QNAP

QuTS hero ZFS

白　　皮　　書

INDEX

頁	標題
01	**概覽**
02	**ZFS 簡介與設定**
02	資料完整性與自我修復
02	建議的共用資料夾 /LUN 區塊大小
03	ZIL 的同步 I/O（標準 / 一律 / 不使用）
04	儲存池預留空間
06	**資料縮減（在線資料壓縮 / 在線重複資料刪除）**
07	在線資料壓縮
08	在線重複資料刪除
09	**SnapSync** 具成本效益的備份、資料保護與災難復原的複製解決方案。
09	簡介
09	SnapSync 的運作方式
11	發生災難時該怎麼辦
12	即時 SnapSync 設定的最佳做法
14	優勢
15	**QNAP 專利技術的 QSAL 適用於全快閃儲存陣列與固態硬碟 RAID**
16	**寫入聚合** 寫入聚合演算法可提升全快閃陣列的隨機寫入效能
17	**建議的讀取快取設定 (L2ARC)**
19	**記憶體大小與儲存池之間的效能相依性**

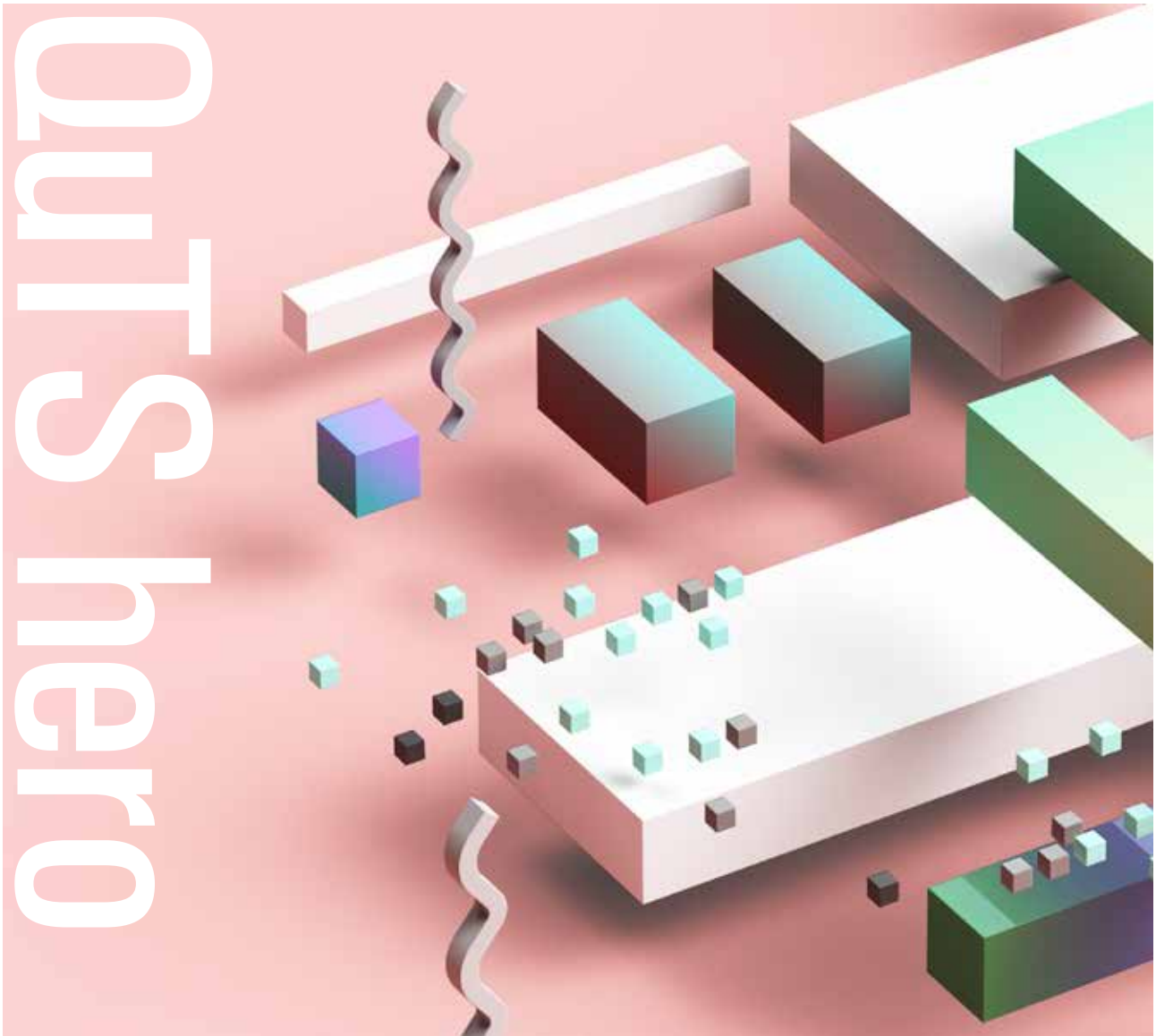
QuTS hero ZFS

白　皮　書

概覽

在現今 IT 密集的世界中，快速應變的需求可能是企業組織和資料中心成功或失敗的運作關鍵。在這種環境下，預算都會受到嚴格的審查，而任何潛在的停機時間－無論是幾分鐘還是幾秒鐘－都可能具有災難性。任何可能的效率及生產力提升都是成功的重要基石。

QuTS hero 作業系統採用 ZFS 檔案系統，其以資料完整性為優先處理，滿足高效能和高穩定性的資料備份需求。QuTS hero 整合了高效能資料保護、資料縮減和虛擬化應用程式，並支援雲端環境和 QoS（服務品質）管理功能，為小型企業、企業級資料中心及虛擬化作業提供更具成本效益的儲存裝置解決方案。

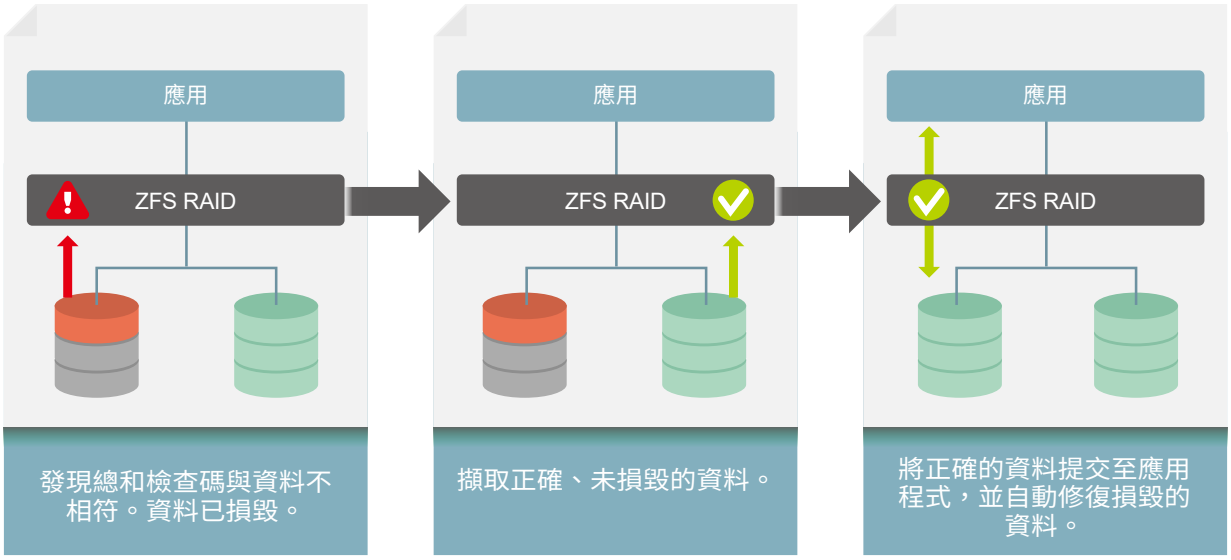


ZFS 簡介與設定

資料完整性與自我修復

資料的完整性及可靠性是確保應用程式和資料庫運作的關鍵。ZFS 能夠防止靜態資料損毀。這是由於硬體問題、纜線損壞或中繼資料錯誤所造成的潛在災難，同時也難以偵測。ZFS 能夠自行修復此類資料，包括檢查所有資料區塊，自動修復不正確的區塊。

