QNAP

# QuTS hero
# ZFS

白　　皮　　书

# INDEX


| 页 | 标题 |
| --- | --- |
| 01 | **概览** |
| 02 | **ZFS 简介与设置** |
| 02 | 数据完整性与自我修复 |
| 02 | 建议的共享文件夹 /LUN 区块大小 |
| 03 | ZIL 的同步 I/O（自动 / 全部 / 无） |
| 04 | 存储池预留空间 |
| 06 | **数据缩减（在线数据压缩 / 在线重复数据删除）** |
| 07 | 在线数据压缩 |
| 08 | 在线重复数据删除 |
| 09 | **SnapSync**<br>具成本效益的备份、数据保护与灾难复原的复制解决方案。 |
| 09 | 简介 |
| 09 | SnapSync 的运作方式 |
| 11 | 发生灾难时该怎么办 |
| 12 | 实时 SnapSync 设置的最佳做法 |
| 14 | 优势 |
| 15 | **QNAP 专利技术的 QSAL 适用于全闪存数组与固态硬盘 RAID** |
| 16 | **写入聚合**<br>写入聚合算法可提升全快闪数组的随机写入性能 |
| 17 | **建议的读取快取设置 (L2ARC)** |
| 19 | **内存大小与存储池之间的性能相依性** |

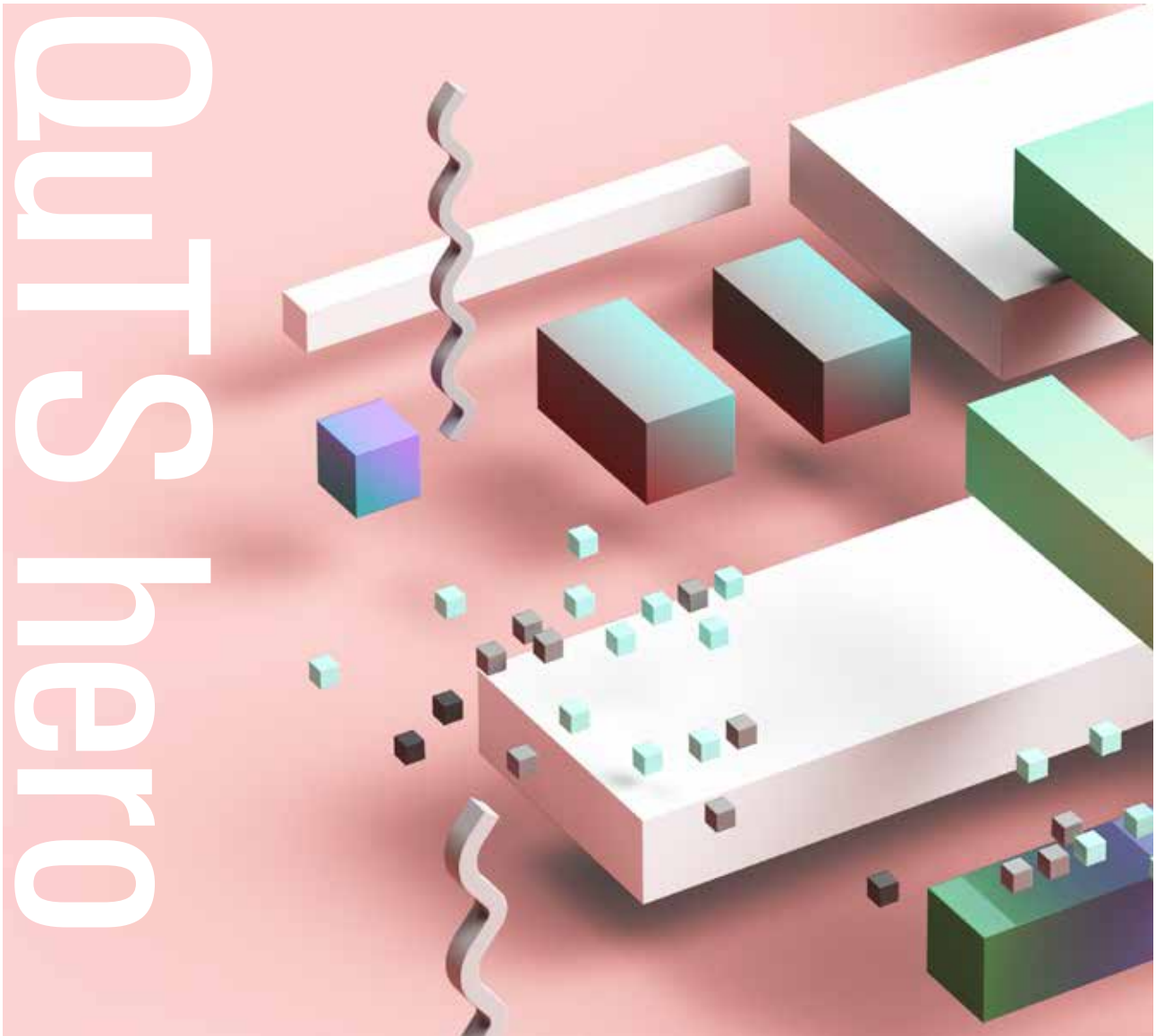
# QuTS hero ZFS

白 皮 书

# 概览

在现今 IT 密集的世界中，快速应变的需求可能是企业组织和数据中心成功或失败的运作关键。在这种环境下，预算都会受到严格的审查，而任何潜在的停机时间 - 无论是几分钟还是几秒钟 - 都可能会有灾难性。任何可能的效率及生产力提升都是成功的重要基石。

QuTS hero 操作系统采用 ZFS 文件系统，其以数据完整性为优先处理，满足高性能和高稳定性的数据备份需求。QuTS hero 整合了高性能数据保护、数据缩减和虚拟化应用程序，并支持云端环境和 QoS（服务质量）管理功能，为小型企业、企业级数据中心及虚拟化作业提供更具成本效益的存储设备解决方案。

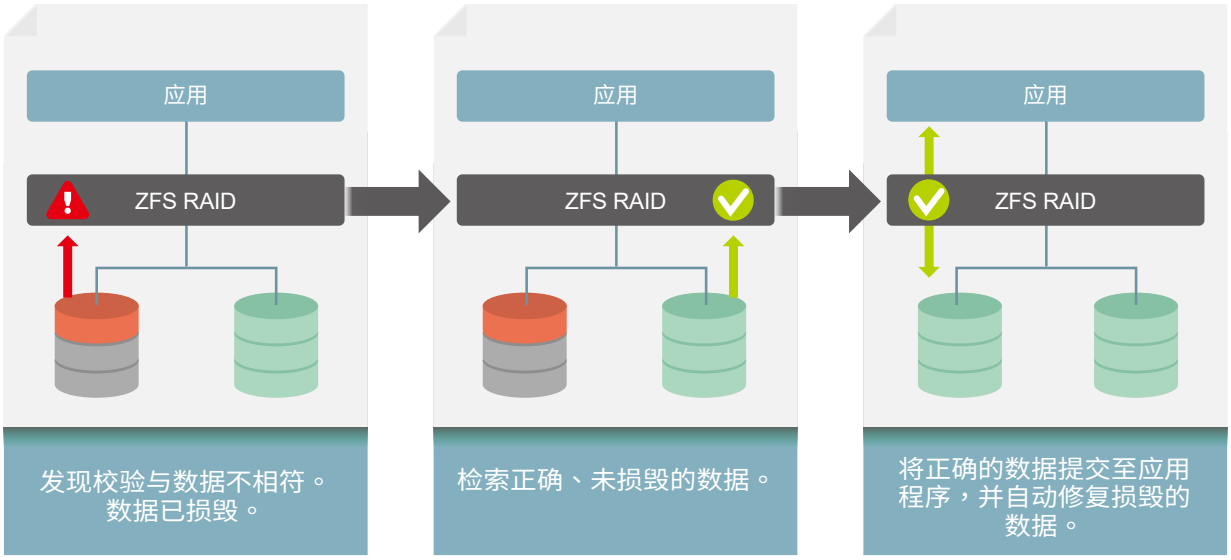


# ZFS 简介与设置

## 数据完整性与自我修复

数据的完整性及可靠性是确保应用程序和数据库运作的关键。ZFS 能够防止静态数据损毁。这是由于硬件问题、缆线损坏或元数据错误所造成的潜在灾难，同时也难以侦测。ZFS 能够自行修复此类数据，包括检查所有数据区块，自动修复不正确的区块。

