

Btrfs vs ZFS 實現 snapshot 的差異



目录

Contents

- Btrfs 的子卷 (subvolume) 和快照 (snapshot)
 - 子卷 (subvolume) 和快照 (snapshot) 的術語
 - 於是子卷在存儲介質中是如何記錄的呢？

- 那麼快照又是如何記錄的呢？
- ZFS 的文件系統 (filesystem)、快照 (snapshot)、克隆 (clone) 及其它
 - ZFS 設計中和快照相關的一些術語和概念
 - 數據集 (dataset)
 - 文件系統 (filesystem)
 - 快照 (snapshot)
 - 克隆 (clone)
 - 書籤 (bookmark)
 - 檢查點 (checkpoint)
 - ZFS 的概念與 btrfs 概念的對比
 - ZFS 中是如何存儲這些數據集的呢
 - ZFS 的塊指針 (block pointer)
 - ZPL 的元對象集 (MOS, Meta Object Set)
- 創建快照這麼簡單麼？那麼刪除快照呢？
 - 日誌結構文件系統中用的垃圾回收 (Garbage Collection) 算法
 - WAFL 早期使用的可用空間位圖數組

zfs 這個東西倒是名不符實。叫 z storage stack 明顯更符合。叫 fs 但不做 fs 自然確實會和 btrfs 有很大出入。

我反而以前還好奇為什麼 btrfs 不弄 zvol，直到我意識到這東西真是一個 fs，名符其實。

—— 某不愿透露姓名的 Ext2FSD 開發者

Btrfs 和 ZFS 都是開源的寫時拷貝（Copy on Write, CoW）文件系統，都提供了相似的子卷管理和快照（snapshot）的功能。網上有不少文章都評價 ZFS 實現 CoW FS 的創新之處，進而想說「Btrfs 只是 Linux/GPL 陣營對 ZFS 的拙劣抄襲」。或許（在存儲領域人盡皆知而在領域外）鮮有人知，在 ZFS 之前就有 NetApp 的商業產品 WAFL(Write Anywhere File Layout) 實現了 CoW 語義的文件系統，並且集成了快照和卷管理之類的功能。描述 btrfs 原型設計的論文和發表幻燈片也明顯提到 WAFL 比提到 ZFS 更多一些，應該說 WAFL 這樣的企業級存儲方案纔是 ZFS 和 btrfs 共同的靈感來源，而無論是 ZFS 還是 btrfs 在其設計中都汲取了很多來自 WAFL 的經驗教訓。

我一開始也帶着「Btrfs 和 ZFS 都提供了類似的功能，因此兩者必然有類似的設計」這樣的先入觀念，嘗試去使用這兩個文件系統，卻經常撞上兩者細節上的差異，導致使用時需要不盡相同的工作流，或者看似相似的用法有不太一樣的性能表現，又或者一邊有的功能（比如 ZFS 的 inband dedup，Btrfs 的 reflink）在另一邊沒有的情況。後來看到了 LWN 的這篇 [《A short history of btrfs》](#) 讓我意識到 btrfs 和 ZFS 雖然表面功能上看起來類似，但是實現細節上完全不一樣，所以需要不一樣的用法，適用於不一樣的使用場景。

爲了更好地理解這些差異，我四處蒐羅這兩個文件系統的實現細節，於是有了這篇筆記，記錄一下我查到的種種發現和自己的理解。~~（或許會寫成一個系列？還是先別亂挖坑不填。）~~ 只是自己的筆記，所有參閱的資料文檔都是二手資料，沒有深挖過源碼，還參雜了自己的理解，於是難免有和事實相違的地方，如有寫錯，還請留言糾正。

Btrfs 的子卷 (subvolume) 和快照 (snapshot)

先從兩個文件系統中（表面上看起來）比較簡單的 btrfs 的子卷（subvolume）和快照（snapshot）說起。關於子卷和快照的常規用法、推薦佈局之類的話題就不細說了，網上能找到很多不錯的資料，比如 [btrfs wiki 的 SysadminGuide 頁](#) 和 [Arch wiki 上 Btrfs#Subvolumes 頁](#) 都有不錯的參考價值。

關於寫時拷貝（CoW）文件系統的優勢，我們為什麼要用 btrfs/zfs 這樣的寫時拷貝文件系統，而不是傳統的文件系統設計，或者寫時拷貝文件系統在使用時有什麼區別之類的，網上同樣也能找到很多介紹，這裏不想再討論。這裏假設你用過 btrfs/zfs 至少一個的快照功能，知道它該怎麼用，並且想知道更多細節，判斷怎麼用那些功能才合理。

子卷（subvolume）和快照（snapshot）的術語

在 btrfs 中，存在於存儲媒介中的只有「子卷」的概念，「快照」只是個創建「子卷」的方式，換句話說在 btrfs 的術語裏，子卷（subvolume）是個名詞，而快照（snapshot）是個動詞。如果脫離了 btrfs 術語的上下文，或者不精確地稱呼的時候，也經常有文檔把 btrfs 的快照命令創建出的子卷叫做一個快照，所以當提到快照的時候，根據上下文判斷這裏是個動詞還是名詞，把名詞的快照當作用快照命令創建出的子卷就可以了。或者我們可以理解為，**互相共享一部分元數據**

(metadata) 的子卷互為彼此的快照 (名詞) ， 那麼按照這個定義的話，在 btrfs 中創建快照 (名詞) 的方式其實有兩種：

1. 用 `btrfs subvolume snapshot` 命令創建快照
2. 用 `btrfs send` 命令並使用 `-p` 參數發送快照，並在管道另一端接收

`btrfs send` 命令的 `-p` 與 `-c`

這裏也順便提一下 `btrfs send` 命令的 `-p` 參數和 `-c` 參數的差異。只看 `btrfs-send(8)` 的描述的話：

`-p <parent>`

send an incremental stream
from parent to subvol

`-c <clone-src>`

use this snapshot as a clone
source for an incremental
send (multiple allowed)