強烈建議定期執行「儲存池清洗」，確保儲存資料的完整性，並於潛在問題出現之前予以修復。您可以設定排程來執行儲存池清洗，例如以每週或每月的頻率。



防止運行中的系統出現靜態資料損毀

建議的共用資料夾 /LUN 區塊大小：

128K - 影像編輯 / 大型檔案 / 備份
64K - 一般 /Hyper-V
32K - VMware
8k - OLTP
4K -VDI/ 資料庫 / 小型檔案

ZIL 同步 I/O（標準 / 一律 / 不使用）：

ZIL（ZFS Intent Log）是寫入日誌機制，可追蹤尚未提交至檔案系統的檔案變更。ZIL 會將這些變更的意圖記錄下來，稱為日誌。如果系統當機或電源故障，QuTS hero 會檢查日誌記錄並套用排程變更，確保快速還原檔案系統，降低資料損毀的機會。

當建立共用資料夾 /LUN 時，必須設定 ZIL 寫入記錄 I/O 模式，以排定資料完整性或效能的優先順序。

標準（預設）：根據應用程式與 I/O 要求類型選擇同步或非同步 I/O 交易。

一律：所有 I/O 交易皆以同步進行，這會改善資料完整性，但會降低效能。

不使用：所有 I/O 交易皆以非同步進行，此選項提供最佳效能，但可能會因為停電而導致資料遺失。

請注意，建議在專用的高速儲存裝置（例如 NVMe 固態硬碟）上設定專用 ZIL 作為 SLOG。對於 ZIL，建議採用超低延遲和高耐用性磁碟（例如 MLC 磁碟機）。對於讀取快取，應優先考慮輸送量。建議分割 ZIL 及讀取快取，以避免因 ZIL 寫入頻率高而導致整個 SSD 群組過早故障。

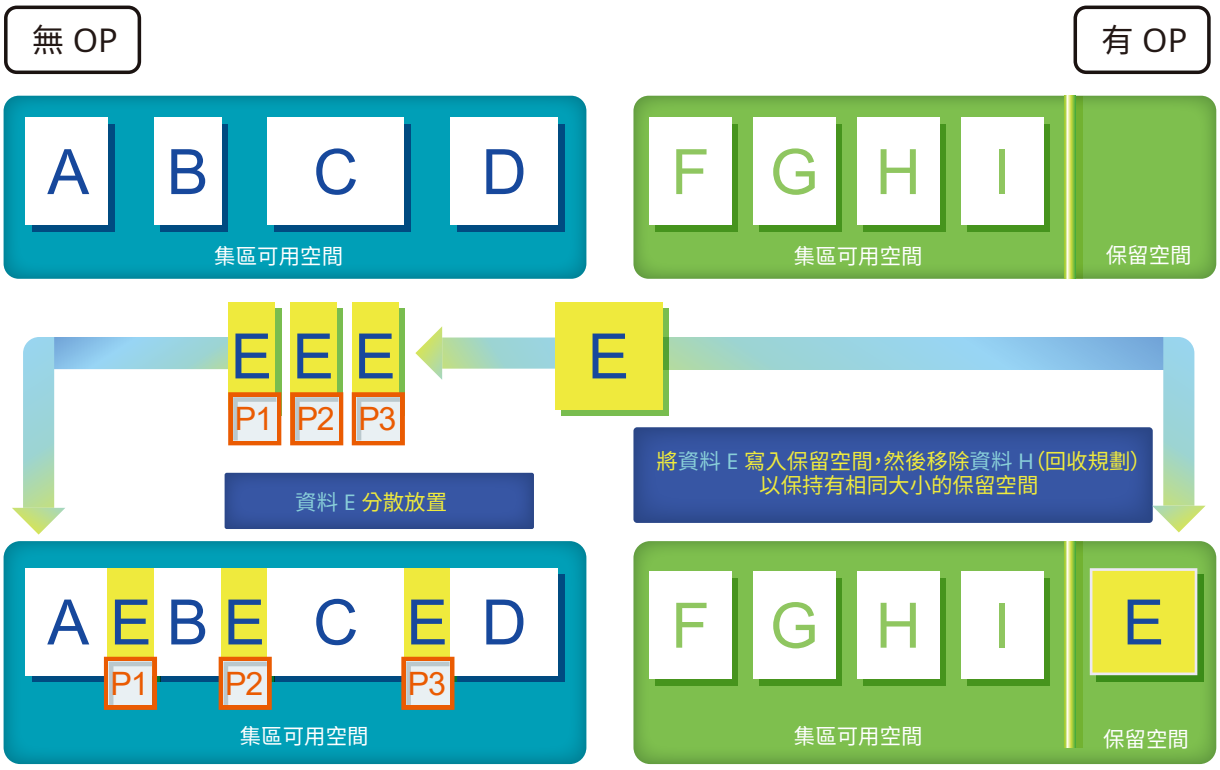
儲存池預留空間

當儲存池幾乎已滿或破碎時，ZFS 效能可能會受到影響。這是由於寫入時的複製行為：ZFS 一律會將新資料寫入不同區塊，更新中繼資料以指向新位置，然後釋放舊區塊，而非直接覆寫舊資料。在多次循環寫入之後，儲存池的資料分佈會變得破碎，這會使得寫入新資料變得更加困難及耗時。

QuTS hero 支援「儲存池預留空間」，可讓您指定並保留儲存池的空間百分比，在儲存池即將用完的時候，確保一致的隨機寫入效能。下圖說明寫入儲存池的資料時，使用或未使用「儲存池預留空間」的情況。如果未使用「儲存池預留空間」時，當儲存池即將用完且破碎時，每次分隔出新的資料寫入區塊將會花費一段時間，導致影響效能。透過「儲存池預留空間」，ZFS 可以找到連續的大區塊來儲存新資料。