强烈建议定期执行「存储池清洗」，确保存储数据的完整性，并在潜在问题出现之前予以修复。您可以设置排程来执行存储池清洗，例如以每周或每月的频率。




防止运行中的系统出现静态数据损毁

## 建议的共享文件夹 /LUN 区块大小：

**128K** - 影像剪辑 / 大型文件 / 备份
**64K** - 一般 /Hyper-V
**32K** - VMware
**8k** - OLTP
**4K** -VDI/ 数据库 / 小型文件

# ZIL 同步 I/O（标准 / 一律 / 不使用）

ZIL（ZFS Intent Log）是写入日志机制，可追踪尚未提交至文件系统的文件变更。ZIL 会将这些变更的意图记录下来，称为日志。如果系统当机或电源故障，QuTS hero 会检查日志记录并套用排程变更，确保快速还原文件系统，降低数据损毁的概率。

当建立共享文件夹 /LUN 时，必须设置 ZIL 写入记录 I/O 模式，以排定数据完整性或性能的优先级。

标准（预设）：根据应用程序与 I/O 要求类型选择同步或异步 I/O 交易。

一律：所有 I/O 交易皆以同步进行，这会改善数据完整性，但会降低性能。

不使用：所有 I/O 交易皆以异步进行，此选项提供最佳性能，但可能会因为停电而导致数据遗失。

请注意，建议在专用的高速存储设备（例如 NVMe 固态硬盘）上设置专用 ZIL 作为 SLOG。

对于 ZIL，建议采用超低延迟和高耐用性磁盘（例如 MLC 磁盘驱动器）。

对于读取快取，应优先考虑吞吐量。

# 存储池预留空间

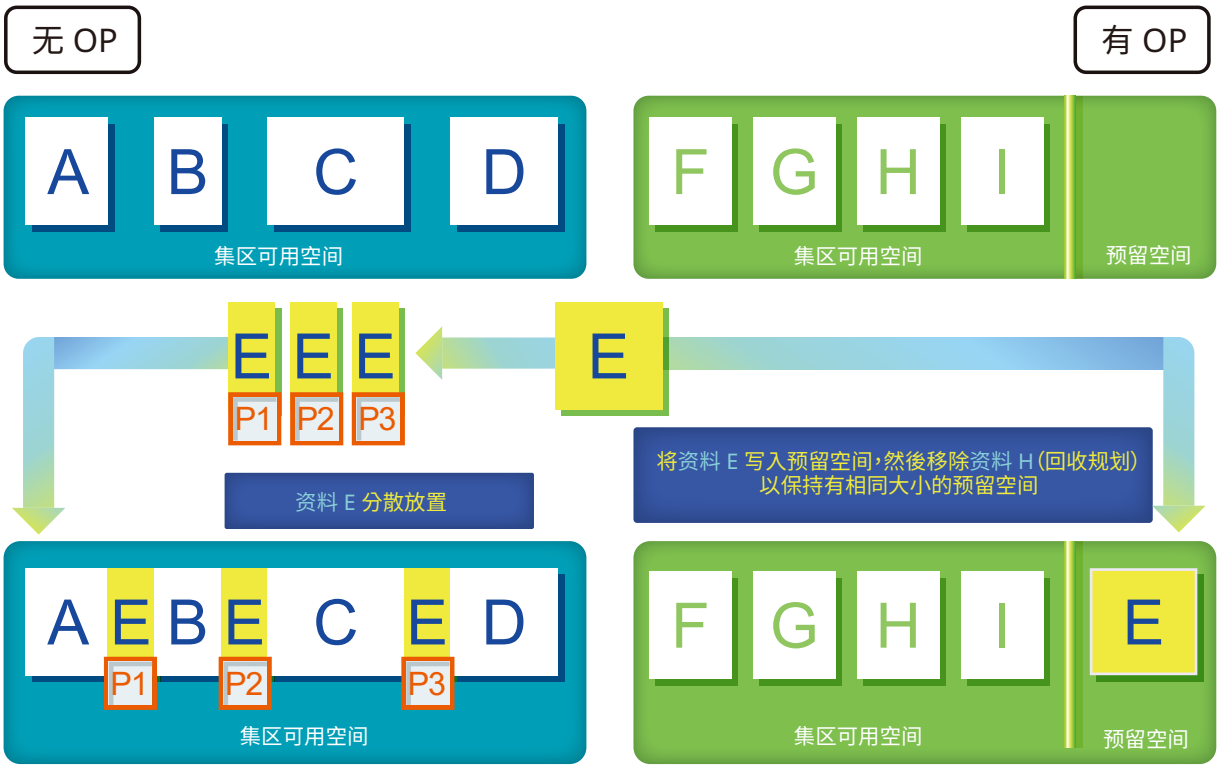
当存储池几乎已满或破碎时，ZFS 性能可能会受到影响。这是由于写入时的复制行为：ZFS 一律会将新数据写入不同区块，更新元数据以指向新位置，然后释放旧区块，而非直接覆写旧数据。在多次循环写入之后，存储池的数据分布会变得破碎，这会使得写入新数据变得更加困难及耗时。

QuTS hero 支持「存储池预留空间」，可让您指定并保留存储池的空间百分比，在存储池即将用完的时候，确保一致的随机写入性能。

下图说明写入存储池的数据时，使用或未使用「存储池预留空间」的情况。如果未使用「存储池预留空间」时，当存储池即将用完且破碎时，每次分隔出新的数据写入区块将会花费一段时间，导致影响性能。通过「存储池预留空间」，ZFS 可以找到连续的大区块来存储新数据。