看起來這兩個都可以用來生成兩個快照之間的差分，只不過 -p 只能指定一個「parent」，而 -c 能指定多個「clone source」。在 unix stackexchange 上有人寫明了這兩個的異同。使用 -p 的時候，產生的差分首先讓接收端用 subvolume snapshot 命令對 parent 子卷創建一個快照，然後發送指令將這個快照修改成目標子卷的樣子，而使用 -c 的時候，首先在接收端用 subvolume create 創建一個空的子卷，隨後發送指令在這個子卷中填充內容，其數據塊儘量共享 clone source 已有的數據。所以 btrfs send -p 在接收端產生是有共享元數據的快照，而 btrfs send -c 在接收端產生的是僅僅共享數據而不共享元數據的子卷。

定義中「互相共享一部分 **元數據**」比較重要，因為除了快照的方式之外，btrfs 的子卷間也可以通過 reflink 的形式共享數據塊。我們可以對一整個子卷（甚至目錄）執行 `cp -r --reflink=always`，創建出一個副本，副本的文件內容通過 reflink 共享原本的數據，但不共享元數據，這樣創建出的就不是快照。

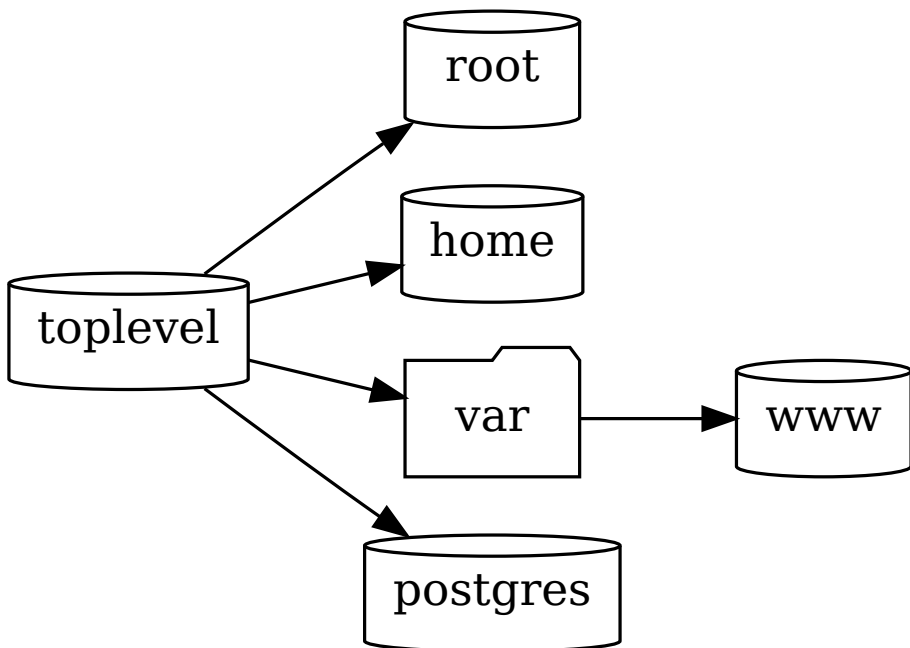
說了這麼多，其實關鍵的只是 btrfs 在傳統 Unix 文件系統的「目錄/文件/inode」這些東西之外只增加了一個「子卷」的新概念，而子卷間可以共享元數據或者數據，用快照命令創建出的子卷就是共享一部分元數據。

於是子卷在存儲介質中是如何記錄的呢？

比如在 SysadminGuide 這頁的 Flat 佈局 有個子卷佈局的例子。

```
toplevel          (volume root direc
tory, not to be mounted by default)
  +-- root        (subvolume root
directory, to be mounted at /)
  +-- home        (subvolume root
directory, to be mounted at /home)
  +-- var         (directory)
    |  \-- www    (subvolume root
directory, to be mounted at /var/ww
w)
      \-- postgres (subvolume root
directory, to be mounted at /var/li
b/postgresql)
```

用圓柱體表示子卷的話畫成圖大概是這個樣子：



首先要說明，btrfs 中大部分長度可變的數據結構都是 CoW B-tree，一種經過修改適合寫時拷貝的B樹結構，所以在 on-disk format 中提到了很多個樹。這裏的樹不是指文件系統中目錄結構樹，而是 CoW B-tree，如果不關心B樹細節的話可以把 btrfs 所說的一棵樹理解為關係數據庫中的一個表，和數據庫的表一樣 btrfs 的樹的長度可變，然後表項內容根據一個 key 排序。有這樣的背景之後，上圖例子中的 Flat 佈局在 btrfs 中大概是這樣的數據結構：



上圖中已經隱去了很多和本文無關的具體細節，所有這些細節都可以通過 `btrfs inspect-internal` 的 `dump-super` 和 `dump-tree` 查看到。btrfs 中的每棵樹都可以看

作是一個數據庫中的表，可以包含很多表項，根據 KEY 排序，而 KEY 是 (object_id, item_type, item_extra) 這樣的三元組。每個對象 (object) 在樹中用一個或多個表項 (item) 描述，同 object_id 的表項共同描述一個對象 (object)。B 樹中的 key 只用來比較大小不必連續，從而 object_id 也不必連續，只是按大小排序。有一些預留的 object_id 不能用作別的用途，他們的編號範圍是 -255ULL 到 255ULL，也就是表中前 255 和最後 255 個編號預留。

ROOT_TREE 中記錄了到所有別的 B 樹的指針，在一些文檔中叫做 tree of tree roots。「所有別的 B 樹」舉例來說比如 2 號 extent_tree，3 號 chunk_tree，4 號 dev_tree，10 號 free_space_tree，這些 B 樹都是描述 btrfs 文件系統結構非常重要的組成部分，但是在本文關係不大，今後有機會再討論它們。在 ROOT_TREE 的 5 號對象有一個 fs_tree，它描述了整個 btrfs pool 的頂級子卷，也就是圖中叫 toplevel 的那個子卷（有些文檔用定冠詞稱 the FS_TREE 的時候就是在說這個 5 號樹，而不是別的子卷的 FS_TREE）。除了頂級子卷之外，別的所有子卷的 object_id 在 256ULL 到 -256ULL 的範圍之間，對子卷而言 ROOT_TREE 中的這些 object_id 也同時是它們的子卷 id，在內核掛載文件系統的時候可以用 subvolid 找到它們，別的一些對子卷的操作也可以直接用 subvolid 表示一個子卷。ROOT_TREE 的 6 號對象描述的不是一棵樹，而是一個名叫 default 的特殊目錄，它指向 btrfs pool 的默認掛載子卷。最初 mkfs 的時候，這個目錄指向 ROOT_ITEM 5，也就是那個頂級子卷，之後

