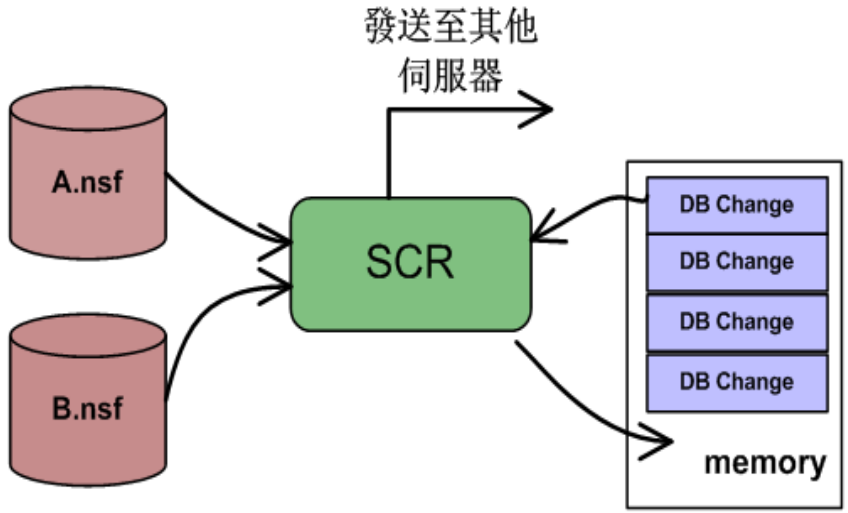
Replication in Cluster(8.0 版本以前)

- 事件驅動
  - 探測到變化，立刻 replicate
  - 將事件加入到記憶體中的佇列
  - 讀出事件，和 cluster 其他伺服器 replicate
- 事件僅保存在記憶體中
  - 若目標伺服器不可達，每隔一段時間 replicate 一次
  - 若關機或重啟，事件丟失
  - 一般還需對 replica 設置標準計畫複製



Streaming Cluster Replicator (SCR)

- 將資料庫的修改存入記憶體
  - 資料庫修改發生後，修改立即被捕捉
  - 修改存入記憶體佇列
- 直接發送修改
  - 向所有包含此副本的 cluster 伺服器發送
  - 其他伺服器收到修改，應用於 replica
- Domino 8.0 以後支持
- 性能提升
  - 時延小 **(269s -> 5s, 4000** 個用戶的測試環境 **)**
  - 源伺服器 CPU 減少 10%

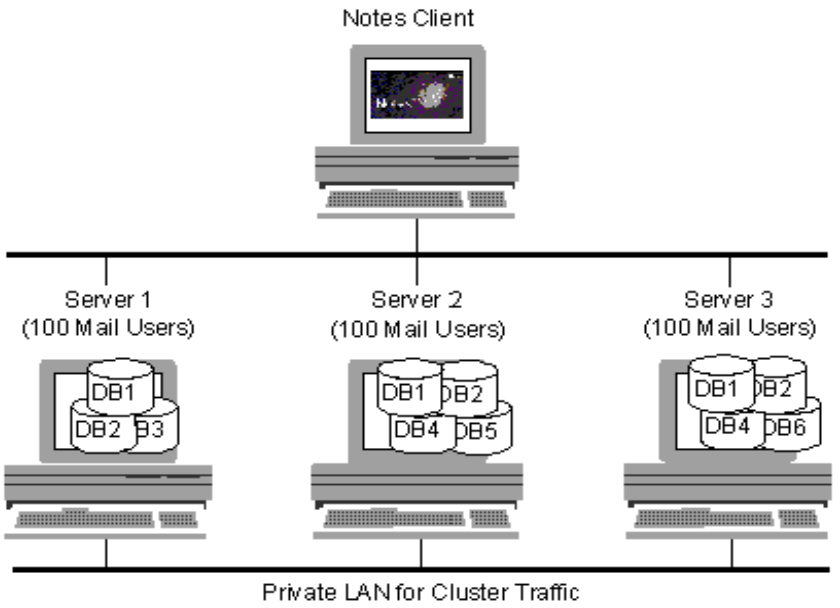


綱要

- Domino Cluster
- 容錯移轉和負載均衡
- Cluster 內的資料庫抄寫
- Cluster 的規劃、創建和監控

Cluster 的規劃

- 高可靠性的資料庫
  - mail 資料庫，重要資料庫，訪問量大的資料庫
- cluster 規模
  - 機器的配置
- 高速專用局域網



建立 cluster

- 創建
  - 設置
- 驗證
- 創建 replica

監控 **cluster**

- 顯示 cluster 成員和可用性
  - show cluster
- 容錯移轉和負載均衡事件
  - log.nsf
- Replication 統計
  - Show stat replica.cluster*
  - Tell clrepl dump

```
sh cluster
 Cluster information:
 Cluster name: cluster_xhp, Server name: mrwin64/851fpx
 Server cluster probe timeout: 1 minute(s)
 Server cluster probe count: 54
 Server cluster default port: *
 Server cluster auxiliary ports:
 Server availability threshold: 0
 Server availability index: 100 (state: AVAILABLE)
 Server availability default minimum transaction time: 3000
 Cluster members (2):
 Server: mraix64/851fpx, availability index: 100
 Server: mrwin64/851fpx, availability index: 100
```

```
Cluster Replicator context at 08/13/2010 02:39:17 AM
Tasks:          1
Work Queue Depth:   0
RetryInProgress:    0
RTR Pool Usage:     4% [42336 of 1048576 bytes used]
   Destination Server database block    73
    Server:         CN=mraix64/O=851fpx
    LastError:      No error
    LastRetryTime:  08/13/2010 02:36:38 AM
08/13/2010 02:39:18 AM  Dumping In-memory context
```