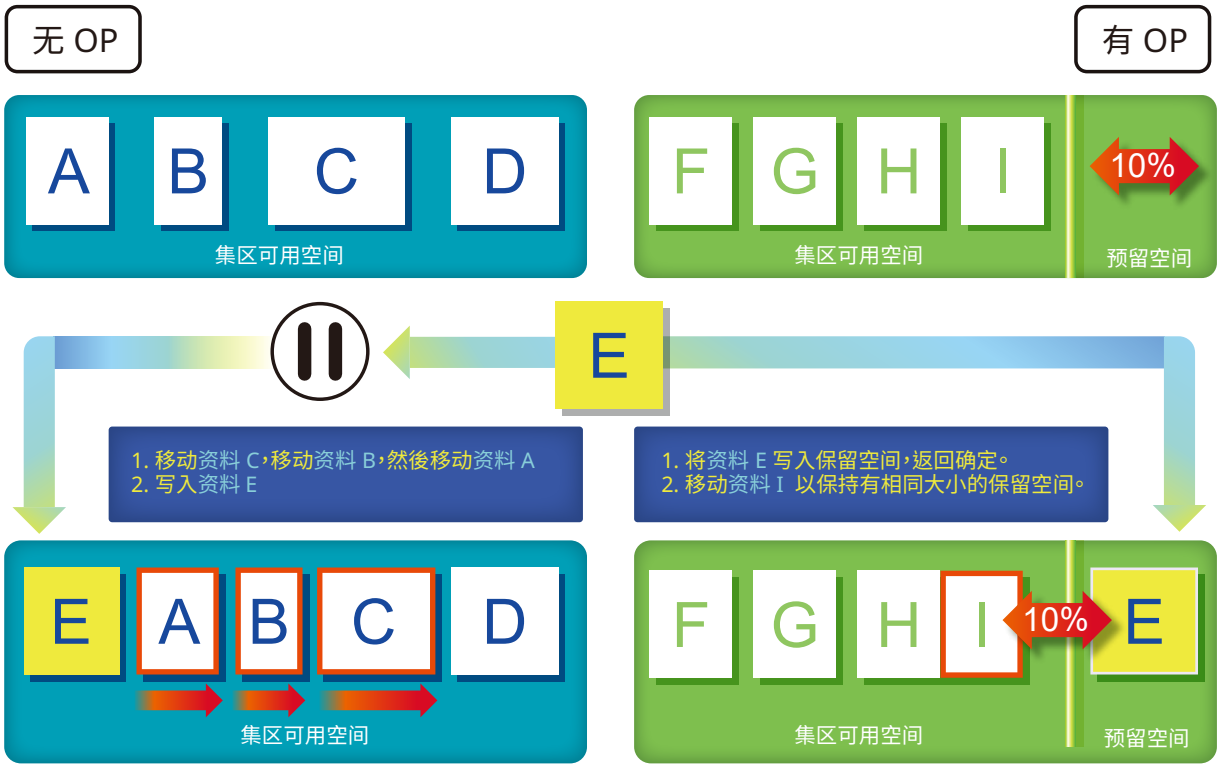
## 预留空间

改善破碎集区的性能（例如：将区块写入 HDD 时）



## 预留空间

改善破碎集区的性能（例如：将区块写入 SSD 时）




改善破碎集区的性能（例如：将区块写入 SSD 时）启用「存储池预留空间」会确保随机写入性能，但同时会影响存储容量。保留的空间越多，性能就越高。确定权衡必要性的关键是考虑读 / 写比率和应用场景。写入越多，「存储池预留空间」提供的性能提升就越高。

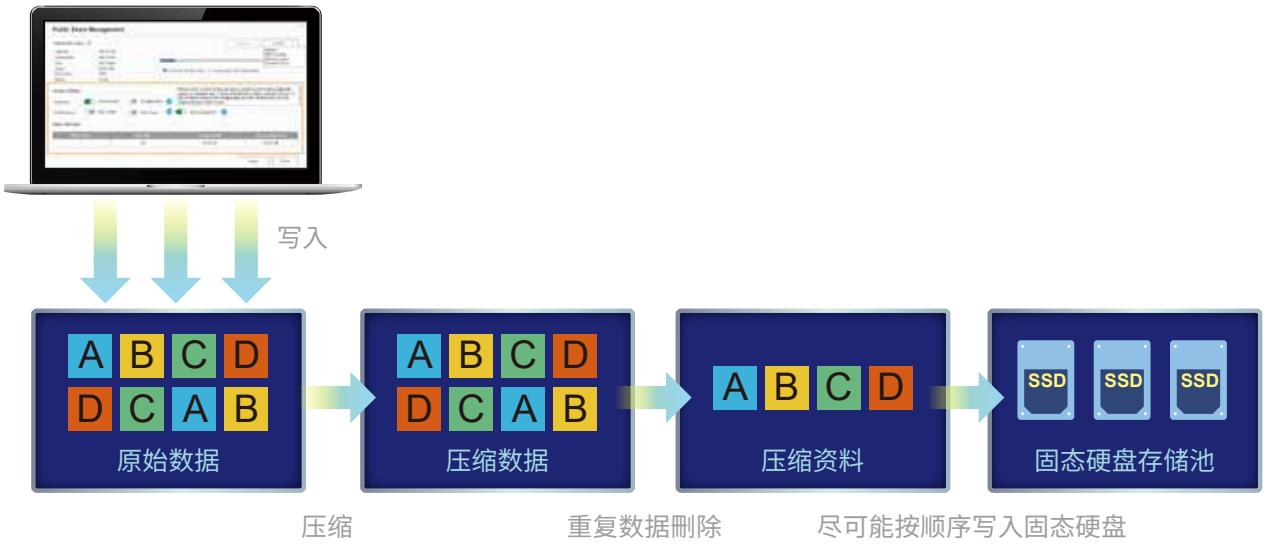
通常建议将「存储池预留空间」值设置为 10-20%。对于写入密集且需要高性能的应用（例如 SQL 记录、备份、监控及医疗影像），建议您设置更高的值。

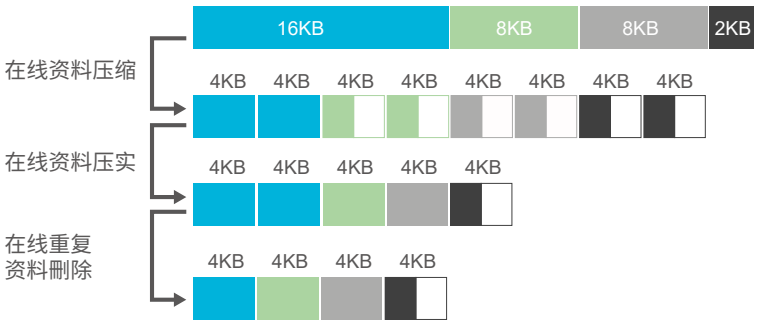
至于读取密集型应用程序（如网页服务器，OLAP 和媒体串流），可能不值得损失存储空间来取得「存储池预留空间」带来的性能提升，而可以设置较低的值。

# 数据缩减（在线数据压缩 / 在线重复数据删除）

「在线重复数据删除」功能可在写入之前删除重复的数据，大幅节省存储空间。在存储大型文件时，也支持「在线数据压缩」来节省空间。这些功能搭配「在线数据压实」技术，提供大幅优化存储使用率的能力。当使用于全闪存组态时，对于存储高重复性数据或大量小文件，降低存储需求的能力极有帮助。这不仅大幅提升了全闪存组态的写入性能，还可延长 SSD 的使用寿命，最大化部署一般成本效益。

对于重复的虚拟化环境，压缩率可达 90%。



# 在线数据压缩

QuTS hero 支持在线数据压缩。当数据写入共享文件夹或 LUN 时，采用较原始表示法更少的位，对文件和区块大小进行编码，以减少文件和区块大小。启用压缩也会提升读取和写入吞吐量，因为需要读取或写入的区块较少。