可以通過命令 `btrfs subvolume set-default` 修改它指向別的子卷，這裏它被改爲指向 `ROOT_ITEM 256` 亦即那個名叫 "root" 的子卷。

每一個子卷都有一棵自己的 `FS_TREE`（有的文檔中叫 file tree），一個 `FS_TREE` 相當於傳統 Unix 文件系統中的一整個 inode table，只不過它除了包含 inode 信息之外還包含所有文件夾內容。在 `FS_TREE` 中，`object_id` 同時也是它所描述對象的 inode 號，所以 `btrfs` 的 **子卷有互相獨立的 inode 編號**，不同子卷中的文件或目錄可以擁有相同的 inode。或許有人不太清楚子卷間 inode 編號獨立意味着什麼，簡單地說，這意味着你不能跨子卷創建 hard link，不能跨子卷 `mv` 移動文件而不產生複製操作。不過因爲 reflink 和 inode 無關，可以跨子卷創建 reflink，也可以用 `reflink + rm` 的方式快速移動文件。

`FS_TREE` 中一個目錄用一個 `inode_item` 和多個 `dir_item` 描述，`inode_item` 是目錄自己的 inode，那些 `dir_item` 是目錄的內容。`dir_item` 可以指向別的 `inode_item` 來描述普通文件和子目錄，也可以指向 `root_item` 來描述這個目錄指向一個子卷。有人或許疑惑，子卷就沒有自己的 inode 麼？其實如果看 數據結構定義 的話 `struct btrfs_root_item` 結構在最開頭的地方包含了一個 `struct btrfs_inode_item` 所以 `root_item` 也同時作爲子卷的 inode，不過用戶通常看不到這個子卷的 inode，因爲子卷在被（手動或自動地）掛載到目錄上之後，用戶會看到的是子卷的根目錄的 inode。

比如上圖 FS_TREE toplevel 中，有兩個對象，第一個 256 是（子卷的）根目錄，第二個 257 是 "var" 目錄，256 有4個子目錄，其中 "root" "home" "postgres" 這三個指向了 ROOT_TREE 中的對應子卷，而 "var" 指向了 inode 257。然後 257 有一個子目錄叫 "www" 它指向了 ROOT_TREE 中 object_id 為 258 的子卷。

那麼快照又是如何記錄的呢？

以上是子卷、目錄、inode 在 btrfs 中的記錄方式，你可能想知道，如何記錄一個快照呢？特別是，如果對一個包含子卷的子卷創建了快照，會得到什麼結果呢？如果我們在上面的佈局基礎上執行：

```
btrfs subvolume snapshot toplevel toplevel/toplevel@1
```

那麼產生的數據結構大概如下所示：



在 ROOT_TREE 中增加了 260 號子卷，其內容複製自 toplevel 子卷，然後 FS_TREE toplevel 的 256 號 inode 也就是根目錄中增加一個 dir_item 名叫 *toplevel@s1* 它指向 ROOT_ITEM 的 260 號子卷。這裏看似是完整複製了整個 FS_TREE 的內容，這是因為 CoW b-tree 當只有一個葉子節點時就複製整個葉子節點。如果子卷內容再多一些，除了葉子之外還有中間節點，那麼只有被修改的葉子和其上的中間節點需要複製。從而創建快照的開銷基本上是 $O(\text{level of FS_TREE})$ ，而 B 樹的高度一般都能維持在很低的程度，所以快照創建速度近乎是常數開銷。

從子卷和快照的這種實現方式，可以看出：**雖然子卷可以嵌套子卷，但是對含有嵌套子卷的子卷做快照的語義有些特別**。上圖中我沒有畫 *toplevel@s1* 下的各個子卷到對應 ROOT_ITEM 之間的虛線箭頭，是因為這時候如果你嘗試直接跳過 *toplevel* 掛載 *toplevel@s1* 到掛載點，會發現那些子卷沒有被自動掛載，更奇怪的是那些子卷的目錄項也不是個普通目錄，嘗試往它們中放東

西會得到無權訪問的錯誤，對它們能做的唯一事情是手動將別的子卷掛載在上面。推測原因在於這些子目錄並不是真的目錄，沒有對應的目錄的 inode，試圖查看它們的 inode 號會得到 2 號，而這是個保留號不應該出現在 btrfs 的 inode 號中。每個子卷創建時會記錄包含它的上級子卷，用 `btrfs subvolume list` 可以看到每個子卷的 top level subvolid，猜測當掛載 A 而 A 中嵌套的 B 子卷記錄的上級子卷不是 A 的時候，會出現上述奇怪行爲。嵌套子卷的快照還有一些別的神奇行爲，大家可以自己探索探索。

建議用平坦的子卷佈局

因爲上述嵌套子卷在做快照時的特殊行爲，我個人建議是 **保持平坦的子卷佈局**，也就是說：

1. 只讓頂層子卷包含其它子卷，除了頂層子卷之外的子卷只做手工掛載，不放嵌套子卷
2. 只在頂層子卷對其它子卷做快照，不快照頂層子卷
3. 雖然可以在頂層子卷放子卷之外的東西（文件或目錄），不過因爲想避免對頂層子卷做快照，所以避免在頂層子卷放普通文件。

btrfs 的子卷可以設置「可寫」或者「只讀」，在創建一個快照的時候也可以通過 `-r` 參數創建出一個只讀快照。通常只讀快照可能比可寫的快照更有用，因爲

`btrfs send` 命令只接受只讀快照作為參考點。子卷可以有兩種方式切換它是否只讀的屬性，可以通過 `btrfs property set <subvol> ro` 直接修改是否只讀，也可以對只讀子卷用 `btrfs subvolume snapshot` 創建出可寫子卷，或者反過來對可寫子卷創建出只讀子卷。