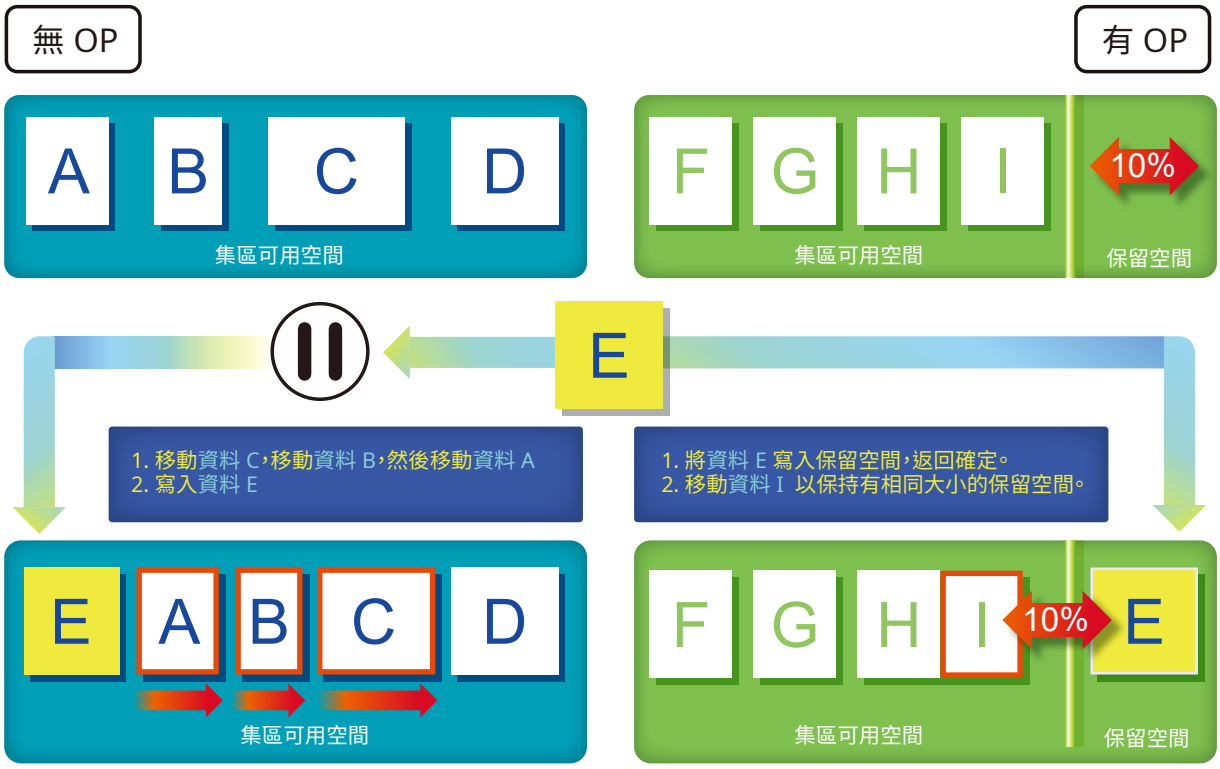
預留空間

改善破碎集區的效能（例如：將區塊寫入 **HDD** 時）



預留空間

改善破碎集區的效能（例如：將區塊寫入 **SSD** 時）



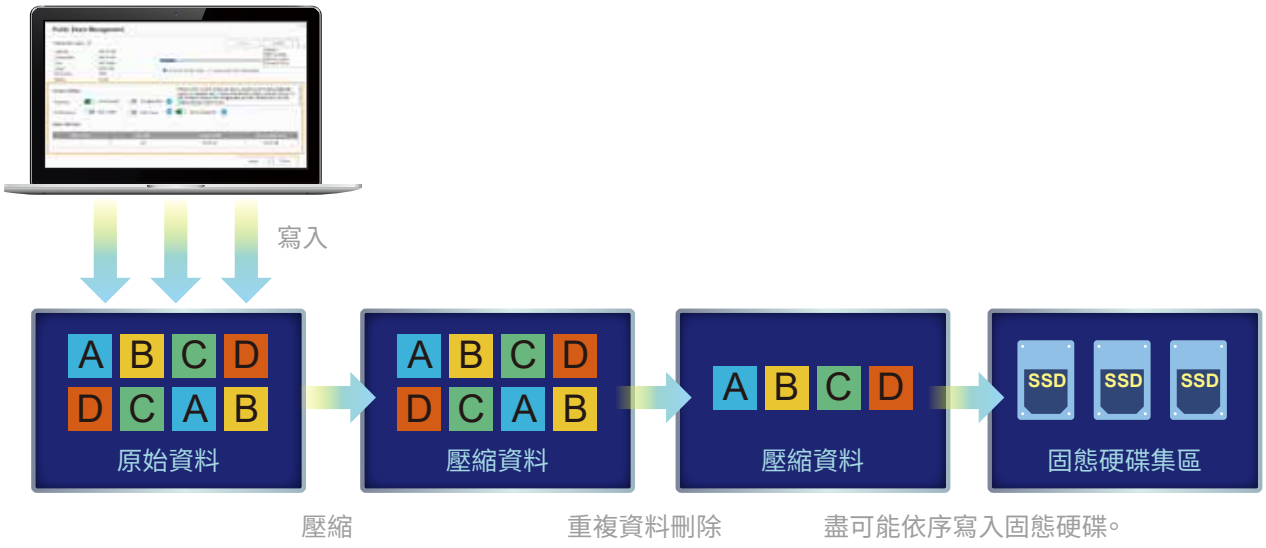
啟用「儲存池預留空間」會確保隨機寫入效能，但同時會影響儲存容量。保留的空間越多，效能就越高。確定權衡必要性的關鍵是考慮讀 / 寫比率和應用情境。寫入越多，「儲存池預留空間」提供的效能提升就越高。

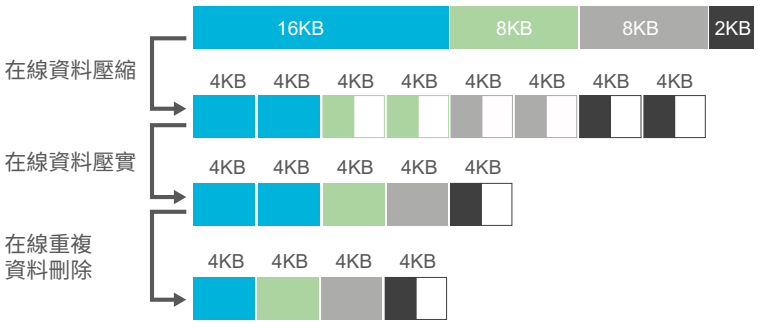
通常建議將「儲存池預留空間」值設定為 10-20%。對於寫入密集且需要高效能的應用（例如 SQL 記錄、備份、監控及醫療影像），建議您設定更高的值。至於讀取密集型應用程式（如網頁伺服器，OLAP 和媒體串流），可能不值得損失儲存空間來取得「儲存池預留空間」帶來的效能提升，而可以設定較低的值。

資料縮減（在線資料壓縮 / 在線重複資料刪除）

「在線重複資料刪除」功能可在寫入之前刪除重複的資料，大幅節省儲存空間。在儲存大型檔案時，也支援「在線資料壓縮」來節省空間。這些功能搭配「在線資料壓實」技術，提供大幅最佳化儲存使用率的能力。當使用於全快閃記憶體組態時，對於儲存高重複性資料或大量小檔案，降低儲存需求的能力極有幫助。這不僅大幅提升了全快閃記憶體組態的寫入效能，還可延長 SSD 的使用壽命，最大化部署一般成本效益。

對於重複的虛擬化環境，壓縮率可達 90%。





在線資料壓縮

QuTS hero 支援在線資料壓縮。當資料寫入共用資料夾或 LUN 時，採用較原始表示法更少的位元，對檔案和區塊大小進行編碼，以減少檔案和區塊大小。啟用壓縮也會提升讀取和寫入輸送量，因為需要讀取或寫入的區塊較少。

QuTS hero 採取無損壓縮，可確保資料一致性。ZFS 預設支援 LZ4 壓縮，因為通常最快。

此壓縮不須佔用到過多的 CPU 資源，建議直接用於共用資料夾與 LUN，並且不會對於已經壓縮過的檔案進行壓縮 (例如 MP3, H.264 視訊和 zip 檔案)，因為再壓縮它們能節省的存儲空間少到可以忽略不計。

LZ4 是種無損的資料壓縮演算法，可提供極快的壓縮和解壓縮速度。當傳輸速度高於 100 MB/s 時是低延時和高 IO 的最佳選擇。它特別適用於「在線資料壓縮」。