QuTS hero 采取无损压缩，可确保数据一致性。ZFS 预设支持 LZ4 压缩，因为通常最快。

此压缩不需占用到过多的 CPU 资源，建议直接用于共享文件夹与 LUN。并且不会对于已经压缩过的文件进行压缩（例如 MP3, H.264 视讯和 zip 文件），因为再压缩它们能节省的存储空间少到可以忽略不计。

LZ4 是种无损的数据压缩算法，可提供极快的压缩和解压缩速度。当传输速度高于 100 MB/s 时是低延迟和高 IO 的最佳选择。它特别适用于「在线数据压缩」。

压缩速度每秒

740 MB

解压缩速度每秒

4530 MB

*经 Izbench 测试

# 在线重复数据删除

QuTS hero 采用针对目标、区块层级以及在线重复数据删除功能。针对目标代表来源端会先将数据传输至目标端，然后在目标端执行重复数据删除程序。区块层级代表可以侦测并移除文件内或不同文件之间的小型数据冗余区块。在线则代表着重复数据删除的程序会在数据传输至存储设备，还尚未写入前就立即起到作用。此一组合（针对目标、区块层级和在线）的优势，相较于其他重复数据删除机制，在于减少客户端所需的 CPU 资源、达到更高的重复数据删除比率，降低硬盘存取需求。

重复数据删除程序包含三个步骤：将数据切割成区块、计算指纹，以及搜索重复数据删除表格。首先，ZFS 将文件切割为小型数据区块。当写入数据区块时，系统会利用密码校验来计算每个数据区块的指纹，以在现有数据中识别是否包含任何相同区块。ZFS 会保有涵盖所有指纹、位置及参考计数的对应表（称为「重复数据删除表格」）。如果写入 NAS 的数据区块跟现有数据区块会有相同指纹，不会写入新位置，ZFS 而会检查「重复数据删除表格」、指定相同位置给此数据区块，并将参考计数加 1。

「在线重复数据删除」功能可减少额外存储空间的需求，但在写入数据时需要更多的运算资源，这可能会影响系统性能。此类性能影响会随「重复数据删除表格」的大小增加而增加 - 由于搜索数据表的难度也会随之增加。为了避免在线重复数据删除期间降低性能，QNAP 采用 SmartDDT（智能型重复数据删除表格）算法，藉由监控性能变化或「重复数据删除表格」大小，抑制日益增加的「重复数据删除表格」。当符合某些条件时，QuTS hero 将会停止重复数据删除服务。如发生这种情况，一种可能是因为数据的重复度不高且使用场景不适合「在线重复数据删除」，或是 NAS 其实需要更多的内存。

对于需要进行重复数据删除的使用场景，我们建议您安装大量的内存，越多越好。最低需求为 16GB，且强烈建议配置为 64GB 以上。

# SnapSync：

具成本效益的备份、数据保护与灾难复原的复制解决方案。

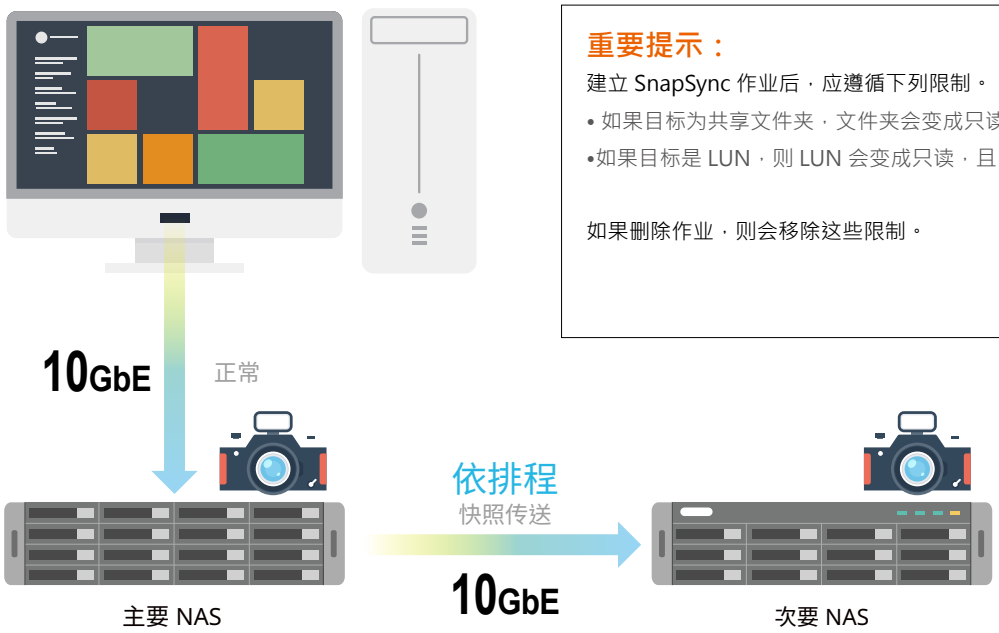
## 简介：

SnapSync 内置于 QuTS hero 系统，是针对于备份、数据保护、灾难复原等使用场景下的一种极具成本效益且易于使用的整合式数据复制解决方案。

SnapSync 是通过主要和次要存储系统上共享 /LUN 之间的镜像关系来设置。它会定期同步复本，使其与已写入主要存储的变更保持同步。此镜像数据是在次要存储系统中建立。您可以在主要站点发生严重损坏时，将数据容错转移至次要站点。通过使用 SnapSync，您可以降低整个灾难复原解决方案的总成本，尤其当灾难复原站点转换为正式业务使用时，更容易证明您投资的合理性。

## SnapSync 的运作方式：

区块层级的 SnapSync 功能可通过快照来备份可变区块，其为轻量且节省带宽方式，可以有效备份重要文件到远程位置。使用增量数据复制时，相同数据永远不会发送两次。SnapSync 的复原时间也比其他方法更快，因为复原只需要完整备份和最后的差异备份。