只讀快照也有些特殊的限制，在 [SysadminGuide#Special_Cases](#) 就提到一例，你不能把只讀快照用 `mv` 移出包含它的目錄，雖然你能用 `mv` 給它改名或者移動包含它的目錄到別的地方。[btrfs wiki](#) 上給出這個限制的原因是子卷中記錄了它的上級，所以要移動它到別的上級需要修改這個子卷，從而只讀子卷沒法移動到別的上級（不過我還沒搞清楚子卷在哪兒記錄了它的上級，記錄的是上級目錄還是上級子卷）。不過這個限制可以通過對只讀快照在目標位置創建一個新的只讀快照，然後刪掉原位置的只讀快照來解決。

ZFS 的文件系統 (filesystem)、快照 (snapshot)、克隆 (clone) 及其它

Btrfs 給傳統文件系統只增加了子卷的概念，相比之下 ZFS 中類似子卷的概念有好幾個，據我所知有這些：

- 數據集 (dataset)
- 文件系統 (filesystem)
- 快照 (snapshot)
- 克隆 (clone)
- 書籤 (bookmark)：從 ZFS on Linux v0.6.4 開始
- 檢查點 (checkpoint)：從 ZFS on Linux v0.8.0 開始

梳理一下這些概念之間的關係也是最初想寫下這篇筆記的初衷。先畫個簡圖，隨後逐一講講這些概念：



上圖中，假設我們有一個 pool，其中有 3 個文件系統叫 fs1~fs3 和一個 zvol 叫 zv1，然後文件系統 fs1 有兩個快照 s1 和 s2，和兩個書籤 b1 和 b2。pool 整體有兩個檢查點 cp1 和 cp2。這個簡圖將作為例子在後面介紹這些概念。

ZFS 設計中和快照相關的一些術語和概念

數據集 (dataset)

ZFS 中把文件系統、快照、克隆、zvol 等概念統稱為數據集 (dataset)。一些文檔和介紹中把文件系統叫做數據集，大概因為在 ZFS 中，文件系統是最先創建並且最有用的數據集。

在 ZFS 的術語中，把底層管理和釋放存儲設備空間的叫做 ZFS 存儲池 (pool)，簡稱 zpool，其上可以容納多個數據集，這些數據集用類似文件夾路徑的語法 `pool_name/dataset_path@snapshot_name` 這樣來稱呼。存儲池中的數據集一同共享可用的存儲空間，每個數據集單獨跟蹤自己所消耗掉的存儲空間。

數據集之間有類似文件夾的層級父子關係，這一點有用的地方在於可以在父級數據集上設定一些 ZFS 參數，這些參數可以被子級數據集基礎，從而通過層級關係可以方便地微調 ZFS 參數。在 btrfs 中目前還沒有類似的屬性繼承的功能。

zvol 的概念和本文關係不大，可以參考我上一篇 ZFS 子系統筆記中 ZVOL 的說明。用 zvol 能把 ZFS 當作一個傳統的卷管理器，繞開 ZFS 的 ZPL (ZFS Posix filesystem Layer) 層。在 Btrfs 中可以用 loopback 塊設備某種程度上模擬 zvol 的功能。

文件系統 (filesystem)

創建了 ZFS 存儲池後，首先要在其中創建文件系統 (filesystem)，才能在文件系統中存儲文件。容易看出 ZFS 文件系統的概念直接對應 btrfs 中的子卷。文件系統 (filesystem) 這個術語，從命名方式來看或許是想要和（像 Solaris 的 SVM 或者 Linux 的 LVM 這樣的）傳統的卷管理器 與其上創建的多個文件系統（Solaris UFS 或者 Linux ext）這樣的上下層級做類比。從 btrfs 的子卷在內部結構中叫作 FS_TREE 這一點可以看出，至少在 btrfs 早期設計中大概也是把子卷稱為 filesystem 做過類似的類比的。和傳統的卷管理器與傳統文件系統的上下層級不同的是，ZFS 和 btrfs 中由存儲池跟蹤和管理可用空間，做統一的數據塊分配和釋放，沒有分配的數據塊算作整個存儲池中所有 ZFS 文件系統或者 btrfs 子卷的可用空間。

與 btrfs 的子卷不同的是，ZFS 的文件系統之間是完全隔離的，（除了後文會講的 dedup 方式之外）不可以共享任何數據或者元數據。一個文件系統還包含了隸屬於其中的快照 (snapshot)、克隆 (clone) 和書籤 (bookmark)。在 btrfs 中一個子卷和對其創建的快照之間雖然有父子關係，但是在 ROOT_TREE 的記錄中屬於平級的關係。

上面簡圖中 pool 裏面包含 3 個文件系統，分別是 fs1~3。

快照 (snapshot)

ZFS 的快照對應 btrfs 的只讀快照，是標記數據集在某一歷史時刻上的只讀狀態。和 btrfs 的只讀快照一樣，ZFS 的快照也兼作 send/receive 時的參考點。快照隸屬於一個數據集，這說明 ZFS 的文件系統或者 zvol 都可以創建快照。

ZFS 中快照是排列在一個時間線上的，因為都是只讀快照，它們是數據集在歷史上的不同時間點。這裏說的時間不是系統時鐘的時間，而是 ZFS 中事務組（TXG, transaction group）的一個序號。整個 ZFS pool 的每次寫入會被合併到一個事務組，對事務組分配一個嚴格遞增的序列號，提交一個事務組具有類似數據庫中事務的語義：要麼整個事務組都被完整提交，要麼整個 pool 處於上一個事務組的狀態，即使中間發生突然斷電之類的意外也不會破壞事務語義。因此 ZFS 快照就是數據集處於某一個事務組時的狀態。

如果不滿於對數據集進行的修改，想把整個數據集恢復到之前的狀態，那麼可以回滾（rollback）數據集到一個快照。回滾操作會撤銷掉對數據集的所有更改，並且默認參數下只能回滾到最近的一個快照。如果想回滾到更早的快照，可以先刪掉最近的幾個，或者可以使用 `zfs rollback -r` 參數刪除中間的快照並回滾。

