StarWind Virtual SAN® L1 and L2 caches. Operational principles

2025

StarWind Documents

Trademarks

"StarWind", "StarWind Software" and the StarWind and the StarWind Software logos are registered trademarks of StarWind Software. "StarWind LSFS" is a trademark of StarWind Software which may be registered in some jurisdictions. All other trademarks are owned by their respective owners.

Changes

The material in this document is for information only and is subject to change without notice. While reasonable efforts have been made in the preparation of this document to assure its accuracy, StarWind Software assumes no liability resulting from errors or omissions in this document, or from the use of the information contained herein. StarWind Software reserves the right to make changes in the product design without reservation and without notification to its users.

Technical Support and Services

If you have questions about installing or using this software, check this and other documents first - you will find answers to most of your questions on the Technical Papers webpage or in StarWind Forum. If you need further assistance, please contact us .

About StarWind

StarWind is a pioneer in virtualization and a company that participated in the development of this technology from its earliest days. Now the company is among the leading vendors of software and hardware hyper-converged solutions. The company's core product is the years-proven StarWind Virtual SAN, which allows SMB and ROBO to benefit from cost-efficient hyperconverged IT infrastructure. Having earned a reputation of reliability, StarWind created a hardware product line and is actively tapping into hyperconverged and storage appliances market. In 2016, Gartner named StarWind "Cool Vendor for Compute Platforms" following the success and popularity of StarWind HyperConverged Appliance. StarWind partners with world-known companies: Microsoft, VMware, Veeam, Intel, Dell, Mellanox, Citrix, Western Digital, etc.

Copyright ©2009-2018 StarWind Software Inc.

Introduction

This document reveals StarWind's approach to L1 and L2 cache memory and the ways it's implemented. The technical paper also contains information on two basic cache policies (Write-Back and Write-Through), as well as gives recommendations on what policy to choose for the highest performance. The document also provides clarity on the interaction of L1 and L2 caches. The technical paper is aimed at experienced virtualization admins and StarWind users, who want to accelerate disk requests processing by using L1 and L2 caches. A full set of up-to-date technical documentation can always be found here, or by pressing the **Help** button in the StarWind Management Console. For any technical inquiries, please visit our online community, Frequently Asked Questions page, or use the support form to contact our technical support department.

General Provisions

The cache is used to speed up the processing of disk requests using a faster intermediate storage memory, which holds frequently used data. The first level cache (RAM Cache) utilizes RAM memory, while the second-level cache (L2, flash cache) uses solid state drives. This section provides information on the disk cache that shouldn't be confused with the CPU cache and other cache types.

Features Of L1 And L2 Caches And Its Implementation In Starwind

As it is known, StarWind uses conventional RAM as a write buffer and L1 cache to adsorb writes, while flash memory serves as a L2 cache. StarWind implements L1 and L2 caches using the same algorithms (shared library). Therefore, most of the notes apply to both types of caches, while differences in their work are mentioned separately. You are well aware that when the cache is full, the algorithm must be chosen in order to discard old items and make room for the new ones. StarWind displaces data using **LRU algorithm**: least recently used. In case when the new data should be written, but the cache memory is exhausted, the least recently used blocks are discarded. In general, three **cache states** can be singled out depending on the level of the cache exhaustion: dirty, clean, and empty. At the beginning, all cache blocks are **"empty"** – the allocated memory holds no data, and cache blocks are not associated with disk blocks. They start being loaded with data during the working process. The cache block is considered **"dirty"** if user data was written to it, but has not been flushed to disk yet. In this case, the cache memory contains actual data while the disk stores irrelevant and older version of the data. The cache block is considered **"clean"** if the data it holds was also flushed to the disk (or the cache block was filled up as a result of disk read operations). In this case, the cache block and the corresponding disk store the same data. Please note that the cache memory is divided into blocks, and now the **block** size is 64kB. The other one thing to remember is the **“cache warm-up”**.