問與答

Legal Disclaimer

The information contained in this publication is provided for informational purposes only. While efforts were made to verify the completeness and accuracy of the information contained in this publication, it is provided AS IS without warranty of any kind, express or implied. In addition, this information is based on IBM's current product plans and strategy, which are subject to change by IBM without notice. IBM shall not be responsible for any damages arising out of the use of, or otherwise related to, this publication or any other materials. Nothing contained in this publication is intended to, nor shall have the effect of, creating any warranties or representations from IBM or its suppliers or licensors, or altering the terms and conditions of the applicable license agreement governing the use of IBM software.

References in this presentation to IBM products, programs, or services do not imply that they will be available in all countries in which IBM operates. Product release dates and/or capabilities referenced in this presentation may change at any time at IBM's sole discretion based on market opportunities or other factors, and are not intended to be a commitment to future product or feature availability in any way. Nothing contained in these materials is intended to, nor shall have the effect of, stating or implying that any activities undertaken by you will result in any specific sales, revenue growth or other results.

Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon many factors, including considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve results similar to those stated here.

All customer examples described are presented as illustrations of how those customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics may vary by customer.

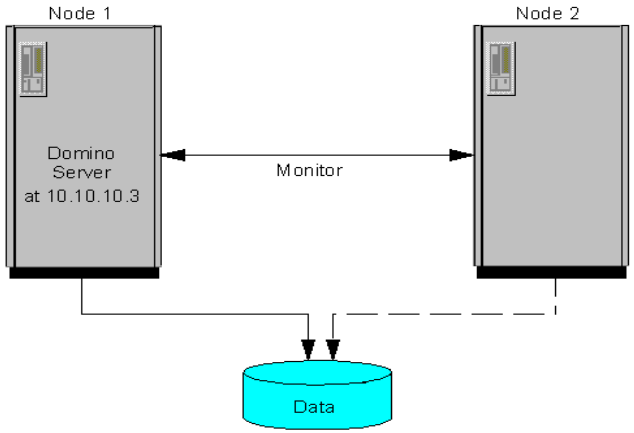
IBM, the IBM logo, Lotus, Lotus Notes, Notes, Domino, and Lotusphere are trademarks of International Business Machines Corporation in the United States, other countries, or both.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.
Other company, product, or service names may be trademarks or service marks of others.

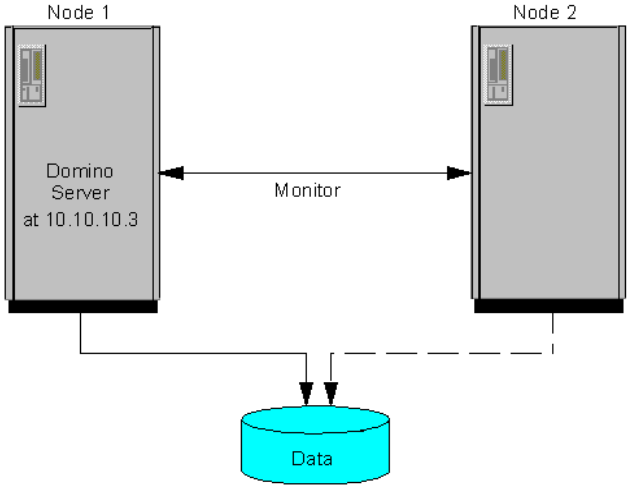
All references to Renovations refer to a fictitious company and are used for illustration purposes only.

Domino Cluster VS HACMP, MSCS

- Domino Cluster 優點
  - 不同作業系統
  - 負載均衡
  - 更智能，容錯移轉至可用性最高的伺服器
  - 多份數據



- OS Cluster 優點
  - Domino 代理容錯移轉
  - Hardcode 的 Domino Application 容錯移轉
  - Administration Process 容錯移轉



群集中的郵件容錯移轉

- 使用者打開不可用的郵件資料庫時
  - 容錯移轉發生，就像打開其他非郵件資料庫
- 用戶編輯郵件時，伺服器不可用
  - 發送時容錯移轉，發送到群集那其他郵件伺服器 MAIL.BOX
- 路由郵件時，目標郵件伺服器不可用
  - 本機伺服器和目標伺服器在同一群集，且本機伺服器有郵件資料庫 replica
  - 本機伺服器和目標伺服器在同一群集，但本機伺服器沒有資料庫 replica
  - 本機伺服器和目標伺服器在同一群集，但群集中沒有可用 replica
    - 週期性向目標伺服器不斷發送
  - 本機伺服器和目標伺服器不在同一群集，發送到目標群集中的一台伺服器

loadmon.ncf

OPEN_DB = 291.8
CLOSE_DB = 52.4166666666667
GET_NOTE_INFO = 95.5294117647059
OPEN_NOTE = 112.125
READ_OBJECT = 84.2
NIF_OPEN_NOTE = 3000000
GET_OBJECT_SIZE = 129
DB_INFO_GET = 69.4
DB_REPLINFO_GET = 64.6666666666667
DB_READ_HIST = 3000000
DB_WRITE_HIST = 3000000
GET_SPECIAL_NOTE_ID = 68.3333333333333