除了回滾操作，還可以直接只讀訪問到快照中的文件。ZFS 的文件系統中有個隱藏文件夾叫 `".zfs"`，所以如果只想回滾一部分文件，可以從

".zfs/snapshots/SNAPSHOT-NAME" 中把需要的文件複製出來。

比如上面簡圖中 fs1 就有 pool/fs1@s1 和 pool/fs1@s2 這兩個快照，那麼可以在 fs1 掛載點下 .zfs/snapshots/s1 的路徑直接訪問到 s1 中的內容。

克隆 (clone)

ZFS 的克隆有點像 btrfs 的可寫快照。因為 ZFS 的快照是只讀的，如果想對快照做寫入，那需要先用 `zfs clone` 從快照中建出一個克隆，創建出的克隆和快照共享元數據和數據，然後對克隆的寫入不影響數據集原本的寫入點。創建了克隆之後，作為克隆參考點的快照會成為克隆的依賴，克隆存在期間無法刪除掉作為其依賴的快照。

一個數據集可以有多個克隆，這些克隆都獨立於數據集當前的寫入點。使用 `zfs promote` 命令可以把一個克隆「升級」成為數據集的當前寫入點，從而數據集原本的寫入點會調轉依賴關係，成為這個新寫入點的一個克隆，被升級的克隆原本依賴的快照和之前的快照會成為新數據集寫入點的快照。

比如上面簡圖中 fs1 有 c1 的克隆，它依賴於 s2 這個快照，從而 c1 存在的時候就不能刪除掉 s2。

書籤 (bookmark)

這是 ZFS 一個比較新的特性，ZFS on Linux 分支從 v0.6.4 開始支持創建書籤的功能。

書籤特性存在的理由是基於這樣的事實：原本 ZFS 在 send 兩個快照間的差異的時候，比如 send S1 和 S2 之間的差異，在發送端實際上只需要 S1 中記錄的時間戳（TXG id），而不需要 S1 快照的數據，就可以計算出 S1 到 S2 的差異。在接收端則需要 S1 的完整數據，在其上根據接收到的數據流創建 S2。因此在發送端，可以把快照 S1 轉變成書籤，只留下時間戳元數據而不保留任何目錄結構或者文件內容。書籤只能作為增量 send 時的參考點，並且在接收端需要有對應的快照，這種方式可以在發送端節省很多存儲。

通常的使用場景是，比如你有一個筆記本電腦，上面有 ZFS 存儲的數據，然後使用一個服務器上 ZFS 作為接收端，定期對筆記本上的 ZFS 做快照然後 send 給服務器。在沒有書籤功能的時候，筆記本上至少得保留一個和服務器上相同的快照，作為 send 的增量參考點，而這個快照的內容已經在服務器上，所以筆記本中存有相同的快照只是在浪費存儲空間。有了書籤功能之後，每次將定期的新快照發送到服務器之後，就可以把這個快照轉化成書籤，節省存儲開銷。

檢查點（checkpoint）

這也是 ZFS 的新特性，ZFS on Linux 分支從 v0.8.0 開始支持創建檢查點。

簡而言之，檢查點可以看作是整個存儲池級別的快照，使用檢查點能快速將整個存儲池都恢復到上一個狀態。這邊有篇文章介紹 ZFS checkpoint 功能的背景、用法和限制，可以看出當存儲池中有檢查點的時候很多存儲池的功能會受影響（比如不能刪除 vdev、不能處於 degraded 狀態、不能 scrub 到當前存儲池中已經釋放而在檢查點還在引用的數據塊），於是檢查點功能設計上更多是給系統管理員準備的用於調整整個 ZFS pool 時的後悔藥，調整結束後日用狀態下應該刪除掉所有檢查點。

ZFS 的概念與 btrfs 概念的對比

先說書籤和檢查點，因為這是兩個 btrfs 目前完全沒有的功能。

書籤功能完全圍繞 ZFS send 的工作原理，而 ZFS send 位於 ZFS 設計中的 DSL 層面，甚至不關心它 send 的快照的數據是來自文件系統還是 zvol。在發送端它只是從目標快照遞歸取數據塊，判斷 TXG 是否老於參照點的快照，然後把新的數據塊全部發往 send stream；在接收端也只是完整地接收數據塊，不加以處理。與之不同的是 btrfs 的 send 的工作原理是工作在文件系統的只讀子卷層面，發送端在內核代碼中根據目標快照的 b 樹和參照點快照的 generation 生成一個 diff（可以通過 btrfs subvolume find-new 直接拿到這個 diff），然後在用戶態代碼中根據 diff 和參照點、目標快照的兩

個只讀子卷的數據產生一連串修改文件系統的指令，指令包括創建文件、刪除文件、讓文件引用數據塊（保持 reflink）等操作；在接收端則完全工作在用戶態下，根據接收到的指令重建目標快照。可見 btrfs send 需要在發送端讀取參照點快照的數據（比如找到 reflink 引用），從而 btrfs 沒法（或者很難）實現書籤功能。

檢查點也是 btrfs 目前沒有的功能。btrfs 目前不能對頂層子卷做遞歸的 snapshot，btrfs 的子卷也沒有類似 ZFS 數據集的層級關係和可繼承屬性，從而沒法實現類似檢查點的功能。

除了書籤和檢查點之外，剩下的概念可以在 ZFS 和 btrfs 之間有如下映射關係：

| | |
|------------------|------------|
| ZFS 文件系統: | btrfs 子卷 |
| ZFS 快照: | btrfs 只讀快照 |
| ZFS 克隆: | btrfs 可寫快照 |

