技術文件

減少 RAID 復原停機時間

序言

依據所選的特定 RAID 層級而定，可配置多硬碟機陣列來獲得最佳資料讀取能力、最高 I/O 效能，抑或同時實現這兩項目標。

每一種 RAID 層級都可以部署獨特的技術 (鏡射、位元層級、位元組層級或區塊階層資料分置、專屬式或分散式同位檢查) 組合來達到這些目標，但是這些 RAID 解決方案都有一個共同點：硬碟機故障時就必須進行冗長且容易出錯的 RAID 復原。

由於儲存系統輸出效能遠不及硬碟機功能的驚人成長，導致 RAID 復原變得越來越耗時間，所以近年來期盼能有解決方案的呼聲有增無減。復原一組企業級 RAID 可能要好幾小時 (通常會是好幾天)，如果在這漫長的復原過程中第二台硬碟機也故障了，RAID 中的資料便會無法讀取，IT 管理人員就必須回頭去找此資料的備份版本。

傳統的 RAID 復原功能通常會以剩餘可用之硬碟機中的同位資料，來復原故障硬碟機上的資料並將資料寫入熱備用裝置，這項程序本質上就既慢且複雜。現在 Seagate 有了更好的解決方案，也就是 Seagate RAID Rebuild™ 技術的部分故障複製功能，此功能可協助主機在求助 RAID 復原功能之前，先以飛快速度盡可能擷取故障硬碟機中的資料；由於要重建的資料大幅減少了，所以此類 RAID 復原功能不僅快上許多，也更不容易出錯。

Seagate RAID Rebuild™ 技術的好處

Seagate 在廣泛的研究中發現，RAID 復原功能最重要的衡量依據為：

- RAID 復原期間又再次發生故障的風險/情況
- RAID 復原期間系統效能下降
- RAID 復原從開始到結束所花時間

減少 RAID 復原停機時間

因為復原必須要進行長時間的讀取-寫入活動，所以 RAID 復原程序會對硬碟機造成巨大的額外負擔。考量到因時下企業硬碟機中所存資料量所需的冗長回復時間，就不會對 RAID 復原期間內又有第二台硬碟機故障感到訝異。

此外，長時間的復原活動也會讓系統效能下降，進而耽擱使用者存取 RAID 剩餘可用硬碟機之資料的時間。雖然讓復原程序有高於 RAID 陣列內主機 I/O 的優先順序能縮短復原時間 (並降低第二台硬碟機故障的機會)，但也會進一步降低系統效能。

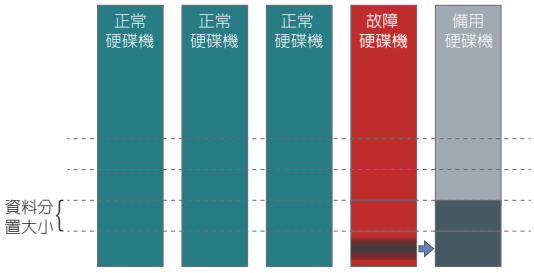
配置 Seagate RAID 重建技術的硬碟機具有部分故障複製功能，前述任何難題皆可迎刃而解。Seagate RAID Rebuild 技術在使用 RAID 復原功能之前，會先透過啓動故障硬碟機，主動來協助主機盡可能地擷取資料，故能確保：

- RAID 復原更快，更不易出錯
- 降低對系統效能的影響
- 快速存取復原資料

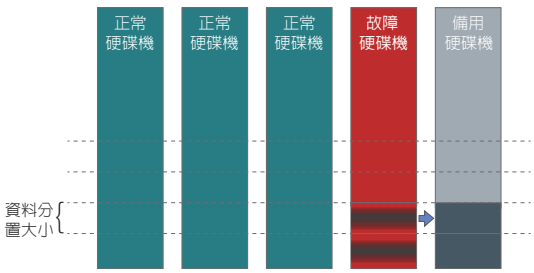
Seagate RAID Rebuild™ 技術：運作方式

為 PI 硬碟機進行格式化

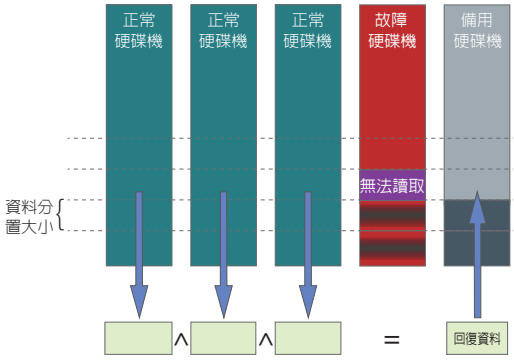
1. 先嘗試讀取故障的硬碟機。
2. 第一個資料分置會透過複製故障硬碟機來進行復原。



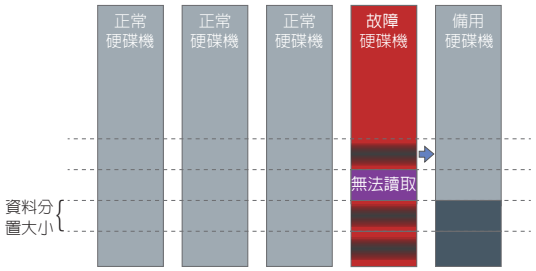
3. 第二個資料分置會透過複製故障硬碟機來進行復原。



4. 第三個資料分置使用故障硬碟機時會發生錯誤，所以會使用正常硬碟機進行復原。



5. 第四個資料分置會透過複製故障硬碟機來進行復原。



減少 RAID 復原停機時間

配置 Seagate RAID Rebuild™ 技術的 SAS 硬碟機

主機會發出 Send Diagnostic 指令，要求故障硬碟機自動區分出來。如果硬碟機沒有在旋轉，就會嘗試透過正常的磁頭旋轉。如果硬碟機無法旋轉，就會傳回「硬體錯誤」，而且無法透過 Seagate RAID 重建部分複製功能擷取該硬碟機中的資料。