壓縮速度每秒	解壓縮速度每秒
740 MB	4530 MB

*經 Izbench 測試

在線重複資料刪除

QuTS hero 採用針對目標、區塊層級以及在線重複資料刪除功能。針對目標代表來源端會先將資料傳輸至目標端，然後在目標端執行重複資料刪除程序。區塊層級代表可以偵測並移除檔案內或不同檔案之間的小型資料冗餘區塊。在線則代表著重複資料刪除的程序會在資料傳輸至儲存裝置，還尚未寫入前就立即起到作用。此一組合（針對目標、區塊層級和在線）的優勢，相較於其他重複資料刪除機制，在於減少用戶端所需的 CPU 資源、達到更高的重複資料刪除比率，降低硬碟存取需求。

重複資料刪除程序包含三個步驟：將資料切割成區塊、計算指紋，以及搜尋重複資料刪除表格。首先，ZFS 將檔案切割為小型資料區塊。當寫入資料區塊時，系統會利用密碼總和檢查碼來計算每個資料區塊的指紋，以在現有資料中識別是否包含任何相同區塊。ZFS 會保有涵蓋所有指紋、位置及參考計數的對應表（稱為「重複資料刪除表格」）。如果寫入 NAS 的資料區塊跟現有資料區塊具有相同指紋，不會寫入新位置，ZFS 而會檢查「重複資料刪除表格」、指定相同位置給此資料區塊，並將參考計數加 1。

「在線重複資料刪除」功能可減少額外儲存空間的需求，但在寫入資料時需要更多的運算資源，這可能會影響系統效能。此類效能影響會隨「重複資料刪除表格」的大小增加而增加－由於搜尋資料表的難度也會隨之增加。為了避免在線重複資料刪除期間降低效能，QNAP 採用 SmartDDT（智慧型重複資料刪除表格）演算法，藉由監控效能變化或「重複資料刪除表格」大小，抑制日益增加的「重複資料刪除表格」。當符合某些條件時，QuTS hero 將會停止重複資料刪除服務。如發生這種情況一種可能是因為資料的重複度不高且使用情境不適合「在線重複資料刪除」，或是 NAS 其實需要更多的記憶體。

對於需要進行重複資料刪除的使用情境，我們建議您安裝大量的記憶體，能越多越好，最低需求為 16GB，且強烈建議能配置達 64GB 以上。

SnapSync：

具成本效益的備份、資料保護與災難復原的複製解決方案。

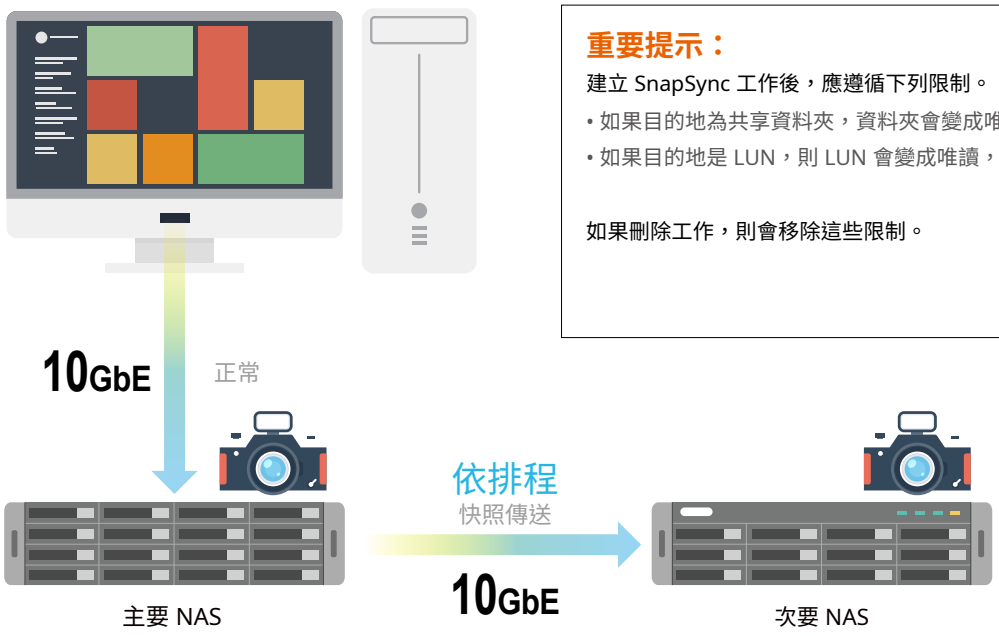
簡介：

SnapSync 內建於 QuTS hero 系統，是針對於備份，資料保護，災難復原等使用情境下的一種極具成本效益且易於使用的整合式資料複製解決方案。

SnapSync 是透過主要和次要儲存系統上共用 /LUN 之間的鏡像關係來設定。它會定期同步複本，使其與已寫入主要儲存的變更保持同步。此鏡像資料是在次要儲存系統中建立。您可以在主要站點發生嚴重損壞時，將資料容錯轉移至次要站點。透過使用 SnapSync，您可以降低整個災難復原解決方案的總成本，尤其當災難復原站點轉換為正式業務使用時，更容易證明您投資的合理性。

SnapSync 的運作方式：

區塊層級的 SnapSync 功能可透過快照來備份可變區塊，其為輕量且節省頻寬方式，可以有效備份重要檔案到遠端位置。使用增量資料複寫時，相同資料永遠不會傳送兩次。SnapSync 的復原時間也比其他方法更快，因為復原只需要完整備份和最後的差異備份。