對 ZFS 數據集的操作，大部分也可以找到對應的對 btrfs 子卷的操作。

| | |
|---------------------|------------------------|
| zfs list: | btrfs subvolume list |
| zfs create: | btrfs subvolume create |
| zfs destroy: | btrfs subvolume delete |
| zfs rename: | mv |

zfs snapshot: `btrfs subvolume snapshot -r`

zfs rollback: 這個在 btrfs 需要對只讀快照創建出可寫的快照（用 snapshot 命令，或者直接修改讀寫屬性），然後改名或者調整掛載點

zfs diff: `btrfs subvolume find-new`

zfs clone: `btrfs subvolume snapshot`

zfs promote: 和 rollback 類似，可以直接調整 btrfs 子卷的掛載點

可見雖然功能上類似，但是至少從管理員管理的角度而言，zfs 對文件系統、快照、克隆的劃分更為清晰，對他們能做的操作也更為明確。這也是很多從 ZFS 遷移到 btrfs，或者反過來從 btrfs 換用 zfs 時，一些人困惑的起源（甚至有人據此說 ZFS 比 btrfs 好在 cli 設計上）。

不過 btrfs 子卷的設計也使它在系統管理上有了更大的靈活性。比如在 btrfs 中刪除一個子卷不會受制於別的子卷是否存在，而在 zfs 中要刪除一個快照必須先保證先摧毀掉依賴它的克隆。再比如 btrfs 的可寫子卷沒有主次之分，而 zfs 中一個文件系統和其克隆之間有明顯的區別，所以需要 promote 命令調整差異。還有比如 ZFS 的文件系統只能回滾到最近一次的快照，要回滾到更久之前的快照需要刪掉中間的快照，並且回滾之後原本的文

件系統數據和快照數據就被丟棄了；而 btrfs 中因為回滾操作相當於調整子卷的掛載，所以不需要刪掉快照，並且回滾之後原本的子卷和快照還可以繼續保留。

加上 btrfs 有 reflink，這給了 btrfs 在使用中更大的靈活性，可以有一些 zfs 很難做到的用法。比如想從快照中打撈出一些虛擬機鏡像的歷史副本，而不想回滾整個快照的時候，在 btrfs 中可以直接 `cp --reflink=always` 將鏡像從快照中複製出來，此時的複製將和快照共享數據塊；而在 zfs 中只能用普通 `cp` 複製，會浪費很多存儲空間。

ZFS 中是如何存儲這些數據集的呢

要講到存儲細節，首先需要瞭解一下 ZFS 的分層設計。不像 btrfs 基於現代 Linux 內核，有許多現有文件系統已經實現好的基礎設施可以利用，並且大體上只用到一種核心數據結構（CoW的B樹）；ZFS 則脫胎於 Solaris 的野心勃勃，設計時就分成很多不同的子系統，逐步提升抽象層次，並且每個子系統都發明了許多特定需求下的數據結構來描述存儲的信息。在這裏和本文內容密切相關的是 ZPL、DSL、DMU 這些 ZFS 子系統。

Sun 曾經寫過一篇 ZFS 的 On disk format 對理解 ZFS 如何存儲在磁盤上很有幫助，雖然這篇文檔是針對 Sun 還在的時候 Solaris 的 ZFS，現在 ZFS 的內部已經

變化挺大，不過對於理解本文想講的快照的實現方式還具有參考意義。這裏藉助這篇 [ZFS On Disk Format](#) 中的一些圖示來解釋 ZFS 在磁盤上的存儲方式。

ZFS 的塊指針 (block pointer)

ZFS 中用的 128 字節塊指針

| | | | | | | | |
|---|-------------|---------|------|-------|------|-------|-------|
| 0 | vdev1 | | | | GRID | ASIZE | |
| 1 | G | offset1 | | | | | |
| 2 | vdev2 | | | | GRID | ASIZE | |
| 3 | G | offset2 | | | | | |
| 4 | vdev3 | | | | GRID | ASIZE | |
| 5 | G | offset3 | | | | | |
| 6 | E | lvl | type | cksum | comp | PSIZE | LSIZE |
| 7 | padding | | | | | | |
| 8 | padding | | | | | | |
| 9 | padding | | | | | | |
| a | birth txg | | | | | | |
| b | fill count | | | | | | |
| c | checksum[0] | | | | | | |
| d | checksum[1] | | | | | | |
| e | checksum[2] | | | | | | |
| f | checksum[3] | | | | | | |

要理解 ZFS 的磁盤結構首先想介紹一下 ZFS 中的塊指針（block pointer, blkptr_t ），結構如右圖所示。ZFS 的塊指針用在 ZFS 的許多數據結構之中，當需要從一個地方指向任意另一個地址的時候都會插入這樣的一個塊指針結構。大多數文件系統中也有類似的指針

結構，比如 btrfs 中有個8字節大小的邏輯地址（logical address），一般也就是個 4字節 到 16字節 大小的整數寫着扇區號、塊號或者字節偏移，在 ZFS 中的塊指針則是一個巨大的128字節（不是 128bit !）的結構體。

128字節塊指針的開頭是3個數據虛擬地址（DVA, Data Virtual Address），每個 DVA 是 128bit，其中記錄這塊數據在什麼設備（vdev）的什麼偏移（offset）上佔用多大（asize），有 3個 DVA 槽是用來存儲最多3個不同位置的副本。然後塊指針還記錄了這個塊用什麼校驗算法（cksum）和什麼壓縮算法（comp），壓縮前後的大小（PSIZE/LSIZE），以及256bit的校驗和（checksum）。

當需要間接塊（indirect block）時，塊指針中記錄了間接塊的層數（lvl），和下層塊指針的數量（fill）。一個間接塊就是一個數據塊中包含一個塊指針的數組，當引用的對象很大需要很多塊時，間接塊構成一棵樹狀結構。

塊指針中還有和本文關係很大的一個值 birth txg，記錄這個塊指針誕生時的整個 pool 的 TXG id。一次 TXG 提交中寫入的數據塊都會有相同的 birth txg，這個相當於 btrfs 中 generation 的概念。實際上現在的 ZFS 塊指針似乎記錄了兩個 birth txg，分別在圖中的9行和a行的位置，一個 physical 一個 logical，用於 dedup 和 device removal。值得注意的是塊指針裏只有 birth txg，沒有引用計數或者別的機制做引用，這對後面要講的東西很關鍵。

ZPL 的元對象集 (MOS, Meta Object Set)