Cache Policies

The cache policy is indicated along with the other parameters while the cache is being created. In V8R5 (StarWind Version 8 Release 5) the cache mode can be changed on the fly for the already existing devices.

Write-Back Policy

This policy writes data initially to the cache and confirms the I/O back to the host immediately. If cache contains clean or empty blocks, the writing speed is similar to the RAM-drive read/write speed. The cache is filled up with data mainly during the write operations. During the read operations, the data enters the cache only if the latter contains either empty memory blocks or the lines that were allocated for these entries earlier and have not been fully exhausted yet. **The following examples show the way the WB policy works in the following cases:**

- If the writes have not been made to the “dirty” cache line for a certain period of time (now the default is 5 seconds), the data is flushed to the disk. The block becomes “clean”, but still contains the copy of the data.
- If all cache blocks are “dirty” during the write operation, the data stored in the oldest blocks is forcibly copied to the disk, and the new data is written to the blocks. Therefore, if the data is continuously written to different blocks and once the WB cache memory is exhausted, the performance falls to values comparable to the speed of the uncached device: new data can be written only after the old data is copied to the disk.
- In case of the program malfunctioning (power outage or service failure), the data which was stored in the “dirty” blocks and, therefore, has not been yet de-staged to the backing store will be lost. For ImageFile-based HA this means device needs full synchronization once it is turned on.
- When you remove the device or turn off the service, all the dirty blocks of the WB cache must be flushed to the disk. In case the cache size is too large (gigabytes), this process may take a long time. You can estimate the approximate time by following next equation:
- cache flushing time = cache size / RAM write speed.

WB policy boosts performance in the following cases:

- load fluctuations: the data is written to the cache during the load spikes and gradually copied to the main storage when the workload goes down.
- continuous rewriting of the same blocks: blocks are rewritten only in memory and several writes to the cache correspond to only one disk write.

Summing up the aforesaid, we can note that the data is copied from the WB cache to the disk if there are no requests and the dirty blocks have no data written to them for more than 5 seconds (this value is changeable via header file); a new write should be made, but all cache blocks are "dirty"; the device is removed or the service stops. Please note that the WB cache policy is preferable for the L1 cache and is currently blocked for L2 cache (L2 cache is always created in WT mode).

Write-Through Policy

First of all, it boosts only read operations. The data is cached when being read from the disk. Therefore, the next time this data is requested, it will be read from the cache memory without accessing the underlying storage. The new data is written synchronously both to the cache and to the underlying storage. If a block with the given address has not been allocated in the cache, that means the cache is filled primarily during the read operations. In WT mode, the data will not be lost if the server or the cache failure occurs before all the data is de-staged to the backing store. So far, the WT mode is the default mode (and the only mode) for the L2 cache. Cached data does not need to be offloaded to the backing store when device is removed or the service is stopped. That's why the shutdown is carried out faster than in case with the same-sized WB cache. Going for WT mode, the cache blocks are always "clean" since the new data is immediately written to the disc.

Recommendations For Starwind Cache Use

L1 Cache

Recommended to be used in Write-Back Mode to cache reads and writes. Solves the problems:

- The requests from the partner node can come with a slight delay that changes the disk access pattern from 1-2-3-4-5-6-7-8 to 1-3-2-4-5-7-6-8 (simplified example). In this case, L1 smoothens the slight mix up with sequential write requests, which occurs when the HA implements the round robin approach.L1 unifies sequential small write requests into larger requests (write coalescing). For example, 16 successive 4k write requests are written to disk in a single 64k unit that is significantly faster. Thus, for this case, a relatively small memory size of 128 MB is enough for L1 cache.

- L1 also compensates overwrites of the same disk locations. The size depends on the frequency of the blocks rewriting and the size of the working data set. **The more often rewriting – the smaller size. The bigger working data set – more size.**L1 cache implements in-memory caching of the working data set that boosts performance due to the ratio of the working data set size to the cache size.Assume that the figure of the cache command processing time is much smaller than the time the disk processes the command. Then, once the cache is flushed, the command processing time is as follows: Tcached = Tdisk * (1 – (cache size) / (working set size)).When the cache size is equal to 0.1 of the data working set size, the average command processing time will be 0.9 to the disc command processing time.When the cache size is equal to the size of the working data set, the average command processing time will be equal to the memory command processing speed.