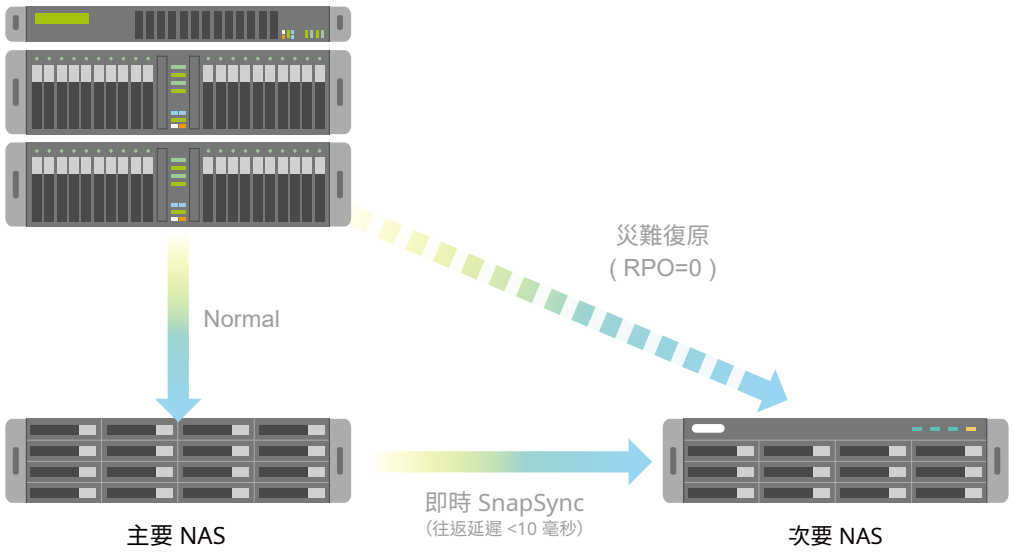
重要提示：

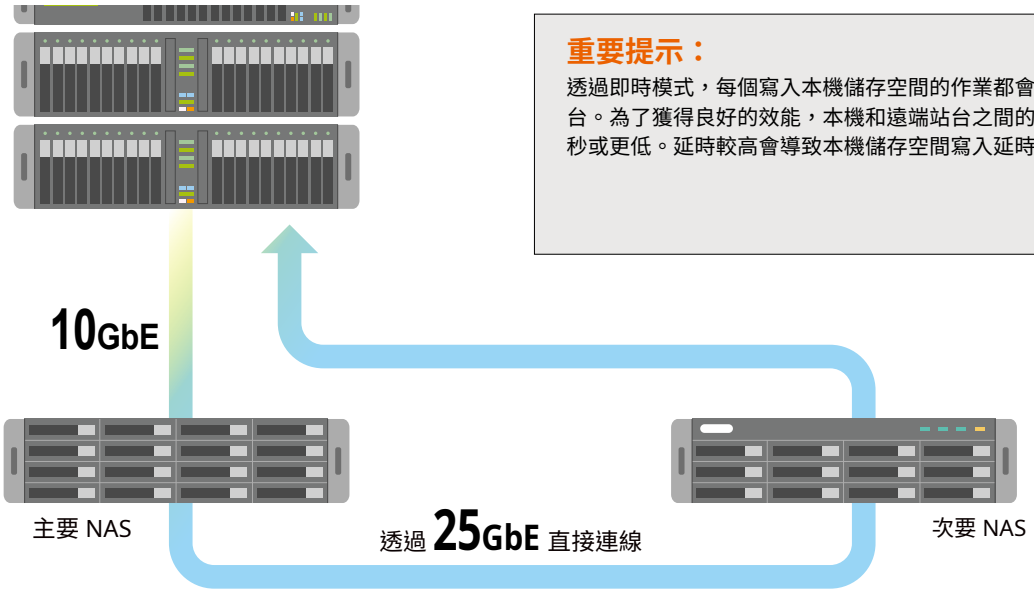
建立 SnapSync 工作後，應遵循下列限制。

- 如果目的地為共享資料夾，資料夾會變成唯讀。
- 如果目的地是 LUN，則 LUN 會變成唯讀，且 iSCSI 啟動器無法存取。

如果刪除工作，則會移除這些限制。

更進階的即時 SnapSync 同步功能，讓主要與次要 NAS 隨時保持同步。





重要提示：
透過即時模式，每個寫入本機儲存空間的作業都會立即複寫到遠端站台。為了獲得良好的效能，本機和遠端站台之間的往返延時必須為 5 毫秒或更低。延時較高會導致本機儲存空間寫入延時。

即時 SnapSync 可確保主要與次要 NAS 的資料完全寫入，然後再回覆主機伺服器。這可確保 RPO=0（復原點目標），也代表次要 NAS 中的資料與主要 NAS 一致。當發生災難時，主機伺服器可以在伺服器容錯移轉時，立即將掛載目標變更為次要 NAS。

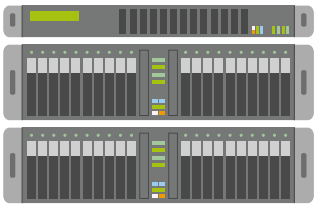
當您建立大型儲存系統時，建議在次要 NAS 設定 SnapSync，然後利用快照備份與 HBS3，將次要 NAS 的多版本備份至另一個備份 NAS。

同時支援 QES NAS（QES 2.1.1 v11 版或更新版本） SnapSync 至 QuTS hero NAS。

發生災難時該怎麼辦

如果發生災難...

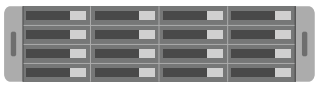
營運伺服器



NAS (共用儲存空間)

首先，刪除 SnapSync 任務 1

主要 NAS
192.168.100.100

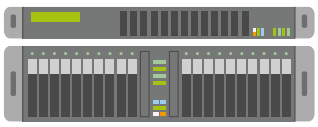


次要 NAS
192.168.100.200

刪除 SnapSync 任務後資料夾的唯讀權限將被移除。

A 重新掛載至次要 NAS IP

營運伺服器



掛載目標 = 192.168.100.200

或

B 將次要 NAS IP 變更為原始主要 NAS IP

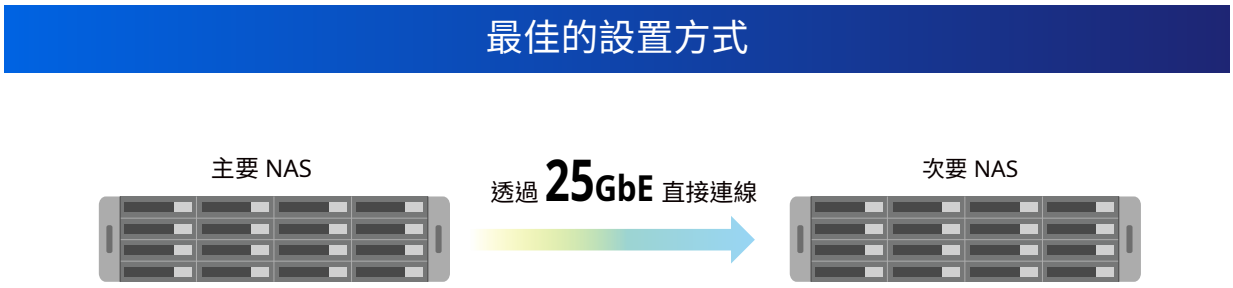
次要 NAS

192.168.100.200 => 192.168.100.100

即時 SnapSync 設定的最佳做法

在即時模式，每個寫入本機儲存空間的作業都會立即複寫到遠端站台。為了獲得良好的效能，本機和遠端站台之間的往返延時必須為 5 毫秒或更低。延時較高會導致本機儲存空間寫入延時。因此，請在設定時注意以下幾點：

- 建議網路環境延時盡可能低於 10 毫秒（小於 5 毫秒最佳）
- 次要 NAS 的 I/O 效能應與主要 NAS 相同。
- 建議您使用 25GbE 直接對接主要與次要 NAS。