理解塊指針和 ZFS 的子系統層級之後，就可以來看看 ZFS 存儲在磁盤上的具體結構了。因為涉及的數據結構種類比較多，所以先來畫一張邏輯上的簡圖，其中箭頭只是某種引用關係不代表塊指針，方框也不是結構體細節：



如上簡圖所示，首先 ZFS pool 級別有個 uberblock，具體每個 vdev 如何存儲和找到這個 uberblock 今後有空再聊，這裏認為整個 zpool 有唯一的一個 uberblock

。從 uberblock 有個指針指向 Meta Object Set ，它是個 DMU 的對象集，它包含整個 pool 的一些配置信息，和根數據集（root dataset）。根數據集再包含整個 pool 中保存的所有頂層數據集，每個數據集有一個 DSL Directory 結構。然後從每個數據集的 DSL Directory 可以找到一系列子數據集和一系列快照等結構。最後每個數據集有個 active 的 DMU 對象集，這是整個文件系統的當前寫入點，每個快照也指向一個各自的 DMU 對象集。

DSL 層的每個數據集的邏輯結構也可以用下面的圖表達（來自 ZFS On Disk Format ）：



ZFS On Disk Format 中 4.1 節的 DSL infrastructure

ZFS On Disk Format 中 4.2 節的 Meta Object Set

需要記得 ZFS 中沒有類似 btrfs 的 CoW b-tree 這樣的統一數據結構，所以上面的這些設施是用各種不同的數據結構表達的。尤其每個 Directory 的結構可以包含一個 ZAP 的鍵值對存儲，和一個 DMU 對象。可以理解為，DSL 用 DMU 對象集 (Objectset) 表示一個整數 (uint64_t 的 dnode) 到 DMU 對象的映射，然後用 ZAP 對象表示一個名字到整數的映射，然後又有很多額外的存儲於 DMU 對象中的 DSL 結構體。如果我們畫出不同的指針和不同的結構體，那麼會得到一個稍顯複雜的圖，見右邊「ZFS On Disk Format 中 4.2 節的 Meta Object Set」，圖中還只畫到了 root_dataset 為止。

看到這裏，大概可以理解在 ZFS 中創建一個 ZFS 快照的操作其實很簡單：找到數據集的 DSL Directory 中當前 active 的 DMU 對象集指針，創建一個表示 snapshot 的 DSL dataset 結構，指向那個 DMU 對象集，然後快照就建好了。因為今後對 active 的寫入會寫時複製對應的 DMU 對象集，所以 snapshot 指向的 DMU 對象集不會變化。

創建快照這麼簡單麼？那麼刪除快照呢？

按上面的存儲格式細節來看，btrfs 和 zfs 中創建快照似乎都挺簡單的，利用寫時拷貝，創建快照本身沒什麼複雜操作。

如果你也聽到過別人介紹 CoW 文件系統時這麼講，是不是會覺得似乎哪兒少了點什麼。創建快照是挺簡單的，**直到你開始考慮如何刪除快照……**

或者不侷限在刪除單個快照上，CoW 文件系統因為寫時拷貝，每修改一個文件內容或者修改一個文件系統結構，都是分配新數據塊，然後考慮是否要刪除這個數據替換的老數據塊，此時如何決定老數據塊能不能刪呢？刪除快照的時候也是同樣，快照是和別的文件系統有共享一部分數據和元數據的，所以顯然不能把快照引

用到的數據塊都直接刪掉，要考察快照引用的數據塊是否還在別的地方被引用着，只能刪除那些沒有被引用的數據。

深究「如何刪快照」這個問題，就能看出 WAFL、btrfs、ZFS 甚至別的 log-structured 文件系統間的關鍵區別，從而也能看到另一個問題的答案：**為什麼 btrfs 只需要子卷的抽象，而 zfs 搞出了這麼多抽象概念？**帶着這兩個疑問，我們來研究一下這些文件系統的塊刪除算法。

日誌結構文件系統中用的垃圾回收（Garbage Collection）算法

講 btrfs 和 zfs 用到的刪除算法之前，先講一下日誌結構（log-structured）文件系統中的 GC 算法。對熟悉編程的人來說，講到空間釋放算法，大概首先會想到 GC，因為這裏要解決的問題乍看起來很像編程語言的內存管理中 GC 想要解決的問題：有很多指針相互指向很多數據結構，找其中沒有被引用的垃圾然後釋放掉。

首先要澄清一下 日誌結構文件系統（log-structured file system） 的定義，因為有很多文件系統用日誌，而用了日誌的不一定是日誌結構文件系統。在維基百科上有個頁面介紹 日誌結構文件系統，還有個 列表列出了一些日誌結構文件系統。通常說，整個文件系統的存儲結構都組織成一個大日誌的樣子，就說這個文

件系統是日誌結構的，這包括很多早期學術研究的文件系統，以及目前 NetBSD 的 LFS、Linux 的 NILFS，用在光盤介質上的 UDF，還有一些專門為閃存優化的 JFFS、YAFFS 以及 F2FS。日誌結構文件系統不包括那些用額外日誌保證文件系統一致性，但文件系統結構不在日誌中的 ext4、xfs、ntfs、hfs+。

簡單來說，日誌結構文件系統就是把存儲設備當作一個大日誌，每次寫入數據時都添加在日誌末尾，然後用寫時複製重新寫入元數據，最後提交整個文件系統結構。從這個特徵上來說，寫時拷貝文件系統（CoW FS）像 btrfs/zfs 這些在一些人眼中也符合日誌結構文件系統的特徵，所以也有人說寫時拷貝文件系統算是日誌結構文件系統的一個子類。不過日誌結構文件系統的另一大特徵是利用 GC 回收空間，這裏是本文要講的區別，所以在我看來不用 GC 的 btrfs 和 zfs 不算是日誌結構文件系統。

舉個例子，比如下圖是一個日誌結構文件系統的磁盤佔用，其中綠色是數據，藍色是元數據（比如目錄結構和 inode），紅色是文件系統級關鍵數據（比如最後的日誌提交點），一開始可能是這樣，有9個數據塊，2個元數據塊，1個系統塊：

| | | | | |
|---|---|----|--|--|
| 1 | 5 | 9 | | |
| 2 | 6 | 10 | | |
| 3 | 7 | 11 | | |
| 4 | 8 | 12 | | |