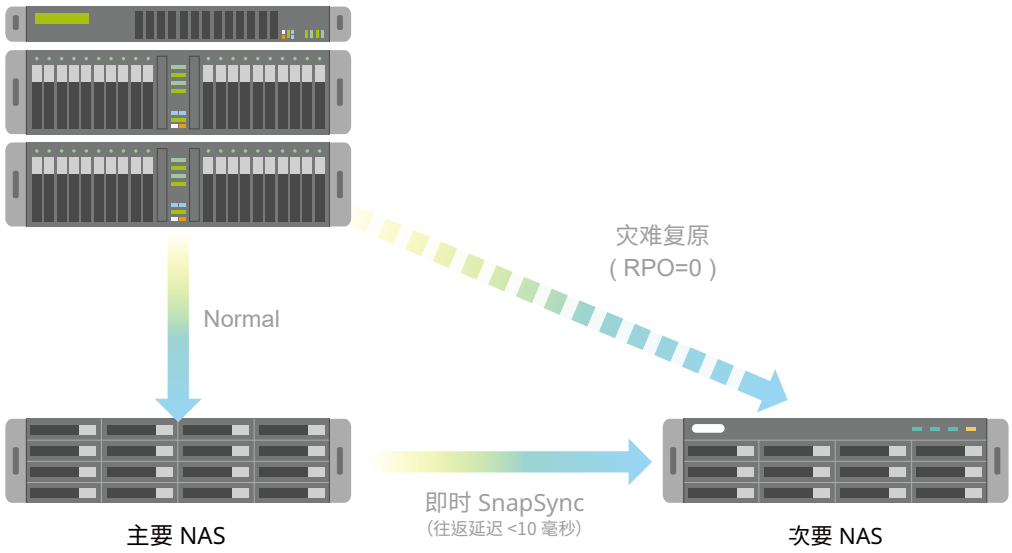

**重要提示：**

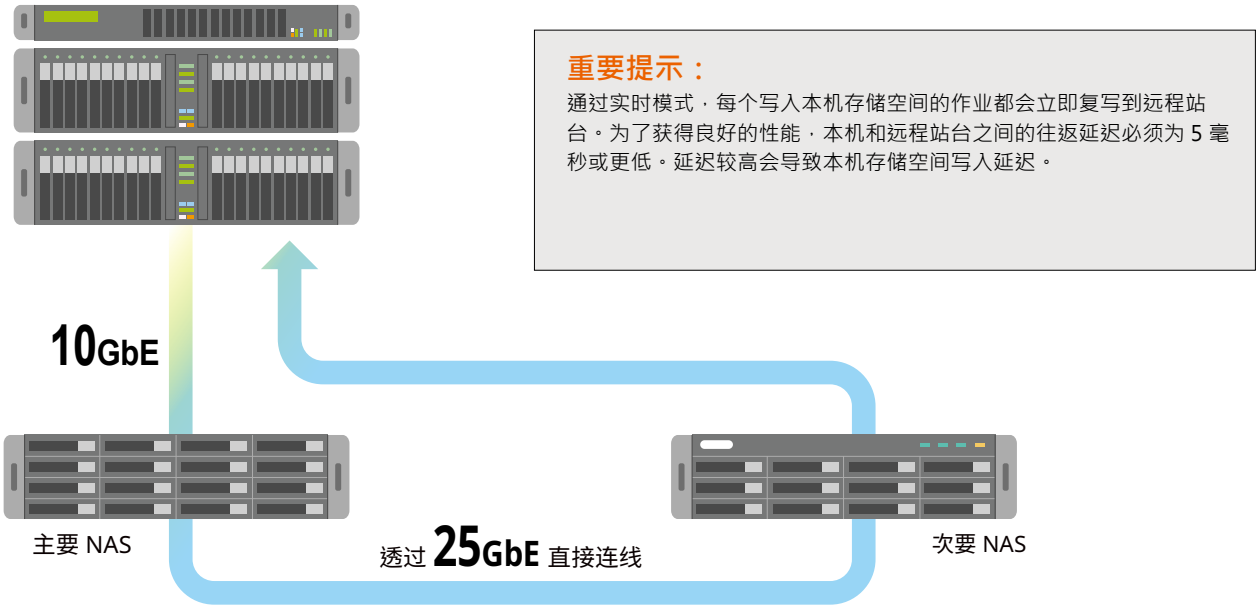
建立 SnapSync 作业后，应遵循下列限制。

- 如果目标为共享文件夹，文件夹会变成只读。
- 如果目标是 LUN，则 LUN 会变成只读，且 iSCSI 启动器无法存取。

如果删除作业，则会移除这些限制。

更进阶的实时 SnapSync 同步功能，让主要与备用 NAS 随时保持同步。




**重要提示：**
通过实时模式，每个写入本机存储空间的作业都会立即复写到远程站台。为了获得良好的性能，本机和远程站台之间的往返延迟必须为 5 毫秒或更低。延迟较高会导致本机存储空间写入延迟。

实时 SnapSync 可确保主要与备用 NAS 的数据完全写入，然后再回复主机服务器。这可确保 RPO=0（恢复点目标），也代表备用 NAS 中的数据与主要 NAS 一致。当发生灾难时，主机服务器可以在服务器故障转移时，立即将挂载目标变更为备用 NAS。

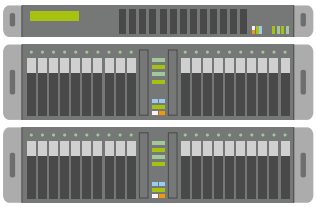
当您建立大型存储系统时，建议在备用 NAS 设置 SnapSync，然后利用快照备份与 HBS3，将备用 NAS 的多版本备份至另一个备份 NAS。

**同时支持 QES NAS（QES 2.1.1 v11 版或更新版本）SnapSync 至 QuTS hero NAS。**

# 发生灾难时该怎么办

## 如果发生灾难...

营运伺服器



NAS（共用储存空间）

## 首先，删除 SnapSync 任务 1

主要 NAS
192.168.100.100

次要 NAS
192.168.100.200

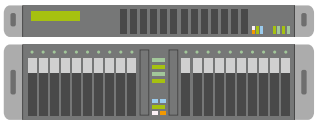
删除SnapSync任务时，文件夹的独占权限将被删除

## A 重新挂载至备用 NAS IP

或

## B 将备用 NAS IP 变更为原始主要 NAS IP

营运伺服器



挂载目标 = 192.168.100.200

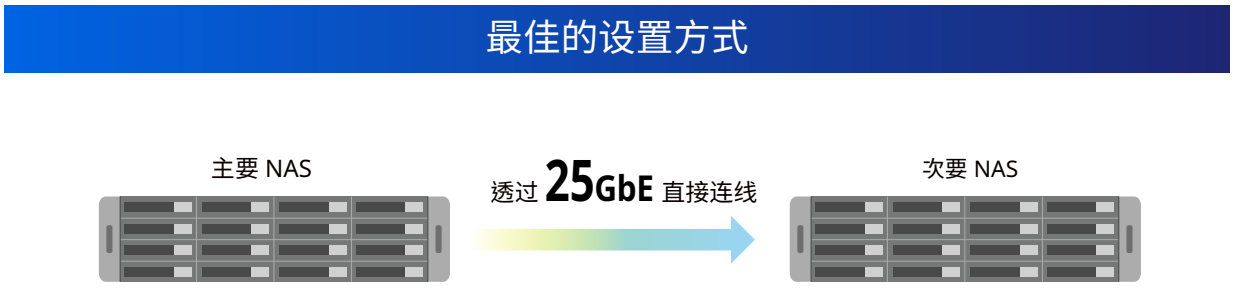
次要 NAS

192.168.100.200 => 192.168.100.100

# 实时 SnapSync 设置的最佳做法

在实时模式，每个写入本机存储空间的作业都会立即复写到远程站台。为了获得良好的性能，本机和远程站台之间的往返延迟必须为 5 毫秒或更低。延迟较高会导致本机存储空间写入延迟。因此，请在设置时注意以下几点：

- 建议网络环境延迟尽可能低于 10 毫秒（小于 5 毫秒最佳）
- 备用 NAS 的 I/O 性能应与主要 NAS 相同
- 建议您使用 25GbE 直接对接主要与备用 NAS。