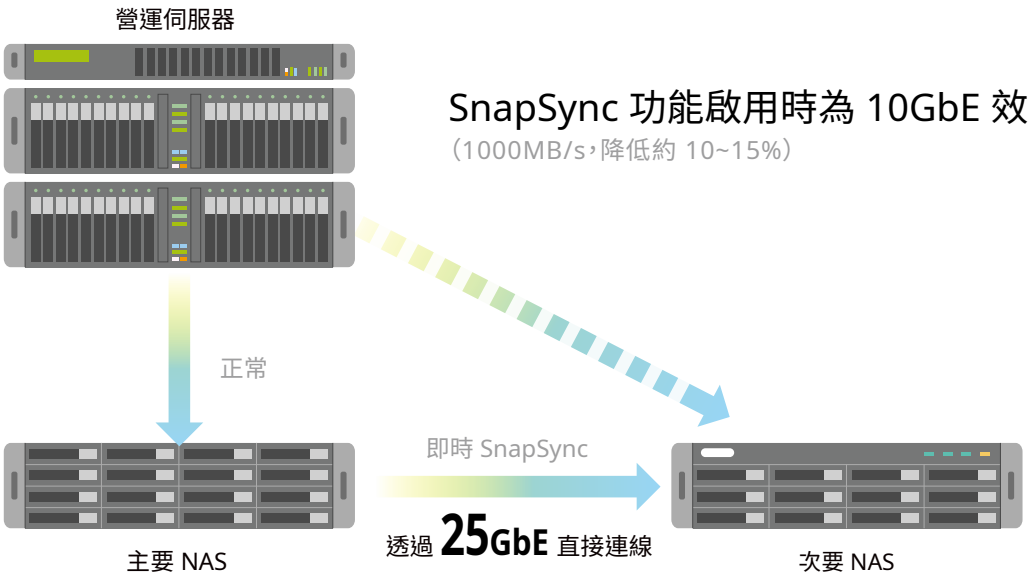


來自 QNAP 實驗室的即時 SnapSync 效能基準測試 ：

- 採用 iSCSI 通訊協定測試：當使用者使用 10GbE 連線至 QuTS 主機 NAS 時，在啟用 SnapSync 前，10GbE 的檔案傳輸速度可達 1127 MB/s，執行即時 SnapSync 的效率為 1005MB/s。大約具有 10~15% 的衰減。
- 採用 SMB 通訊協定測試：當使用者使用 10GbE 連線至 QuTS 主機 NAS 時，在啟用 SnapSync 前，10GbE 的檔案傳輸速度可達 1025 MB/s，執行即時 SnapSync 的效率為 930MB/s。大約具有 10~15% 的衰減。
- 採用 NFS 通訊協定測試：當使用者使用 10GbE 連線至 QuTS 主機 NAS 時，在啟用 SnapSync 前，10GbE 的檔案傳輸速度可達 961 MB/s，執行即時 SnapSync 的效率為 897MB/s。大約具有 7~15% 的衰減。

未受到 SnapSync 保護

啟用即時 SnapSync 後



測試環境：

- IO 模式：同步 = 標準 & 同步 = 無
- 區塊大小 = 128K
- Jumbo Frame (MTU) = 9000
- 當上述測試在相同環境下以 10GbE 連線時，由於網路連線和延時的影響，啟用即時 SnapSync 後的衰減將再次增加，達約 20%。

優點：

減少停機時間並防止資料遺失

SnapSync 提供排程和即時複寫功能。這能使復原點目標 (RPO) 達到秒級。即使對於最大的寫入密集工作負載，您也可以滿足嚴格的復原目標。

降低網路頻寬使用率：

SnapSync 利用壓縮 / 重複資料刪除功能，透過網路傳送增量區塊，以加速資料傳輸並降低頻寬使用率，藉此發揮儲存效率。透過 SnapSync，您可以利用精簡複製資料流來建立單一儲存庫並維護使用中的鏡像。

輕鬆散布大型資料：

有時候您必須傳送大量資料：移轉伺服器機房陣列、合併遠端辦公室或設定新分支。SnapSync 提供快速、高效且靈活的方法來移動資料。如果您的企業分散於各處，並且所有地點都需要存取相同的資料包（例如訓練影片或銷售資料），您可以透過 SnapSync 快速將相同資料散布到所有位置。

CDM（複製資料管理）和資料分析：

執行廣泛的分析對您的企業來說可能很重要，但可能會干擾到營運環境的效能。透過 SnapSync 和快照，您可以利用複製的資料在次要資料副本上執行複雜的分析。

資料保留、法規遵循和多版本需求：

許多使用案例需要較長的資料保留期間。透過結合 SnapSync 和快照，您可以符合法規要求以便保護資料。快照備份可協助降低勒索軟體的影響，並透過即時 SnapSync 提高可用性和災難復原。

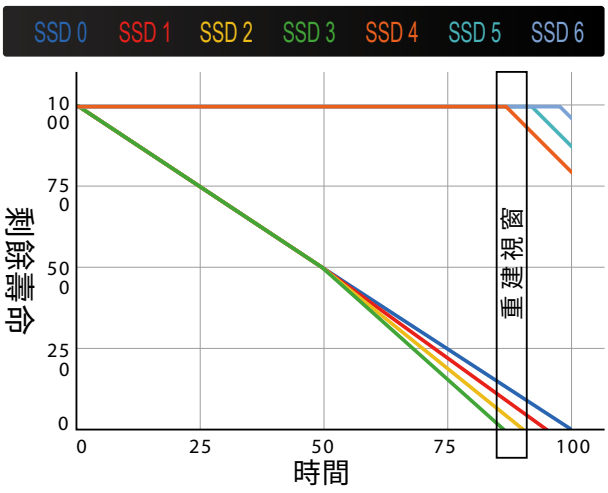
QNAP 專利技術的 QSAL 適用於全快閃儲存陣列與固態硬碟 RAID

專利的 QSAL（QNAP SSD Anti-wear Leveling）演算法會自動啟動來防止多個固態硬碟同時發生故障。

對應於 SSD RAID 5 / 6 / 50 / 60 / TP（三同位元）會預設啟用。

M0 OP = 0

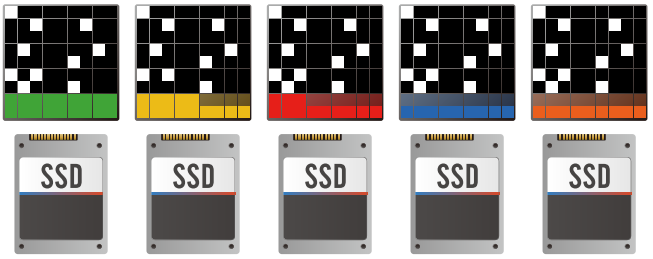
Mi OP = i x W / Mi-1, if (Mi-1 - Mi) > W



像是競爭對手產品所使用的 SSD RAID（特別是僅針對最舊 SSD 提升負載的產品）在後期往往會有特別高的資料遺失風險，因為半數 SSD 太接近使用壽命而來不及重建資料。此外，此演算法無法用於寫入大量大型檔案，而且僅支援 RAID 5，都是一種風險。

QSAL 定期偵測 SSD 的使用壽命及耐用性，當 SSD 使用壽命低於 50% 時，它會動態分配超額配置保留空間的大小，確保每個 SSD 都有足夠的時間在使用壽命結束前進行重建。它在使用率和超額配置保留空間之間達到最佳平衡。QSAL 在啟用後不會影響空間使用率，所導致的效能降低可忽略不計。整體而言，將可大幅提升全快閃儲存陣列的資料保護力。

QSAL 可以隨時啟動，並相容於先前尚未設定的 SSD 同位檢查 RAID。建議您在 SSD 使用壽命達到 50% 之前啟用 QSAL。如果啟用時間太晚，可能會遇到重建時間不足的問題，導致 SSD RAID 損壞的風險。