現在要覆蓋 2 和 3 的內容，新寫入 n2 和 n3，然後修改 10 裏面的 inode 變成 n10 引用這些新數據，然後寫入一個新提交 n12，用黃色表示不再被引用的垃圾，提交完大概是這樣：

| | | | | |
|----|---|-----|-----|--|
| 1 | 5 | 9 | n2 | |
| o2 | 6 | o10 | n3 | |
| o3 | 7 | 11 | n10 | |
| 4 | 8 | o12 | n12 | |

日誌結構文件系統需要 GC 比較容易理解，寫日誌嘛，總得有一個「添加到末尾」的寫入點，比如上面圖中的 n12。空盤上連續往後寫而不 GC 總會遇到空間末尾，這時候就要覆蓋寫空間開頭，就很難判斷「末尾」在什麼地方了。這時候 GC 的話，也不能只回收那些不再需要的空間（圖中的 o2 和 o3 和 o10），因為這樣的話剩餘空間會很零碎。

和內存管理時的 GC 不同的一點在於，文件系統的 GC 肯定不能停下整個世界跑 GC，也不能把可用空間對半分然後 Mark-and-Sweep，這些在內存中還尚可的簡單策略直接放到文件系統中絕對是性能災難。

通常文件系統的 GC 是，先把整個盤分成幾個段 (segment) 或者區域(zone)，術語不同不過表達的概念類似，然後 GC 時挑一個老段，掃描文件系統找出目標段中還被引用的數據塊，搬運到日誌末尾，然後整個釋放一段。搬運數據塊時，也要調整文件系統別的地方對被搬運的數據塊的引用。

物理磁盤上有扇區的概念，通常是512字節到4KiB大小，在文件系統中一般把連續幾個物理塊作為一個數據塊，大概 4KiB 到 1MiB 的數量級，然後日誌結構文件系統中一個段通常是連續的很多塊，數量級來看大約是 4MiB 到 64MiB 這樣的數量級。在數據塊之上，ufs/ext4/btrfs/zfs 通常還有 block group 的概念，大概是 128MiB 到 1GiB 的大小。可見日誌結構文件系統的段，是位於數據塊和其它文件系統 block group 中間的一個大小。

繼續上面的例子，假設上面文件系統的圖示中每一列是一個段，想要回收最開頭那個段，那麼需要搬運還在用的 1 和 4 到空閒空間，順帶修改引用它們的 n10，最後提交：

| | | | | |
|----|---|-----|-----|-----|
| 1 | 5 | 9 | n2 | 1 |
| o2 | 6 | o10 | n3 | 4 |
| o3 | 7 | 11 | o10 | n10 |
| 4 | 8 | o12 | o12 | n12 |

要掃描並釋放一整段，需要掃描整個文件系統中別的地方來確定有沒有引用到目標段中的地址，可見釋放一個段是一個 $\mathcal{O}(N)$ 的操作，其中 N 是段的數量，按文件系統的大小增長，於是刪除快照之類可能要連續釋放很多段的操作在日誌文件系統中是個 $\mathcal{O}(N^2)$ 甚至更昂貴的操作。在文件系統相對比較小而系統內存相對比較大的時候，比如手機上或者PC讀寫SD卡，大部分元數據塊（其中包含塊指針）都能放入內存緩存起來的話，這個掃描操作的開銷還是可以接受的。但是對大型存儲系統顯然掃描並釋放空間就不合適了。

段的抽象用在閃存類存儲設備上的一點優勢在於，閃存通常也有擦除塊的概念，比寫入塊的大小要小，日誌結構的文件系統中一個段可以直接對應到閃存的一個擦除塊上。所以閃存設備諸如U盤或者SSD通常在底層固件中用日誌結構文件系統模擬一個塊設備，來做寫入平衡。

基於段的GC還有一個顯著缺陷，需要複製搬運仍然被引用到的塊，這不光會增加設備寫入，還需要調整現有數據結構中的指針，調整指針需要更多寫入，同時又釋放更多數據塊，F2FS等一些文件系統設計中把這個問題叫 Wandering Tree Problem，在F2FS設計中是通過近乎「作弊」的NAT轉換表放在存儲設備期待的FAT所在位置減少這種指針調整導致的寫入。

不過基於段的GC也有一些好處，它不需要複雜的文件系統設計，不需要特殊構造的指針，就能很方便地支持大量快照。一些日誌結構文件系統比如NILFS用這一點支持了「連續快照（continuous snapshots）」，每

次文件系統提交都是創建一個快照，用戶可以手動標記需要保留哪些快照，GC 算法則排除掉用戶手動標記的快照之後，根據快照創建的時間，先從最老的未標記快照開始回收。即便如此，GC 的開銷（CPU時間和磁盤讀寫帶寬）仍然是 NILFS 最爲被人詬病的地方，是它難以被廣泛採用的原因。

WAFL 早期使用的可用空間位圖數組

從日誌結構文件系統的 GC 的困境中可以看出，文件系統級別實際更合適的不是運行期的 GC，而是在創建快照時或者寫入數據時就記錄下快照使用的數據塊，這也是 WAFL 早期實現快照的設計思路。

WAFL 早期記錄快照佔用數據塊的思路也很「暴力」，傳統文件系統一般有個「位圖（bitmap）」用一個 bit 記錄一個數據塊是否佔用，靠掃描位圖來尋找可用空間和已用空間。WAFL 的設計早期考慮既然需要支持快照，那就把記錄數據塊佔用情況的位圖，變成一個快照的數組。於是整個文件系統有個 256 大小的快照利用率數組，數組中每個快照記錄自己佔用的數據塊位圖，文件系統中最多能容納 255 個快照。

| | block1 | block2 | block3 | block4 | block5 | block6 | block7 |
|-------------|--------|--------|--------|--------|--------|--------|--------|
| filesystem | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| snapshot1 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| ... | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| snapshot255 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |

上面每個單元格都是一個 bit ，表示某個快照有沒有引用某個數據塊。