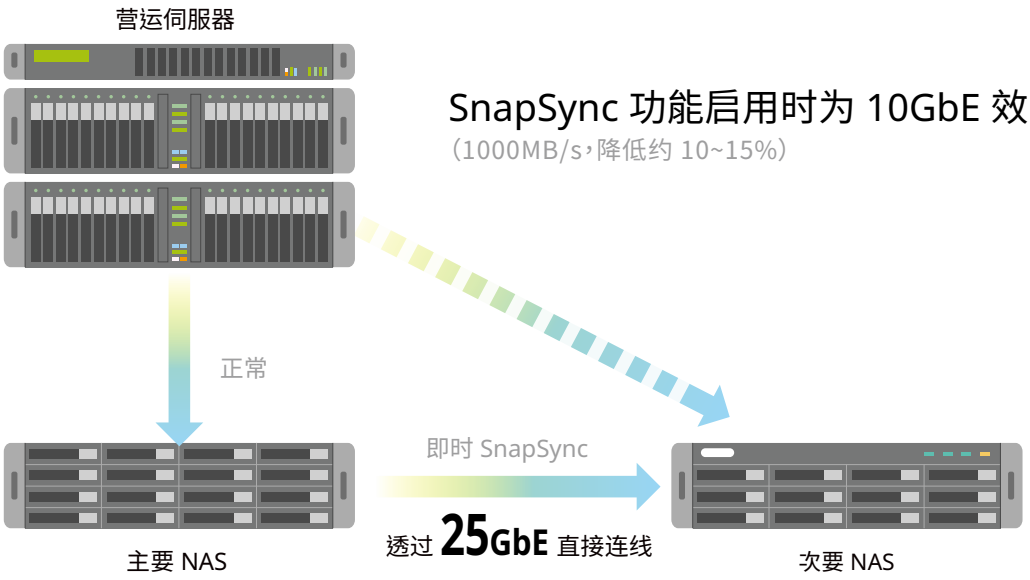



## 来自 QNAP 实验室的实时 SnapSync 性能基准检验 ：

- 采用 iSCSI 通讯协议测试：当用户使用 10GbE 联机至 QuTS 主机 NAS 时，在启用 SnapSync 前，10GbE 的文件传输速度可达 1127 MB/s，执行实时 SnapSync 的效率为 1005MB/s。大约会有 10~15% 的衰减。
- 采用 SMB 通讯协议测试：当用户使用 10GbE 联机至 QuTS 主机 NAS 时，在启用 SnapSync 前，10GbE 的文件传输速度可达 1025 MB/s，执行实时 SnapSync 的效率为 930MB/s。大约会有 10~15% 的衰减。
- 采用 NFS 通讯协议测试：当用户使用 10GbE 联机至 QuTS 主机 NAS 时，在启用 SnapSync 前，10GbE 的文件传输速度可达 961 MB/s，执行实时 SnapSync 的效率为 897MB/s。大约会有 7~15% 的衰减。

## 未受到 SnapSync 保护

## 启用即时 SnapSync 後




**测试环境：**

- IO 模式：同步 = 标准 & 同步 = 无
- 区块大小 = 128K
- Jumbo Frame (MTU) = 9000
- 当上述测试在相同环境下以 10GbE 联机时，由于网络联机和延迟的影响，启用实时 SnapSync 后的衰减将再次增加，约为 20%。

# 优点：

## 减少停机时间并防止数据遗失

SnapSync 提供排程和实时复写功能。这能使恢复点目标 (RPO) 达到秒级。即使对于最大的写入密集作业负载，您也可以满足严格的复原目标。

## 降低网络带宽使用率：

SnapSync 利用压缩 / 重复数据删除功能，通过网络传送增量区块，以加速数据传输并降低带宽使用率，借此发挥存储效率。通过 SnapSync，您可以利用精简复制数据流来建立单一存储库并维护使用中的镜像。

## 轻松发送大型数据：

有时候您必须传送大量数据：转移服务器机房数组、合并远程办公室或设置新分支。SnapSync 提供快速、高效且灵活的方法来移动数据。如果您的企业分散于各处，并且所有地点都需要存取相同的数据报（例如训练影片或销售数据），您可以通过 SnapSync 快速将相同数据发送到所有位置。

## CDM（复制数据管理）和数据分析：

运行广泛的分析对您的企业来说可能很重要，但可能会干扰到营运环境的性能。通过 SnapSync 和快照，您可以利用复制的数据在次要数据副本上执行复杂的分析。

## 数据保留、法规遵循和多版本需求：

许多使用案例需要较长的数据保留期间。通过结合 SnapSync 和快照，您可以符合法规要求以便保护数据。快照备份可协助降低勒索软件的影响，并通过实时 SnapSync 提高可用性和灾难复原。

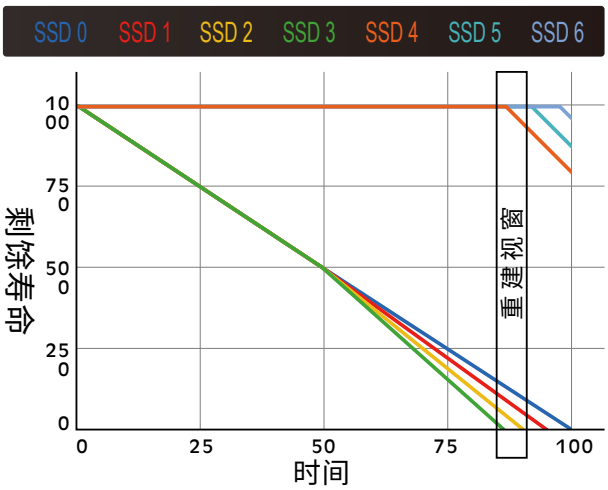
# QNAP 专利技术的 QSAL 适用于全快闪存储数组与固态硬盘 RAID

获得专利的 QSAL（QNAP SSD Anti-wear Leveling）算法会自动启动来防止多个固态硬盘同时发生故障。

对应于 SSD RAID 5 / 6 / 50 / 60 / TP（三同位）会默认启用

**M0 OP = 0**

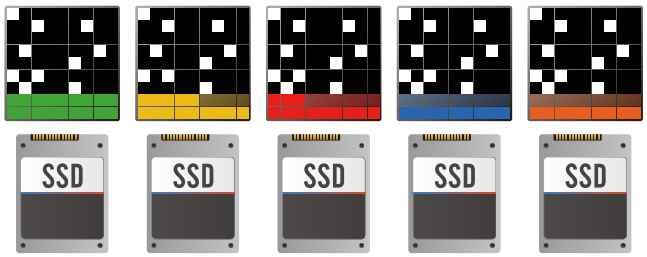
**Mi OP = i x W / Mi-1, if (Mi-1 - Mi) > W**




像是竞争对手产品所使用的 SSD RAID（特别是仅针对最旧 SSD 提升负载的产品）在后期往往会有特别高的数据遗失风险，因为半数 SSD 太接近使用寿命而来不及重建数据。此外，此算法无法用于写入大量大型文件，而且仅支持 RAID 5，这也是一种风险。

QSAL 定期侦测 SSD 的使用寿命及耐用性，当 SSD 使用寿命低于 50% 时，它会动态分配超额配置预留空间的大小，确保每个 SSD 都有足够的时间在使用寿命结束前进行重建。它在使用率和超额配置预留空间之间达到最佳平衡。QSAL 在启用后不会影响空间使用率，所导致的性能降低可忽略不计。整体而言，将可大幅提升全快闪存储数组的数据保护力。

QSAL 可以随时启动，并兼容于先前尚未设置的 SSD 同位检查 RAID。建议您在 SSD 使用寿命达到 50% 之前启用 QSAL。如果启用时间太晚，可能会遇到重建时间不足的问题，导致 SSD RAID 损坏的风险增加。