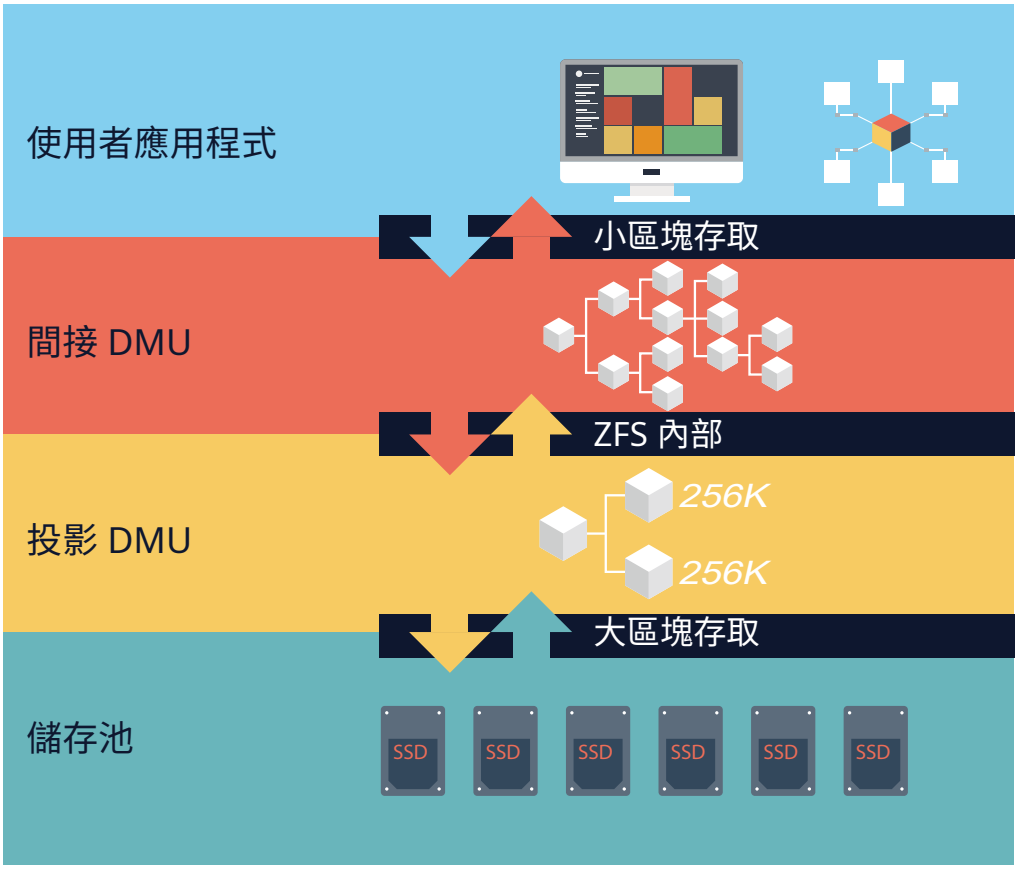


專利編號：US10705744B2 / US20200333959A1。同時包括幾項保密專利。

寫入聚合

「寫入聚合」演算法可提升全快閃陣列的隨機寫入效能

「寫入放大」是 SSD 的一種不良現象，而會對快閃記憶體的效能和耐用性帶來挑戰。QuTS hero 擁有 QNAP 獨家的「寫入聚合」演算法，專為快閃記憶體最佳化而設計，可將所有隨機寫入轉換為連續寫入，同時減少 I/O。它不僅能夠有效提升全快閃記憶體環境的隨機寫入效能，還可以改善 SSD 使用壽命。



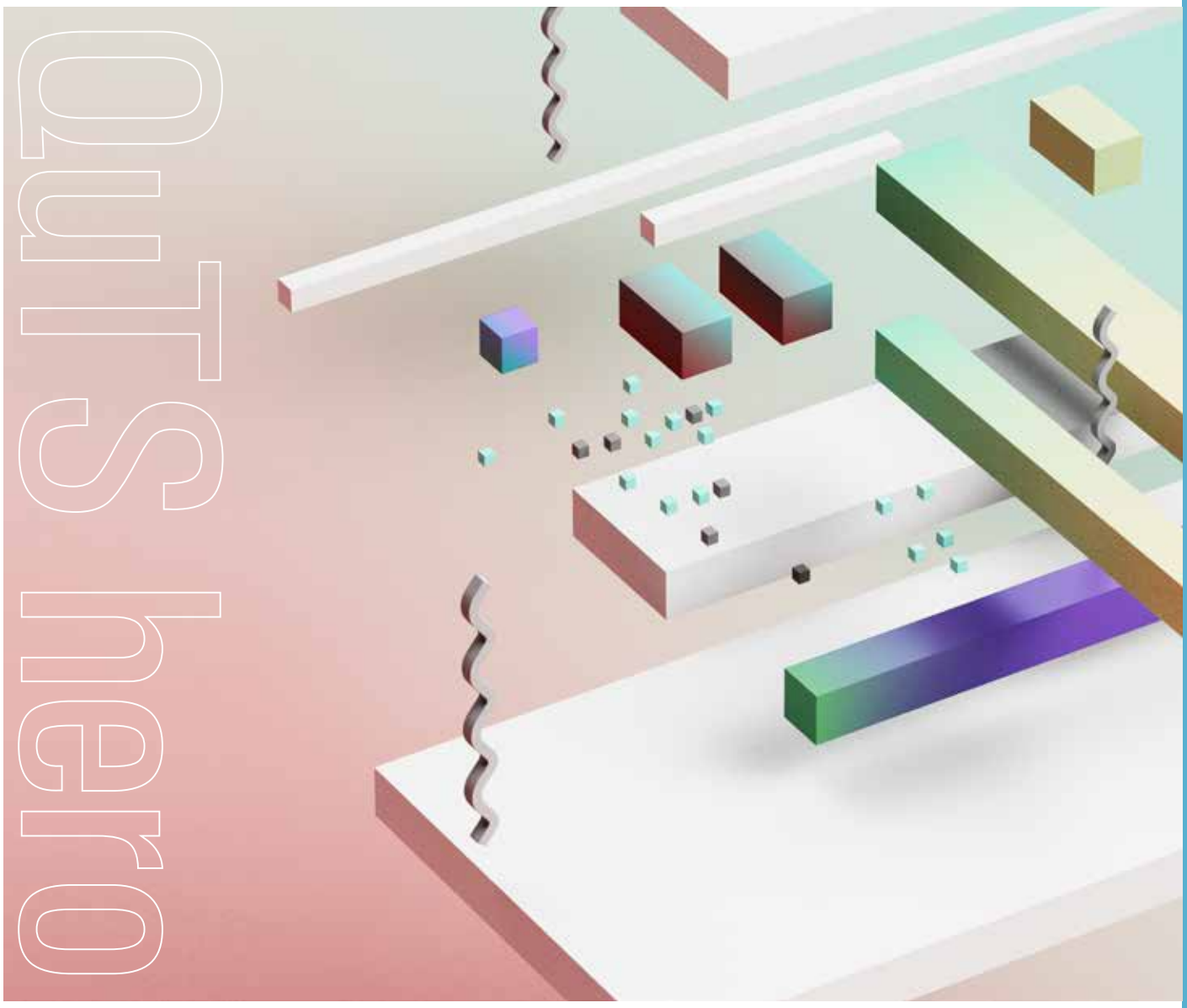
建議的讀取快取設定 (L2ARC)

記憶體在 ZFS 效能中扮演著極為重要的角色，尤其是在高速資料傳輸、重複資料刪除、ARC 和快取方面。建議您盡可能安裝更多記憶體，以獲得 ZFS 效能及最佳化商務負載的最大優勢。

根據 QuTS hero NAS 安裝的記憶體，以及讀取快取的使用比例，我們建議您將讀取快取的大小設定如下：

- **32 GB RAM** 支援 1TB SSD 大小的讀取快取
- **64 GB RAM** 支援 2TB SSD 大小的讀取快取
- **128GB RAM** 支援 4TB SSD 大小的讀取快取
- **256GB RAM** 支援 8TB SSD 大小的讀取快取

- **1TB RAM** 支援約 30TB SSD 的讀取快取。
- **4TB RAM** 支援約 120TB SSD 的讀取快取。


QuTS hero

記憶體大小與儲存池之間的效能相依性

重要提示：本章僅介紹 RAM 和儲存池之間的基本依存性。如果系統正在執行其他專屬服務（例如 Virtualization Station），建議您分開使用量，並另外考量分配給這些虛擬機器的記憶體。

NAS 記憶體－虛擬機器指派的記憶體 = ZFS 已使用記憶體

這篇是分析記憶體與儲存效能的一般概念，以及如何針對高效能儲存識別記憶體大小的最佳作法。接著，我們會說明如何選擇足夠的記憶體來達到最佳效能。

記憶體容量在 ZFS 中扮演著極為重要的角色，對於 QuTS hero 儲存池的表現影響極大。當 ZFS 啟用資料縮減功能（例如壓實、快速複製和重複資料刪除）時，系統需要一些表格來記錄這些資料區塊的參考計數。因此，當資料量變大時，ZFS 必須在記憶體中保留盡可能多的參考資料表，以便能夠以更有效的方式處理 IO 要求。