L1 Cache Implementation Recommendations

StarWind L1 cache in write-back or write-through mode can be implemented to boost the performance for:

- Highly Available StarWind image file devices on HDD-based storage in RAID 10 configuration.
- Highly Available StarWind image file devices on all-flash storage if L1 cache is significantly faster than the main storage array (for example, in case of using read-intensive SSDs).

In the majority of use cases, there is no need to assign L1 cache for all-flash storage arrays. The size of L1 cache should be equal to the amount of the average working data set. The total amount of L1 cache assigned influences the time required for system shutdown. Please make sure not to overprovision the L1 cache amount to avoid the service interruption and the loss of cached data. **NOTE:** In case of using L1 cache in write-back mode, UPS units have to be installed to ensure the correct shutdown of StarWind Virtual SAN nodes. If a power outage occurs, this will prevent the loss of cached data. The UPS capacity must cover the time required for flushing the cached data to the underlying storage.

Flash Cache (L2 Cache)

Recommended for use in Write-Through mode (the only available mode for now) to accelerate the read operations. It comes in handy when the main storage is based on spindle drives, while the L2 cache resides on a SSD. The basic pattern is the Random Read. In this case, the spindles work much slower than when doing a sequential read. At the same time the SSD performance during the random read access pattern can be compared to the sequential read access pattern. L2 is very effective in case the working data set fully fits the L2 cache size. Then, after the cache warm-up, the random read speed can be compared to the random read speed of the SSD, used for L2 cache. If the working set data doesn’t fit into the cache, the performance goes down and can be estimated according to the formula given above for the L1 cache.

L2 Cache In Starwind

The simple ImageFile device is created in order to store L2 cache data. This device has no target and is connected to the stack of the cacheable device similar to the case when the device with data (ImageFile) is connected to the stack of HA-device. The system with StarWind device with L2 cache configured, should have more RAM available, since L2 cache needs 6,5 GB of RAM per 1 TB of L2 cache to store metadata. You can find more information, as well as the configuration guidance for the L2 cache in our Knowledge Base.

L2 Cache Implementation Recommendations

StarWind L2 cache can be used in write-through mode only
to increase the performance on read operations for:

- Highly Available StarWind image file devices on HDD-based storage in RAID 10 configuration.
- Highly Available StarWind image file devices on all-flash storage in specific scenarios. Please find StarWind KB article for assigning L2 cache for all-flash storage arrays here.

In the majority of use cases, if L1 cache is already assigned, there is no need to configure L2 cache.   The size of L2 cache should be equal to the average amount of unique data accessed on a regular basis.

Conclusion

StarWind uses RAM for L1 cache and flash memory for L2 cache to speed up the processing of disk requests. When it comes to cache policies, StarWind recommends to use L1 cache in writeback mode and L2 in the write-through mode respectively in order to achieve high performance and additional protection. Using L1 in WB mode unifies small random writes into one big sequential write, compensates workload fluctuations. As for the L2 cache, it accelerates the read operations when used in WT mode and is also very effective in case when the working data set fully fits the L2 cache size.

Contacts

US Headquarters	EMEA and APAC
☎ +1 617 829 44 95	☎ +44 2037 691 857 (United Kingdom)
✉ +1 617 507 58 45	☎ +49 800 100 68 26 (Germany)
✉ +1 866 790 26 46	☎ +34 629 03 07 17 (Spain and Portugal)
	☎ +33 788 60 30 06 (France)

Customer Support Portal:	https://www.starwind.com/support
Support Forum:	https://www.starwind.com/forums
Sales:	sales@starwind.com
General Information:	info@starwind.com

StarWind Software, Inc. 100 Cummings Center Suite 224-C Beverly MA 01915, USA