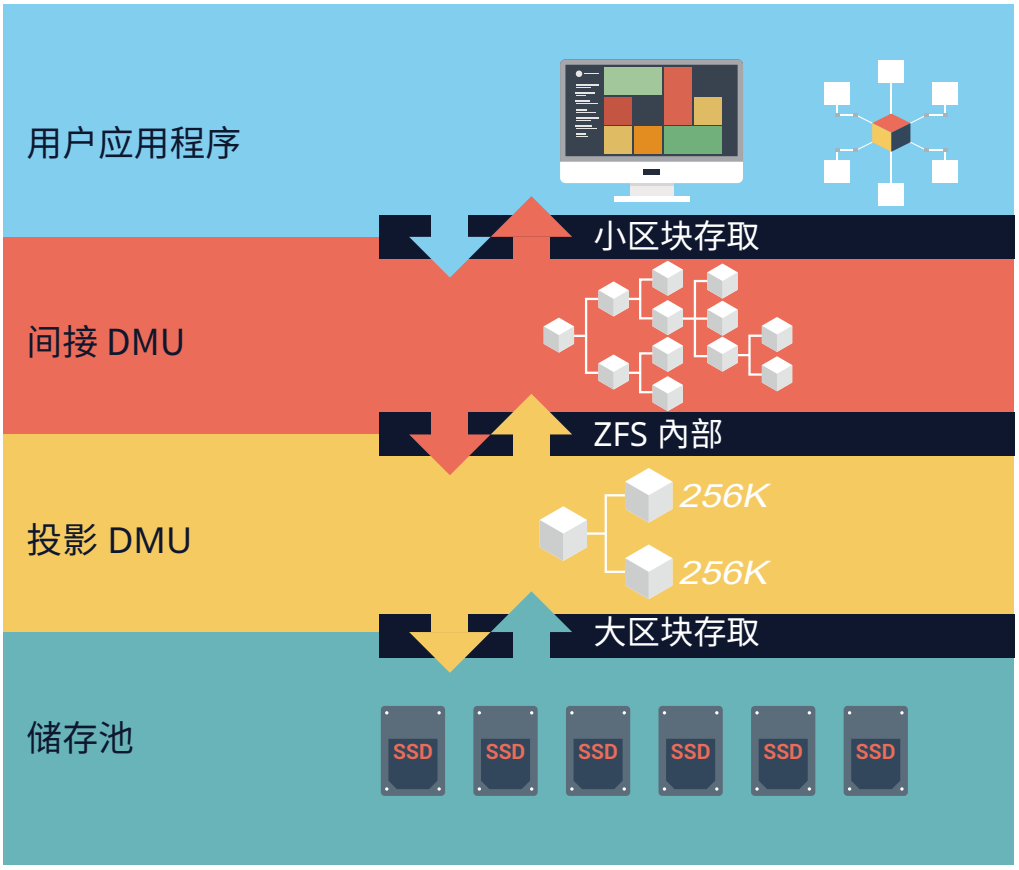



**专利编号：**US10705744B2 / US20200333959A1。同时包括几项保密专利。

# 写入聚合

## 「写入聚合」算法可提升全快闪数组的随机写入性能

「写入放大」是 SSD 的一种不良现象，而会给闪存的性能和耐用性带来挑战。QuTS hero 拥有 QNAP 独家的「写入聚合」算法，专为闪存优化而设计，可将所有随机写入转换为连续写入，同时减少 I/O。它不仅能够有效提升全闪存环境的随机写入性能，还可以改善 SSD 使用寿命。



# 建议的读取快取设置 (L2ARC)

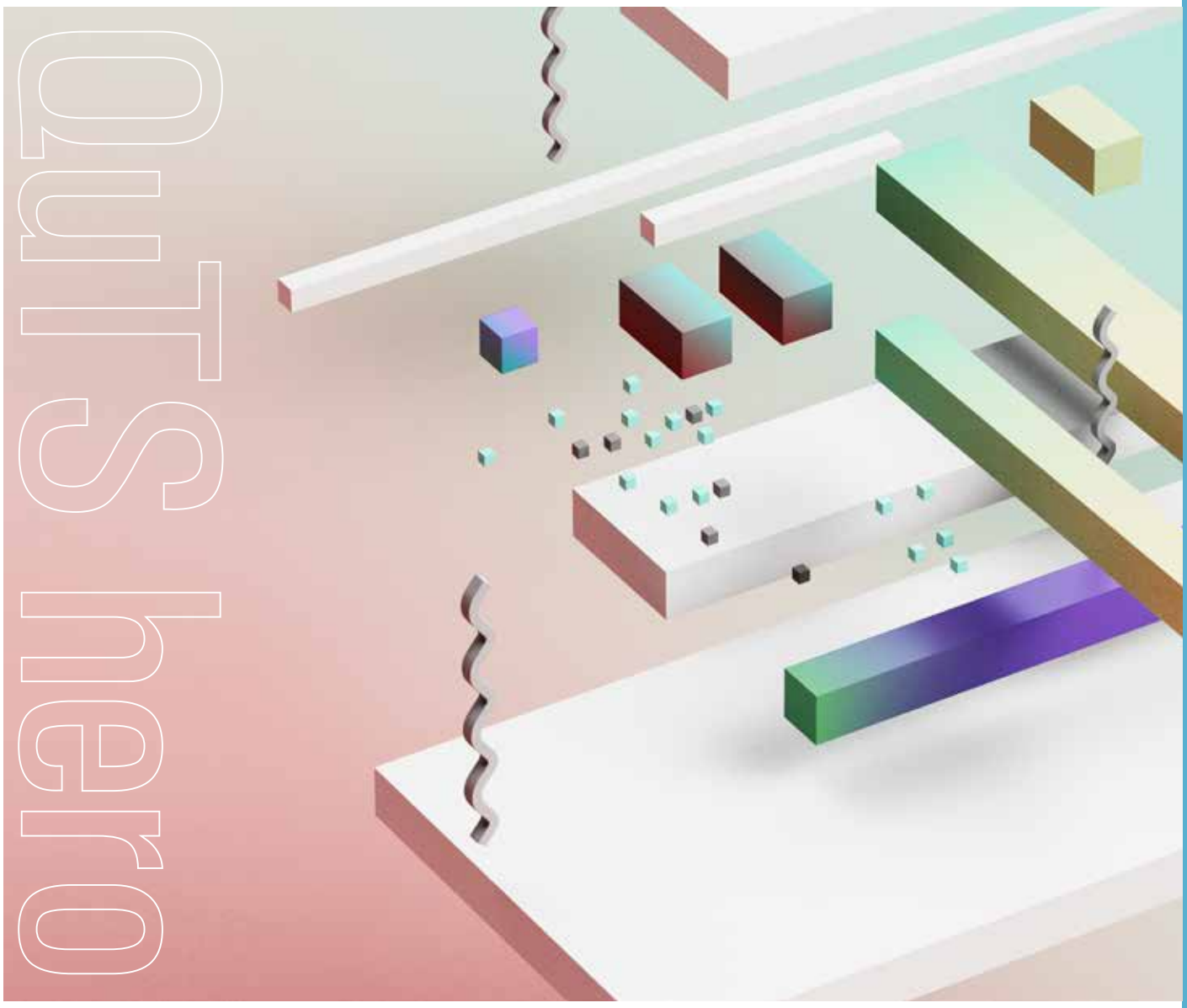
内存在 ZFS 性能中扮演着极为重要的角色，尤其是在高速数据传输、重复数据删除、ARC 和快取方面。建议您尽可能安装更多内存，以获得 ZFS 性能及优化商务负载的最大优势。