其次，ZFS 支援資料保護的交易群組功能。每個檔案修改（例如寫入）都與某個交易群組（TXG）關聯。每個 TXG 會在固定間隔關閉，並且儲存池會針對該群組要求同步作業。當 ARC 表示目前已快取過多的「已變更」記憶體時，也可能會關閉 TXG。當一個 TXG 關閉時，另一個 TXG 隨即立即開啟，檔案修改於是跟新的作用中 TXG 產生關聯。

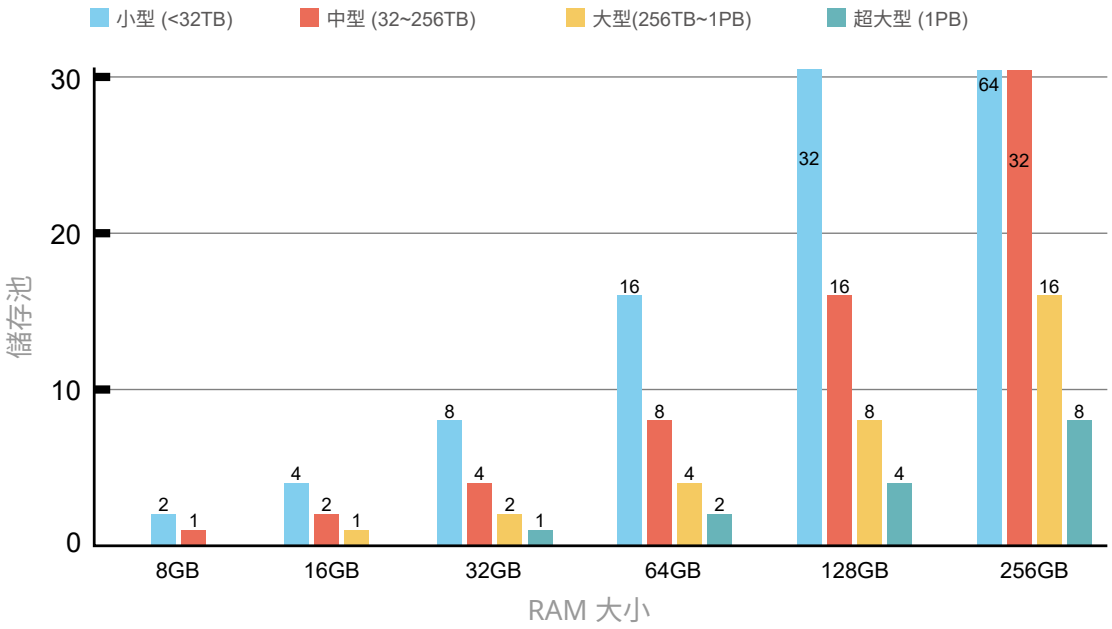
如果作用中的 TXG 關閉時，先前的 TXG 仍在同步資料至儲存裝置，則應用程式會受到限制，直到執行同步完成為止。在此情況下，TXG 正在同步，而 TXG + 1 因記憶體限制而關閉且正在等待同步；應用程式會限制等待寫入 TXG + 2。儲存空間需要持續耗用或是限制記憶體來抑制應用程式。

儲存池的同步包含將所有層級 0 的資料區塊傳送至硬碟。當完成時，會傳送所有層級 1 的間接區塊（諸如此類），直到代表檔案系統新狀態的每個區塊已認可為止。此時 uberblock 會更新為指向儲存池中新的一致狀態。

因此，ZFS 必須在完成每個交易後更新 uberblock，這是個極大的額外負荷。如果系統能夠盡可能提供更多記憶體，交易群組可以更大，並且更新 uberblock 的額外負荷將變得相對較小。因此，更多的記憶體可以有效幫助 ZFS 改善寫入 IO 的要求。

您可以參考以下建議，並為不同的儲存設定選擇足夠的記憶體。

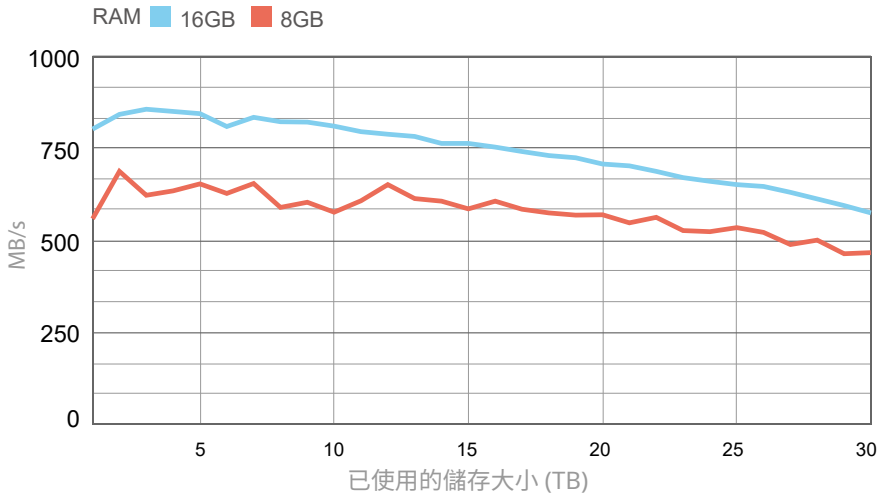
RAM 大小與儲存池配置之間的最佳效能作法



例如，如果您的 ZFS NAS 具有 16GB 記憶體，則建議使用的儲存池組態為 1 個大型儲存池（256TB~1PB）或 2 個中型儲存池，或是 4 個小型儲存池。這種類型的儲存池組態可以讓您的系統達到良好的效能。如果您建立的儲存池數量超過這些建議，NAS 仍然能夠可靠運作，但效能可能會降低。在這裡我們以入門級的 ZFS 機種 TS-973AX 作為參考，以顯示 8GB（最低需求）和 16GB RAM 之間的不同效能。

（我們已經先建立了一個 SSD 儲存池跟一個約 35TB 的 HDD 儲存池，來呈現 HDD 儲存池的效能差異）。

TS-h973AX-16GB 對比 8GB - IO 效能 (5xHDD-R0)



QuTS hero
ZFS
白　皮　書

QNAP SYSTEMS, INC.

TEL : +886-2-2641-2000 FAX: +886-2-2641-0555 Email: qnapsales@qnap.com

Address : 3F, No.22, Zhongxing Rd., Xizhi Dist., New Taipei City, 221, Taiwan

QNAP may make changes to specification and product descriptions at any time, without notice.
Copyright © 2021 QNAP Systems, Inc. All rights reserved.
QNAP® and other names of QNAP Products are proprietary marks or registered trademarks of QNAP Systems, Inc.
Other products and company names mentioned herein are trademarks of their respective holders.

Netherlands (Warehouse Services)
Email: nlsales@qnap.com
TEL: +31(0)107600830

China
Email: cnsales@qnap.com
TEL: +86-400-028-0079

Thailand
Email: thsales@qnap.com
TEL: +66-2-5415988

Japan
Email: jpsales@qnap.com
FAX: 03-6435-9686

US
Email: usasales@qnap.com
TEL: +1-909-595-2782

India
Email: indiasales@qnap.com

Germany
Email: desales@qnap.com

France
Email: Frsales@qnap.com