如果硬碟機成功地旋轉起來，就會透過刪去不必要的背景活動、判斷硬碟機是否包含沒有使用的磁頭、提供媒體寫入防護以及進入將錯誤復原限制為任意重試等特別模式，以備使用部分複製功能。此模式會在硬碟機完成功率循環前保持作用中狀態，且不受匯流排重設的影響。

而主機應發出後續讀取工作負載，來擷取可用資料。在後續讀取期間，硬碟機可選擇根據重試速率加入沒有使用的磁頭清單。如果某指令包含的磁頭位於沒有使用的磁頭清單中，硬碟機就會傳回可識別失敗之 LBA 所在位置及下個可接受 LBA 所在位置的感應資料；主機應在下個可接受的 LBA 重新啓動後續讀取作業。(如果主機要發出佇列的讀取指令，其中數個指令就可能會失敗，並指向相同的下一個可接受 LBA。)

主機會負責維護仍須重建的 LBA 清單，即那些無法從該硬碟機讀取到的 LBA。[1]

配置 Seagate RAID Rebuild™ 技術的 SATA 硬碟機

主機會向故障硬碟機發出 S.M.A.R.T. Offline Immediate 指令。(如果硬碟機無法旋轉，就無法透過 Seagate RAID 重建部分複制功能來擷取該硬碟機的資料。) 執行此指令時，硬碟機會刪去不必要的背景活動、判斷是否包含沒有使用的磁頭、提供媒體寫入保護以及進入將錯誤復原限制為任意重試等特別模式。此模式會在硬碟機完成功率循環前保持作用中狀態。

而主機應發出後續讀取工作負載，來使用 Read FPDMA Queued 指令擷取可用資料。在後續讀取期間，硬碟機可選擇根據重試速率加入沒有使用的磁頭清單。如果某指令包含的磁頭位於沒有使用的磁頭清單中，硬碟機就會傳回對應的狀態值及錯誤值，接著主機就會從記錄頁面 0×10 讀取錯誤 LBA 與下一個可用 LBA，以便繼續。(在 SATA 通訊協定中，裝置不會在發生 NCQ 錯誤後接受任何新指令，除非主機讀取此記錄。) 而主機應在下個可接受的 LBA 重新開始後續讀取作業。

主機會負責維護仍須重建的 LBA 清單，即那些無法從該硬碟機讀取到的 LBA。[2]

[1] 搭載 SAS 硬碟機的 Force Unit Access (FUA) 位元適用 Read 指令，可重新啓動完整錯誤復原功能，且每個指令都可忽略故障的磁頭清單。

[2] 搭載 SATA 硬碟機的 Force Unit Access (FUA) 位元適用 Read FPDMA Queued 指令，可重新啓動完整錯誤復原功能，且每個指令都可忽略故障的磁頭清單。

減少 RAID 復原停機時間

結論

搭載 Seagate RAID Rebuild™ 部分故障複製功能的硬碟機，其價值明確亮眼：能從故障硬碟機直接擷取的資料越多，要在耗時易錯之 RAID 復原期間內復原的資料就越少。

Seagate RAID 重建技術的復原速度極快，能充分降低第二台硬碟機故障的風險/情況，進而保護 RAID 環境中的資料完整性。這項獨家的 Seagate 技術可減少系統效能下降的情況，並確保可快速救回故障硬碟機中的資料。

公司機構

身為領先業界標準組織的一員大將，Seagate 已將其發佈之標準規格中的內容，利用*重建協助*這個標題向 T10[3] (SAS 提案：11-298) 及 SATA-IO[4] (SATA 提案：SATA31_TPR_D144) 委員會提交了一份 RAID Rebuild™ 功能的開放式標準提案。

[3] T10 是隸屬資訊技術標準國際委員會 (International Committee on Information Technology Standards，INCITS) 的一個技術委員會，經美國國際標準學會 (American National Standards Institute，ANSI) 認定並遵循該學會所核可之規範運作。這些規範的設計目的，是為確保業界團體能開發出一套自發性標準。INCITS 開發出資訊處理系統標準，而 ANSI 則會核准這些標準的開發過程，然後將其發佈。ANSI 同時也在國際標準組織 (International Standards Organization，ISO) 和國際電工委員會 (International Electrotechnical Commission，IEC) 的聯合技術委員會 - 1 (JTC-1) 中，擔任美國代表。如需更多資訊，請前往 http://www.t10.org/

[4] Serial ATA 國際組織 (SATA-IO) 是頂尖業界公司為同等級公司所開發的獨立、非營利組織。SATA-IO 可為業界提供導入 SATA 規格的指導方針及支援。標準化 SATA 規格以預期未來 10 年內不會被淘汰的高速序列匯流排，來取代歷時 15 年的技術。SATA-IO 的成員都能影響或直接促成 SATA 規格開發。 如需更多資訊，請前往 http://www.sata-io.org/

www.seagate.com

RAID Rebuild™

美洲地區　Seagate Technology LLC　10200 South De Anza Boulevard, Cupertino, California 95014, United States, +1 408 658 1000
亞太地區　Seagate Singapore International Headquarters Pte. Ltd.　7000 Ang Mo Kio Avenue 5, Singapore 569877, +65 6485 3888
歐洲、中東和非洲　Seagate Technology SAS　16–18, rue du Dôme, 92100 Boulogne-Billancourt, France, +33 1 41 86 10 00