根据 QuTS hero NAS 安装的内存，以及读取快取的使用比例，我们建议您将读取快取的大小设置如下：

- **32 GB RAM** 支持 1TB SSD 大小的读取快取
- **64 GB RAM** 支持 2TB SSD 大小的读取快取
- **128GB RAM** 支持 4TB SSD 大小的读取快取
- **256GB RAM** 支持 8TB SSD 大小的读取快取

---

- **1TB RAM** 支持约 30TB SSD 的读取快取
- **4TB RAM** 支持约 120TB SSD 的读取快取


QuTS hero

# 内存大小与存储池之间的性能相依性

本章仅介绍 RAM 和存储池之间的基本相关性。如果系统正在执行其他专属服务（例如 Virtualization Station），建议您分开使用量，并另外考虑分配给这些虚拟机的内存。

**NAS 内存 - 虚拟机指派的内存 = ZFS 已使用内存**

这篇是分析内存与存储性能的一般概念，以及如何针对高性能存储识别内存大小的最佳作法。接着，我们会说明如何选择足够的内存来达到最佳性能。

内存容量在 ZFS 中扮演着极为重要的角色，对于 QuTS hero 存储池的表现影响极大。当 ZFS 启用数据缩减功能（例如压实、快速复制和重复数据删除）时，系统需要一些表格来记录这些数据区块的参考计数。因此，当数据量变大时，ZFS 必须在内存中保留尽可能多的参考数据表，以便能够以更有效的方式处理 IO 要求。

其次，ZFS 支持数据保护的交易群组功能。每个文件修改（例如写入）都与某个交易群组（TXG）关联。每个 TXG 会在固定间隔关闭，并且存储池会针对该群组要求同步作业。当 ARC 表示目前已快取过多的「已变更」内存时，也可能会关闭 TXG。当一个 TXG 关闭时，另一个 TXG 随即立即开启，文件修改于是跟新的作用中 TXG 产生关联。

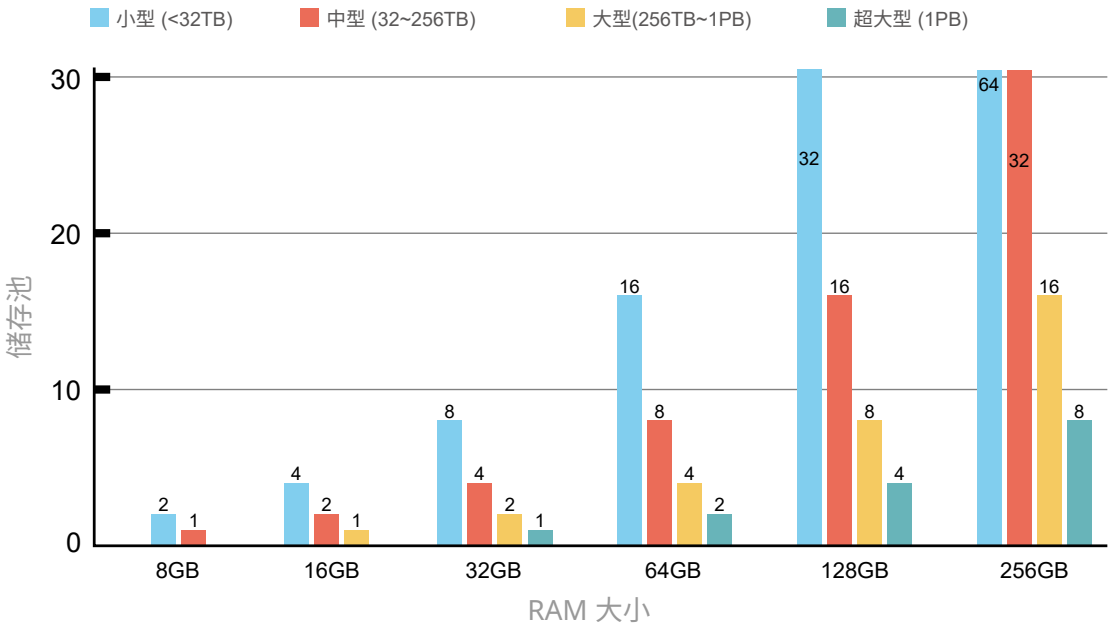
如果作用中的 TXG 关闭时，先前的 TXG 仍在同步数据至存储设备，则应用程序会受到限制，直到执行同步完成为止。在此情况下，TXG 正在同步，而 TXG + 1 因内存限制而关闭且正在等待同步；应用程序会限制等待写入 TXG + 2。存储空间需要持续耗用或是限制内存来抑制应用程序。

存储池的同步包含将所有层级 0 的数据区块传送至硬盘。当完成时，会传送所有层级 1 的间接区块（诸如此类），直到代表文件系统新状态的每个区块已认可为止。此时 uberblock 会更新为指向存储池中新的一致状态。

因此，ZFS 必须在完成每个交易后更新 uberblock，这是个极大的额外负荷。如果系统能够尽可能提供更多内存，交易群组可以更大，并且更新 uberblock 的额外负荷将变得相对较小。因此，更多的内存可以有效帮助 ZFS 改善写入 IO 的要求。

您可以参考以下建议，并为不同的存储设置选择足够的内存。

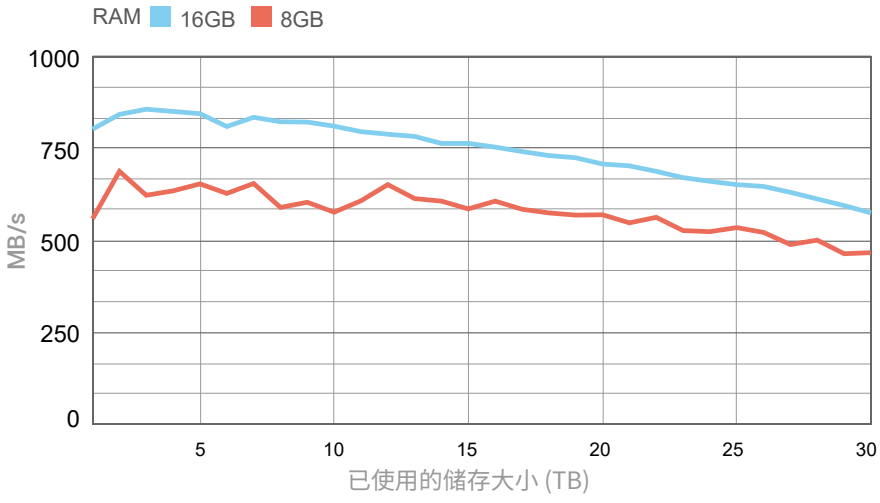
RAM 大小与存储池配置之间的最佳性能作法




例如，如果您的 ZFS NAS 会有 16GB 内存，则建议使用的存储池组态为 1 个大型存储池（256TB~1PB）或 2 个中型存储池或是 4 个小型存储池。这种类型的存储池组态可以让您的系统达到良好的性能。如果您建立的存储池数量超过这些建议，NAS 仍然能够可靠运作，但性能可能会降低。在这里我们以入门级的 ZFS 机种 TS-973AX 作为参考，以显示 8GB（最低需求）和 16GB RAM 之间的不同性能。

（我们已经先建立了一个 SSD 存储池跟一个约 35TB 的 HDD 存储池，来呈现 HDD 存储池的性能差异）。

TS-h973AX-16GB 对比 8GB - IO 性能 (5xHDD-R0)



# QuTS hero ZFS 白皮书


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