Storage Subsystem Configuration Recommendations

There are many storage technologies available today from traditional magnetic spinning HDD devices to SSD and PCIe Flash based devices. There are several technologies for accessing the storage as well such as Network Attached Storage (NAS), Storage Area Network (SAN), Fibre Channel over Ethernet (FCoE), direct attached, PCIe, and virtual or software-defined storage with hyper-converged infrastructure.

The storage technology best for your application depends on the application access patterns. For example, for applications that are predominantly random reads, SSD or Flash based storage would be an ideal solution, and for applications that are mostly write intensive traditional HDD devices may be a better solution. However new flash technologies promote the benefits of both high durability and very fast writes.

The following guidelines are provided as a general suggestion. Specific storage products may have separate and even contradicting best practices that should be consulted and followed accordingly.

Storage Connectivity

Storage Area Network (SAN) Fibre Channel

- Use multiple paths from each host to the SAN switches or storage controllers. This can be in the form of multi-port Host Bus Adapters (HBAs), or multiple HBAs. The level of protection increases with multiple HBAs to protect from a single card failure, however a minimum recommendation is to use at least a dual-port HBA.
- To provide resiliency at the storage array layer, an array with at least dual-controllers in either an active-active or active-passive configuration is recommended to protect from a storage controller failure and also provide continued access even during maintenance periods for activities such as firmware updates. Higher-end storage arrays support 4-node controllers and even multiple “engines” which have multiple controllers per engine.
- If using multiple SAN switches for redundancy, a good general practice is to make each switch a separate SAN fabric to protect from errant configuration changes on a single switch impacting both switches and impeding all storage access.

Network Attached Storage (NAS)

- With 10Gb (and now 40Gb) Ethernet commonly available, for best performance a minimum of 10Gb switches and host network interface cards (NICs) are recommended.
- Having a dedicated NAS infrastructure is also advised to isolate traffic from normal network traffic on the LAN. This will help ensure predictable NAS performance between the hosts and the storage.
- Jumbo frame support should be included to provide efficient communication between the hosts and storage.
- Many NICs provide support for a TCP Offload Engine (TOE). TOE support has mixed reviews and opinions. The overhead and gains greatly depend on the server's CPU itself for available cycles (or lack there of). Additionally TOE support has a limited lifetime of usefulness because system-processing power rapidly catches up to TOE performance levels of a given NIC, or in many cases exceed it.

Storage Configuration

The storage array landscape is ever changing in technology features, functionality, and performance options, and multiple options will provide optimal performance and resiliency for Caché. The following guidelines provide a general best practice for optimal Caché performance and data resiliency.

- For performance and recoverability, the primary journal directory, alternate journal directory, databases, and write image journal files should reside on separate sets of physical devices. However, storage arrays today – especially SSD/Flash based arrays do not always allow for this type of segregation. In those cases consult and follow the respective storage vendor's recommendations for performance and resiliency.
- In the past RAID10 was recommended for maximum protection and performance. However storage controller capacities, RAID types and algorithm efficiencies, and controller features such as inline compression and de-duplication provide for more options than ever before. Additionally your specific application's IO patterns will help you decide with your storage vendor which storage RAID levels and configuration are best suited.
- Where possible it is best to use block sizes similar to that of the file type. Most storage arrays have a limit on how small of a block size that is configurable - for example a CACHE.DAT file with 8KB block format. In those cases a 32KB or 64KB storage block size is usually a viable option on the storage array to effectively support the 8KB database block. The goal here is to avoid excessive/wasted IO on the storage array based on your application's needs.

The following table is provided as a general overview of the storage IO profiles within a Caché installation.

IO Type	When	How	Notes
Database Reads	Continuous by user processes	User process initiates a disk I/O to read the data	Database reads are performed by either the daemons serving web pages, SQL queries, or direct user processes.
Database Writes	Burst approx. every 80 seconds or percentage of database cache pending updates	Database write daemons (8 processes)	Database writes are performed by a set of database system processes known as write daemons. User processes update database cache and a trigger (time or activity threshold) will send the updates to disk using the write daemons. Typically expect anywhere from a few MBs to several GBs that need to be written during the write cycle depending on update rates.
Journal Writes	<2secs, full journal buffers, or sync request	Database journal daemon (1 process)	Journal writes are sequential and variable in size from 4KB to 4MB. There can be as low as a few dozen writes per second to several thousand per second for very large deployments using ECP and separate application servers.
Write Image Journal Writes	Burst approx. every 80 seconds or percentage of database cache pending updates	Database master write daemon (1 process)	This journal is used to protect physical database file integrity from system failure during a database write cycle. Writes are approximately 256KB each in size.

Bottlenecks in storage are one of the most common problems affecting database system performance. A common problem is sizing storage simply for GB capacity, rather than allocating a high enough number of discrete disks to support expected Input/output Operations Per Second (IOPS).

The requirements vary slightly depending on whether separate application (ECP) servers are used. As noted in the architecture section, most sites do not require separate ECP servers.

The following table details the suggested storage response times for a given IO type.

IO Type	Average Response Time	Maximum Response Time	Notes
8KB Database Random Read (non-cached)	<=6 ms	<=10 ms	Database blocks are a fixed 8K, 16K, 32K, or 64K – most reads to disk will not be cached because of large database cache on the host.
8KB Database Random Write (cached)	<=1 ms	<2 ms	All database file writes are expected to be cached by the storage controller cache memory if available – in the case of SSD storage it is best to disable write cache.
4KB to 4MB Journal Write (without ECP)	<=2 ms	<=5 ms	Journal writes are sequential and variable in size from 4KB to 4MB. Write volume is relatively low when no application servers are used.
4KB to 4MB Journal Write (with ECP)	<=1 ms	<=2 ms	Journal synchronization requests generated from ECP impose a stringent response time requirement to maintain scalability. The synchronization requests may trigger writes to last block in the journal to ensure data durability.

Please note these figures are just a guideline and any given application have higher or lower tolerances and thresholds for ideal performance. Some high preforming applications may require the use of all-flash and maybe even the extreme low latency of PCIe flash-based storage.

These figures and IO profiles are to be used as a starting point for your discussions with your storage vendor.